Российская академия наук
Институт вычислительной математики

Марчук Г. И.
Избранные труды

Том 1

Методы вычислительной математики

Москва 2018

УДК 519.6
ББК 22.19
С56

Ответственный редактор: д.ф.-м.н., профессор Агошков В.И.

Марчук Г.И. Избранные труды: в 5 т. / Российская академия наук, Институт вычислительной математики. – М.: РАН, 2018.

Т.1.: Методы вычислительной математики / [отв. ред. В.И. Агошков]. – 765с.

В настоящее собрание избранных трудов вошли монографии и статьи, наиболее ярко отображающие многолетнюю научную деятельность Г. И. Марчука в вычислительной математике и математическом моделировании. Подготовка к изданию данного собрания велась в ИВМ РАН ближайшими учениками и соратниками Г.И.Марчука. Тома содержат комментарии, в которых проанализирован вклад работ Г.И. Марчука в современную науку.

В Томе 1 дано изложение численных методов решения задач математической физики. Основное внимание уделяется сложным задачам математической физики, которые в процессе решения сводятся, как правило, к более простым, допускающим реализацию алгоритмов на ЭВМ. Рассмотрены многие современные подходы к численным методам.

Для специалистов в области вычислительной математики и математического моделирования, аспирантов и студентов старших курсов.

ISBN 978-5-906906-26-7

Оглавление

Предисловие к Избранным трудам академика Г.И. Марчука

Академик Российской академии наук Г.И. Марчук (1925-2013) - выдающийся ученый России, широко известный во всем мире, крупный организатор науки. Блестящий разносторонний исследователь в области естествознания Г.И.Марчук внес огромный вклад в развитие вычислительной и прикладной математики, физики ядерных реакторов, математического моделирования, динамической метеорологии, информатики, иммунологии.

Гурий Иванович Марчук родился 8 июня 1925 года в поселке Петро-Херсонец Оренбургской области в семье сельского учителя. После окончания математико-механического факультета Ленинградского государственного университета Г.И.Марчук поступил в аспирантуру и в 1952 г. защитил кандидатскую диссертацию "Динамика крупномасштабных полей метеорологических элементов в бароклинной атмосфере". С 1953 по 1962 гг. Г.И.Марчук работал в Физико-энергетическом институте (г.Обнинск), где заведовал лабораторией, а затем математическим отделом. В этот период он предложил новые методы расчета ядерных реакторов, которые до настоящего времени составляют основу моделирования имитационных расчетов промышленных реакторов. Большую известность получили его работы по теории переноса излучения. Результаты этих исследований обобщены в монографии "Численные методы расчета ядерных реакторов"и в его докторской диссертации (1965 г.). В 1959-1961 гг. он принял участие в разработке требований к ядерной безопасности для заводов и других предприятий атомной промышленности, проводившейся по инициативе И.В.Курчатова. В 1961 г. за работы в области теории ядерных реакторов ему была присуждена Ленинская премия.

В 1962г. Г.И.Марчук был избран членом-корреспондентом АН СССР (специальность "атомная энергетика"), а в 1968г. он был избран действительным членом АН СССР (специальность "физика атмосферы").

Г. И. Марчук - автор более 350 научных работ, в том числе 25 монографий. Научные труды Г. И. Марчука посвящены созданию и исследованию эффективных алгоритмов вычислительной математи-

ки, методов расчета ядерных реакторов, исследованию и моделированию процессов физики атмосферы и океана, математическому моделированию в проблемах охраны окружающей среды, в проблемах иммунологии и медицины, изучению актуальных задач информатики и вычислительной техники.

В области вычислительной математики Г. И. Марчуком сделан существенный вклад в развитие разностных схем. Им построены и исследованы разностные схемы для классов уравнений, возникающих в теории ядерных реакторов, предложен метод построения разностных схем на основе интегральных тождеств, который получил развитие в работах советских и зарубежных ученых. Г. И. Марчуком и его учениками решен ряд проблем в теории разностных и вариационно-разностных схем для различных задач математической физики.

Г. И. Марчук внес большой вклад в разработку методов расщепления, алгоритмов возмущений, построенных на основе использования сопряженных уравнений. За применение этих подходов к развитию методов статистического моделирования (метода Монте-Карло) Г. И. Марчук в составе коллектива авторов был удостоен Государственной премии СССР.

В методах расчета ядерных реакторов Г. И. Марчуком на основе теории сопряженных уравнений и алгоритмов возмущений разработаны принципы построения эффективных малогрупповых моделей ядерного реактора, созданы математические модели реактора в различных приближениях метода сферических гармоник и предложены численные схемы реализации возникающих уравнений. Эти модели широко использовались для расчетов критических масс промышленных реакторов.

Большой вклад Г. И. Марчук внес в решение задач численного прогноза погоды, моделирования общей циркуляции атмосферы и океана и проблему моделирования климата и его изменений. Его совместная работа с Н. И. Булеевым, в которой была сформулирована система квазигеострофических уравнений для трехмерной атмосферы и построена функция Грина для ее решения, давно стала классической. В 60-е годы он сформулировал направление численного прогноза погоды, основой которого было использование полных недиабатических уравнений динамики атмосферы, а в качестве ме-

тода решения предложен метод расщепления по физическим процессам и геометрическим переменным - в то время, пожалуй, единственный метод, с помощью которого можно было решить такую сложную задачу. Важно отметить, что задача численного прогноза по полным уравнениям была доведена до оперативного использования в Западно-Сибирской Гидрометслужбе. В 70-е годы Г.И.Марчук сформулировал новый подход к решению задачи долгосрочного прогноза погоды, основанный на использовании так называемых сопряженных уравнений для нелинейных уравнений термогидродинамики атмосферы и океана, дающих возможность построить функцию чувствительности для нестационарных нелинейных задач. Этот подход стал основным при выделении энергоактивных зон Мирового океана, изучению которых была посвящена программа "Разрезы которую сформулировал и организовал Гурий Иванович. В те же 70-е годы большое внимание Г. И. Марчуком уделялось созданию совместной модели общей циркуляции атмосферы и океана, которая должна была стать основой для моделирования климата и его изменений. За работы по численному прогнозу погоды Г. И. Марчуку была присуждена премия им. А.А.Фридмана АН СССР, а за работы по решению задач физики атмосферы и океана Государственная премия Российской Федерации.

Г. И. Марчуком создана теория математического моделирования оптимизационных проблем в охране окружающей среды. Им поставлены и предложены алгоритмы решения общей задачи определения допустимой области размещения промышленных предприятий, планирования строительства с учетом допустимых доз загрязнения экономически значимых зон.

Марчук Г. И. является одним из авторов нового направления прикладной математики – математического моделирования в иммунологии и медицине. Он построил систему нелинейных дифференциальных уравнений, описывающих иммунные реакции человеческого организма, возникающих в результате вирусных и бактериальных инфекций.

Г. И. Марчук - крупный организатор науки. С 1964 по 1979 год он — директор Вычислительного центра СО АН СССР; с 1969 по 1975 год - заместитель, а затем Председатель Сибирского отделения АН СССР; с 1980 по 1986 год — Председатель Государственного комите-

та СССР по науке и технике в ранге заместителя Председателя Совета Министров СССР; с 1975 по 1980 год - Вице-президент, а с 1986 по 1991 год Президент АН СССР. В 1980 году Г. И. Марчук создал и возглавил Отдел вычислительной математики АН СССР. В 1991 году Отдел был преобразован в Институт вычислительной математики РАН. До 2000 года Г. И. Марчук - директор ИВМ РАН. С 2000 по 2013 год Г. И. Марчук - советник Президиума РАН и почетный директор ИВМ РАН.

За научные заслуги и вклад во внедрение научных достижений в народном хозяйстве Г. И. Марчук удостоен звания Героя Социалистического Труда, награжден четырьмя орденами Ленина, орденами "За заслуги перед Отечеством" IV и II степеней. Ему присуждены Ленинская и Государственные премии. Он обладатель Большой золотой медали Российской академии наук им. М. В. Ломоносова - главной научной награды РАН, Золотой медали им. М. В. Келдыша, а также Золотой медали им. П. Л. Чебышева.

Научные успехи Г. И. Марчука высоко оценены и за рубежом. Он является почетным доктором Хьюстонского, Орегонского, Тулузского, Дрезденского, Тель-Авивского, Калькутского, Карлова и Будапештского университетов, членом Европейской академии наук, иностранным членом Академии наук Франции, Финляндии, Индии, Польши, Болгарии и др.; лауреатом Международной премии им. А.П.Карпинского; Командором Ордена Почетного Легиона – государственной награды Франции. Ему присуждена медаль Вильгельма Бьеркнеса Европейского союза наук о Земле. Г.И.Марчук был членом редколлегий пяти иностранных (США, Германия, Италия, Франция, Швеция), а также отечественных журналов, главным редактором журнала Russian Journal of Numerical Analysis and Mathematical Modelling.

На протяжении всей своей научной биографии Марчук Г. И. уделял большое внимание подготовке научных кадров, возглавлял кафедры в ФЭИ (Обнинск), НГУ (Новосибирск), МФТИ (Долгопрудный), МГУ (Москва). Под его руководством защищено свыше 35 кандидатских диссертаций. Среди учеников Г. И. Марчука 25 докторов наук.

В настоящее собрание избранных трудов вошли монографии и статьи, наиболее ярко отображающие многолетнюю научную деятельность Г. И. Марчука в вычислительной математике и матема-

тическом моделировании. Подготовка к изданию данного собрания велась в ИВМ РАН ближайшими учениками и соратниками Г. И. Марчука. Все тома содержат комментарии, в которых проанализирован вклад работ Г. И. Марчука в современную науку. Избранные труды будут значительным вкладом в академическое собрание научной и учебной литературы по широкому спектру прикладной математики и будут служить увековечиванию памяти академика Г. И. Марчука.

Собрание научных трудов Г. И. Марчука несомненно будет востребовано широким кругом научных работников, исследования которых связаны с решением сложных задач современной науки и техники. Оно будет также представлять интерес для преподавателей, аспирантов и студентов университетов, для подготовки специализированных курсов по различным научным дисциплинам.

Академик В. П. Дымников

Предисловие к тому 1

Предлагаемая книга является результатом обработки курса лекций по вычислительной математике, который в течение ряда лет читался автором для студентов математического факультета Новосибирского государственного университета. Автор стремился акцентировать внимание на сложных задачах математической физики, которые в процессе решения, как правило, редуцируются к более простым, хорошо изученным теоретически и допускающим эффективную реализацию алгоритмов на современных вычислительных машинах. Именно с такими сложными задачами зачастую сталкивается молодой исследователь в своей практической работе после окончания высшего учебного заведения. Поэтому данная книга прежде всего рассчитана на тех, кто впервые встречается с необходимостью решения больших задач математической физики и хочет получить рекомендации о рациональных подходах к решению.

Автором избрана такая форма изложения, которая, по его мнению, способствует привлечению внимания к проблемам прикладной и вычислительной математики более или менее широкого круга исследователей. Эта форма потребовала известных уступок в изложении, позволив сосредоточить внимание лишь на основных идеях и подходах к решению задач. Что касается деталей, иногда существенных, и возможных обобщений, например таких, как минимальные требования к гладкости функций, ограничения на входные данные задач и т. п., то для специалистов они в большинстве случаев очевидны, а для начинающего исследователя предоставляют хорошие возможности для полезных упражнений.

Двенадцатая глава основана на материалах доклада автора на Международном математическом конгрессе в Ницце (1970 г.), дополненных новыми материалами. Эта глава дает некоторое представление не только о методах и проблемах вычислительной математики, рассмотренных в курсе, но и о тех направлениях, которые не вошли в книгу, но имеют существенное значение как в теоретическом плане, так и для приложений.

Часть материала книги была изложена в монографии под тем же названием, вышедшей в 1973 г. Так, в первое издание (1977 г.) включен ряд новых идей и алгоритмов, которые представляют методиче-

ский и практический интерес. В частности, в книгу были включены новые алгоритмы оптимизации на основе вариационных методов, вопросы автоматизации вычислительного процесса на основе так называемого метода фиктивных областей, рассмотрен итерационный алгоритм расщепления задачи в случае некоммутирующих операторов, метод неполной факторизации и др. Раздел книги, посвященный интерполяции функций с помощью сплайнов, был расширен и выделен в самостоятельную главу. Также в отдельную главу выделен круг идей, связанных с экстраполяцией по Ричардсону, для решения задач с повышенной точностью.

Во втором издании данного учебного пособия (1980 г.) были исправлены неточности и опечатки и была включена часть нового материала, что расширило круг рассматриваемых методов. В книгу была также включена новая глава по теории оптимизации, становящейся в наши дни неотъемлемой частью формирования математических моделей и методов их реализации.

Третье издание книги являлось существенно переработанным вариантом второго издания. Изменения внесены в ряд глав. Так, значительная часть примеров применения разностных методов к простым, но широко распространенным задачам математической физики перенесена в первую главу, после чего эта глава может быть использована для знакомства с основными понятиями теории разностных схем.

В третье издание были также включены две новые главы. Одна из них посвящена алгоритмам возмущений, а другая – методам Шварца и разделения области. Алгоритмы возмущений давно используются в прикладной математике, но в учебной литературе они изложены недостаточно полно. Появление другой новой главы обусловлено тем обстоятельством, что в настоящее время возрос интерес к известному методу Шварца, а также появилось целое направление алгоритмов, которые можно объединить одним названием – метод разделения области.

Четвертое издание являлось стереотипным третьему.

В современной практике вычислений многие решаемые задачи являются существенно многомерными, и в них возникают массивы с большим количеством данных. Это потребовало новых методов и подходов к исследованию подобных задач. Такими методами являются

вычислительные тензорные методы, интенсивно развивающиеся в настоящее время и образующие новое направление в вычислительной математике. Этим методам посвящена новая глава в настоящем издании. Также в двенадцатой главе приведен дополнительный материал с кратким обзором истории и научной литературы по вычислительным тензорным методам.

При работе над данной книгой автору постоянно помогали обсуждения и научные контакты со многими учеными различных учреждений нашей страны: Института вычислительной математики РАН, Института математической геофизики и вычислительной математики СО РАН (г. Новосибирск), Института вычислительного моделирования СО РАН (г. Красноярск), Математического института РАН им. В. А. Стеклова, Института прикладной математики РАН им. М. В. Келдыша, Вычислительного центра РАН, Московского и Новосибирского государственных университетов, Института математики СО РАН и многих других. Их замечания и пожелания в значительной степени способствовали усовершенствованию книги. Фамилии этих ученых читатели найдут в тексте при изложении соответствующего раздела или в списке литературы. Всем им автор выражает свою искреннюю благодарность.

Г. И. Марчук
Москва, 2012

Указатель обозначений

$\mathbf{R}^n$ — евклидово пространство n-мерных вещественных векторов
D — область в пространстве $\mathbf{R}^n$
∂D — граница области D
$L_2(D)$ — подпространство вещественных измеримых функций, суммируемых с квадратом в области D
$(\cdot,\cdot)$ — скалярные произведения в $L_2(D)$ и в $\mathbf{R}^n$
$\|\cdot\|$ — нормы в $L_2(D)$ и в $\mathbf{R}^n$
A — оператор, действующий в пространстве $L_2(D)$, или матрица, действующая в $\mathbf{R}^n$
$A \geq 0$ — положительно полуопределенный оператор
$A > 0$ — положительный оператор
$\Phi(A)$ — область определения оператора A
A^* — сопряженный оператор
$A = A^*$ — самосопряженный оператор
$\lambda(A)$ — собственное число оператора A
$\|A\|$ — норма оператора A
$\alpha(A) = \lambda_{min}(A)$ — минимальное по модулю собственное число оператора A
$\beta(A)$ — спектральный радиус оператора A
$W_2^l(D)$ — пространство Соболева
$(\cdot,\cdot)_{W_2^l(D)}$ — скалярное произведение в $W_2^l(D)$
$\overset{\circ}{W}{}_2^1(D)$ — пространство Соболева функций пространства $W_2^l(D)$, обращающихся в нуль на ∂D
Δ — оператор Лапласа
Δ^h — разностный аналог оператора Лапласа на равномерной сетке
D_h — сеточная область
∂D_h — граница сеточной области D_h
$(\varphi)_h$ — проекция функции φ на сетку
$\Delta_x, \Delta_y, \nabla_x, \nabla_y$ — разностные операторы
A^h, a^h — сеточные операторы
F_h, G_h, Φ_h — пространства сеточных функций
$\|\cdot\|_{F_h}, \|\cdot\|_{G_h}, \|\cdot\|_{\Phi_h}$ — нормы в пространствах F_h, G_h, Φ_h
$H_N^m(a,b)$ — пространство кусочно-полиномиальных функций на отрезке $[a,b]$
$C^m(a,b)$ — пространство функций, непрерывных вместе с производны-

ми до порядка m включительно, на отрезке $[a, b]$
T — оператор шага
$supp(\varphi)$ — носитель функции $\varphi(x)$ (замыкание множества точек x, в которых $\varphi(x) \neq 0$)
$\varnothing$ — пустое множество

Введение

Современные электронные вычислительные машины дали в руки исследователей эффективное средство для математического моделирования сложных задач науки и техники. Именно поэтому количественные методы исследования в настоящее время проникают практически во все сферы человеческой деятельности, а математические модели становятся средством познания.

Роль математических моделей далеко не исчерпывается проблемой познания закономерностей. Их значение непрерывно возрастает в связи с естественной тенденцией к оптимизации технических устройств и технологических схем планирования эксперимента. В процессе познания и в стремлении создать детальную картину исследуемых процессов мы приходим к необходимости строить все более сложные математические модели, которые в свою очередь требуют универсального тонкого математического аппарата. Реализация математических моделей на ЭВМ осуществляется с помощью методов вычислительной математики, которая непрерывно совершенствуется вместе с прогрессом в области электронно-вычислительной техники.

Всякая редукция задач математической физики или техники в конечном итоге обычно сводится к алгебраическим уравнениям той или иной структуры. Поэтому предмет вычислительной математики, как правило, связан с методами сведения задач к системам алгебраических уравнений и их последующему решению.

Построение систем алгебраических уравнений, соответствующих той или иной задаче с непрерывно меняющимися аргументами, обычно существенно опирается на априорную информацию, связанную с исходной задачей. Такой информацией может быть принадлежность решения к тому или иному классу функций, обладающих определенными свойствами гладкости, свойства операторов задачи, свойства входных данных и т. д. Априорная информация во многих случаях оказывает решающее влияние на выбор методов вычислительной математики, используемых для решения указанных алгебраических уравнений. При этом, как правило, должно иметь место соответствие между априорными требованиями для исходной задачи и свойствами ее алгебраического аналога. Это прежде всего отно-

сится к операторам задач, свойства которых должны быть по возможности сохранены при редукции задачи от непрерывных аргументов к дискретным.

Такой принцип, по-видимому, является основополагающим при решении многих задач. Одновременно следует отметить, что преемственность свойств операторов задач при редукции дает возможность опираться на хорошо разработанные методы функционального анализа, что обычно позволяет простым и универсальным путем проводить исследования эффективности алгоритмов вычислительной математики.

Теперь мы переходим к краткому обзору книги, с тем чтобы отразить главные моменты и новые идеи, предлагаемые читателю.

Первая глава посвящается общим вопросам теории разностных схем. Наряду с уже ставшими классическими понятиями в теории разностных схем, такими как аппроксимация, счетная устойчивость и сходимость решений разностных уравнений, в этой главе приведены некоторые важные результаты, связанные с общими свойствами основных и сопряженных задач, которые будут использованы во многих главах книги. Нам хотелось бы особо выделить п. **??**, в котором приведены современные алгоритмы для вычисления границ неотрицательного спектра матриц. Как известно, верхняя граница спектра находится с помощью хорошо разработанных итерационных процессов, и эта проблема, как правило, не вызывает трудностей в реализации. Что касается минимального собственного числа — нижней границы спектра, то его вычисление обычно является сложной проблемой.

Теоретически наиболее простой подход, связанный с оценкой максимального собственного числа обратного оператора, оказывается алгоритмически малоэффективным. В книге изложен другой подход, связанный со сдвигом спектра операторов, который позволяет достаточно просто находить нижнюю границу спектра. На этом вопросе мы остановились подробно, поскольку многие вычислительные алгоритмы, особенно связанные с оптимизацией итерационных процессов, существенно опираются на априорную информацию о границах спектра.

Во второй главе рассмотрены методы построения разностных схем. При этом мы сконцентрировали свое внимание на двух под-

ходах: методе интегральных соотношений и вариационных способах построения разностных схем. Каждый из этих подходов имеет свои определенные преимущества и некоторые недостатки. Отметим лишь, что эти подходы не являются независимыми и при определенных условиях приводят к тождественным разностным схемам, аппроксимирующим исходные дифференциальные задачи.

Тем не менее следует отметить, что вариационный подход к построению разностных схем во многих случаях бывает более предпочтительным, поскольку он приводит к сохранению свойств определенности исходных операторов при переходе к разностным. Важно отметить, что для широкого класса задач это происходит автоматически.

В книге мы ограничились рассмотрением трех методов построения разностных схем на основе вариационных принципов: метода Ритца, метода Галеркина и метода наименьших квадратов. Конечно, они не исчерпывают всего многообразия вариационных подходов, однако они позволяют познакомиться с общими принципами конструкции разностных схем, которые весьма просто могут быть распространены и на другие случаи.

Несколько слов о методе конечных элементов (вариационно–разностный метод, проекционно–сеточный метод). Можно сказать, что этот метод является удобным средством для построения разностных схем на основе вариационных принципов. В своей методологической основе метод конечных элементов тесно связан с методом рядов Фурье, но вместо привычных нам базисных функций (например, тригонометрических функций, многочленов Лежандра, Эрмита и т. д.) здесь мы имеем дело с многочленами, отличными от нуля только в сравнительно небольшой области изменения аргументов, т. е. с финитными функциями.

Применение вариационных принципов к построению разностных схем не случайно. Из теории следует, что вариационный функционал, адекватно отражающий определенные закономерности механики, математической физики, динамики и т. д., достигает своего экстремального значения на решении интересующей нас задачи. Поэтому если нам задан вариационный функционал и определен класс функций, на которых следует минимизировать функционал, то

дальнейшая задача состоит в алгоритмическом отыскании функци, доставляющей экстремум функционалу.

Если класс допустимых функций сужать, налагая на них дополнительные ограничения, то минимизирующая функция не обязательно будет решением исходной задачи, а будет только приближаться к точному решению.

В будущем, когда средства вычислительной техники станут еще более мощными, роль вариационных функционалов в построении решений задач математической физики будет непрерывно возрастать.

Появятся методы целенаправленного перебора пробных функций, принадлежащих широким классам, позволяющие эффективно находить экстремальные решения. Таким образом, использование вариационных функционалов для решения задач все более смыкается с проблемой оптимальной организации алгоритма получения решения задачи с заданной точностью, т. е. с теорией оптимизации.

Наряду с классически поставленными задачами, при решении задач науки и техники зачастую приходится иметь дело с задачами, поставленными неклассически. К ним, например, относится задача с ограничениями. Правда, простейшие задачи с ограничениями являются классическими. Ограничениями, например, являются краевые условия для дифференциальных задач.

Более сложные задачи с ограничениями требуют для своего решения и более сложного математического аппарата. Например, если требуется решить задачу о прогибе мембраны под действием различных сил, если ее положение сверху и снизу ограничивается заданными функциями координат, то обычный классический подход оказывается бессильным. Тем не менее если такой задаче поставить в соответствие некоторый вариационный функционал и отыскивать его минимум на классе функций, каждая из которых удовлетворяет заданному ограничению, то минимизирующая функция будет доставлять решение нашей задачи.

В главе третьей рассмотрены вопросы интерполяции сеточных функций. Проблема интерполяции возникает всякий раз, когда требуется заданную на сетке функцию восполнить непрерывными функциями на всю область. Сюда относятся задача продолжения приближенного решения на всю область по его значениям в узлах сетки

и задача обработки экспериментальных данных, известных на дискретном множестве точек.

Задача интерполяции становится фундаментальным звеном в системе автоматизации проектно-конструкторских работ, где в самом существе проблемы заложены способы графического отображения информации. Проблема интерполяции не является новой, и в математической литературе классические методы изложены достаточно полно. Новым в последние десятилетия направлением в теории интерполяции является использование так называемых сплайновых интерполяций, описанию которых в основном и посвящена третья глава.

Сплайновые интерполяции являются наилучшим средством построения гладких восполнений сеточных функций на заданных классах функций. Оптимальность сплайна связана с его специальным экстремальным свойством. Сплайновые аппроксимации все более широко применяются во всех областях науки и техники, поэтому знакомство с ними читателя, по нашему мнению, является необходимым.

Четвертая глава в основном посвящена итерационным методам решения линейных алгебраических уравнений. Здесь изложены как общие подходы к решению алгебраических систем, так и специфические методы, связанные с особенностями аппроксимаций задач математической физики, с помощью разностных и вариационно-разностных методов. Хотя литература по итерационным методам весьма обширна и содержит описание многих эффективных алгоритмов, тем не менее в настоящей книге мы, наряду с рассмотрением классических процессов, основное внимание уделили итерационным методам, оптимизируемым с помощью квадратичных функционалов. В этом состоит выражение нашего общего подхода к вопросам оптимизации как при построении вычислительных алгоритмов, так и при их реализации.

Касаясь специфических проблем, связанных с частным видом матриц, возникающих при численном решении задач математической физики, мы ориентируемся на методы расщепления матриц на простейшие в общей схеме итерационного процесса. Метод расщепления является естественным развитием метода попеременных направлений, сыгравшим исключительную роль в численном решении

задач математической физики. Метод расщепления имеет различные модификации и обобщения, в том числе с использованием вариационных принципов.

Особого внимания заслуживают прямые методы решения конечноразностных уравнений, изложенные в конце этой главы. Это прежде всего быстрое преобразование Фурье и метод циклической редукции. Эти методы предложены недавно и их популярность непрерывно возрастает.

Методам решения нестационарных задач посвящена пятая глава книги. Эти методы в основном связаны с использованием идеи расщепления сложных операторов задач на более простые. Здесь не только проанализированы хорошо утвердившиеся в практике методы, такие как метод стабилизации и метод предиктор–корректор, но и детально описан наиболее эффективный, по нашему мнению, метод покомпонентного расщепления, идея которого изложена в 5.3.3 и 5.4.

Метод покомпонентного расщепления позволяет на каждом временном шаге сводить сложную задачу математической физики к последовательности простейших однокомпонентных задач. В результате мы приходим к эффективному алгоритму реализации на ЭВМ, абсолютно устойчивому и обеспечивающему второй порядок аппроксимации решения как по пространственным переменным, так и по времени. Этот метод применяется для широкого класса нестационарных задач математической физики.

В шестой главе рассматриваются методы уточнения приближенных решений, восходящие к Ричардсону и Рунге. Как известно, уточнение приближенных решений можно производить различными методами. Обычно для этой цели используются более точные аппроксимации дифференциальных или интегральных уравнений с помощью схем высокого порядка точности. Ричардсон предложил для этой цели использовать разностную аппроксимацию сравнительно невысокого порядка точности, но примененную для различных сеток. Так, если исходное разностное уравнение соответствует аппроксимации на сетке с шагом h, то следующее соответствует шагу $h/2$ и т. д. В результате мы приходим к разностным уравнениям, записанным для последовательности сеток. Оказывается, что при выполнении ряда требований относительно операторов, шага сетки и исход-

ных данных задачи линейная комбинация приближенных решений на последовательности сеток позволяет получить решение более высокого порядка точности по сравнению с исходными решениями.

Метод экстраполяции Ричардсона, первоначально предложенный для обыкновенных дифференциальных уравнений, удалось применить к решению краевых задач для уравнений эллиптического и параболического типов. Здесь, естественно, возникают различные особенности, которые отмечены в схемах реализации. Важно подчеркнуть, что метод Ричардсона может быть применен к решению задач с малым параметром или для решения условно корректных задач на основе методов регуляризации. В этом случае метод Ричардсона основывается на решении задач с различными параметрами, сходящимися к предельному их значению. Таким образом, метод экстраполяции позволяет ввести в вычислительную математику новые идеи, которые с успехом используются для оптимизации различных алгоритмов решения задач.

Следует также подчеркнуть особое место, которое отводится этому методу при решении задач вариационно-разностными методами. В самом деле, здесь имеет место обычно следующая альтернатива: либо получать решение разностных уравнений с очень мелким шагом на основе довольно грубых разностных аппроксимаций, либо с помощью схем высокого порядка точности, но при более крупных шагах разностных схем.

Первый метод прост, но требует большого объема вычислений, второй логически значительно труднее, но требует меньшего числа арифметических операций. Таким образом, ни тот, ни другой метод не оказывается эффективным для задач математической физики, если требуется высокая точность результатов. Поэтому возникла мысль использовать наиболее простые вариационно-разностные схемы первого или второго порядка аппроксимаций, но реализованные на последовательности сеток. Линейная комбинация решений таких задач, как указано выше, во многих случаях позволяет получить решение требуемого порядка точности.

В седьмой главе рассматриваются методы, к которым в последнее время значительно возрос интерес. Эти методы позволяют в ряде случаев свести решение задач в области со сложной границей к решению последовательности задач в более простых областях. От-

метим здесь метод разделения области, исследования по которому ведутся во многих странах. Интерес к данному методу обусловлен и тем обстоятельством, что он часто допускает крупноблочное распараллеливание процесса решения исходной задачи. Это обстоятельство является важным в связи с внедрением в практику вычислений многопроцессорных ЭВМ, работающих в параллельном режиме.

В восьмой главе формулируются алгоритмы теории регулярных возмущений. Данные алгоритмы излагаются в применении к неоднородным функциональным уравнениям, задачам на собственные значения, вычислению линейных функционалов. Рассмотрение здесь проводится с использованием как основных, так и сопряженных уравнений, осуществляется иллюстрация алгоритмов на примере конкретных прикладных задач.

Девятая глава посвящена постановке и численному решению обратных задач. Методы математического моделирования сложных задач науки и техники постоянно выдвигают перед исследователем проблемы, связанные с восстановлением решения задачи по некоторым функционалам от решения или с восстановлением вида оператора задачи. Этот класс обратных задач оказывается наиболее трудным с точки зрения вычислительной математики, поскольку он, как правило, связан с решением некорректных по Адамару задач. В математике возникло целое направление исследований некорректных задач, основные результаты которых были получены советской школой математиков. А. Н. Тихонов ввел в рассмотрение процесс регуляризации таких задач, и они вскоре нашли свое обоснование и алгоритмическое оформление. В этой главе делается акцент на постановку обратных задач по восстановлению структуры дифференциальных операторов и входных данных. Хотя вид дифференциального оператора фиксируется, но его коэффициенты предполагаются неизвестными, требующими определения. Теория обратных задач тесным образом связана с использованием основных и сопряженных уравнений. Разработанный автором математический аппарат оказывается эффективным для оценки малых возмущений функционалов от решений задач в зависимости от вариаций входных параметров. Следует особо отметить, что разработанные здесь методы могут быть применены как для линейных, так и для нелинейных задач математической физики.

В десятой главе изложены методы оптимизации, которые активно вторгаются в математическое моделирование технологических процессов, экономики и управления. Это прежде всего методы линейного и квадратичного программирования, которые с алгоритмической точки зрения продвинуты наиболее существенно. В этой главе даны также описания подходов к нелинейному программированию для выпуклых функций и выпуклых областей. Далее читатель знакомится с общими идеями метода динамического программирования и принципа максимума Понтрягина. В заключение рассматриваются проблемы оптимизации задач математической физики с ограничениями на основе метода вариационных неравенств, получившие широкое развитие в трудах французских математиков.

В настоящее время исследователям приходится встречаться с прикладными задачами, являющимися существенно многомерными. Для их численного решения активно применяются вычислительные тензорные методы. Данным методам посвящается одиннадцатая глава.

Двенадцатая глава посвящена обзору методов вычислительной математики. Настоящая книга, разумеется, не могла включать в себя огромного объема алгоритмов вычислительной математики, разработанных к настоящему времени. Многие из них вообще выпали из рассмотрения, поскольку хорошо описаны в учебниках по вычислительной математике или в специальной литературе.

Мы не говорим здесь о таких классических основах численного анализа, как кубатурные формулы, методы численного решения обыкновенных дифференциальных уравнений, простейшие методы интерполяции и др. Речь идет о новых методах современной вычислительной математики, таких как метод крупных частиц, метод интегральных соотношений, универсальные методы линейной алгебры и др. Они отражены в обзоре.

Обзор методов вычислительной математики, данный в настоящей главе, сопровождается постановкой ряда проблем вычислительной математики и анализом тенденций их развития. По нашему мнению, это поможет читателю не только сориентироваться в проблемах вычислительной математики, но и определить наиболее активно развивающиеся ее области.

Поскольку книга является учебным пособием, мы старались в основном тексте избегать ссылок на библиографию, которые могли бы отвлечь читателя от систематического ознакомления с материалом. Этот пробел частично восполняется одиннадцатой главой, где, кроме обзора методов вычислительной математики, даются ссылки на соответствующие источники.

Особое место в книге отводится списку литературы. Этот список систематизирован по различным вопросам вычислительной математики, что позволит читателю быстро войти в круг интересующих его проблем.

Вся книга имеет общую цель — подготовить читателя к решению сложных задач вычислительной и прикладной математики.

При изучении предлагаемой книги автор рекомендует читателю использовать задачник В. И. Дробышевича, В. П. Дымникова, Г. С. Ривина «Задачи по вычислительной математике» (М.: Наука, 1980), который соответствует изложению большей части материала книги.

В заключение следует отметить, что в книге широко использованы новые результаты исследований советских и зарубежных авторов, и в первую очередь сотрудников Института вычислительной математики РАН, Института математической геофизики и вычислительной математики СО РАН и Института вычислительного моделирования СО РАН. Это прежде всего результаты, полученные В. И. Агошковым, В. А. Булавским, А. Л. Бухгеймом, В. А. Василенко, В. П. Ильиным, Ю. А. Кузнецовым, В. И. Лебедевым, А. М. Мацокиным, И. В. Оселедцем, В. В. Смеловым, Е. Е. Тыртышниковым, В. А. Цецохо, В. В. Шайдуровым, В. П. Шутяевым и др. Раздел книги по вариационным неравенствам был написан по материалам, любезно представленным автору французскими математиками Ж.-Л. Лионсом и Р. Гловинским.

Глава 1.

Общие сведения из теории разностных схем

В настоящей главе приводятся краткие сведения по фундаментальным вопросам теории разностных схем, которые существенно использованы в последующих главах книги. Поскольку нашей основной задачей является знакомство с некоторыми современными принципами построения вычислительных алгоритмов для решения задач математической физики, то при рассмотрении вопросов теории мы ограничимся только наиболее простыми случаями.

В данной главе мы также приведем конечно-разностные аппроксимации некоторых простейших (но вместе с тем широко распространенных) краевых задач.

1.1. Основные понятия и определения

Рассмотрим в n-мерном евклидовом пространстве $\mathbf{R}^n$ некоторую область D. Обозначим через $L_2(D)$ гильбертово пространство всех вещественных измеримых функций $f(x)$, суммируемых с квадратом, т. е. таких, что

$$\int\limits_D f^2(x)\,dx < \infty,$$

со скалярным произведением

$$(f,g) = \int\limits_D f(x)g(x)\,dx. \tag{1.1.1}$$

Как обычно, норму функции f из $L_2(D)$ определим равенством

$$\|f\| = (f, f)^{1/2}. \tag{1.1.2}$$

Выделим теперь из гильбертова пространства $L_2(D)$ некоторое множество элементов $\Phi \subset L_2(D)$, содержащее вместе с φ и ψ их линейную комбинацию $\alpha\varphi + \beta\psi$, где $\alpha, \beta \in \mathbf{R}$, причем возможно, что каждый элемент $\varphi \in \Phi$ удовлетворяет некоторым дополнительным условиям. Такими условиями в зависимости от рассматриваемой задачи могут быть, например, требования заданной гладкости, удовлетворение предельным соотношениям на границе области D и т. д. Указанные условия, однако, должны быть достаточными для того, чтобы оператор A задачи переводил элемент $\varphi \in \Phi$ в элемент $A\varphi \in L_2(D)$.

Линейный оператор A, определенный на линейном многообразии Φ, называется *положительно полуопределенным*, если для всех $\varphi \in \Phi$ выполняется неравенство

$$(A\varphi, \varphi) \geq 0, \tag{1.1.3}$$

причем равенство нулю скалярного произведения $(A\varphi, \varphi)$ допускается на элементе φ, тождественно не равном нулю. Обычно это записывается так: $A \geq 0$. Если равенство исключается, т. е.

$$(A\varphi, \varphi) > 0 \tag{1.1.4}$$

при $\varphi \neq 0$, то оператор A называют *положительным* и пишут $A > 0$.

Наконец, в случае более сильного неравенства

$$(A\varphi, \varphi) \geq \gamma(\varphi, \varphi), \tag{1.1.5}$$

где $\gamma > 0$ — некоторая постоянная, общая для всех $\varphi \in \Phi$, оператор A называют *положительно определенным*.

Заметим, что если оператором A является квадратная матрица конечного порядка, то для нее из условия положительности следует положительная определенность.

Множество Φ будем называть *множеством определения оператора* A и обозначать его $\Phi(A)$.

Введем, далее, в рассмотрение *сопряженный оператор* A^* с помощью тождества Лагранжа

$$(Ag, h) = (g, A^*h). \tag{1.1.6}$$

Множества $\Phi(A)$ и $\Phi(A^*)$ гильбертова пространства $L_2(D)$, вообще говоря, не совпадают друг с другом, хотя функции, являющиеся их элементами, определены на одной и той же области D.

В том случае, когда $Ah = A^*h$ при всех $h \in \Phi(A)$ и $\Phi(A) \equiv \Phi(A^*)$, оператор A называют *самосопряженным*.

Отметим одно важное следствие, связанное со свойствами сопряженных операторов, а именно: если $\Phi(A) \equiv \Phi(A^*)$, то из условия $A > 0$ следует $A^* > 0$.

Для анализа алгоритмов большое значение имеют разложения функций в ряды Фурье по собственным функциям основных и сопряженных операторов.

Пусть $A \geq 0$. Рассмотрим две спектральные задачи

$$Au = \lambda(A)u, \quad A^*u^* = \lambda(A^*)u^*. \tag{1.1.7}$$

Предположим, что собственные функции $\{u_n\}$ и $\{u_m^*\}$ этих задач образуют полную систему и нормированы следующим образом:

$$(u_n, u_m^*) = \begin{cases} 1, & n = m, \\ 0, & n \neq m, \end{cases} \tag{1.1.8}$$

а соответствующие собственные числа $\lambda_n(A)$ и $\lambda_m(A^*)$ вещественны. В этом случае, как известно, $\lambda_n(A) = \lambda_n(A^*)$. Пусть собственные числа спектральных задач принадлежат отрезку $[a, b]$:

$$\alpha \leq \lambda_n(A) \leq \beta.$$

Эту полную систему функций будем называть *биортонормированной*. Тогда в предположении полноты системы u_n и u_m^* любые функции f из Φ и f^* из Φ^* (где $\Phi = \Phi(A), \Phi^* = \Phi(A^*)$) могут быть представлены в виде рядов Фурье:

$$f = \sum_n f_n u_n, \; f^* = \sum_n f_n^* u_n^*, \tag{1.1.9}$$

где

$$f_n = (f, u_n^*),\ f_n^* = (f^*, u_n). \tag{1.1.10}$$

Важное значение для анализа вычислительных алгоритмов имеют оценки норм операторов. *Норму оператора* A определим следующим образом:

$$\|A\|^2 = \sup_{\substack{\varphi\in\Phi \\ \varphi\neq 0}} \frac{(A\varphi, A\varphi)}{(\varphi, \varphi)} \tag{1.1.11}$$

(в дальнейшем для простоты записи ограничение $\varphi \neq 0$ указываться не будет). Принимая во внимание соотношение (конечно, при условии, что $A\varphi \in \Phi(A^*)$)

$$(A\varphi, A\varphi) = (\varphi, A^*A\varphi),$$

формулу (1.1.11) можно записать в виде

$$\|A\|^2 = \sup_{\varphi\in\Phi} \frac{(\varphi, A^*A\varphi)}{(\varphi, \varphi)}. \tag{1.1.12}$$

Оператор A^*A — симметричный и положительно полуопределенный. Рассмотрим спектральную задачу

$$A^*A\Omega = \lambda(A^*A)\Omega, \tag{1.1.13}$$

которая определяет систему ортонормированных собственных функций $\{\Omega_n\}$ и собственные числа $\lambda_n(A^*A) \geq 0$. Будем предполагать, что система $\{\Omega_n\}$ полна. Представим функцию φ в виде ряда Фурье

$$\varphi = \sum_n \varphi_n \Omega_n, \tag{1.1.14}$$

где

$$\varphi_n = (\varphi, \Omega_n). \tag{1.1.15}$$

Подставив ряд (1.1.14) в (1.1.12) и используя условия ортонормировки функций Ω_n, получим

$$\|A\|^2 = \sup_{\{\varphi_n\}\in Q} \frac{\sum\limits_n \lambda_n(A^*A)\varphi_n^2}{\sum\limits_n \varphi_n^2}, \tag{1.1.16}$$

где Q — пространство коэффициентов Фурье. Нетрудно убедиться, что

$$\frac{1}{\|A^{-1}\|^2} = \lambda_{min}(A^*A),\ \|A\|^2 = \lambda_{max}(A^*A), \tag{1.1.17}$$

где через λ_{min} и λ_{max} обозначены соответственно точная нижняя и точная верхняя грани множества собственных чисел $\{\lambda_n(A^*A)\}$ спектральной задачи (1.1.13). Величину $\beta(A^*A) = \lambda_{max}(A^*A)$ обычно называют спектральным радиусом оператора A^*A.

Спектральным радиусом оператора A называют $\beta(A) = \sup\{|\lambda(A)|\}$.

Пусть A — самосопряженный оператор. Рассмотрим спектральную задачу

$$Au = \lambda u. \tag{1.1.18}$$

Справедливо равенство

$$\|A\| = \beta(A). \tag{1.1.19}$$

Нетрудно видеть, что для самосопряженного оператора имеет место соотношение

$$\beta(A^2) = [\beta(A)]^2. \tag{1.1.20}$$

Пусть в гильбертовом пространстве $L_2(D)$ задан некоторый оператор C. Предположим, что его область определения $\Phi = \Phi(C)$ *всюду плотна* в $L_2(D)$, т. е. для любого элемента $f \in L_2(D)$ найдется такой элемент $g \in \Phi$, что будет выполняться соотношение $\|f - g\| \le \varepsilon$, где ε — произвольно малая положительная константа. Если оператор C — положителен, то

$$(C\varphi, \varphi) > 0 \tag{1.1.21}$$

для всех $\varphi \neq 0$ из Φ. Пусть Φ^* — область определения сопряженного оператора C^* — совпадает с Φ и, таким образом, для всех $\varphi \in \Phi$ существует $C^*\varphi$. Тогда

$$(C^*\varphi, \varphi) = (\varphi, C\varphi) = (\overline{C}\varphi, \varphi),\ \text{где}\ \overline{C} = \frac{1}{2}(C + C^*).$$

Оператор $\overline{C}$ является симметричным положительным оператором, что позволяет ввести в Φ новое скалярное произведение

$$(f, g)_{\overline{C}} = (\overline{C}f, g)$$

и норму

$$\|\varphi\|_C^2 = (C\varphi, \varphi) = (\overline{C}\varphi, \varphi).$$

Эту норму будем называть *энергетической* или C*-нормой*. Можно получить следующую важную оценку:

$$\|\varphi\|_C^2 = \|\varphi\|_{\overline{C}}^2 \leq \|\overline{C}\|\|\varphi\|^2 = \beta(\overline{C})\|\varphi\|^2, \tag{1.1.22}$$

где $\beta(\overline{C})$ — максимальное собственное число оператора $\overline{C}$.

В заключение отметим, что при рассмотрении основных и сопряженных задач математической физики бывает удобно пользоваться функциями *пространства Соболева* $W_2^l(D)$. Это пространство состоит из функций пространства $L_2(D)$, которые имеют в D суммируемые с квадратом обобщенные производные до порядка l включительно, и является гильбертовым. Скалярное произведение в пространстве W_2^l определяется следующей формулой:

$$(u, v)_{W_2^l} = \int\limits_D \sum_{|k|=0}^{l} \sum_{(k)} \frac{\partial^{|k|} u}{\partial x_1^{\alpha_1} \dots \partial x_n^{\alpha_n}} \frac{\partial^{|k|} v}{\partial x_1^{\alpha_1} \dots \partial x_n^{\alpha_n}}\, dD. \tag{1.1.23}$$

Здесь $|k| = \alpha_1 + \ldots + \alpha_n$, а символ $\sum\limits_{(k)}$ означает суммирование по всем производным порядка $|k|$. Норма в пространстве $W_2^l(D)$ определяется соотношением

$$\|\varphi\|_{W_2^l} = (\varphi, \varphi)_{W_2^l}^{1/2}. \tag{1.1.24}$$

Если функции φ принадлежат пространству Соболева $W_2^1(D)$ и, кроме того, удовлетворяют условию

$$\varphi = 0 \text{ на границе } \partial D \text{ области } D,$$

то такое пространство будем обозначать через $\overset{\circ}{W}{}_2^1(D)$.

Пусть D — ограниченная область из $\mathbf{R}^n$. Тогда для любой функции $\varphi(x) \equiv \varphi(x_1, \dots, x_n) \in \overset{\circ}{W}{}_2^1(D)$ справедливо неравенство Стеклова

$$\int\limits_D |\varphi|^2\, dx \leq C_0^2 \sum_{i=1}^{N} \int\limits_D \left|\frac{\partial \varphi}{\partial x_i}\right|^2 dx, \tag{1.1.25}$$

где C_0 есть постоянная, зависящая лишь от области D (заметим, что $C_0 \leq \mathrm{const}(\mathrm{mesD})^{1/n}$). С помощью этого неравенства несложно показать,

что в $\overset{\circ}{W}{}_2^1(D)$ можно ввести новое скалярное произведение

$$[\varphi,\psi]=\sum_{i=1}^{N}\int\limits_D \frac{\partial\varphi}{\partial x_i}\frac{\partial\psi}{\partial x_i}\,dx, \tag{1.1.26}$$

которое порождает норму $[\varphi]=[\varphi,\varphi]^{1/2}$, эквивалентную исходной норме $\|\varphi\|_{W_2^1(D)}$ при $\varphi\in\overset{\circ}{W}{}_2^1(D)$.

Для широкого класса областей D и функций $\varphi\in W_2^1(D)$ справедливо неравенство Пуанкаре

$$\int\limits_D |\varphi|^2\,dx\le C_1^2\left(\left(\int\limits_D\varphi\,dx\right)^2+\sum_{i=1}^{N}\int\limits_D\left|\frac{\partial\varphi}{\partial x_i}\right|^2\,dx\right), \tag{1.1.27}$$

где $C_1=\mathrm{const}<\infty$. Это неравенство позволяет уже при рассмотрении всего пространства $W_2^1(D)$ ввести норму

$$[\varphi]_1=\left(\left(\int\limits_D\varphi\,dx\right)^2+\sum_{i=1}^{N}\int\limits_D\left|\frac{\partial\varphi}{\partial x_i}\right|^2\,dx\right)^{1/2}, \tag{1.1.28}$$

эквивалентную норме $\|\varphi\|_{W_2^1(D)}$.

Пространства $W_2^1(D),\overset{\circ}{W}{}_2^1(D)$ играют основную роль при изучении краевых задач для уравнений второго порядка различных типов.

1.1.1. Оценки норм некоторых матриц

Рассмотрим вещественную положительно полуопределенную матрицу $A\ge 0$, действующую на векторы из евклидова пространства $\mathbf{R}^n$. Имеет место следующее соотношение (E — единичная матрица):

$$\|(E+\sigma A)^{-1}\|\le 1 \tag{1.1.29}$$

для любых значений параметра $\sigma\ge 0$. Доказательство этого важного утверждения проведем с помощью формулы

$$\|(E+\sigma A)^{-1}\|^2=\sup_{\varphi}\frac{((E+\sigma A)^{-1}\varphi,(E+\sigma A)^{-1}\varphi)}{(\varphi,\varphi)}.$$

Положим

$$\psi = (E + \sigma A)^{-1}\varphi.$$

Тогда

$$\|(E + \sigma A)^{-1}\|^2 = \sup_{\psi} \frac{(\psi, \psi)}{((E + \sigma A)\psi, (E + \sigma A)\psi)} =$$
$$= \frac{1}{\inf\limits_{\psi} \left[1 + 2\sigma \frac{(A\psi, \psi)}{(\psi, \psi)} + \sigma^2 \frac{(A\psi, A\psi)}{(\psi, \psi)}\right]}.$$

Поскольку $A \geq 0$, то из последнего соотношения следует оценка (1.1.29).

Если $A > 0$, то при $\sigma > 0$ выполняется неравенство

$$\|(E + \sigma A)^{-1}\| < 1.$$

Лемма (Келлог). *Если* $A \geq 0$ *и* $\sigma \geq 0$, *то*

$$\|(E - \sigma A)(E + \sigma A)^{-1}\| \leq 1.$$

В самом деле, введем обозначение

$$T = (E - \sigma A)(E + \sigma A)^{-1}$$

и рассмотрим выражение для $\|T\|^2$. Имеем

$$\|T\|^2 = \sup_{\varphi} \frac{((E - \sigma A)(E + \sigma A)^{-1}\varphi, (E - \sigma A)(E + \sigma A)^{-1}\varphi)}{(\varphi, \varphi)} =$$
$$= \sup_{\psi} \frac{((E - \sigma A)\psi, (E - \sigma A)\psi)}{((E + \sigma A)\psi, (E + \sigma A)\psi)} = \sup_{\psi} \frac{(\psi, \psi) - 2\sigma(A\psi, \psi) + \sigma^2(A\psi, A\psi)}{(\psi, \psi) + 2\sigma(A\psi, \psi) + \sigma^2(A\psi, A\psi)} \leq 1.$$

Здесь существенно использовано свойство положительной полуопределенности матрицы A.

В случае когда матрица A положительна и $\sigma > 0$, получим

$$\|(E - \sigma A)(E + \sigma A)^{-1}\| < 1.$$

1.1.2. Вычисление границ спектра положительной матрицы

Рассмотрим задачу на отыскание максимального и минимального собственных чисел матрицы $A > 0$, имеющей положительный спектр. С этой целью воспользуемся методом Люстерника [4] (часто этот метод называют степенным).

Предположим, что спектральная задача $Au = \lambda u$ определяет полную систему собственных векторов $u_k \in \Phi$ и набор положительных собственных чисел $\lambda_k(A)$. Рассмотрим итерационный процесс

$$\varphi^{(n+1)} = \frac{1}{c_n} A\varphi^{(n)}, \tag{1.1.30}$$

$$\varphi^{(0)} = g,$$

где g — произвольный ненулевой вектор, а c_n — нормировочный множитель, который удобно выбрать в виде

$$c_n = \|\varphi^{(n)}\|.$$

Тогда имеем

$$\varphi^{(n+1)} = A\frac{\varphi^{(n)}}{\|\varphi^{(n)}\|}. \tag{1.1.31}$$

Пусть

$$0 < \alpha(A) = \lambda_1 \le \lambda_2 \le \cdots \le \lambda_{m-1} \le \lambda_m = \beta(A).$$

Справедливо соотношение

$$\beta(A) = \lim_{n\to\infty} \|\varphi^{(n)}\|. \tag{1.1.32}$$

В самом деле, вследствие полноты системы векторов u_n имеет место разложение

$$\varphi^{(0)} = \sum_k g_k u_k,$$

где $g_k = (g, u_k^*)$, а $\{u_k^*\}$ — собственные векторы матрицы A^*. С учетом рекуррентного соотношения

$$\varphi^{(n)} = A\frac{\varphi^{(n-1)}}{\|\varphi^{(n-1)}\|} = \cdots = \frac{A^n g}{\|A^{n-1} g\|}$$

имеем

$$\lim_{n\to\infty}\|\varphi^{(n)}\| = \lim_{n\to\infty}\frac{\|A^n g\|}{\|A^{n-1}g\|}.$$

Поскольку

$$A^n g = \sum_k [\lambda_k(A)]^n g_k u_k,$$

то при достаточно больших n получаем

$$A^n g = \beta^n(A)\left\{g_m u_m + O\left[\left(\frac{\lambda_{m-1}}{\lambda_m}\right)^n\right]\right\},$$

где $\beta(A) = \lambda_m$ — минимальное собственное число матрицы A.

Если в качестве начального приближения g случайно выбран вектор, являющийся линейной комбинацией собственных векторов, соответствующих собственным числам, отличным от $\beta(A)$, то процесс последовательных приближений и в этом случае, как правило, позволяет получить $\beta(A)$ за счет появления (из-за ошибок округления) всех компонент базисных векторов в разложении $\varphi^{(n)}$.

Из последнего соотношения имеем

$$\frac{\|A^n g\|}{\|A^{n-1}g\|} = \beta(A) + O\left[\left(\frac{\lambda_{m-1}}{\lambda_m}\right)^n\right]$$

и, следовательно,

$$\beta(A) = \lim_{n\to\infty}\frac{\|A^n g\|}{\|A^{n-1}g\|}. \tag{1.1.33}$$

Перейдем теперь к вычислению минимального собственного числа матрицы A, предполагая, что $\lambda_1 < \lambda_2$. Рассмотрим новую матрицу[1)]

$$B = \beta(A)E - A \tag{1.1.34}$$

и спектральную задачу

$$Bu = \lambda(B)u. \tag{1.1.35}$$

Очевидно, что $B \geq 0$. Из формулы (1.1.34) вытекает, что матрицы A и B имеют общий базис $\{u_k\}$. Аналогично предыдущему рассмотрим итерационный процесс

$$\psi^{(n+1)} = B\frac{\psi^{(n)}}{\|\psi^{(n)}\|}. \tag{1.1.36}$$

[1)]Заметим, что вместо $\beta(A)$ можно взять любое число, большее $\beta(A)$.

В результате получим, что

$$\beta(B) = \lim_{n\to\infty} \|\psi^{(n)}\|. \tag{1.1.37}$$

Заметим, что из (1.1.34) и общности базисов матриц A и B следует соотношение

$$\beta(B) = \beta(A) - \alpha(A).$$

Отсюда

$$\alpha(A) = \beta(A) - \beta(B). \tag{1.1.38}$$

Таким образом, для матриц указанного вида минимальное собственное чисто находится в виде разности максимальных собственных чисел соответственно матриц A и $B = \beta(A)E - A$. Поскольку максимальные собственные числа $\beta(A)$ и $\beta(B)$ получаются с помощью приведенного выше итерационного процесса, то задача нахождения минимального собственного числа матрицы A в принципе решена.

Однако следует отметить, что в случае плохо обусловленных матриц A минимальное собственное число $\alpha(A)$ определяется как разность больших чисел $\beta(A)$ и $\beta(B)$. Поэтому при реализации этого алгоритма возможны ошибки не только в величине $\alpha(A)$, но и даже в знаке. Чтобы избежать таких ошибок, несколько изменим процесс нахождения $\alpha(A)$. Вновь рассмотрим итерационный процесс

$$\psi^{(n+1)} = B\frac{\psi^{(n)}}{\|\psi^{(n)}\|}, \quad \psi^{(0)} = h.$$

Заметим, что упорядоченная по возрастанию собственных чисел система собственных векторов u_k матрицы A переходит в упорядоченную систему собственных векторов v_k матрицы B, так что $v_k = u_{m-k+1}$ $(k = 1, 2, \ldots, m)$. Рассмотрим выражения

$$\psi^{(n)} = \frac{B\psi^{(n-1)}}{\|\psi^{(n-1)}\|} = \cdots = \frac{B^n\psi^{(0)}}{\|B^{n-1}\psi^{(0)}\|} = \frac{\sum\limits_{k=1}^{m} \lambda_k^n(B)h_k v_k}{\|\sum\limits_{k=1}^{m} \lambda_k^{n-1}(B)h_k v_k\|},$$

$$A\psi^{(n)} = \frac{\sum\limits_{k=1}^{m} \lambda_k^n(B)h_k Av_k}{\|\sum\limits_{k=1}^{m} \lambda_k^{n-1}(B)h_k v_k\|},$$

где $h_k = (h, v_k^*)$. Переходя к пределу в предыдущих равенствах при $n \to$ $\to \infty$, имеем

$$\lim_{n\to\infty} \psi^{(n)} = \beta(B)\frac{h_m v_m}{\|h_m v_m\|},\ \lim_{n\to\infty} A\psi^{(n)} = \beta(B)\frac{h_m A v_m}{\|h_m v_m\|}.$$

Поскольку $Av_m = Au_1 = \alpha(A)u_1 = \alpha(A)v_m$, то приходим к алгоритму

$$\psi^{(n+1)} = B\frac{\psi^{(n)}}{\|\psi^{(n)}\|}, \tag{1.1.39}$$

$$\alpha(A) = \lim_{n\to\infty} \frac{\|A\psi^{(n)}\|}{\|\psi^{(n)}\|}. \tag{1.1.40}$$

При использовании последней формулы, как правило, уже не приходится иметь дело с разностью больших чисел, а весь процесс нахождения границ спектра A, как правило, эффективно реализуется на ЭВМ.

Следует, однако, отметить, что итерационные процессы (1.1.31), (1.1.36), (1.1.39) сходятся медленно. Для ускорения сходимости можно применять различные методы, наиболее употребительными из которых являются чебышевское ускорение или методы сдвига спектра.

Заметим, что для симметричных матриц при вычислении $\alpha(A)$ и $\beta(A)$ лучше пользоваться энергетической нормой.

При оптимизации вычислительных процессов и при всевозможных теоретических оценках алгоритмов зачастую необходимо знание нормы оператора A и нормы обратного оператора A^{-1}. Имеют место следующие соотношения, справедливые для любых обратимых операторов:

$$\|A\|^2 = \sup_{\varphi\in\Phi} \frac{(A\varphi, A\varphi)}{(\varphi,\varphi)} = \sup_{\varphi\in\Phi} \frac{(A^*A\varphi,\varphi)}{(\varphi,\varphi)} = \beta(A^*A),$$

$$\|A^{-1}\|^2 = \sup_{\varphi\in\Phi} \frac{(A^{-1}\varphi, A^{-1}\varphi)}{(\varphi,\varphi)} = \sup_{\varphi\in\Phi} \frac{((AA^*)^{-1}\varphi,\varphi)}{(\varphi,\varphi)} = [\alpha(A^*A)]^{-1}.$$

Отсюда

$$\|A\| = \sqrt{\beta(A^*A)}, \tag{1.1.41}$$

$$\|A^{-1}\| = \left(\sqrt{\alpha(A^*A)}\right)^{-1}. \tag{1.1.42}$$

Величины $\alpha(A^*A)$ и $\beta(A^*A)$ вычисляются с помощью описанного метода последовательных приближений.

Отметим, что изложенный выше алгоритм вычисления границ спектра положительных матриц открывает возможности для оптимизации итерационных процессов решения задач математической физики на основе хорошо разработанных методов (они будут рассмотрены в гл. 4). Такие процессы становятся конструктивными и позволяют эффективно решать различные задачи математической физики.

Помимо метода Люстерника, полезную оценку границ спектра дает теорема Гершгорина[2)]: все собственные значения $\lambda(A)$ (вообще говоря, комплексные) произвольной матрицы A порядка m с элементами a_{kl} принадлежат объединению кругов

$$|z - a_{kk}| \leq R_k,\ k = 1, 2, \ldots, m,$$

где

$$R_k = \sum_{l=1, l \neq k}^{m} |a_{kl}|.$$

Заметим кстати, что одним из следствий теоремы Гершгорина является следующая оценка для спектрального радиуса матрицы:

$$\beta(A) \leq \max_k \sum_{l=1}^{m} |a_{kl}|.$$

В заключение обратим внимание на то, что в настоящее время в связи с возрастающей мощностью вычислительных машин при решении задач математической физики все чаще используется метод вычисления ряда компонент решения, соответствующих нескольким первым собственным числам спектральных задач

$$A\varphi = \lambda\varphi.$$

Сначала алгоритмом, изложенным выше, находится собственное число $\lambda_1 = \alpha$ и соответствующий ему собственный вектор u_1. Пусть для простоты матрица A симметричная и имеет простой спектр λ_k. Для вычисления λ_2 и u_2 воспользуемся следующим ите-

[2)] Полное доказательство теоремы Гершгорина можно найти, например, в книге Д. К. Фаддеева и В. Н. Фаддеевой «Вычислительные методы линейной алгебры» [8].

рационным процессом:

$$\psi^{(n+1)} = B\frac{\psi^{(n)}}{\|\psi^{(n)}\|}, \quad B = \beta(A)E - A,$$

предполагая, что ни нулевое приближение $\varphi^{(0)} = h$, ни последующие приближения $\varphi^{(n)}$ не содержат компонент, соответствующих собственному вектору u_1. Для того чтобы это предположение было выполнено, необходимо предусмотреть в итерационном процессе (1.1.39) ортогонализацию по отношению к первому собственному вектору в форме

$$\overline{\psi}^{(n)} = \psi^{(n)} - a_1 u_1, \tag{1.1.43}$$

где a_1 — константа, которая выбирается из условия ортогональности $\overline{\psi}^{(n)}$ и u_1:

$$(\overline{\psi}^{(n)}, u_1) = 0. \tag{1.1.44}$$

Умножая (1.1.43) скалярно на u_1 и учитывая (1.1.44), получим

$$(\overline{\psi}^{(n)}, u_1) = (\psi^{(n)}, u_1) - a_1(u_1, u_1) = 0.$$

Отсюда

$$a_1 = \frac{(\psi^{(n)}, u_1)}{(u_1, u_1)}. \tag{1.1.45}$$

В результате приходим к алгоритму

$$\begin{aligned} \overline{\psi}^{(n)} &= \psi^{(n)} - a_1 u_1, \\ \psi^{(n+1)} &= B\frac{\overline{\psi}^{(n)}}{\|\overline{\psi}^{(n)}\|}. \end{aligned} \tag{1.1.46}$$

При $n \to \infty$ предельным элементом в процессе (1.1.46) является u_2, а соответствующее ему собственное число λ_2 вычисляется по формуле

$$\lambda_2 = \lim_{n\to\infty} \frac{\|A\overline{\psi}^{(n)}\|}{\|\overline{\psi}^{(n)}\|}.$$

Аналогичным образом вычисляем последующие векторы и соответствующие им собственные числа.

В самом деле, предположим, что уже найдены k первых собственных векторов $u_1, u_2, \ldots, u_k$ и требуется найти вектор u_{k+1}. Для это-

го воспользуемся следующим итерационным процессом:

$$\overline{\psi}^{(n)} = \psi^{(n)} - a_1 u_1 - a_2 u_2 - \ldots - a_k u_k,$$

$$\overline{\psi}^{(n+1)} = B \frac{\overline{\psi}^{(n)}}{\|\overline{\psi}^{(n)}\|},$$

а $a_1, a_2, \ldots, a_k$ вычисляются из условия ортогональности $\overline{\psi}^{(n)}$ ко всем векторам $u_1, u_2, \ldots, u_k$. В результате приходим к соотношениям

$$\begin{aligned} a_1(u_1, u_1) &= (\psi^{(n)}, u_1), \\ a_2(u_2, u_2) &= (\psi^{(n)}, u_2), \\ &\vdots \\ a_k(u_k, u_k) &= (\psi^{(n)}, u_k). \end{aligned}$$

Отсюда следует

$$a_1 = \frac{(\psi^{(n)}, u_1)}{(u_1, u_1)}, \quad a_2 = \frac{(\psi^{(n)}, u_2)}{(u_2, u_2)}, \quad \ldots, \quad a_k = \frac{(\psi^{(n)}, u_k)}{(u_k, u_k)}.$$

Предельным при $n \to \infty$ элементом построенного процесса будет элемент

$$u_{k+1} = \lim_{n \to \infty} \overline{\psi}^{(n)},$$

а для соответствующего ему собственного числа λ_{k+1} справедлива формула

$$\lambda_{k+1} = \lim_{n \to \infty} \frac{\|A\overline{\psi}^{(n)}\|}{\|\overline{\psi}^{(n)}\|}.$$

Мы здесь лишь отметим, что изложенный метод последовательной ортогонализации можно применять не на каждом шаге итерационного процесса, а через сравнительно большое число шагов, чтобы ошибки округления, соответствующие «подавляемым» векторам, не сильно возросли. Желаемым условием, разумеется, является ортогонализация «начальных» приближений $\psi^{(0)}$ в указанных процессах.

Изложенный метод иллюстрирует принципиальную схему подхода к отысканию первых собственных чисел и соответствующих им собственных векторов для симметричных матриц с простым спектром.

Предположим теперь, что матрица A несимметрична. Это значит, что мы одновременно имеем дело с двумя спектральными задачами:

$$A\varphi = \lambda\varphi, \quad A^*\varphi^* = \lambda\varphi^*. \tag{1.1.47}$$

По-прежнему будем предполагать, что собственные векторы и спектр задач (1.1.47) вещественны и собственные числа простые. Вследствие биортогональности базисов $\{\varphi_n\}$ и $\{\varphi_n^*\}$, т. е.

$$(\varphi_l, \varphi_k^*) = 0, \text{ если } l \neq k,$$

алгоритм построения первых собственных элементов выглядит следующим образом. Сначала находим первые собственные элементы

$$u_1 = \lim_{n\to\infty} B\frac{\psi^{(n)}}{\|\psi^{(n)}\|}, \quad u_1^* = \lim_{n\to\infty}, B^*\frac{\psi^{*(n)}}{\|\psi^{*(n)}\|}$$

$$B = \beta(A)E - A, \quad B^* = \beta(A^*)E - A^*$$

и минимальное собственное число

$$\lambda_1 = \lim_{n\to\infty}\frac{\|A\psi^{(n)}\|}{\|\psi^{(n)}\|} = \lim_{n\to\infty}\frac{\|A^*\psi^{*(n)}\|}{\|\psi^{*(n)}\|}.$$

Затем последовательно находим другие собственные элементы и соответствующие им собственные числа. С этой целью используется следующий итерационный процесс:

$$\psi^{(n+1)} = B\frac{\overline{\psi}^{(n)}}{\|\overline{\psi}^{(n)}\|}, \quad \psi^{*(n+1)} = B^*\frac{\overline{\psi}^{*(n)}}{\|\overline{\psi}^{*(n)}\|},$$

где

$$\overline{\psi}^{(n)} = \psi^{(n)} - a_1u_1 - a_2u_2 - \ldots - a_ku_k,$$

$$\overline{\psi}^{*(n)} = \psi^{*(n)} - b_1u_1^* - b_2u_2^* - \ldots - b_ku_k^*.$$

Константы $a_1, \ldots, a_k$ и $b_1, \ldots, b_k$ вычисляются из условий ортогональности

$$\begin{matrix}(\overline{\psi}^{(n)}, u_i^*) = 0, \\ (\overline{\psi}^{(n)}, u_i) = 0,\end{matrix} \quad i = 1, 2, \ldots, k.$$

С учетом свойства биортогональности приходим к соотношениям

$$\begin{matrix}(\psi^{(n)}, u_i^*) - a_i(u_i, u_i^*) = 0, \\ (\psi^{*(n)}, u_i^*) - b_i(u_i, u_i^*) = 0,\end{matrix} \quad i = 1, 2, \ldots, k.$$

В результате константы a_i, b_i определяются из равенств

$$a_i = \frac{(\psi^{(n)}, u_i^*)}{(u_i, u_i^*)}, \quad b_i = \frac{(\psi^{*(n)}, u_i)}{(u_i, u_i^*)}, \quad i = 1, 2, \ldots, k.$$

Необходимо иметь в виду, что в настоящее время формулируется и другая, более мощная система алгоритмов решения полной проблемы собственных чисел на основе QR-алгоритма. В результате на базе QR-алгоритма появился ряд первоклассных программ и процедур по нахождению собственных элементов и собственных чисел задач линейной алгебры. Они успешно конкурируют с итерационными процессами решения спектральных задач и, по-видимому, в дальнейшем станут основными для решения широких классов задач, в том числе и для матриц высокого порядка. Эффективность этих алгоритмов для больших матриц существенно зависит от мощности ЭВМ, которая непрерывно возрастает.

В порядке обсуждения тенденций вычислительной математики можно высказать предположение, что в ближайшем будущем на основе решения полной проблемы собственных чисел снова приобретает былое могущество классический метод Фурье в применении к тем проблемам математической физики, которые редуцируются к началам линейной алгебры.

В самом деле, если предположить, что имеется задача линейной алгебры

$$A\varphi = f, \tag{1.1.48}$$

которую требуется решать многократно с одним и тем же оператором A при различных правых частях f, и учесть, что нам известны решения спектральных задач

$$Aw = \lambda w,$$

$$A^* w^* = \lambda w^*$$

в виде полных базисов $\{w_n\}$ и $\{w_n^*\}$, то решение задачи (1.1.48) и правую часть удобно представить в виде сумм Фурье

$$\begin{aligned} \varphi &= \sum_{n=1}^{m} \varphi_n w_n, \\ f &= \sum_{n=1}^{m} f_n w_n, \end{aligned} \tag{1.1.49}$$

где

$$f_n = (f, w_n^*).$$

Подставляя формулы (1.1.49) в (1.1.48) и умножая результат скалярно на w_n^*, приходим к алгебраической системе

$$\lambda_n \varphi_n = f_n, \ n = 1, 2, \ldots, m,$$

из которой следует (при $\lambda \neq 0$), что

$$\varphi_n = \frac{f_n}{\lambda_n}.$$

Таким образом, решение задачи (1.1.48) находится в виде

$$\varphi = \sum_{n=1}^{m} \frac{f_n}{\lambda_n} w_n. \tag{1.1.50}$$

Метод Фурье в случае, когда мы располагаем лишь набором нескольких (например, $k < m$) собственных векторов, позволяет представить решение задачи (1.1.48) в виде

$$\varphi = \sum_{n=1}^{k} \frac{f_n}{\lambda_n} w_n + \xi. \tag{1.1.51}$$

Здесь ξ удовлетворяет уравнению

$$A\xi = \eta, \tag{1.1.52}$$

где

$$\eta = f - \sum_{n=1}^{k} f_n w_n.$$

На первый взгляд кажется, что задача (1.1.48) свелась к аналогичной задаче (1.1.52). Формально это так. Но если порядок матрицы

A настолько велик, что для решения задач (1.1.48) и (1.1.52) можно использовать только итерационные методы (см. гл. 4), то оказывается, что эти методы применительно к задаче (1.1.52) являются значительно более эффективными в реализации на ЭВМ. (По сравнению с тем, если бы их применили к решению (1.1.48).) Именно в этом и состоит важность алгоритма вычисления первых собственных элементов спектральной задачи, описанного в настоящем параграфе. В дальнейшем эти идеи неоднократно будут нами использованы при построении вычислительных алгоритмов.

1.1.3. Собственные числа и функции оператора Лапласа

Пусть

$$A = -\Delta, \tag{1.1.53}$$

где $\Delta = \dfrac{\partial^2}{\partial x^2} + \dfrac{\partial^2}{\partial y^2}$ — оператор Лапласа. Оператор $A = -\Delta$ определим на множестве Φ, элементами которого являются функции $\varphi(x, y) \in W_2^2(D)$, удовлетворяющие условию

$$\varphi = 0 \text{ на } \partial D, \tag{1.1.54}$$

где ∂D — граница $D \subset \mathbf{R}^2$. Оператор A будем считать действующим в $L_2(D)$ (т. е. и область его определения $\Phi(A)$, и область его значений $R(A)$ рассматриваются как множества из пространства $L_2(D)$). Покажем, что оператор A симметричен. Рассмотрим некоторую функцию $\varphi^* \in L_2(D)$ и функционал

$$(A\varphi, \varphi^*) = -\int\limits_D \varphi^* \Delta\varphi \, dD. \tag{1.1.55}$$

Необходимо отметить, что условия на функции φ обеспечивают ограниченность функционала $(A\varphi, \varphi^*)$ для любых φ^*. Предположим теперь, что φ^* принадлежит $W_2^2(D)$, и с помощью второй формулы Грина получим, что

$$(A\varphi, \varphi^*) = -\int\limits_{\partial D} \left(\varphi^* \frac{\partial \varphi}{\partial n} - \varphi \frac{\partial \varphi^*}{\partial n} \right) \, ds - \int\limits_D \varphi \Delta \varphi^* \, dD, \tag{1.1.56}$$

где n — внешняя нормаль по отношению к D. Если функция φ^* удовлетворяет граничному условию

$$\varphi^* = 0 \text{ на } \partial D, \tag{1.1.57}$$

то с учетом условий (1.1.54), (1.1.57) получим

$$(A\varphi, \varphi^*) = -\int\limits_D \varphi \Delta \varphi^* \, dD = (\varphi, A\varphi^*). \tag{1.1.58}$$

Это значит, что исследуемый оператор A является симметричным.

Изучим далее вопрос о знакоопределенности A. С этой целью рассмотрим функционал

$$(A\varphi, \varphi) = -\int\limits_D \varphi \Delta \varphi \, dD. \tag{1.1.59}$$

С помощью первой формулы Грина получим

$$(A\varphi, \varphi) = -\int\limits_{\partial D} \varphi \frac{\partial \varphi}{\partial n} \, ds + \int\limits_D \left[\left(\frac{\partial \varphi}{\partial x} \right)^2 + \left(\frac{\partial \varphi}{\partial y} \right)^2 \right] \, dD. \tag{1.1.60}$$

Поскольку φ удовлетворяет условию (1.1.54), то

$$(A\varphi, \varphi) = \int\limits_D \left[\left(\frac{\partial \varphi}{\partial x} \right)^2 + \left(\frac{\partial \varphi}{\partial y} \right)^2 \right] \, dD > 0 \tag{1.1.61}$$

для любой не равной тождественно нулю функции $\varphi \in \Phi$. Более того, на основе (1.1.25) заключаем, что оператор A положительно определенный.

Проиллюстрируем на этом примере проблему собственных чисел для случая $D = \{(x, y) : 0 < x < 1, \ 0 < y < 1\}$. Известно, что ортонормированная система собственных функций задачи

$$Au = \lambda u \text{ в } D \tag{1.1.62}$$

при условии

$$u = 0 \text{ на } \partial D \tag{1.1.63}$$

будет полной. Она имеет вид

$$u_{mp} = 2\sin(m\pi x)\sin(p\pi y), \tag{1.1.64}$$

где $m = 1, 2, \ldots$ и $p = 1, 2, \ldots$. При этом собственные числа оператора A имеют вид

$$\lambda_{mp}(A) = (m^2 + p^2)\pi^2 > 0. \tag{1.1.65}$$

Отсюда следует, что

$$2\pi^2 \leq \lambda_{mp}(A).$$

Таким образом,

$$\alpha(A) = 2\pi^2, \quad \beta(A) = \infty \tag{1.1.66}$$

и $(A\varphi, \varphi) \geq 2\pi^2(\varphi, \varphi)$. Следовательно, оператор A является положительно определенным и неограниченным. Поскольку система собственных функций $\{u_{mp}\}$ полна, то любую функцию из Φ можно представить в виде ряда Фурье

$$\varphi(x, y) = \sum_{m,\,p} \varphi_{mp} u_{mp}(x, y) = \sum_{i} \varphi_i u_i(x, y), \tag{1.1.67}$$

а так как, кроме того, система $\{u_{mp}\}$ ортонормирована, то

$$\varphi_i = (\varphi, u_i), \tag{1.1.68}$$

где i — новый индекс упорядоченного ряда.

Заметим, что в случае областей более сложной формы и операторов с переменными коэффициентами явное вычисление собственных значений становится уже невозможным и приходится с помощью специальных методов оценивать границы спектра.

1.1.4. Сетки и сеточные функции. Собственные числа и векторы конечно-разностного аналога оператора Лапласа

Приведем несколько понятий из теории конечно-разностных методов.

Пусть D есть ограниченная область из $\mathbf{R}^n$ с границей ∂D. Пространство $\mathbf{R}^n$ разобьем на элементарные ячейки $\{x \equiv (x_1, x_2, \ldots, x_n) : k_i h_i < x_i < (k_i + 1)h_i,\ i = 1, 2, \ldots, n\}$, где k_i — целые числа, $h_i = \mathrm{const} > 0$. Обозначим через $\overline{D}_h$ объединение элементарных ячеек (вместе с их границами), таких, что они и их границы принадлежат $\overline{D} = D \cup \partial D$. Пусть ∂D_h есть граница области $\overline{D}_h$. Множество вершин ячеек, принадлежащих $\overline{D}_h$, обозначим также символом $\overline{D}_h$ и назовем его *сеткой*, введенной на $\overline{D}$, а сами вершины — *узлами сетки*. Множество узлов сетки, принадлежащих D, обозначим D_h, а сами такие узлы назовем *внутренними*. Множество $\overline{D}_h \backslash D_h$ составлено из *граничных* узлов сетки, и его для простоты также обозначим через ∂D_h. Величина h_i называется шагом сетки по переменной x_i. Если $h_i = \mathrm{const}$, то сетка по x_i называется *равномерной* (таким образом, мы здесь ограничиваемся рассмотрением равномерной по каждой из переменных сеткой).

В узлах сеток $D_h, \partial D_h, \overline{D}_h = D_h \cup \partial D_h$ можно определить некоторые функции. Функции, областью определения которых является сетка, будем называть *сеточными функциями* и обозначать через $\varphi^h, \psi^h, \ldots$. Значение сеточной функции φ^h в узле $(x_{k_1}, x_{k_2}, \ldots, x_{k_n})$, где $x_{k_i} = k_i \times h_i$, часто обозначают $\varphi^h_{k_1 \ldots k_n}$, т. е. $\varphi^h(x_{k_1}, x_{k_2}, \ldots, x_{k_n}) = \varphi^h_{k_1 \ldots k_n}$.

Множество сеточных функций φ^h, определенных на D_h, обозначим Φ_h. На этом множестве можно ввести скалярное произведение $(\varphi^h, \psi^h)_{\Phi_h}$ и норму $\|\varphi^h\|_{\Phi_h} = (\varphi^h, \varphi^h)^{1/2}_{\Phi_h}$, превратив тем самым Φ_h в конечномерное гильбертово пространство сеточных функций. Примером такого пространства является пространство $L_{2,h}$, состоящее из вещественных сеточных функций φ^h, определенных на D_h. Скалярное произведение и норму в $L_{2,h}$ можно задать в виде

$$(\varphi^h, \psi^h)_{L_{2,h}} = \sum_{k_1, k_2, \ldots, k_n} h_1 h_2 \ldots h_n \varphi^h_{k_1 \ldots k_n} \psi^h_{k_1 \ldots k_n},$$

$$\|\varphi^h\|_{L_{2,h}} = \left(\sum_{k_1, k_2, \ldots, k_n} h_1 h_2 \ldots h_n |\varphi^h_{k_1 \ldots k_n}|^2 \right)^{1/2},$$

где суммирование осуществляется по индексам, соответствующим узлам сетки D_h. Однако Φ_h можно превратить в конечномерное банахово пространство, если ввести на Φ_h норму, которая не порождается скалярным произведением. Примером банахова пространства является пространство C_h с нормой $\|\varphi^h\|_{C_h} = \max\limits_{k_1 k_2 \ldots k_n} |\varphi^h_{k_1 \ldots k_n}|$.

Аналогично тому, как это сделано для случая D_h, вводятся соответствующие пространства сеточных функций на $\overline{D}_h$ и ∂D_h.

Пусть Φ есть линейное множество функций $\varphi(x)$, определенных на $D \subset \mathbf{R}^n$. Считаем, что эти функции обладают определенной степенью гладкости и имеет смысл рассматривать их значения в узлах сетки D_h. Тогда каждой функции $\varphi \in \Phi$ можно поставить в соответствие сеточную функцию, которую обозначим $(\varphi)_h$ по следующему правилу: значение $(\varphi)_h$ в узле $(x_{k_1}, x_{k_2}, \ldots, x_{k_n})$ равно $\varphi(x_{k_1}, x_{k_2}, \ldots, x_{k_n})$. Указанное соответствие задает линейный оператор P_h, область определения которого есть Φ, а область значений принадлежит Φ_h. Этот оператор назовем *оператором проектирования* функции $\varphi(x)$ на сетку D_h. Таким образом, имеем $P_h\varphi \equiv (\varphi)_h(x_{k_1}, x_{k_2}, \ldots, x_{k_n}) = \varphi(x_{k_1}, x_{k_2}, \ldots, x_{k_n})$. Заметим, что оператор проектирования можно вводить различными способами. Так, другим оператором проектирования будет оператор, который каждой функции $\varphi \in \Phi$ ставит в соответствие сеточную функцию $(\varphi)_h$, значения которой в узле $(x_{k_1}, x_{k_2}, \ldots, x_{k_n})$ есть

$$\int\limits_{x_{k_1}-h_1/2}^{x_{k_1}+h_1/2} dx_1 \ldots \int\limits_{x_{k_n}-h_n/2}^{x_{k_n}+h_n/2} \varphi(x_1, x_2, \ldots, x_n)\, dx_n (h_1, h_2, \ldots, h_n)^{-1}$$

— *среднее значение* в этом узле. Легко заметить, что оператор отличается от введенного ранее оператора P_h: область его определения может быть значительно расширена и может включать функции $\varphi(x)$, для которых $\int\limits_D |\varphi(x)|\, dx < \infty$.

Если предположить, что функции $\varphi(x) \in \Phi$ достаточно гладкие, то, применяя простейшие квадратурные формулы, несложно показать, что $\|(\varphi)_h\|_{L_{2,h}} \to \|\varphi\|_{L_2(D)}$ при $h = \max\limits_i h_i \to 0$. Нормы $\|\cdot\|_{\Phi_h}, \|\cdot\|_{\Phi}$, для которых имеем $\|(\varphi)_h\|_{\Phi_h} \to \|\varphi\|_{\Phi}$ при $h \to 0$, называются *согласованными*. Использование именно согласованных норм нередко оказывается важным моментом при исследовании различных вопросов конечно-разностных методов.

Подобно тому как это сделано выше, можно осуществить проектирование функций $\varphi(x)$ (определенных на $\overline{D}$) на $\overline{D}_h$ и ∂D_h, ввести соответствующие операторы проектирования.

Пусть, далее, A — линейный оператор, заданный на функциях $\varphi \in \Phi$. Тогда функцию $\psi = A\varphi$ также можно спроектировать на сетку D_h, положив $(\psi)_h = (A\varphi)_h$. Если на ∂D определена функция $a\varphi$ (где a

есть линейный оператор — оператор граничного условия), то также можно осуществить проектирование $a\varphi$ на ∂D_h, получая при этом сеточную функцию $(a\varphi)_h$. Отмеченные выше проектирования функций $\varphi, A\varphi, a\varphi, \ldots$ на соответствующие сетки широко используются при построении конечно-разностных аналогов уравнений ($A\varphi = f$ в D; $a\varphi = g$ на ∂D и др.), методы построения которых, а также вопросы аппроксимации, счетной устойчивости и сходимости приближенных решений задачи к решению точной будут рассматриваться в дальнейшем. А здесь мы проиллюстрируем некоторые из введенных понятий на примере оператора $A = -\Delta$, рассмотренного в предыдущем пункте.

Пусть $(x_1, x_2) \equiv (x, y)$, а D есть единичный квадрат из $\mathbf{R}^2$: $D = \{(x, y) : 0 < x < 1,\ 0 < y < 1\}$. Рассмотрим задачу

$$\begin{aligned} -\Delta\varphi &= f \text{ в } D, \\ \varphi &= 0 \text{ на } \partial D, \end{aligned} \tag{1.1.69}$$

где $f = f(x, y)$ — гладкая функция, $\Delta = \dfrac{\partial^2}{\partial x^2} + \dfrac{\partial^2}{\partial y^2}$. Выберем в качестве пространства F, в котором будет действовать оператор $A = -\Delta$, пространство вещественных функций $L_2(D)$ со скалярным произведением (1.1.1) и нормой (1.1.2). Через Φ обозначим множество непрерывных в $\overline{D} = D \cup \partial D$ функций, обладающих непрерывными в D первыми и вторыми производными. Множество Φ принимаем в качестве области определения оператора $A = -\Delta$, т. е. $\Phi = \Phi(A)$. В качестве области определения оператора граничного значения $a\varphi \equiv \varphi|_{\partial D}$ берем то же множество Φ, а сам оператор a рассматриваем действующим в гильбертовом пространстве $L_2(\partial D)$, состоящим из вещественных функций, определенных на ∂D и имеющих конечную норму $\left(\int\limits_{\partial D} |\varphi|^2\, d\Gamma\right)^{1/2} < \infty$. Теперь рассматриваемую задачу можно записать в операторной форме:

$$\begin{aligned} A\varphi &= f \text{ в } D, \\ a\varphi &= 0 \text{ на } \partial D. \end{aligned} \tag{1.1.70}$$

Введем конечномерную аппроксимацию этой задачи. Для этого квадрат $\overline{D} = D \cup \partial D$ покроем равномерно по x и y сеткой с шагом $h = 1/n$. Узлы сетки будем обозначать (x_k, y_l), где $x_k = hk = k/n, y_l = hl = l/n, 0 \le k, l \le n$. Пусть $\varphi_h, (\varphi)_h$ есть сеточные функции соответственно с компонентами вида $\varphi^h_{k,l} = (\varphi^h)(x_k, y_l), (\varphi)_h(x_k, y_l) = \varphi(x_k, y_l)$, где

$\varphi \in \Phi$. Множество таких сеточных функций обозначим $\Phi_h = \Phi_h(\overline{D}_h)$. Рассмотрим следующие аппроксимации производных $\dfrac{\partial^2\varphi}{\partial x^2}$, $\dfrac{\partial^2\varphi}{\partial y^2}$:

$$-\frac{\partial^2\varphi}{\partial x^2} \to \frac{2\varphi(x_k, y_l) - \varphi(x_{k-1}, y_l) - \varphi(x_{k+1}, y_l)}{h^2},$$
$$-\frac{\partial^2\varphi}{\partial y^2} \to \frac{2\varphi(x_k, y_l) - \varphi(x_k, y_{l-1}) - \varphi(x_k, y_{l+1})}{h^2}, \quad 1 \le k, l \le n-1,$$

которые позволяют ввести Δ^h-разностный аналог оператора Лапласа следующим равенством:

$$(\Delta^h\varphi^h)_{k,l} = \frac{\varphi^h_{k+1,l} + \varphi^h_{k-1,l} + \varphi^h_{k,l+1} + \varphi^h_{k,l-1} - 4\varphi^h_{k,l}}{h^2}. \tag{1.1.71}$$

Предположим, что сеточная функция $\varphi^h \in \Phi_h$ обращается в нуль на границе сеточной области, т. е.

$$(\varphi^h)_{k,l} = 0 \text{ на } \partial D_h. \tag{1.1.72}$$

Введем разностные операторы

$$(\Delta_x\varphi^h)_{k,l} = \frac{1}{h}(\varphi^h_{k+1,l} - \varphi^h_{k,l}),\ (\nabla_x\varphi^h)_{k,l} = \frac{1}{h}(\varphi^h_{k,l} - \varphi^h_{k-1,l})$$

и аналогично операторы

$$(\Delta_y\varphi^h)_{k,l} = \frac{1}{h}(\varphi^h_{k,l+1} - \varphi^h_{k,l}),\ (\nabla_y\varphi^h)_{k,l} = \frac{1}{h}(\varphi^h_{k,l} - \varphi^h_{k,l-1}).$$

Рассмотрим новые разностные операторы A_x, A_y и A^h, определенные следующими соотношениями:

$$A_x = -\Delta_x\nabla_x, \quad A_y = -\Delta_y\nabla_y, \quad A^h = A_x + A_y. \tag{1.1.73}$$

Тогда будем иметь

$$-\Delta^h = A_x + A_y = A^h.$$

Совокупность узлов, для которых $k = 0, n$ или $l = 0, n$, образует ∂D_h. Напомним, что в этих узловых точках φ^h в соответствии с (1.1.72) обращается в нуль. Теперь с учетом введенных обозначений и операторов конечномерный аналог рассматриваемой задачи можно за-

писать в виде

$$A^h\varphi^h = f^h \text{ в } D_h,$$
$$\varphi^h = 0 \text{ на } \partial D_h,$$

где D_h — множество внутренних узлов, а сеточная функция f^h в узлах характеризуется значениями $f^h_{k,l} = f(x_k, y_l)$. В дальнейшем индекс h при сеточных функциях φ и φ^* ради простоты будем опускать.

Рассмотрим скалярное произведение

$$(\varphi, \psi) = h^2 \sum_{k=1}^{n-1} \sum_{l=1}^{n-1} \varphi_{k,l}\psi_{k,l},$$
$$\|\varphi\| = \sqrt{(\varphi, \varphi)}.$$

Построим функционал

$$(A^h\varphi, \varphi^*) = -h^2 \sum_{k=1}^{n-1} \sum_{l=1}^{n-1} [(\Delta_x \nabla_x \varphi)_{k,l} + (\Delta_y \nabla_y \varphi)_{k,l}]\varphi^*_{k,l}.$$

Имеют место следующие тождества, аналогичные первой и второй формулам Грина:

$$\begin{aligned} -\sum_{k=1}^{n-1} (\Delta_x \nabla_x \varphi)_{k,l}\varphi^*_{k,l} &= \sum_{k=1}^{n} (\nabla_x \varphi)_{k,l}(\nabla_x \varphi^*)_{k,l}, \\ -\sum_{k=1}^{n-1} (\Delta_x \nabla_x \varphi)_{k,l}\varphi^*_{k,l} &= -\sum_{k=1}^{n-1} (\Delta_x \nabla_x \varphi^*)_{k,l}\varphi_{k,l}. \end{aligned} \tag{1.1.74}$$

Формулы (1.1.74) справедливы только для функций $\varphi \in \Phi_h$, удовлетворяющих условию (1.1.72), и $\varphi^* \in \Phi^*_h$, удовлетворяющих соотношению

$$\varphi^*_{k,l} = 0 \text{ на } \partial D_h. \tag{1.1.75}$$

Аналогичные равенства имеют место и для сумм по индексу l. С помощью второго соотношения (1.1.74) получим

$$(A^h\varphi, \varphi^*) = (\varphi, A^h\varphi^*).$$

Отсюда следует самосопряженность A^h, т. е.

$$A^h = (A^h)^* \text{ и } \Phi(A^h) = \Phi^*((A^h)^*).$$

Рассмотрим функционал

$$(A^h\varphi,\varphi) = -h^2\sum_{k=1}^{n-1}\sum_{l=1}^{n-1}[(\Delta_x\nabla_x\varphi)_{k,l} + (\Delta_y\nabla_y\varphi)_{k,l}]\varphi_{k,l}.$$

С помощью первого тождества (1.1.74) для k и l получим

$$(A^h\varphi,\varphi) = h^2\sum_{k=1}^{n}\sum_{l=1}^{n}[((\nabla_x\varphi)_{k,l})^2 + ((\nabla_y\varphi)_{k,l})^2],$$

откуда следует, что

$$(A^h\varphi,\varphi) > 0,$$

если φ не нуль-вектор.

Наконец, рассмотрим спектральную задачу

$$\begin{aligned} A^h u &= \lambda u \text{ в } D_h, \\ u &= 0 \text{ на } \partial D_h. \end{aligned} \tag{1.1.76}$$

Компоненты ортонормированных собственных векторов, соответствующих задаче (1.1.76), имеют вид

$$u^{kl}_{mp} = 2\sin(m\pi kh)\sin(p\pi lh), \tag{1.1.77}$$

$$m = 1,2,\ldots,n-1;\ p = 1,2,\ldots,n-1.$$

Напомним, что

$$(u_{m_1p_1}, u_{m_2p_2}) = h^2\sum_{k=1}^{n-1}\sum_{l=1}^{n-1} u^{kl}_{m_1p_1}u^{kl}_{m_2p_2}.$$

Индексы k,l в соотношениях (1.1.77) указывают компоненты, а m и p — номера собственных элементов, которые можно упорядочить, записав

$$u_{mp} = u_i,\ i = 1,2,\ldots.$$

Поскольку имеют место очевидные соотношения

$$-(\Delta_x\nabla_x u_i)_{k,l} = 2\left(\frac{4}{h^2}\sin^2\frac{m\pi h}{2}\sin(m\pi kh)\right)\sin(p\pi lh),$$
$$-(\Delta_y\nabla_y u_i)_{k,l} = 2\sin(m\pi kh)\left(\frac{4}{h^2}\sin^2\frac{p\pi h}{2}\sin(p\pi lh)\right),$$

то собственные числа будут иметь вид

$$\lambda_{mp}(A^h) = \frac{4}{h^2}\left(\sin^2\frac{m\pi h}{2} + \sin^2\frac{p\pi h}{2}\right). \tag{1.1.78}$$

Заметим, что m и p изменяются от единицы до $n-1$. Следовательно, $h = 1/n \le mh \le (n-1)h = 1-h$ и $h \le ph \le 1-h$, поэтому

$$\frac{8}{h^2}\sin^2\frac{\pi h}{2} \le \lambda_i(A^h) \le \frac{8}{h^2}\cos^2\frac{\pi h}{2}.$$

Здесь $\lambda_i(A^h)$ — упорядоченные $\lambda_{mp}(A^h)$. Поскольку, как правило, $\pi h/2 \ll 1$, то можно приближенно записать

$$\sin^2\frac{\pi h}{2} = \frac{\pi^2 h^2}{4} - O(h^4), \quad \cos^2\frac{\pi h}{2} = 1 - O(h^2).$$

Следовательно,

$$\alpha(A^h) \le \lambda_i \le \beta(A^h), \tag{1.1.79}$$

где

$$\alpha(A^h) = \frac{1}{\|(A^h)^{-1}\|} \approx 2\pi^2, \quad \beta(A^h) = \|A^h\| \approx \frac{8}{h^2}. \tag{1.1.80}$$

Заметим, что оператор A^h при любом $h \neq 0$ является оператором ограниченным, в то время как исходный дифференциальный оператор был неограничен. Таким образом, редукция дифференциальной задачи к конечно-разностной «улучшила» спектральные свойства оператора.

Базис собственных векторов (1.1.77) может быть использован для разложения функции φ в ряд. Получим

$$\varphi = \sum_i \varphi_i u_i, \tag{1.1.81}$$

где

$$\varphi_i = (\varphi, u_i). \tag{1.1.82}$$

1.2. Аппроксимация

Рассмотрим некоторую задачу математической физики в операторной форме:

$$\begin{aligned} A\varphi &= f \text{ в } D, \\ a\varphi &= g \text{ на } \partial D, \end{aligned} \tag{1.2.1}$$

где A — линейный оператор, $\varphi \in \Phi$ и $f \in F$. Здесь Φ и F — гильбертовы пространства с областями определения элементов $D + \partial D$ и D соответственно, a — линейный оператор граничного условия, $g \in G$, где G — гильбертово пространство функций с областью определения ∂D.

Наряду с уравнением (1.2.1) рассмотрим уравнение в конечномерном пространстве сеточных функций

$$\begin{aligned} A^h\varphi^h &= f^h \text{ в } D_h, \\ a^h\varphi^h &= g^h \text{ на } \partial D_h, \end{aligned} \tag{1.2.2}$$

где A^h — линейный оператор, зависящий от шага сетки h, $\varphi^h \in \Phi_h$, $f^h \in F_h$, а Φ_h и F_h — пространства сеточных функций. Здесь D_h — множество внутренних узловых точек области D, а ∂D_h — множество узловых точек, на которых аппроксимируется граничное условие задачи, a^h — линейный оператор, $g^h \in G_h$, G_h — евклидово пространство сеточных векторов с областью определения ∂D_h.

Введем в сеточных пространствах F_h, G_h, Φ_h соответственно нормы $\|\cdot\|_{F_h}$, $\|\cdot\|_{G_h}$, $\|\cdot\|_{\Phi_h}$. Пусть $(\cdot)_h$ — линейный оператор, который элементу $\varphi \in \Phi$ ставит в соответствие элемент $(\varphi)_h \in \Phi_h$ так, что $\lim\limits_{h\to 0} \|(\varphi)_h\|_{\Phi_h} = \|\varphi\|_{\Phi}$.

Будем говорить, что задача (1.2.2) *аппроксимирует* задачу (1.2.1) *с порядком* n на решении φ, если существуют такие положительные константы $\overline{h}$, M_1, M_2, что для всех $h < \overline{h}$ выполняются неравенства

$$\begin{aligned} \|A^h(\varphi)_h - f^h\|_{F_h} &\le M_1 h^{n_1}, \\ \|a^h(\varphi)_h - g^h\|_{G_h} &\le M_2 h^{n_2} \end{aligned} \tag{1.2.3}$$

и $n = \min(n_1, n_2)$.

В тех случаях, когда решение φ задачи (1.2.1) обладает достаточной гладкостью, порядок аппроксимации удобно находить с помощью нормы, естественной для пространства непрерывных и дифференцируемых функций. С этой целью обычно пользуются разло-

жением решения и других функций, участвующих в задаче, в ряды Тейлора.

В дальнейшем будем полагать, что редукция задачи (1.2.1) к задаче (1.2.2) осуществлена и, более того, граничное условие из (1.2.2) использовано для исключения значений решения в граничных точках области $D_h + \partial D_h$. В результате приходим к эквивалентной задаче

$$\widetilde{A}^h \widetilde{\varphi}^h = \widetilde{f}^h. \tag{1.2.4}$$

При этом значения решения в граничных точках найдутся из уравнения (1.2.2) после того, как будет построено решение уравнения (1.2.4). В некоторых случаях удобно пользоваться записью аппроксимационной задачи в форме (1.2.4), а в некоторых — в форме (1.2.2). Итак, в результате проведенной редукции и с учетом требуемой аппроксимации задача с непрерывным аргументом (1.2.1) приводится к задаче линейной алгебры (1.2.4), заключающейся в решении системы алгебраических уравнений.

В дальнейшем в основном будут использованы гильбертовы пространства сеточных функций, а соответствующая норма $\|\varphi^h\|$ будет определяться (если не оговорено особо) соотношением $\|\varphi^h\| = (\varphi^h, \varphi^h)^{1/2}$. Однако необходимо отметить, что многие из вводимых понятий (аппроксимация и т. д.) можно перенести на случай банаховых пространств и в ряде утверждений и иллюстрационных примеров будут введены нормы сеточных функций, которые не связаны со скалярным произведением указанным выше соотношением.

Пример. Рассмотрим задачу

$$-\Delta\varphi = f \text{ в } D, \varphi = 0 \text{ на } \partial D. \tag{1.2.5}$$

Здесь предполагается, что областью определения D является множество $\{(x, y) : 0 < x < 1,\ 0 < y < 1\}$, а f — гладкая функция. Квадрат $D = D + \partial D$ покроем равномерной по x и по y сеткой с шагом h. Узлы области будем отмечать двумя индексами (k, l), где первый индекс k $(0 \le k \le n)$ соответствует точкам деления по координате x, а второй индекс l $(0 \le l \le n)$ — по y. Рассмотрим следующие аппроксимации:

$$\varphi_{xx} \to \Delta_x \nabla_x (\varphi)_h, \quad \varphi_{yy} \to \Delta_y \nabla_y (\varphi)_h,$$

где Δ_x, Δ_y, ∇_x и ∇_y — разностные операторы, определенные в **??**. Тогда задача (1.2.5) может быть аппроксимирована следующей:

$$\begin{gathered} -[\Delta_x \nabla_x \varphi^h + \Delta_y \nabla_y \varphi^h] = f^h \text{ в } D_h, \\ \varphi^h = 0 \text{ на } \partial D_h, \end{gathered} \tag{1.2.6}$$

где ∂D_h — множество узлов, принадлежащих границе. С учетом изложенного задача (1.2.6) может быть приведена к виду

$$\begin{aligned} -\Delta^h \varphi^h &= f^h \text{ в } D_h, \\ \varphi^h &= 0 \text{ на } \partial D_h, \end{aligned} \tag{1.2.7}$$

где φ^h и f^h — векторы с компонентами $\varphi^h_{k,l}$ и $f^h_{k,l}$, а

$$(\Delta^h \varphi^h)_{k,l} = \frac{1}{h^2}(\varphi^h_{k+1,l} + \varphi^h_{k-1,l} + \varphi^h_{k,l+1} + \varphi^h_{k,l-1} - 4\varphi^h_{k,l}),$$

$$f^h_{k,l} = \frac{1}{h^2} \int\limits_{x_{k-1/2}}^{x_{k+1/2}} \int\limits_{y_{l-1/2}}^{y_{l+1/2}} f dx dy,$$

$$x_{k\pm 1/2} = x_k \pm \frac{h}{2},\ y_{l\pm 1/2} = y_l \pm \frac{h}{2}.$$

В приведенных здесь и ниже схемах в качестве $f^h_{k,l}$ берется некоторое усреднение $f(x,y)$, вычисленное по приведенной выше формуле. Это обстоятельство, вообще говоря, позволяет рассматривать разностные схемы при $f(x,y)$, не обладающей предполагаемой в данном параграфе достаточной гладкостью. В последних случаях можно также получить соответствующие оценки ошибок аппроксимаций.

Введем в рассмотрение пространство решений Φ_h. За область определения элементов из Φ_h примем $D_h + \partial D_h = \{(x_k, y_l) : 0 \le k \le \le n, 0 \le l \le n\}$. Вектор f^h принадлежит пространству F_h с областью определения $D_h = \{(x_k, y_l) : 1 \le k \le n-1, 1 \le l \le n-1\}$. Разлагая решение по формуле Тейлора в окрестности точки $\{x_k, y_l\}$ и предполагая ограниченность производных по x и y вплоть до четвертого порядка, будем иметь

$$\varphi(\bar{x}, \bar{y}) = \sum_{n=0}^{3} \frac{1}{n!} \left\{ \left[(\bar{x} - x_k)\frac{\partial}{\partial x} + (\bar{y} - y_l)\frac{\partial}{\partial y} \right]^n \varphi \right\}_{k,l} +$$

$$+ \quad \frac{1}{4!} \left\{ \left[(\bar{x} - x_k)\frac{\partial}{\partial x} + (\bar{y} - y_l)\frac{\partial}{\partial y} \right]^4 \varphi \right\}_{k+\theta_1, l+\theta_2},$$

где $(\bar{x}, \bar{y})$ — произвольная точка области

$$\{(x,y) : x_{k-1} \le x \le x_{k+1},\ y_{l-1} \le y \le y_{l+1}\},$$

$$|\theta_1| < |\bar{x} - x_k|/h,\ |\theta_2| < |\bar{y} - y_l|/h$$

и

$$x_{k+\theta_1} = x_k + \theta_1 h, x_{l+\theta_2} = y_l + \theta_2 h.$$

Аналогичное разложение будем иметь и для функции $f(x,y)$.

Введем в пространстве F^h норму

$$\|f^h\|_{F^h} = \max_{k,l} |f^h_{k,l}|.$$

Аналогично вводится норма в пространстве G_h. В качестве $(\varphi)_h$ возьмем вектор, компонентами которого являются значения функции φ в соответствующем узле сетки. Тогда, используя указанные выше разложения для φ и f, получим

$$\| - \Delta^h(\varphi)_h - f^h\|_{F^h} \le M_1 h^2, \tag{1.2.8}$$

где

$$M_1 = \text{const} < \infty.$$

Аппроксимация граничных условий в этом случае является точной. Из последнего утверждения и оценки (1.2.8) следует, что задача (1.2.7) аппроксимирует задачу (1.2.5) со вторым порядком на решениях задачи (1.2.5), имеющих ограниченные четвертые производные.

До сих пор рассматривалась аппроксимация задачи по пространственным переменным. Аналогичным образом может быть рассмотрена задача аппроксимации эволюционного уравнения[3)]

$$\begin{aligned} \frac{\partial \varphi}{\partial t} + A\varphi &= f \text{ в } D \times D_t, \\ a\varphi &= g \text{ на } \partial D \times D_t, \\ \varphi &= \varphi^0 \text{ в } D \text{ при } t = 0. \end{aligned} \tag{1.2.9}$$

[3)]Так будем называть уравнение, которое явно разрешается относительно первой производной по времени и не содержит в A производных по времени.

Аппроксимацию задачи (1.2.9) проведем в два этапа. Сначала эту задачу аппроксимируем в области $(D_h \cup \partial D_h) \times D_t$ по пространственным переменным. В результате приходим к дифференциальному уравнению по времени и разностному по пространственным переменным.

В полученной дифференциально-разностной задаче в ряде случаев легко исключить решения в граничных точках области $(D_h \cup \partial D_h) \times D_t$ на базе разностных краевых условий. Предполагая, что это проделано, приходим к эволюционному уравнению вида

$$\frac{d\varphi^h}{dt} + \Lambda\varphi^h = f^h, \tag{1.2.10}$$

где Λ, f^h и φ^h — функции времени t. В дальнейшем индекс h в задаче (1.2.10) будем опускать, предполагая, что мы имеем дело с разностным аналогом по пространственным переменным исходной задачи математической физики.

Уравнение (1.2.10) является системой обыкновенных дифференциальных уравнений для компонент вектора φ^h.

Рассмотрим следующую задачу Коши:

$$\begin{gathered} \frac{d\varphi}{dt} + \Lambda\varphi = f, \\ \varphi = g \text{ при } t = 0. \end{gathered} \tag{1.2.11}$$

Предположим, что оператор Λ не зависит от времени. Рассмотрим простейшие методы аппроксимации задачи (1.2.11) по времени. Наиболее употребительными разностными схемами в настоящее время являются схемы первого и второго порядков аппроксимации по t.

Сначала рассмотрим простейшую явную схему первого порядка аппроксимации на сетке D_τ (D_τ есть множество узлов t_j по переменной t):

$$\frac{\varphi^{j+1} - \varphi^j}{\tau} + \Lambda\varphi^j = f^j, \quad \varphi^0 = g, \tag{1.2.12}$$

где $\tau = t_{j+1} - t_j$, f — некоторая проекция функции f. Ради простоты здесь можно принять $f = f(t_j)$.

Если рассматривается простейшая неявная схема, то имеем

$$\frac{\varphi^{j+1} - \varphi^j}{\tau} + \Lambda\varphi^{j+1} = f^j, \quad \varphi^0 = g \tag{1.2.13}$$

и f выбираем в виде $f(t_{j+1})$. Схемы (1.2.12) и (1.2.13) — первого порядка аппроксимации по времени. В этом легко убедиться с помощью разложения по формуле Тейлора по времени, допустив, например, существование ограниченных производных (по времени) второго порядка от решения.

Разрешая схемы (1.2.12) и (1.2.13) относительно неизвестного, приходим к рекуррентному соотношению

$$\varphi^{j+1} = T\varphi^j + \tau S f^j, \tag{1.2.14}$$

где T — *оператор шага*, а S — *оператор источника*, определяемые следующим образом: для схемы (1.2.12) $T = E - \tau\Lambda$, $S = E$; для схемы (1.2.13) $T = (E + \tau\Lambda)^{-1}$, $S = T$.

Разностные схемы типа (1.2.14) для эволюционных уравнений будем называть *двухслойными*.

Большое применение в приложениях имеет схема второго порядка аппроксимации — схема Кранка — Николсона:

$$\frac{\varphi^{j+1} - \varphi^j}{\tau} + \Lambda\frac{\varphi^{j+1} + \varphi^j}{2} = f^j, \quad \varphi^0 = g, \tag{1.2.15}$$

где $f^j = f(t_{j+1/2})$. Схему (1.2.15) также можно представить в форме (1.2.14) при

$$T = \left(E + \frac{\tau}{2}\Lambda\right)^{-1}\left(E - \frac{\tau}{2}\Lambda\right), \quad S = \left(E + \frac{\tau}{2}\Lambda\right)^{-1}.$$

В некоторых случаях разностные уравнения (1.2.12), (1.2.13) и (1.2.15) удобно записывать в форме системы двух уравнений, из которых одно аппроксимирует само уравнение в $D_{h\tau}$. В этом случае разностный аналог задачи (1.2.9) имеет вид

$$\begin{gathered} L^{h\tau}\varphi^{h\tau} = f^{h\tau} \text{ в } D_{h\tau}, \\ l^{h\tau}\varphi^{h\tau} = g^{h\tau} \text{ на } \partial D_{h\tau}, \end{gathered} \tag{1.2.16}$$

где

$$\begin{gathered} D_{h\tau} = D_h \times D_\tau,\ \partial D_{h\tau} = D_h \times \{0\} \cup \partial D_h \times D_\tau, \\ \|L^{h\tau}(\varphi)_{h\tau} - f^{h\tau}\|_{F_{h\tau}} \le M_1 h^n + N_1\tau^p, \\ \|l^{h\tau}(\varphi)_{h\tau} - g^{h\tau}\|_{G_{h\tau}} \le M_2 h^n + N_2\tau^p. \end{gathered} \tag{1.2.17}$$

В этих неравенствах, как и в (1.2.3), $(\cdot)_{h\tau}$ есть оператор проектирования на соответствующее сеточное пространство.

Разностное уравнение в канонической форме (1.2.14) с помощью введения вектор-функции и новых операторов также можно записать в виде

$$\widetilde{L}^{h\tau}\widetilde{\varphi}^{h\tau} = \widetilde{f}^{h\tau}. \tag{1.2.18}$$

Таким образом, эволюционное уравнение с учетом граничных условий и начальных данных может быть редуцировано к задаче линейной алгебры (1.2.18).

Пример. Рассмотрим задачу

$$\begin{aligned} A\varphi \equiv \frac{\partial\varphi}{\partial t} - \Delta\varphi &= f \text{ в } D\times D_t, \\ \varphi &= 0 \text{ на } \partial D\times D_t, \\ \varphi &= g \text{ в } D \text{ при } t=0. \end{aligned} \tag{1.2.19}$$

Областью определения решения будем считать $(D\cup\partial D)\times D_t$, где D, как и прежде, квадрат, а $D_t = \{0\le t\le T\}$. Перейдем от D к D_h, от ∂D к ∂D_h и от D_t к D_τ. Пусть D_τ — множество точек t_j и $t_{j+1}-t_j=\tau$. В качестве аппроксимации задачи (1.2.19) примем следующую:

$$\begin{aligned} A^{h\tau}\varphi &= f^{h\tau} \text{ в } D_h\times D_\tau, \\ \varphi^{h\tau} &= 0 \text{ на } \partial D_h\times D_\tau, \\ \varphi^0 &= g \text{ в } D_h\times\{0\}. \end{aligned} \tag{1.2.20}$$

Рассмотрим простейшую явную аппроксимацию

$$(A^{h\tau}\varphi^{h\tau})_{k,l} \equiv \frac{\varphi^{j+1}_{k,l}-\varphi^j_{k,l}}{\tau} - (\Delta^h\varphi^j)_{k,l}, \tag{1.2.21}$$

$$f^j_{k,l} = \frac{1}{h^2}\int\limits_{x_{k-1/2}}^{x_{k+1/2}}\int\limits_{y_{l-1/2}}^{y_{l+1/2}} f(x,y,t_j)\,dx\,dy, \tag{1.2.22}$$

$$g_{k,l} = \frac{1}{h^2}\int\limits_{x_{k-1/2}}^{x_{k+1/2}}\int\limits_{y_{l-1/2}}^{y_{l+1/2}} g(x,y)\,dx\,dy, \tag{1.2.23}$$

где $\varphi^j_{k,l} = \varphi^{h\tau}(x_k,y_l,t_j)$. Тогда

$$\varphi^{j+1}_{k,l} = \varphi^j_{k,l} + \tau(\Delta^h\varphi^j)_{k,l} + \tau f^j_{k,l} \text{ в } D_h\times D_\tau. \tag{1.2.24}$$

Кроме того,

$$\begin{aligned} \varphi^j_{k,l} &= 0 \text{ на } \partial D_h \times D_\tau, \\ \varphi^0_{k,l} &= g_{k,l} \text{ в } D_h \times \{0\}. \end{aligned} \tag{1.2.25}$$

Подставим рекуррентное соотношение (1.2.24) в следующем виде:

$$\varphi^{j+1} = T\varphi^j + \tau f^j, \tag{1.2.26}$$

где φ^j — сеточная функция с областью определения $D_h \times \partial D_h$ и

$$\varphi^j(x_k, y_l) = \varphi^{h\tau}(x_k, y_l, t_j),$$

$T = E + \tau\Delta^h = E - \tau(A_x + A_y)$ — оператор шага, а A_x и A_y определены формулами (1.1.73).

Пусть ради простоты в данном примере $F_h = \Phi_h$ и $\|\varphi^h\|_{\Phi_h} = \sqrt{\sum\limits_{k,l} |\varphi_{k,l}|^2 h^2}$. Оценим норму оператора T. Для этой цели найдем его максимальное собственное число:

$$\begin{aligned} Tu &= \lambda(T)u \text{ в } D_h, \\ u &= 0 \text{ на } \partial D_h. \end{aligned} \tag{1.2.27}$$

Справедливо очевидное соотношение

$$\lambda_h(T) = 1 + \tau\lambda_n(\Delta^h).$$

Следовательно, норма оператора T имеет вид

$$\|T\| = \max\left\{\left|1 - \frac{8\tau}{h^2}\cos^2\frac{\pi h}{2}\right|, \left|1 - \frac{8\tau}{h^2}\sin^2\frac{\pi h}{2}\right|\right\}, \tag{1.2.28}$$

и если $\tau/h^2 < 1/4$, то $\|T\| < 1$.

Наряду с явной аппроксимацией первого порядка по τ можно рассматривать неявную аппроксимацию первого порядка по τ и второго порядка по h. Тогда вместо выражения (1.2.21) будем иметь следующее:

$$(A^{h\tau}\varphi^{h\tau})_{k,l} \equiv \frac{\varphi^{j+1}_{k,l} - \varphi^j_{k,l}}{\tau} - (\Delta^h\varphi^{j+1})_{k,l}; \tag{1.2.29}$$

величины $f^j_{k,l}$ и $g_{k,l}$ определяются формулами (1.2.22), (1.2.23). В этом случае уравнение (1.2.20) уже не разрешается явно, и мы приходим

к операторному уравнению

$$((E-\tau\Delta^h)\varphi^{j+1})_{k,l} = \varphi^j_{k,l} + \tau f^{j+1}_{k,l} \text{ в } D_h \times D_\tau, \tag{1.2.30}$$

решение которого должно удовлетворять следующим условиям:

$$\begin{aligned} \varphi^j_{k,l} &= 0 \text{ на } \partial D_h \times D_\tau, \\ \varphi^0_{k,l} &= g_{k,l} \text{ в } D_h \times \{0\}. \end{aligned} \tag{1.2.31}$$

Запишем уравнение (1.2.30) в форме

$$\varphi^{j+1}_{k,l} = (T(\varphi^j + \tau f^j))_{k,l}, \tag{1.2.32}$$

где $T = (E-\tau\Delta^h)^{-1}$. В этом случае норма оператора T будет равна

$$\|T\| = \max\left\{\frac{1}{1+\frac{8\tau}{h^2}\cos^2\frac{\pi h}{2}}, \frac{1}{1+\frac{8\tau}{h^2}\sin^2\frac{\pi h}{2}}\right\}; \tag{1.2.33}$$

следовательно, $\|T\| < 1$ при любых τ и h.

Наконец, рассмотрим аппроксимацию по схеме Кранка — Николсона. Определим операторы и функции в задаче (1.2.20) следующим образом:

$$(A^{h\tau}\varphi^{h\tau})_{k,l} \equiv \frac{\varphi^{j+1}_{k,l} - \varphi^j_{k,l}}{\tau} - \left(\Delta^h \frac{\varphi^{j+1} + \varphi^j}{2}\right)_{k,l}, \tag{1.2.34}$$

$$\begin{aligned} f^j_{k,l} &= \frac{1}{h^2}\int\limits_{x_{k-1/2}}^{x_{k+1/2}}\int\limits_{y_{l-1/2}}^{y_{l+1/2}} f(x,y,t_{j+1/2})\,dx\,dy, \\ g^j_{k,l} &= \frac{1}{h^2}\int\limits_{x_{k-1/2}}^{x_{k+1/2}}\int\limits_{y_{l-1/2}}^{y_{l+1/2}} g(x,y)\,dx\,dy. \end{aligned} \tag{1.2.35}$$

Тогда приходим к задаче

$$\left(\left(E-\frac{\tau}{2}\Delta^h\right)\varphi^{j+1}\right)_{k,l} = \left(\left(E+\frac{\tau}{2}\Delta^h\right)\varphi^j\right)_{k,l} + \tau f^j_{k,l} \text{ в } D_h \times D_\tau, \tag{1.2.36}$$

$$\begin{aligned} \varphi^j_{k,l} &= 0 \text{ на } \partial D_h \times D_\tau, \\ \varphi^0_{k,l} &= g_{k,l} \text{ в } D_h \times \{0\}. \end{aligned} \tag{1.2.37}$$

В этом случае (1.2.36) формально разрешается относительно неизвестной $\varphi_{k,l}^{j+1}$ в виде

$$\varphi_{k,l}^{j+1} = (T\varphi^{j+1})_{k,l} + \tau(Sf^j)_{k,l}, \tag{1.2.38}$$

где

$$T = \left(E - \frac{\tau}{2}\Delta^h\right)^{-1}\left(E + \frac{\tau}{2}\Delta^h\right), \quad S = \left(E - \frac{\tau}{2}\Delta^h\right)^{-1}.$$

Норма оператора шага будет равна

$$\|T\| = \max\left\{\left|\frac{1 - \frac{4\tau}{h^2}\cos^2\frac{\pi h}{2}}{1 + \frac{4\tau}{h^2}\cos^2\frac{\pi h}{2}}\right|, \left|\left|\frac{1 - \frac{4\tau}{h^2}\sin^2\frac{\pi h}{2}}{1 + \frac{4\tau}{h^2}\sin^2\frac{\pi h}{2}}\right|\right\}. \tag{1.2.39}$$

Откуда следует $\|T\| < 1$.

1.3. Счетная устойчивость

Мы не будем стремиться к возможной общности определения понятия счетной устойчивости разностных схем, поскольку нас в основном будут интересовать простейшие алгоритмические подходы к анализу качества разностных схем, аппроксимирующих задачи математической физики.

Для выяснения основных определений и понятий теории устойчивости рассмотрим сначала явную разностную схему (1.2.12):

$$\varphi^{j+1} = (E - \tau\Lambda)\varphi^j + \tau f^j, \quad \varphi^0 = g. \tag{1.3.1}$$

Решение φ_j ищется для $0 \le \tau j \le T$.

Предположим, что оператор Λ положительный и что он имеет полную систему собственных функций $\{u_n\}$ и множество собственных чисел $\{\lambda_n > 0\}$, соответствующих спектральной задаче

$$Au = \lambda u.$$

Введем в рассмотрение следующие ряды Фурье:

$$\varphi^j = \sum_n \varphi_n^j u_n, \quad f^j = \sum_n f_n^j u_n, \quad g = \sum_n g_n u_n, \tag{1.3.2}$$

где $\varphi_n^j = (\varphi^j, u_n^*)$, $f_n^j = (f^j, u_n^*)$, $g_n = (g, u_n^*)$; u_n^* — собственные функции сопряженной спектральной задачи. Подставим (1.3.2) в (1.3.1) и результат скалярно умножим на u_n^*. Тогда придем к следующим выражениям для коэффициентов Фурье:

$$\varphi_n^{j+1} = (1 - \tau\lambda_n)\varphi_n^j + \tau f_n^j. \tag{1.3.3}$$

Поскольку

$$\varphi^0 = \sum_n g_n u_n,$$

то начальное условие имеет вид

$$\varphi_n^0 = g_n. \tag{1.3.4}$$

Последовательным исключением неизвестных получим

$$\varphi_n^j = r_n^j g_n + \tau \sum_{i=1}^{j} r_n^{j-i} f_n^{i-1}, \tag{1.3.5}$$

где

$$r_n = 1 - \tau\lambda_n. \tag{1.3.6}$$

Из равенства (1.3.5) следует, что при $\tau > 0$

$$|\varphi_n^j| \le |r_n|^j |g_n| + \tau \sum_{i=1}^{j} |r_n|^{j-i} |f_n^{i-1}|.$$

Усредним последнее неравенство, заменив $|f_n^{i-1}|$ под знаком суммы на $|f_n| = \max\limits_j |f_n^j|$. Получим

$$|\varphi_n^j| \le |r_n|^j |g_n| + \frac{1 - |r_n|^j}{1 - |r_n|}\tau |f_n|. \tag{1.3.7}$$

Джон Нейман ввел в рассмотрение так называемый *спектральный критерий устойчивости*, смысл которого состоит в следующем. Если для каждого коэффициента φ_n^j ряда Фурье (1.3.2) имеет место соотношение

$$|\varphi_n^j| \le C_{1n}|g_n| + C_{2n}|f_n|, \quad n = 1, 2, \ldots, \tag{1.3.8}$$

где C_{1n}, C_{2n} — константы, равномерно ограниченные при $0 \le j\tau \le T$, то разностная схема (1.3.1) объявляется *счетно-устойчивой*. Посмот-

рим, какие условия достаточно наложить на параметры разностной схемы (1.2.12), чтобы выполнялось соотношение (1.3.8). Анализ соотношения (1.3.7) показывает, что критерий устойчивости (1.3.8) будет выполнен, если на параметр r_n наложить ограничение[4)]

$$|r_n| \leq 1, \ n = 1, 2, \ldots \tag{1.3.9}$$

Предположим, что спектр оператора Λ расположен в промежутке

$$0 < \alpha(\Lambda) \leq \lambda_n(\Lambda) \leq \beta(\Lambda).$$

Тогда при условии

$$\tau \leq \frac{2}{\beta(\Lambda)} \tag{1.3.10}$$

в соответствии с (1.3.6) соотношение (1.3.9) будет выполняться. Неравенство (1.3.10) и будет конструктивным условием устойчивости разностной схемы (1.3.1). Заметим, что условие (1.3.10) является достаточным условием устойчивости. В этом случае, очевидно, соотношение (1.3.7) переходит в следующее:

$$|\varphi_n^j| \leq |g_n| + j\tau|f_n|. \tag{1.3.11}$$

Но $j\tau < T$, где T фиксировано. Это значит, что при малом τ рассматривается большое число шагов j и $j \to \infty$ при $\tau \to 0$, но так, чтобы верхняя граница временного промежутка T оставалась фиксированной. Тогда снова приходим к схемам, устойчивым по Нейману.

Рассмотрим теперь другие разностные схемы, основанные на неявных разностных аппроксимациях. В случае неявной схемы первого порядка аппроксимации (1.2.13) получим выражение, аналогичное (1.3.7):

$$|\varphi_n^j| \leq |r_n|^j|g_n| + \frac{1 - |r_n|^j}{1 - |r_n|}\tau|r_n||f_n|, \tag{1.3.12}$$

где

$$r_n = \frac{1}{1 + \tau\lambda_n(\Lambda)}.$$

[4)] В дальнейшем будет введено более слабое ограничение (1.3.27) на норму оператора шага.

Для данной разностной схемы при $\lambda_n(\Lambda) > 0$ имеет место устойчивость для любого значения $\tau > 0$, поскольку

$$|r_n| < 1,\ n = 1, 2, \ldots.$$

В таких случаях устойчивость будем называть *абсолютной*.

Для схемы Кранка — Николсона (1.2.15) оценка для решения коэффициентов Фурье имеет вид

$$|\varphi_n^j| \le |r_n|^j |g_n| + \frac{1 - |r_n|^j}{1 - |r_n|} \tau |\mu_n| |f_n|, \tag{1.3.13}$$

$$r_n = \frac{1 - \frac{\tau}{2}\lambda_n(\Lambda)}{1 + \frac{\tau}{2}\lambda_n(\Lambda)},\ \mu_n = \frac{1}{1 + \frac{\tau}{2}\lambda_n(\Lambda)}.$$

Отсюда следует, что $|r_n| < 1$ при любых $\tau > 0$, если $\lambda_n(\Lambda) > 0$.

Необходимо отметить, во-первых, что устойчивость по Нейману основана на анализе спектра операторной задачи. Это означает, что при таком подходе вычисление максимального собственного числа задачи или его оценка сверху является необходимым элементом алгоритма. Во-вторых, спектральный критерий, очевидно, устанавливает устойчивость решения по отношению к каждой гармонике ряда Фурье, но иногда ничего не говорит об устойчивости решения в энергетической норме. А между тем норма решения φ^j зачастую оказывается единственной характеристикой решения задачи. Все это побудило исследователей дать иные определения устойчивости, связанные с нормами операторов задачи. Следует подчеркнуть, что анализ устойчивости по Нейману до сих пор играет исключительную роль в приложениях.

Перейдем теперь к более общему определению понятия счетной устойчивости. С этой целью рассмотрим задачу

$$\begin{aligned} \frac{\partial \varphi}{\partial t} + A\varphi &= f \quad \text{в} \qquad D \times D_t, \\ \varphi &= g \quad \text{при} \qquad t = 0, \end{aligned} \tag{1.3.14}$$

которая аппроксимируется разностной задачей

$$\begin{aligned} \varphi^{j+1} &= T\varphi^j + \tau S f^j \text{ на } D_h \times D_\tau, \\ \varphi^{(0)} &= g. \end{aligned} \tag{1.3.15}$$

Будем говорить, что разностная схема (1.3.15) *устойчива*, если при любом параметре h, характеризующем данную аппроксимацию, и $j \le T/\tau$ имеет место соотношение

$$\|\varphi^j\|_{\Phi_h} \le C_1\|g\|_{G_h} + C_2\|f^{h\tau}\|_{F_h\tau}, \tag{1.3.16}$$

где константы C_1 и C_2 равномерно ограничены на $0 \le t \le T$ и не зависят от τ, h, g и f; через G_h обозначено пространство, которому принадлежит g в (1.3.15).

Определение счетной устойчивости тесно связано с непрерывным аргументом. Можно сказать, что счетная устойчивость устанавливает непрерывную зависимость решения от входных данных в случае задач с дискретным аргументом. Действительно, пусть в качестве входных данных задачи (1.3.15) выбраны $f = f_*^j$, $g = g_*$. Соответствующее им решение задачи обозначим через φ_*^j. Далее, в качестве входных данных выберем $f^j = f_*^j + \xi$, $g = g_* + \delta$ и обозначим новое решение через φ_{**}^j. Тогда для разности решений $\varepsilon = \varphi_{**}^j - \varphi_*^j$ получим следующую задачу:

$$\varepsilon^{j+1} = T\varepsilon^j + \tau S\xi^j,\ \varepsilon^0 = \delta.$$

При этом условие устойчивости примет вид

$$\|\varepsilon^{j+1}\|_{\Phi_h} \le C_1\|\delta\|_{G_h} + C_2\|\xi^{h\tau}\|_{F_{h\tau}}.$$

Отсюда следует, что малым вариациям входных данных f и g соответствуют малые вариации решения φ.

Легко видеть, что определение устойчивости в форме (1.3.16) уже связывает само решение с априорными сведениями о входных данных задачи. Для анализа устойчивости многих задач такое определение более удобно, чем определение устойчивости по Нейману, и более информативно. Рассмотрим устойчивость схемы (1.2.12). Для этого рекуррентное соотношение (1.3.1) перепишем в форме

$$\varphi^{j+1} = T\varphi^j + \tau f^j, \quad \varphi^0 = g, \tag{1.3.17}$$

где

$$T = E - \tau\Lambda, \tag{1.3.18}$$

а Λ — оператор, аппроксимирующий A. Формальное решение задачи (1.3.17) имеет вид

$$\varphi^j = T^j g + \tau \sum_{i=1}^{j} T^{j-i} f^{i-1}. \tag{1.3.19}$$

Принимая $G_h = F_h$ [5)] и оценивая решение (1.3.19) по норме, получим

$$\|\varphi^j\|_{\Phi_h} = \|T\|^j \|g\|_{G_h} + \tau \sum_{i=1}^{j} \|T\|^{j-i} \|f^{i-1}\|_{F_h}. \tag{1.3.20}$$

Заменим под знаком суммы $\|f^{i-1}\|$ максимальным значением по всем j из фиксированного временного интервала. Пусть

$$\|f^{h\tau}\| = \max_j \|f^j\|_{F_h},$$

тогда

$$\|\varphi^j\|_{\Phi_h} = \|T\|^j \|g\|_{G_h} + \frac{1 - \|T\|^j}{1 - \|T\|} \tau \|f^{h\tau}\|. \tag{1.3.21}$$

Если предположить, что

$$\|T\| \le 1, \tag{1.3.22}$$

то схема (1.2.12) будет устойчивой в смысле определения (1.3.16). Естественно, что условие (1.3.22) является достаточным условием устойчивости. Можно было бы получить более тонкие критерии через нормы степеней операторов шага $\|T^i\|$ $(i = 1, 2, \ldots, j)$. Однако ослабление условия затрудняет конструктивную процедуру установления критерия устойчивости. В практических расчетах, как правило, используется достаточное условие (1.3.22). Рассмотрим случай, когда в (1.3.18) оператор $\Lambda = \Lambda^* > 0$. Положим

$$J[\varphi] = \frac{(T\varphi, T\varphi)}{(\varphi, \varphi)}. \tag{1.3.23}$$

[5)] Это предположение делается для упрощения изложения. В противном случае вместо одной нормы $\|T\| \equiv \|T\|_{F_h \to \Phi_h} = \sup\limits_{\varphi \in F_h} \frac{\|T\varphi\|_{\Phi_h}}{\|\varphi\|_{F_h}}$ нужно было бы ввести две:

$$\|T\|_{F_h \to \Phi_h} \text{ и } \|T\|_{G_h \to \Phi_h} = \sup_{\varphi \in G_h} \frac{\|T\varphi\|_{\Phi_h}}{\|\varphi\|_{G_h}}.$$

Тогда

$$J[\varphi] = 1 - 2\tau\frac{(\Lambda\varphi,\varphi)}{(\varphi,\varphi)} + \tau^2\frac{(\Lambda\varphi,\Lambda\varphi)}{(\varphi,\varphi)}.$$

Пусть

$$\varphi = \sum_n \varphi_n u_n,$$

где $\{u_n\}$ — базис оператора Λ. Тогда

$$J[\varphi] = 1 - 2\tau\widetilde{\lambda} + \tau^2\widetilde{\widetilde{\lambda}}, \tag{1.3.24}$$

где

$$\widetilde{\lambda} = \frac{\sum\limits_n \lambda_n(\Lambda)\varphi_n^2}{\sum\limits_n \varphi_n^2}, \quad \widetilde{\widetilde{\lambda}} = \frac{\sum\limits_n [\lambda_n(\Lambda)]^2\varphi_n^2}{\sum\limits_n \varphi_n^2}.$$

Легко проверить, что неравенство $J[\varphi] \leq 1$ выполняется при условии, что

$$\tau \leq 2\frac{\widetilde{\lambda}}{\widetilde{\widetilde{\lambda}}} = 2\frac{\sum\limits_n \lambda_n\varphi_n^2}{\sum\limits_n \lambda_n^2\varphi_n^2}.$$

Отсюда вытекает, что если $\beta(\Lambda) = \|\Lambda\| = \max\limits_n \lambda_n(\Lambda) = \lambda_1(\Lambda)$, то

$$\tau \leq \frac{2}{\lambda_1(\Lambda)}\frac{\varphi_1^2 + \sum\limits_{n>1}\dfrac{\lambda_n(\Lambda)}{\lambda_1(\Lambda)}\varphi_n^2}{\varphi_1^2 + \sum\limits_{n>1}\dfrac{[\lambda_n(\Lambda)]^2}{[\lambda_1(\Lambda)]^2}\varphi_n^2}. \tag{1.3.25}$$

Так как

$$\frac{\varphi_1^2 + \sum\limits_{n>1}\dfrac{\lambda_n(\Lambda)}{\lambda_1(\Lambda)}\varphi_n^2}{\varphi_1^2 + \sum\limits_{n>1}\dfrac{[\lambda_n(\Lambda)]^2}{[\lambda_1(\Lambda)]^2}\varphi_n^2} \geq 1,$$

то приходим к следующему достаточному условию выполнения неравенства $J[\varphi] \leq 1$:

$$\tau \leq \frac{2}{\beta(\Lambda)}.$$

В этом случае, согласно определению (1.1.11), нормы оператора

$$\|T\|^2 = \sup_\varphi(J[\varphi]) \leq 1,$$

и, следовательно, счет будет устойчив в смысле определения (1.3.16). Заметим, что в случае самосопряженного оператора достаточные условия счетной устойчивости по Нейману (1.3.10) и в смысле определения (1.3.16) совпадают.

Аналогичным образом можно рассмотреть устойчивость неявных разностных уравнений (1.2.13) и (1.2.15). В этих случаях имеем

$$\|\varphi^j\| \le \|T\|^j\|g\| + \frac{1-\|T\|^j}{1-\|T\|}\tau\|S\|\|f\|,$$

где для схемы (1.2.13)

$$T = (E+\tau\Lambda)^{-1},\ S = (E+\tau\Lambda)^{-1},$$

а для схемы (1.2.15)

$$T = (E+\frac{\tau}{2}\Lambda)^{-1}(E-\frac{\tau}{2}\Lambda), \quad S = (E+\frac{\tau}{2}\Lambda)^{-1}.$$

Нетрудно показать, что построенные разностные схему будут абсолютно устойчивы в смысле определения (1.3.16), если $\Lambda > 0$ и

$$\|\varphi^j\|_{\Phi_h} = \sqrt{\sum_{k,l}|\varphi^j_{k,l}|^2h^2}.$$

Несколько слов о предельных переходах. При решении разностных аналогов эволюционных задач математической физики нам приходится иметь дело с аппроксимацией как по времени с шагом τ, так и по пространству с характерным шагом h. Это значит, что оператор перехода $T = T(\tau, h)$ зависит как от τ, так и от h.

Проблема конструирования устойчивого алгоритма при заданном способе аппроксимации обычно сводится к установлению связи между τ и h, обеспечивающей счетную устойчивость. Если разностная схема оказывается устойчивой при любых значениях $\tau > 0$ и $h > 0$, то она объявляется *абсолютно устойчивой*. Если же схема оказывается устойчивой только при определенной связи между τ и h, то такая схема называется *условно устойчивой*.

Предположим, что связь параметров τ и h задается в форме

$$\tau \le Ch^p, \tag{1.3.26}$$

где C и p — заданные константы, не зависящие от τ и h.

Предположим, что нам требуется повысить точность решения задачи формальным уменьшением шага h. Тогда одновременно мы должны уменьшить τ так, чтобы снова выполнялось указанное выше неравенство, но уже с новыми параметрами сетки. Это значит, что можно допустить и предельный переход при $\tau \to 0$ и $h \to 0$, обеспечив выполнение условия (1.3.26), например в виде

$$\frac{\tau}{h^p} = \text{const} \leq C.$$

Наряду с изложенными выше определениями счетной устойчивости в литературе используются и другие определения, которые позволяют расширить класс разностных схем, интересных для приложений. Например, схема называется *устойчивой*, если

$$\|T\| \leq 1 + O(\tau). \tag{1.3.27}$$

Такое определение устойчивости при малых τ допускает экспоненциальное возрастание по времени погрешностей округлений.

Мы рассмотрели принципиальную схему исследования счетной устойчивости разностной схемы в предположении, что оператор Λ не зависит от времени. Такое предположение для ряда задач математической физики является единственным. Вместе с тем оно позволяет ввести в рассмотрение ряд дальнейших конструктивных приемов, широко используемых в вычислительной математике. В самом деле, исследование устойчивости сводится к оценке нормы оператора шага T. Как было указано в 1.1, квадрат нормы оператора T совпадает со спектральным радиусом самосопряженного положительного оператора T^*T, а для определения спектрального радиуса может быть использован итерационный процесс

$$\|T\|^2 = \lim_{k\to\infty} \frac{(T^*T\varphi^{(k)}, \varphi^{(k)})}{(\varphi^{(k)}, \varphi^{(k)})},$$

где $\varphi^{(k)}$ — элементы следующего процесса:

$$\varphi^{(k+1)} = \frac{1}{\|\varphi^{(k)}\|} T^*T\varphi^{(k)}. \tag{1.3.28}$$

Таким образом, задача определения нормы оператора T сводится к последовательной реализации рекуррентного соотношения (1.3.28). Именно этот путь является конструктивно наиболее разработанным применительно к ЭВМ. В случае самосопряженного оператора T

$$\|T\| = \beta(T).$$

Сделаем теперь некоторые частные замечания. При исследовании устойчивости разностных схем иногда используют метод определения спектрального радиуса для задачи, периодической по пространственным координатам. Для задач с непериодическими граничными условиями оценку спектрального радиуса обязательно следует производить (с помощью метода Люстерника) для операторов T, при конструировании которых уже учтены реальные граничные условия.

Если оператор Λ со временем меняется, то задача исследования устойчивости неизмеримо затрудняется, так как норма оператора T также будет изменяться со временем и спектральный радиус необходимо находить, вообще говоря, на каждом шаге, поскольку он будет зависеть от номера временного шага. В этом случае целесообразно идти по пути построения абсолютно устойчивых разностных аналогов задач. Такие схемы будут специально рассмотрены в главе 5.

В заключение отметим, что если аппроксимация эволюционного уравнения исследуется в пространствах сеточных функций, определенных на $D_h \times D_\tau$, то и определение устойчивости полезно давать в терминах тех же пространств. В самом деле, пусть разностная задача имеет вид

$$\begin{array}{rcll} L^{h\tau}\varphi^{h\tau} & = & f^{h\tau} & \text{в} \quad D_h \times D_\tau, \\ l^{h\tau}\varphi^{h\tau} & = & g^{h\tau} & \text{на} \quad \partial D_h \times D_\tau. \end{array} \tag{1.3.29}$$

Введем критерий устойчивости в следующем виде:

$$\|\varphi^{h\tau}\|_{\Phi_{h\tau}} \le C_1 \|f^{h\tau}\|_{F_{h\tau}} + C_2 \|g^{h\tau}\|_{G_{h\tau}}, \tag{1.3.30}$$

где C_1 и C_2 на отрезке $0 \le t \le T$ — константы, не зависящие от h, τ, $f^{h\tau}$, $g^{h\tau}$.

Пусть исходная задача математической физики аппроксимируется с помощью разностного уравнения так, что граничные и начальные условия уже учтены при его построении. Тогда критерий устой-

чивости удобно ввести в следующей форме:

$$\|\varphi^{h\tau}\|_{\Phi_{h\tau}} \leq C\|f^{h\tau}\|_{F_{h\tau}}, \tag{1.3.31}$$

где C ограничена на отрезке $0 \leq t \leq T$.

1.4. Теорема сходимости

Сформулируем основной результат теории конечно-разностных алгоритмов — теорему сходимости. Доказано несколько вариантов этой теоремы. Так, А. Ф. Филиппов, определив устойчивость для произвольных разностных схем

$$L^{h\tau}\varphi^{h\tau} = f^{h\tau}$$

как равномерную ограниченность семейства операторов $(L^{h\tau})^{-1}$, доказал, что из аппроксимации и устойчивости следует сходимость решения разностной задачи к решению дифференциальной. П. Лакс предложил для корректно поставленных эволюционных задач такую систему определений аппроксимации и устойчивости, при которых устойчивость имеет место одновременно со сходимостью, если имеет место аппроксимация. Эта теорема известна под названием теоремы Лакса.

Исследование сходимости разностного решения к решению исходной задачи как для стационарных, так и для эволюционных задач математической физики осуществляется на основе одних и тех же принципов. Это позволяет проследить основную идею доказательства на примере стационарной задачи

$$\begin{gathered} A\varphi = f \text{ в } D, \\ a\varphi = g \text{ на } \partial D, \end{gathered} \tag{1.4.1}$$

которая аппроксимируется разностной схемой

$$\begin{aligned} A^h\varphi^h &= f^h \quad \text{в} \quad D_h, \\ a^h\varphi^h &= g^h \quad \text{на} \quad \partial D_h. \end{aligned} \tag{1.4.2}$$

Справедлива следующая теорема сходимости:

Пусть

1) разностная схема (1.4.2) *аппроксимирует исходную задачу* (1.4.1) *на решении* φ *с порядком* n*;*

2) A^h *и* a^h *— линейные операторы;*

3) разностная схема (1.4.2) *устойчива в смысле* (1.3.30)*, т. е. существуют такие положительные константы* $\overline{h}$*,* C_1*,* C_2*, что для всех* $h < \overline{h}$*,* $f^h \in F_h$*,* $g^h \in G_h$ *существует, и притом единственное, решение* φ^h *задачи* (1.4.2)*, удовлетворяющее неравенству*

$$\|\varphi^h\|_{\Phi_h} \le C_1\|f^h\|_{F_h} + C_2\|g^h\|_{G_h}. \tag{1.4.3}$$

Тогда решение разностной задачи φ^h *сходится к решению* φ *исходной задачи, т. е.*

$$\lim_{h\to 0}\|(\varphi)_h - \varphi^h\|_{\Phi_h} = 0,$$

причем имеет место следующая оценка скорости сходимости:

$$\|(\varphi)_h - \varphi^h\|_{\Phi_h} \le (C_1M_1 + C_2M_2)h^n, \tag{1.4.4}$$

где M_1 *и* M_2 *— константы из* (1.2.3)

Докажем это утверждение. Пусть $\overline{h}$ — минимальное из $\overline{h}$, введенных в определениях аппроксимации и устойчивости. Тогда в силу устойчивости для любых правых частей f^h и g^h при $h < \overline{h}$ существует единственное решение φ^h, т. е. для $h < \overline{h}$ мы имеем право рассматривать разность $(\varphi)_h - \varphi^h$. В силу линейности A^h получаем

$$A^h[(\varphi)_h - \varphi^h] = A^h(\varphi)_h - A^h\varphi^h = A^h(\varphi)_h - f^h.$$

Аналогично

$$a^h[(\varphi)_h - \varphi^h] = a^h(\varphi)_h - g^h.$$

Так как $h < \overline{h}$, то в силу устойчивости и аппроксимации из соотношений

$$\begin{gathered} A^h[(\varphi)_h - \varphi^h] = A^h(\varphi)_h - f^h, \\ a^h[(\varphi)_h - \varphi^h] = a^h(\varphi)_h - g^h \end{gathered} \tag{1.4.5}$$

следует, что

$$\begin{gathered} \|(\varphi)_h - \varphi^h\|_{\Phi_h} \le C_1\|A^h(\varphi)_h - f^h\|_{F_h} + C_2\|a^h(\varphi)_h - g^h\|_{G_h} \le \\ \le C_1M_1h^{n_1} + C_2M_2h^{n_2} \le (C_1M_1 + C_2M_2)h^n. \end{gathered}$$

При получении последнего неравенства, не нарушая общности, можно считать, что $h < 1$.

Доказательство завершено.

Отметим, что при доказательстве теоремы использовано свойство линейности лишь для операторов A^h и a^h.

В случае эволюционной задачи рассмотрим

$$\begin{array}{lll} \delta f^{h\tau} & \equiv L^{h\tau}[(\varphi)_{h\tau} - \varphi^{h\tau}] & \equiv \quad L^{h\tau}(\varphi)_{h\tau} f^{h\tau}, \\ \delta g^{h\tau} & \equiv l^{h\tau}[(\varphi)_{h\tau} - \varphi^{h\tau}] & \equiv \quad l^{h\tau}(\varphi)_{h\tau} g^{h\tau}. \end{array} \tag{1.4.6}$$

Из (1.4.6) и условия устойчивости (1.3.30) имеем

$$\|(\varphi)_{h\tau} - \varphi^{h\tau}\|_{\Phi_h} \le C_1 \|\delta f^{h\tau}\|_{F_{h\tau}} + C_2 \|\delta g^{h\tau}\|_{G_{h\tau}},$$

или, с учетом (1.2.17),

$$\|(\varphi)_{h\tau} - \varphi^{h\tau}\|_{\Phi_h} \le K_1 h^n + K_2 \tau^p, \tag{1.4.7}$$

где

$$K_1 = C_1 M_1 + C_2 M_2, \; K_2 = C_1 N_1 + C_2 N_2.$$

Оценка (1.4.7) доказывает сходимость разностного решения к точному и дает четкое представление о сходимости как по отношению к пространственному шагу сетки h, так и по отношению к шагу τ по временной оси.

В предположении теоремы было включено весьма жесткое условие, что C_1 и C_2 не зависят от h и τ. Особенно неприятным является требование независимости этих констант от h. Между тем при $h \to 0$ в некоторых случаях C_1 и C_2 могут стремиться к бесконечности.

Пусть

$$C_1^h = \frac{C_1}{h_m} \quad \text{и} \quad C_2^h = \frac{C_2}{h^m},$$

где $m \ge 0$. Тогда скорость сходимости приближенного решения к точному будет оцениваться следующим образом:

$$\|(\varphi)_{h\tau} - \varphi^{h\tau}\|_{\Phi_h} \le M h^{n-m} + N \tau^p h^{-m}.$$

Если $n > m$ и $\tau^p h^{-m} \to 0$ при $\tau \to 0$, $h \to 0$, то сходимость имеет место. Естественно, что теорема сходимости может быть сформулирована и в тех случаях, когда C_1 и C_2 зависят как от h, так и от τ.

1.5. Конечно-разностные аналоги некоторых задач математической физики

Рассмотрим ряд простых и вместе с тем типичных задач математической физики, на которых будет проиллюстрирован метод конечных разностей в применении к данным задачам.

1.5.1. Задача Дирихле для одномерного уравнения Пуассона

Сначала рассмотрим задачу Дирихле для одномерного уравнения Пуассона

$$\begin{aligned} \frac{d^2\varphi}{dx^2} &= f, \quad 0 < x < 1, \\ \varphi(0) &= a, \quad \varphi(1) = b, \end{aligned} \tag{1.5.1}$$

где $f(x)$ — функция источников; a, b — заданные константы. Известно, что если $f(x)$ непрерывна на $[0,1]$ и имеет на $[0,1]$ непрерывные производные $d^i f/dx^i$ $(i = 1,2)$ (т. е. $f \in C^2[0,1]$), то существует единственное решение задачи (1.5.1), которое принадлежит классу $C^4[0,1]$.

Для построения разностного аналога задачи (1.5.1) разобьем промежуток $0 \le x \le 1$ с точками x_k на $N \ge 2$ равных промежутков $x_k \le x \le x_{k+1}$ с постоянным шагом $h = x_{k+1} - x_k$. Введем разностный оператор A, который во внутренних точках сетки задается выражениями

$$(A\varphi)_k = \frac{2\varphi_k - \varphi_{k+1} - \varphi_{k-1}}{h^2}, \quad k = 1, 2, \ldots, N-1.$$

В граничных точках принимаем

$$\varphi_0 = a, \quad \varphi_N = b.$$

Тогда разностная схема для задачи (1.5.1) записывается в виде

$$\begin{aligned} (A\varphi)_k &= f_k, \quad k = 1, 2, \ldots, N-1, \\ \varphi_0 &= a, \quad \varphi_N = b, \end{aligned} \tag{1.5.2}$$

где $f_k = f(x_k)$. Если $\varphi(x) \in C^4[0,1]$, то, используя разложения в ряды Тейлора, получаем

$$\left| \left(-\frac{d^2\varphi}{dx^2}(x_k) \right) - \frac{2\varphi(x_k) - \varphi(x_{k+1}) - \varphi(x_{k-1})}{h^2} \right| \leq$$
$$\leq Ch^2 \max_{x\in[0,1]} \left| \frac{d^4\varphi(x)}{dx^4} \right|, \ k = 1, 2, \ldots, N-1.$$

Таким образом, во внутренних узлах сетки схема (1.5.2) обладает вторым порядком относительно h как в метрике $\|\varphi\|_{L_{2,h}} = \left(\sum_{i=1}^{N-1} h|\varphi_i|^2 \right)^{1/2}$, так и в метрике $\|\varphi\|_{C_h} = \max_{i=1,2,\ldots,N-1} |\varphi_i|$. Заметим, что аппроксимация граничных условий здесь осуществляется точно.

Если из разностных уравнений (1.5.2) исключить граничные точки с помощью краевых условий, изменив таким образом оператор задачи (для простоты оставим за ним прежнее обозначение A), то приходим к следующей задаче:

$$\frac{2\varphi_1 - \varphi_2}{h^2} = \frac{a}{h^2} + f_1,$$

$$\frac{-\varphi_{k-1} + 2\varphi_k - \varphi_{k+1}}{h^2} = f_k, \ k = 2, \ldots, N-2, \tag{1.5.3}$$

$$\frac{-\varphi_{N-2} + 2\varphi_{N-1}}{h^2} = \frac{b}{h^2} + f_{N-1}.$$

В результате получим систему $N-1$ уравнений с неизвестными $\varphi_1, \varphi_2, \ldots, \varphi_{N-1}$, которую можно записать в матричной форме

$$A\varphi = g, \tag{1.5.4}$$

где

$$A = \frac{1}{h^2} \begin{Vmatrix} 2 & -1 & 0 & 0 & \ldots & 0 & 0 \\ -1 & 2 & -1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 2 & -1 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & \ldots & -1 & 2 \end{Vmatrix}, \ \varphi = \begin{Vmatrix} \varphi_1 \\ \varphi_2 \\ \ldots \\ \varphi_{N-1} \end{Vmatrix}, \ g = \begin{Vmatrix} g_1 \\ g_2 \\ \ldots \\ g_{N-1} \end{Vmatrix},$$

$$g_k = \begin{cases} f_1 + \dfrac{a}{h^2}, & k = 1, \\ f_k, & k = 2, 3, \ldots, N-2, \\ f_{N-1} + \dfrac{b}{h^2}, & k = N-1. \end{cases}$$

Нетрудно установить, что при $\varphi \neq 0$

$$(A\varphi, \varphi) \geq \gamma^2(\varphi, \varphi),\ \gamma^2 \cong \pi^2.$$

Это обеспечивает однозначность разрешимости задачи и гарантирует устойчивость схемы, а, как следствие, устойчивости и аппроксимации справедлива соответствующая теорема сходимости. Задачу (1.5.3) будем решать с помощью *метода факторизации*. Изложим данный метод в применении к системе

$$\begin{aligned} b_1\varphi_1 + c_1\varphi_2 &= g_1, \\ a_k\varphi_{k-1} + b_k\varphi_k + c_k\varphi_{k+1} &= g_k, \quad k = 2, \ldots, N-2, \\ a_{N-1}\varphi_{N-2} + b_{N-1}\varphi_{N-1} &= g_{N-1}. \end{aligned}$$

Ее решение будем искать в виде $\varphi_k = \beta_{k+1}\varphi_{k+1} + z_{k+1}$. Подставляя эти соотношения в уравнение системы, получаем

$$\begin{gathered} \beta_2 = -c_1/b_1, \quad z_2 = g_1/b_1, \quad \beta_{k+1} = -c_k/(a_k\beta_k + b_k), \\ z_{k+1} = (g_k - a_k z_k)/(a_k\beta_k + b_k), \quad k = 2, \ldots, N-2 \end{gathered} \tag{1.5.5}$$

— *прямой ход метода факторизации*. Затем отыскание решения осуществляется по формулам

$$\begin{aligned} \varphi_{N-1} &= (g_{N-1} - a_{N-1}z_{N-1})/(a_{N-1}\beta_{N-1} + b_{N-1}), \\ \varphi_k &= \beta_{k+1}\varphi_{k+1} + z_{k+1}, \quad k = N-2, \ldots, 1 \end{aligned} \tag{1.5.6}$$

— *обратный ход метода факторизации*.

Если теперь умножить обе части системы (1.5.4) на h^2 и принять

$$\begin{gathered} \varphi_0 \equiv \varphi_N \equiv 0, \quad \beta_2 = 1/2, \quad z_2 = (a + h^2 f_1)/2, \\ \beta_{k+1} = 1/(2 - \beta_k), \quad z_{k+1} = \beta_{k+1}(z_k + h^2 g_k), \end{gathered} \tag{1.5.7}$$

то получим решение системы (1.5.3).

1.5.2. Одномерная задача Неймана

Рассмотрим теперь задачу Неймана

$$-\frac{d^2\varphi}{dx^2} = f,$$

$$\frac{d\varphi}{dx} = a \quad \text{при} \quad x = 0, \tag{1.5.8}$$

$$\frac{d\varphi}{dx} = b \quad \text{при} \quad x = 1,$$

где a и b — заданные константы.

Уравнение из задачи (1.5.8) проинтегрируем по всей области определения решения и используем граничные условия. Тогда приходим к соотношению

$$a - b = \int_0^1 f(x)\, dx, \tag{1.5.9}$$

которое является необходимым условием разрешимости задачи (1.5.8). Это условие в случае $a = b$ означает, что суммарное количество источников субстанции равно нулю, т. е. если где-то имеются источники субстанции, то в другом месте субстанция должна поглощаться.

Итак, пусть задача (1.5.8) разрешима и $\varphi_0(x)$ — ее решение. Можно убедиться, что $\varphi(x) = \varphi_0(x) + C$ (C — произвольная константа) также будет решением, причем этой формулой исчерпывается все семейство решений.

Для получения разностного аналога второго порядка аппроксимации решение задачи при остаточной его гладкости удобно продолжить вне области определения решения $0 \le x \le 1$ еще на один интервал h слева и справа от границ. Это значит, что мы имеем дело с сеточной областью

$$x_k = kh, \quad k = -1, 0, \ldots, N, N+1, \quad h = 1/N.$$

На этой сетке определим аппроксимацию задачи в виде

$$\begin{gathered} \frac{-\varphi_{k-1} + 2\varphi_k - \varphi_{k+1}}{h^2} = f_k, \quad k = 0, 1 \ldots, N, \\ \frac{\varphi_1 - \varphi_{-1}}{2h} = a, \quad \frac{\varphi_{N+1} - \varphi_{N-1}}{2h} = b. \end{gathered} \tag{1.5.10}$$

Подготовительную работу к решению задачи (1.5.8) начнем с исключения граничных условий. С этой целью из задачи (1.5.10) исключим две новые неизвестные φ_{-1} и φ_{N+1} следующим образом. Разрешив граничные условия из (1.5.10) относительно φ_{-1}, φ_{N+1}, находим

$$\varphi_{-1} = \varphi_1 - 2ha, \quad \varphi_{N+1} = \varphi_{N-1} + 2hb. \tag{1.5.11}$$

Полученные значения подставим в разностные уравнения из (1.5.10). Тогда будем иметь

$$\frac{\varphi_0 - \varphi_1}{h^2} = \frac{f_0}{2} - \frac{a}{h},$$

$$\frac{-\varphi_{k-1} + 2\varphi_k - \varphi_{k+1}}{h^2} = f_k, \quad k = 1, 2, \ldots, N-1, \tag{1.5.12}$$

$$\frac{-\varphi_{N-1} + \varphi_N}{h^2} = \frac{f_N}{2} + \frac{b}{h}.$$

Введем в рассмотрение матрицу

$$A = \frac{1}{h^2} \begin{Vmatrix} 1 & -1 & 0 & 0 & \ldots & 0 & 0 \\ -1 & 2 & -1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 2 & -1 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & \ldots & 2 & -1 \\ 0 & 0 & 0 & 0 & \ldots & -1 & 1 \end{Vmatrix} \tag{1.5.13}$$

и векторы

$$\varphi = \begin{Vmatrix} \varphi_0 \\ \varphi_1 \\ \varphi_2 \\ \ldots \\ \varphi_N \end{Vmatrix}, \; g = \begin{Vmatrix} g_0 \\ g_1 \\ g_2 \\ \ldots \\ g_N \end{Vmatrix}, \; g_k = \begin{cases} \dfrac{f_0}{2} - \dfrac{a}{h}, & k = 0, \\ f_k, & k = 1, 2, \ldots, N-1, \\ \dfrac{f_N}{2} + \dfrac{b}{h}, & k = N. \end{cases}$$

Тогда приходим к задаче

$$A\varphi = g. \tag{1.5.14}$$

Первый вопрос, который следует поставить при анализе матрицы A, — это определенность. Матрица A в разностной задаче Неймана сингулярна, поскольку ее наименьшее собственное число обращается в нуль. В этом убедиться совсем просто, рассмотрев спектральную задачу

$$Au = \lambda u,$$

которая имеет своим собственным вектором вектор u_0 с равными компонентами и соответствующее собственное число $\lambda = 0$. Нетрудно показать, далее, что $(A\varphi, \varphi) \geq 0$. Такая задача линейной алгебры обстоятельно изучена (см. также в § 4.5), и установлено необходимое и достаточное условие разрешимости, которое в нашем конкретном случае означает ортогональность g вектору u_0, т. е.

$$\sum_{k=0}^{N} g_k = 0. \tag{1.5.15}$$

Левая часть (1.5.15) может быть записана в виде

$$\sum_{k=0}^{N} g_k = \frac{1}{h}\left[\frac{h}{2}(f_0 + 2f_1 + \ldots + 2f_{N-1} + f_N) - a + b\right]. \tag{1.5.16}$$

Но

$$\frac{h}{2}(f_0 + 2f_1 + \ldots + 2f_{N-1} + f_N)$$

есть известная квадратурная формула трапеций для интеграла

$$\int_0^1 f(x)\, dx.$$

Следовательно, квадратная скобка в (1.5.16) либо равна нулю, либо близка к нулю, что обусловлено погрешностью квадратурной формулы. Чтобы (1.5.14) привести в соответствие с условием разрешимости (1.5.15), нужно подправить возможную «неувязку» и взять, например, в качестве g_k величины $\widetilde{g}_k = g_k - \widetilde{g}$, где

$$\widetilde{g} = \frac{1}{N+1}\sum_{k=0}^{N} g_k.$$

Мы добились того, что

$$\sum_{k=0}^{N} \widetilde{g}_k = 0. \tag{1.5.17}$$

Совместную систему

$$A\varphi = \widetilde{g} \tag{1.5.18}$$

с трехдиагональной матрицей можно решать методом факторизации, если в алгоритм ввести небольшое изменение, обусловленное вырожденностью матрицы A.

На примере рассматриваемой модельной задачи можно легко провести весь анализ метода факторизации и увидеть основные моменты алгоритма.

Итак, обратимся к формулам факторизации. В рассматриваемом конкретном случае

$$\beta_0 = 0, \quad \beta_1 = \beta_2 = \ldots = \beta_N = 1,$$
$$z_0 = 0, \quad z_{k+1} = z_k + \widetilde{g}_k, \quad k = 0, 1, \ldots, N-1.$$

Из последнего равенства легко находим

$$z_{m+1} = \sum_{k=0}^{m} \widetilde{g}_k$$

и, в частности,

$$z_N = \sum_{k=0}^{N-1} \widetilde{g}_k.$$

Наконец,

$$z_{N+1} = \frac{\widetilde{g}_N + z_N}{b_N + a_N\beta_N} = \frac{\widetilde{g}_N + \sum\limits_{k=0}^{N-1} \widetilde{g}_k}{b_N + a_N\beta_N} = \frac{\sum\limits_{k=0}^{N} \widetilde{g}_k}{1-1} = \frac{0}{0}.$$

Таким образом, в алгоритме факторизации встречается неопределенность. Это вполне естественно, если учесть, что $\varphi_N = z_{N+1}$ в силу неоднозначности решения может быть произвольным числом. Итак, в качестве φ_N можно принять любое число и получить искомое решение. Разумеется, об особенности такого сорта следует помнить

при решении задачи Неймана с более сложным оператором, нежели $-d^2/dx^2$.

1.5.3. Двумерное уравнение Пуассона

Приступая к рассмотрению многомерных задач математической физики, с самого начала следует отметить, что проблема аппроксимации таких задач представляет собой задачу нетривиальную. Возникающие при этом проблемы лучше всего пояснить на случае двумерной задачи для уравнения Пуассона. Если граница области определения решения гладкая и достаточную гладкость имеют функции, заданные на границах, то второй порядок аппроксимации уравнения и граничных условий при наличии устойчивости будет гарантировать второй порядок точности решения. Если, однако, либо граница области, либо функции, заданные на границах, оказываются негладкими, то в решении задачи в окрестности особых точек возникают весьма существенные погрешности. В самом деле, пусть областью определения решения является прямоугольник или L-образная область. Тогда, как хорошо известно, в окрестности угловых точек в задачах с эллиптическими операторами обычно возникают особенности либо логарифмического, либо дробного характера. Поэтому равномерная сетка и второй порядок аппроксимации задачи по обеим переменным внутри области еще не обеспечивают решения задачи со вторым порядком точности. В таких случаях используют либо метод сгущения сеток в окрестности особенностей решения, либо метод предварительного выделения особенностей с последующим численным решением регулярной задачи, уже обеспечивающей второй порядок точности. Заметим, однако, что при определенной согласованности граничных условий и правой части уравнения в особых точках границы решение может оказаться гладким, и в этом случае дополнительных проблем аппроксимации уже не возникнет. Эти вопросы будут затронуты в главе 6, и их следует иметь в виду при численной алгоритмизации задач математической физики.

Рассмотрим двумерное уравнение Пуассона

$$\begin{aligned} -\left(\frac{\partial^2\varphi}{\partial x^2}+\frac{\partial^2\varphi}{\partial y^2}\right) &= f \quad \text{в} \quad D, \\ \varphi &= g \quad \text{на} \quad \partial D, \end{aligned} \tag{1.5.19}$$

где областью определения решения D является квадрат $\{0 \le x \le \le 1,\ 0 \le y \le 1\}$. Будем считать, что упомянутая согласованность здесь имеет место и решение задачи φ обладает достаточной гладкостью.

В области D выберем сеть узловых точек (x_k, y_l) на пересечении координатных линий $x = x_k$ и $y = y_l$. Тогда приходим к разностному уравнению

$$\begin{gathered} \frac{\varphi_{k-1,l} - \varphi_{k+1,l} - \varphi_{k,l-1} - \varphi_{k,l+1} + 4\varphi_{k,l}}{h^2} = f_{k,l}, \\ \varphi_{0,l} = a_l, \quad \varphi_{N,l} = b_l, \\ \varphi_{k,0} = c_k, \quad \varphi_{k,N} = d_k, \\ k,\ l = 1, 2, \ldots, N-1, \end{gathered} \tag{1.5.20}$$

где

$$a_l = g_{0,l}, \quad b_l = g_{N,l}, \quad c_k = g_{k,0}, \quad d_k = g_{k,N}$$

(здесь $x_k = kh$, $y_l = lh$).

В разностном уравнении (1.5.20) исключим граничные условия и результат запишем в векторно-матричной форме. С этой целью сначала введем в рассмотрение матрицу A и векторы $\{\varphi_l; g_l\}_{l=1}^{N-1}$:

$$A = \frac{1}{h^2} \begin{Vmatrix} 2 & -1 & 0 & 0 & \ldots & 0 & 0 \\ -1 & 2 & -1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 2 & -1 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & \ldots & -1 & 2 \end{Vmatrix},$$

$$\varphi_l = \begin{Vmatrix} \varphi_{1l} \\ \varphi_{2l} \\ \ldots \\ \varphi_{N-1,l} \end{Vmatrix}, \quad g_l = \begin{Vmatrix} f_{1l} + \dfrac{a_l}{h^2} \\ f_{2l} \\ \ldots \\ f_{N-1,l} + \dfrac{b_l}{h^2} \end{Vmatrix}.$$

Определив матрицу

$$B = h^2 A + 2E,$$

запишем систему уравнений (1.5.20) в виде

$$\frac{1}{h^2}(-\varphi_{l-1} + B\varphi_l - \varphi_{l+1}) = g_l, \quad l = 1, 2, \ldots, N-1, \tag{1.5.21}$$

при условии, что

$$\varphi_0 = c, \quad \varphi_N = d, \tag{1.5.22}$$

где

$$c = \begin{Vmatrix} c_1 \\ c_2 \\ \ldots \\ c_{N-1} \end{Vmatrix}, \; d = \begin{Vmatrix} d_1 \\ d_2 \\ \ldots \\ d_{N-1} \end{Vmatrix}.$$

Из уравнения (1.5.21) исключим граничные условия (1.5.22). Тогда будем иметь

$$\begin{gathered} \frac{1}{h^2}(B\varphi_1 - \varphi_2) = g_1 + \frac{1}{h^2}c, \\ \frac{1}{h^2}(-\varphi_{l-1} + B\varphi_l - \varphi_{l+1}) = g_l, \quad l = 2, 3, \ldots, N-2, \\ \frac{1}{h^2}(-\varphi_{N-2} + B\varphi_{N-1}) = g_{N-1} + \frac{1}{h^2}d. \end{gathered} \tag{1.5.23}$$

Систему уравнений (1.5.23) запишем с помощью блочных матриц и векторов

$$A = \frac{1}{h^2}\begin{Vmatrix} B & -E & 0 & 0 & \ldots & 0 & 0 \\ -E & B & -E & 0 & \ldots & 0 & 0 \\ 0 & -E & B & -E & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & \ldots & -E & B \end{Vmatrix},$$

$$\varphi = \begin{Vmatrix} \varphi_1 \\ \varphi_2 \\ \ldots \\ \varphi_{N-1} \end{Vmatrix}, \; F = \begin{Vmatrix} \frac{1}{h^2} + g_1 \\ g_2 \\ \ldots \\ \frac{1}{h^2} + g_{N-1} \end{Vmatrix},$$

где E — единичная матрица. Тогда система уравнений (1.5.23) примет вид

$$\Lambda\varphi = F. \tag{1.5.24}$$

Матрицу Λ представим в виде суммы $\Lambda = \Lambda_1 + \Lambda_2$ двух матриц

$$\Lambda_1 = \left\| \begin{array}{ccccc} A & 0 & 0 & \dots & 0 \\ 0 & A & 0 & \dots & 0 \\ 0 & 0 & A & \dots & 0 \\ \dots & \dots & \dots & \dots & 0 \\ 0 & 0 & 0 & \dots & A \end{array} \right\|, \; \Lambda_2 = \frac{1}{h^2} \left\| \begin{array}{cccccc} 2E & -E & 0 & 0 & \dots & 0 \\ -E & 2E & -E & 0 & \dots & 0 \\ 0 & -E & 2E & -E & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 2E \end{array} \right\|.$$

Нетрудно установить, что

$$\Lambda_1 \varphi = \left\| \begin{array}{c} A\varphi_1 \\ A\varphi_2 \\ \dots \\ A\varphi_{N-1} \end{array} \right\|, \; \Lambda_2 \varphi = \frac{1}{h^2} \left\| \begin{array}{c} 2\varphi_1 - \varphi_2 \\ \dots \\ -\varphi_{l-1} + 2\varphi_l - \varphi_{l+1} \\ -\varphi_{N-2} + 2\varphi_{N-1} \end{array} \right\|. \tag{1.5.25}$$

Введем новые обозначения векторов:

$$(\Lambda_1 \varphi)_l = A\varphi_l, \quad l = 1, 2, \dots, N-1,$$

$$(\Lambda_2 \varphi)_l = \begin{cases} \frac{1}{h^2}(2\varphi_1 - \varphi_2), & l = 1, \\ \frac{1}{h^2}(-\varphi_{l-1} + 2\varphi_l - \varphi_{l+1}), & l = 2, 3, \dots, N-2, \\ \frac{1}{h^2}(-\varphi_{N-2} + 2\varphi_{N-1}), & l = N-1. \end{cases} \tag{1.5.26}$$

При такой записи очевидно, что выражения $(\Lambda_1 \varphi)_l$ и $(\Lambda_2 \varphi)_k$ являются компонентами векторов $\Lambda_1 \varphi$ и $\Lambda_2 \varphi$, и при этом имеют место весьма удобные покомпонентные их представления

$$(\Lambda_1 \varphi)_l = \frac{1}{h^2} \left\| \begin{array}{c} 2\varphi_{1,l} - \varphi_{2,l} \\ \dots \\ -\varphi_{k-1,l} + 2\varphi_{k,l} - \varphi_{k+1,l} \\ \dots \\ -\varphi_{N-2,l} + 2\varphi_{N-1,l} \end{array} \right\|,$$

$$((\Lambda_2\varphi)_l)_k = \begin{cases} \dfrac{1}{h^2}(2\varphi_{k,l} - \varphi_{k,l+1}), & l = 1, \\ \dfrac{1}{h^2}(-\varphi_{k,l-1} + 2\varphi_{k,l} - \varphi_{k,l+1}), & l = 2, 3, \ldots, N-2, \\ \dfrac{1}{h^2}(-\varphi_{k,l-1} + 2\varphi_{k,l}), & l = N-1, \end{cases}$$
$$k = 1, 2, \ldots, N-1.$$

Аналогичным образом вектор F представляется в форме

$$F = \begin{Vmatrix} F_1 \\ F_2 \\ \ldots \\ F_{N-1} \end{Vmatrix}, \quad F_1 = \begin{Vmatrix} f_{11} + \dfrac{c_1}{h^2} + \dfrac{a_1}{h^2} \\ f_{21} + \dfrac{c_2}{h^2} \\ \\ f_{31} + \dfrac{c_3}{h^2} \\ \ldots \\ f_{N-1,1} + \dfrac{c_{N-1}}{h^2} + \dfrac{b_1}{h^2} \end{Vmatrix}, \quad F_l = \begin{Vmatrix} f_{1l} + \dfrac{a_l}{h^2} \\ f_{2l} \\ f_{3l} \\ \ldots \\ f_{N-1,l} + \dfrac{b_l}{h^2} \end{Vmatrix},$$
$$l = 2, \ldots, N-2,$$
$$F_{N-1} = \begin{Vmatrix} f_{1,N-1} + \dfrac{d_1}{h^2} + \dfrac{a_{N-1}}{h^2} \\ f_{2,N-1} + \dfrac{d_2}{h^2} \\ f_{3,N-1} + \dfrac{d_3}{h^2} \\ \ldots \\ f_{N-1,N-1} + \dfrac{d_{N-1}}{h^2} + d^{\frac{b_{N-1}}{h^2}} \end{Vmatrix}.$$

В результате приходим к покомпонентной записи задачи:

$$(\Lambda_1\varphi)_l + (\Lambda_2\varphi)_l = F_l. \tag{1.5.27}$$

Если теперь потребовать, чтобы индекс l в (1.5.27) принимал все значения $l = 1, 2, \ldots, N-1$, то приходим к окончательной векторно-матричной форме записи задачи (1.5.20):

$$(\Lambda_1 + \Lambda_2)\varphi = F. \tag{1.5.28}$$

В этом случае, как нетрудно убедиться, каждая из компонент уравнения (1.5.28) будет соответствовать разностному уравнению из

(1.5.20), в котором уже учтены заданные граничные значения решения.

Найдем верхнюю и нижнюю границы спектра оператора Λ. Поскольку собственные функции матрицы Λ имеют вид (см. 1.1.4)

$$u_{kl}^{mp} = \sin(m\pi kh)\sin(p\pi lh), \quad k,l,m,p = 1,2,\ldots,N-1, \tag{1.5.29}$$

то

$$\lambda_{mp} = \frac{4}{h^2}\left(\sin^2\frac{m\pi h}{2} + \sin^2\frac{p\pi h}{2}\right). \tag{1.5.30}$$

Откуда следует, что

$$\alpha = \frac{8}{h^2}\sin^2\frac{\pi h}{2}, \quad \beta = \frac{8}{h^2}\cos^2\frac{\pi h}{2}, \tag{1.5.31}$$

где

$$\alpha = \alpha(\Lambda), \quad \beta = \beta(\Lambda).$$

Если мы имеем дело с двумерной задачей Неймана, то аналогично предыдущему приходим к задаче (1.5.28), которая отличается от рассматриваемой тем, что число компонент решения, подлежащих определению, будет не $(N-1)^2$, как это имело место в задаче Дирихле, а $(N+1)^2$. Спектр задачи Неймана для разностного оператора $A = -\Delta$ находится по формуле

$$\lambda_{mp} = \frac{4}{h^2}\left(\sin^2\frac{m\pi}{2(N+1)} + \sin^2\frac{p\pi}{2(N+1)}\right), \quad m,p = 0,1,\ldots,N, \tag{1.5.32}$$

и в качестве границ спектра имеем

$$\alpha^*(\Lambda) = \frac{8}{h^2}\sin^2\frac{\pi}{2(N+1)}, \quad \beta(\Lambda) = \frac{8}{h^2}\cos^2\frac{\pi}{2(N+1)}. \tag{1.5.33}$$

1.5.4. Проблема граничных условий

На основе рассмотренных выше подходов к решению уравнения Пуассона можно сделать важное общее заключение по поводу построения эффективных алгоритмов решения краевых задач математической физики. Это прежде всего относится к проблеме граничных условий. Выше была последовательно проведена идея исключения граничных условий, налагаемых в качестве дополнительных связей на решение задачи, и модификации их с учетом разност-

ных аналогов исследуемых задач. В таком подходе заложен глубокий смысл: если при решении задач математической физики разностными методами граничные условия из рассмотрения исключены, то для формирования того или иного вычислительного алгоритма уже не приходится специально заботиться об удовлетворении граничных условий, поскольку они автоматически учитываются в модифицированных разностных уравнениях. Это важно при решении стационарных задач и особенно при решении задач нестационарных, проблема граничных условий в которых требует тщательного анализа. Именно поэтому в главе 5 мы откажемся от расщепления нестационарных задач на простейшие в дифференциальной формулировке, поскольку это потребовало бы дополнительных исследований постановки граничных условий, согласованных с расщепленной системой. Более просто, с нашей точки зрения, поставить в соответствие исходной задаче математической физики систему разностных уравнений по пространственным переменным и из этой системы исключить граничные значения функций, используя разностные аналоги краевых условий задачи, согласованных по точности с самим разностным уравнением. И только после этого можно проводить аппроксимацию уравнения по времени на основе методов расщепления или других алгоритмов. Этот прием позволяет избежать согласования граничных условий на каждом этапе построения вычислительного алгоритма с помощью схем расщепления.

Обратим внимание на следующий факт, тесно связанный с проблемой граничных условий. При решении некоторых задач математической физики с помощью разностных схем оказывается целесообразным воспользоваться методом представления решения в виде ряда Фурье по собственным элементам оператора разностной задачи. Этим приемом мы уже пользовались неоднократно при изучении свойств вычислительных алгоритмов. Однако для применения этого метода необходимо, чтобы разностная задача замыкалась с помощью однородных граничных условий. Если граничные условия неоднородны, то требуется предварительная трансформация задачи к виду, где граничные условия являлись бы однородными. Такую трансформацию удобно провести методом, построение которого проиллюстрируем на простейшем примере задачи (1.5.1). С этой целью рассмотрим систему разностных уравнений (1.5.2). Область определе-

ния решения $\{x_k\}$ $(k = 1, 2, \ldots, N-1)$ расширим на две узловые точки $x = 0$ $(k = 0)$ и $x = 1$ $(k = N)$ и доопределим решение в этих точках следующим образом:

$$\varphi_0 = 0, \quad \varphi_N = 0. \tag{1.5.34}$$

Такое доопределение, конечно, носит только формальный характер и не имеет никакого отношения к фактическому значению решения в точках $x = 0$ и $x = 1$, определяемых неоднородными граничными условиями (1.5.1). Это означает, что после решения задачи мы должны снова сузить область решения, исключая из нее точки $k = 0$ и $k = N$, в которых решение определено в форме (1.5.1). Такой метод позволяет вместо задачи (1.5.2) рассмотреть следующую, эквивалентную ей задачу:

$$\frac{-\varphi_{k-1} + 2\varphi_k - \varphi_{k-1}}{h^2} = g_k, \quad k = 1, 2, \ldots, N-1, \tag{1.5.35}$$

$$\varphi_0 = 0, \quad \varphi_N = 0, \tag{1.5.36}$$

где

$$g_k = \begin{cases} \dfrac{a}{h^2} + f_1, & k = 1, \\ f_k, & k = 2, 3, \ldots, N-2, \\ \dfrac{b}{h^2} + f_{N-1}, & k = N-1. \end{cases} \tag{1.5.37}$$

Таким образом, задача (1.5.1) с неоднородными граничными условиями ценой некоторого изменения правой части разностных уравнений свелась к задаче однородной, подготовленной к решению с помощью ряда Фурье.

В случае неоднородной задачи Неймана (1.5.10) область определения решения следует расширить, присоединив к ней узловые точки x_{-1} и x_{N+1}. В результате приходим к задаче

$$\frac{-\varphi_{k-1} + 2\varphi_k - \varphi_{k-1}}{h^2} = g_k, \quad k = 0, 1, 2, \ldots, N, \tag{1.5.38}$$

$$\varphi_{-1} = \varphi_1, \quad \varphi_{N+1} = \varphi_{N-1}, \tag{1.5.39}$$

где

$$g_k = \begin{cases} f_0 - \dfrac{2a}{h}, & k = 0, \\ f_k, & k = 1, 2, \ldots, N-1, \\ f_N + \dfrac{2b}{k}, & k = N. \end{cases} \tag{1.5.40}$$

После того как решение задачи (1.5.38), (1.5.39) найдено, следует из него исключить вспомогательные компоненты φ_{-1} и φ_{N+1}.

Аналогичный прием может быть применен и для многомерных стационарных и нестационарных задач.

Конечно, при необходимости с системами уравнений (1.5.35)—(1.5.39) можно поступать так же, как и с неоднородными задачами (1.5.2) и (1.5.10), исключив из уравнений граничные значения функций. В этом случае мы снова приходим к рассмотренным задачам (1.5.3) и (1.5.12).

В дальнейшем мы не будем рассматривать неоднородные краевые условия, поскольку описанные выше алгоритмы позволяют либо совсем исключить их из рассмотрения, либо свести их к однородным.

1.5.5. Уравнение теплопроводности

Задача теплопроводности является одной из типичных нестационарных задач математической физики. Исторически сложилось так, что именно благодаря уравнению теплопроводности были поставлены и решены многие принципиальные вопросы теории вычислений и построены первоклассные алгоритмы решения задач математической физики. Поскольку проблеме численного решения уравнения теплопроводности посвящена серия специальных монографий и оригинальных статей, мы ограничимся рассмотрением некоторых методов, получивших наибольшее применение в приложениях.

Сначала рассмотрим простейшую задачу о распространении тепла в однородном ограниченном стержне, нагреваемом за счет внутренних источников, с источниками и стоками тепла на границах. В этом случае имеем следующую задачу:

$$\frac{1}{c^2}\frac{\partial \varphi}{\partial t} = \frac{\partial^2 \varphi}{\partial x^2} + f(x,t), \tag{1.5.41}$$

$$\varphi(0,t) = a(t), \quad \varphi(1,t) = b(t), \tag{1.5.42}$$

$$\varphi(x,0) = \varphi^0(x), \tag{1.5.43}$$

где f, a, b и φ^0 — заданные достаточно гладкие функции, $c = \text{const}$, переменные x и t пробегают все значения из области определения решения $D \times D_t = \{0 \le x \le 1, 0 \le t \le T\}$.

Решение задачи (1.5.41)—(1.5.43) будем искать с помощью метода конечных разностей. С этой целью, прежде всего, аппроксимируем эту задачу по x со вторым порядком точности. Пусть, как и для задачи (1.5.1), интервал $0 \le x \le 1$ разбит точками x_k $(k = 1, 2, \ldots, N-1)$ на N интервалов длины $h = 1/N$. Тогда приходим к следующей задаче, аналогичной (1.5.2):

$$\frac{1}{c^2}\frac{d\varphi_k}{dt} + \frac{-\varphi_{k-1} + 2\varphi_k - \varphi_{k+1}}{h^2} = f_k(t), \tag{1.5.44}$$

$$\varphi_0 = a(t), \quad \varphi_N = b(t), \tag{1.5.45}$$

$$\varphi_k = \varphi_k^0 \quad \text{при} \quad t = 0, \tag{1.5.46}$$

$$k = 1, 2, \ldots, N-1,$$

или в векторно-матричном виде

$$\frac{1}{c^2}\frac{d\varphi}{dt} + A\varphi = g, \tag{1.5.47}$$

$$\varphi = \varphi^0 \quad \text{при} \quad t = 0, \tag{1.5.48}$$

где A — положительная матрица, а $\varphi(t)$ и $g(t)$ — вектор-функции, определенные для любого t в задаче (1.5.4).

Используя соображения, высказанные в 1.5.4, систему уравнений (1.5.49)—(1.5.51) можно записать в виде

$$\frac{1}{c^2}\frac{d\varphi_k}{dt} + \frac{-\varphi_{k-1} + 2\varphi_k - \varphi_{k+1}}{h^2} = g_k, \tag{1.5.49}$$

$$\varphi_0 = 0, \quad \varphi_N = 0, \tag{1.5.50}$$

$$\varphi_k = \varphi_k^0 \quad \text{при} \quad t = 0, \tag{1.5.51}$$

где g_k определены формулами (1.5.37).

В этом случае, как было отмечено в 1.5.4, решение задачи (1.5.49)—(1.5.51) имеет смысл только в узловых точках $x_1, x_2, \ldots, x_{N-1}$.

Займемся теперь решением системы обыкновенных дифференциальных уравнений (1.5.49) по времени на интервале $t_j \le t \le t_{j+1}$. Получим

$$\frac{\varphi_k^{j+1} - \varphi_k^j}{\tau} = \frac{\overline{\varphi}_{k-1}^j - 2\overline{\varphi}_k^j + \overline{\varphi}_{k+1}^j}{h^2} + \overline{g}_k^j, \tag{1.5.52}$$

где

$$\overline{\varphi}_k^j = \frac{1}{\Delta t}\int\limits_{t_j}^{t_{j+1}} \varphi_k \, dt, \quad \overline{g}_k^j = \frac{1}{\Delta t}\int\limits_{t_j}^{t_{j+1}} g_k \, dt. \tag{1.5.53}$$

Кроме того, здесь приняты обозначения

$$\varphi_k^j = \varphi_k(t_j), \quad \tau = c^2\Delta t, \quad \Delta t = t_{j+1} - t_j, \quad j = 0, 1, \ldots. \tag{1.5.54}$$

Различные разностные уравнения будем получать на основе тех или иных аппроксимаций в (1.5.53). Предположим, что имеет место одна из трех следующих простейших формул:

$$\frac{1}{\Delta t}\int\limits_{t_j}^{t_{j+1}} \varphi_k \, dt \simeq \begin{cases} \varphi_k^j, \\ \varphi_k^{j+1}, \\ \frac{1}{2}(\varphi_k^j + \varphi_k^{j+1}); \end{cases} \qquad \frac{1}{\Delta t}\int\limits_{t_j}^{t_{j+1}} g_k \, dt \simeq \begin{cases} g_k^j, \\ g_k^{j+1}, \\ \dfrac{1}{2}(g_k^j + g_k^{j+1}). \end{cases} \tag{1.5.55}$$

Тогда приходим к наиболее распространенным разностным схемам:

явной схеме треугольника (. : .)

$$\frac{\varphi_k^{j+1} - \varphi_k^j}{\tau} = \frac{\varphi_{k-1}^j - 2\varphi_k^j + \varphi_{k+1}^j}{h^2} + g_k^j; \tag{1.5.56}$$

неявной схеме треугольника (· : ·)

$$\frac{\varphi_k^{j+1} - \varphi_k^j}{\tau} = \frac{\varphi_{k-1}^{j+1} - 2\varphi_k^{j+1} + \varphi_{k+1}^{j+1}}{h^2} + g_k^{j+1}; \tag{1.5.57}$$

схеме Кранка — Николсона (:::)

$$\frac{\varphi_k^{j+1} - \varphi_k^j}{\tau} = \frac{\varphi_{k-1}^{j+1} - 2\varphi_k^{j+1} + \varphi_{k+1}^{j+1}}{2h^2} + \frac{\varphi_{k-1}^j - 2\varphi_k^j + \varphi_{k+1}^j}{2h^2} + \frac{g_k^{j+1} + g_k^j}{2}. \tag{1.5.58}$$

К системам уравнений (1.5.56)—(1.5.58) необходимо присоединить граничные условия

$$\varphi_0^j = 0, \quad \varphi_N^j = 0. \tag{1.5.59}$$

Схема треугольника (. : .) явно разрешается относительно неизвестной

$$\varphi_k^{j+1} = \varphi_k^j + \mu(\varphi_{k-1}^j - 2\varphi_k^j + \varphi_{k+1}^j) + \tau g_k^j, \tag{1.5.60}$$

$$\varphi_0^j = 0, \quad \varphi_N^j = 0, \quad \mu = \frac{\tau}{h^2}. \tag{1.5.61}$$

Неявная схема треугольника ($\cdot$: $\cdot$) имеет более сложную реализацию, сводящуюся к решению разностного уравнения

$$\varphi_{k-1}^{j+1} + (2 + \frac{1}{\mu})\varphi_k^{j+1} - \varphi_{k+1}^{j+1} = \frac{1}{\mu}\varphi_k^j + h^2 g_k^{j+1}, \tag{1.5.62}$$

$$\varphi_0^{j+1} = 0, \quad \varphi_N^{j+1} = 0. \tag{1.5.63}$$

Наконец, схема Кранка — Николсона (::) имеет следующий вид:

$$-\xi_{k-1}^{j+1} + \frac{2(1+\mu)}{\mu}\xi_k^{j+1} - \xi_{k+1}^{j+1} = \frac{2}{\mu}\varphi_k^j + h^2 g_k^{j+1/2}, \tag{1.5.64}$$

$$\xi_0^{j+1} = 0, \quad \xi_N^{j+1} = 0, \tag{1.5.65}$$

$$\varphi_k^{j+1} = 2\xi_k^{j+1} - \varphi_k^j, \tag{1.5.66}$$

где

$$g_k^{j+1/2} = \frac{1}{2}(g_k^{j+1} + g_k^j).$$

Задачи (1.5.62), (1.5.63) эффективно решаются с помощью метода факторизации.

Исследуем устойчивость разностных схем (1.5.56)—(1.5.58) при условиях (1.5.59). С этой целью разложим решение в ряд Фурье по полной системе функций $\{\sin(n\pi kh)\}$, где $h = 1/N$, удовлетворяющих условиям (1.5.59). Пусть

$$\varphi_k^j = \sum_{n=1}^{N-1} \Phi_n^j \sin(n\pi kh), \quad g_k^j = \sum_{n=1}^{N-1} G_n^j \sin(n\pi kh), \tag{1.5.67}$$

где

$$\Phi_n^j = \frac{1}{q_n}\sum_{k=1}^{N-1} \varphi_k^j \sin(n\pi kh), \quad G_n^j = \frac{1}{q_n}\sum_{k=1}^{N-1} g_k^j \sin(n\pi kh),$$

$$q_n = \sum_{k=1}^{N-1} \sin^2(n\pi kh).$$

Подставим (1.5.67) в (1.5.56)—(1.5.58), полученные соотношения умножим на $\sin(n\pi lh)$ и просуммируем по l. В результате приходим к рекуррентным соотношениям для коэффициентов Фурье. Для схемы $(.:.)$

$$\frac{\Phi_n^{j+1} - \Phi_n^j}{\tau} + \lambda_n \Phi_n^j = G_n^j; \tag{1.5.68}$$

для схемы $(\cdot:\cdot)$

$$\frac{\Phi_n^{j+1} - \Phi_n^j}{\tau} + \lambda_n \Phi_n^{j+1} = G_n^{j+1}, \tag{1.5.69}$$

$$\frac{\Phi_n^{j+1} - \Phi_n^j}{\tau} + \lambda_n \frac{\Phi_n^{j+1} + \Phi_n^j}{2} = \frac{G_n^{j+1} + G_n^j}{2}. \tag{1.5.70}$$

Здесь

$$\lambda_n = \frac{4}{h^2} \sin^2 \frac{n\pi h}{2}. \tag{1.5.71}$$

Решая уравнения (1.5.68)—(1.5.70), получим соответственно

$$\Phi_n^{j+1} = (1 - \tau\lambda_n)\Phi_n^j + \tau G_n^j,$$

$$\Phi_n^{j+1} = \frac{1}{1 + \tau\lambda_n}\Phi_n^j + \frac{\tau}{1 + \tau\lambda_n} G_n^{j+1}, \tag{1.5.72}$$

$$\Phi_n^{j+1} = \frac{1 - \dfrac{\tau}{2}\lambda_n}{1 + \dfrac{\tau}{2}\lambda_n}\Phi_n^j + \frac{\tau}{2}\frac{G^{j+1} + G_n^j}{1 + \dfrac{\tau}{2}\lambda_n}.$$

Отметим, что

$$\frac{4}{h^2} \sin^2 \frac{\pi h}{2} \le \lambda_n \le \frac{4}{h^2} \cos^2 \frac{\pi h}{2} < \frac{4}{h^2}. \tag{1.5.73}$$

При $\pi h/2 \ll 1$ неравенства (1.5.73) превращаются в асимптотические:

$$\pi^2 \le \lambda_n \le \frac{4}{h^2}. \tag{1.5.74}$$

Учитывая (1.5.73), приходим к выводу, что при

$$\tau \le \frac{h^2}{2\cos^2 \dfrac{\pi h}{2}} \tag{1.5.75}$$

схема треугольника (. : .) будет устойчивой, поскольку для всех n имеет место неравенство

$$|1 - \tau\lambda_n| \leq 1. \tag{1.5.76}$$

Условие (1.5.75) можно заменить достаточным:

$$\tau \leq \frac{h^2}{2}. \tag{1.5.77}$$

В случае неявной схемы треугольника и схемы Кранка — Николсона соответствующие неравенства

$$\left|\frac{1}{1+\tau\lambda_n}\right| < 1 \quad \left|\frac{1-\frac{\tau}{2}\lambda_n}{1+\frac{\tau}{2}\lambda_n}\right| < 1,$$

выполняются при любых h и $\tau > 0$. Это значит, что схема (.:.) устойчива при условии (1.5.75), а схемы ($\cdot : \cdot$) и (:::) абсолютно устойчивы.

Наибольшее распространение в расчетах получила схема Кранка — Николсона, которая является схемой второго порядка аппроксимации.

В случае задачи Неймана для уравнения теплопроводности оператор A и вектор-функция g определяются так же, как и в уравнении (1.5.14). В результате приходим к уравнениям, аналогичным (1.5.52), (1.5.53), решение которых не представляет труда. Вместе с тем необходимо сделать следующее замечание. Если при решении задачи Неймана для уравнения Лапласа гармоника решения, соответствующая собственному числу $\lambda = 0$, была паразитической и «отфильтровывалась» от приближенного решения, то в задаче распространения тепла эта гармоника уже необходима в решении, поскольку она описывает общее повышение или понижение температуры стержня за счет внешних источников и стоков.

1.5.6. Уравнение колебаний

В приложении существенное место занимают уравнения гиперболического типа, численные методы решения которых изучены достаточно полно. Отличительной чертой уравнений гиперболического типа является то, что область зависимости решения таких уравнений ограничена характеристическим конусом, так что область про-

странства $D \times D_t$, расположенная вне этого конуса, не влияет на решение в рассматриваемой точке. Другой особенностью решений гиперболических уравнений является то, что решения исходной задачи, как правило, допускают существование и негладких решений. Последнее обстоятельство особенно важно иметь в виду при разработке численных схем.

Рассмотрим одномерную задачу, описывающую малые колебания однородной струны:

$$\frac{1}{c^2}\frac{\partial^2 \varphi}{\partial t^2} = \frac{\partial^2 \varphi}{\partial x^2} + f(x,t),$$

$$\varphi(0,t) = a(t), \quad \varphi(1,t) = b(t), \tag{1.5.78}$$

$$\varphi(x,0) = p(x), \quad \frac{\partial \varphi}{\partial t}(x,0) = q(x),$$

где c — скорость распространения возмущений вдоль однородной струны; $a(t)$, $b(t)$, $f(x,t)$, $p(x)$ и $q(x)$ — заданные функции.

Одномерная теория колебаний изучена достаточно полно. Нашей задачей является обсуждение вопросов численного решения (1.5.78). С этой целью, прежде всего, проведем редукцию задачи к системе обыкновенных дифференциальных уравнений по времени с помощью разностных соотношений по переменной x. Методом, изложенным в 1.5.5, задачу (1.5.78) аппроксимируем следующей:

$$\frac{1}{c^2}\frac{\partial^2 \varphi_k}{\partial t^2} + \frac{-\varphi_{k-1} + 2\varphi_k - \varphi_{k+1}}{h^2} = g_k(t),$$

$$\varphi_0 = 0, \quad \varphi_N = 0, \tag{1.5.79}$$

$$\varphi_k = p_k, \quad \frac{\partial \varphi_k}{\partial t} = q_k \quad \text{при} \quad t = 0,$$

где функция $g_k(t)$, как и в (1.5.57), определена формулами (1.5.4).

Решение задачи (1.5.79) $\varphi_k(t)$, как подчеркнуто ранее, имеет смысл только в узловых точках $k = 1, 2, \ldots, N-1$. В точках $k = 0$ и $k = N$ решение определяется граничными условиями из (1.5.78), а однородные условия в (1.5.79) являются результатом специального расширения области определения решения и не являются аппрокси-

мациями функций $\varphi(x,t)$ в точках x_0 и x_N. Заметим, что задача (1.5.79) имеет второй порядок аппроксимации по $h = \Delta x$.

Для конечно-разностной аппроксимации задачи (1.5.79) по времени введем систему узловых точек $t = t_j$ (причем $t_{j+1} = t_j + \Delta t$) и рассмотрим две наиболее употребительные схемы:

явную схему «крест» ($\cdot\vdots\cdot$)

$$\frac{\varphi_k^{j+1} - 2\varphi_k^j + \varphi_k^{j-1}}{\tau^2} = \frac{\varphi_{k-1}^j - 2\varphi_k^j + \varphi_{k+1}^j}{h^2} + g_k^j,$$

$$\varphi_0^j = 0, \quad \varphi_N^j = 0, \quad \varphi_k^0 = p_k, \tag{1.5.80}$$

$$\varphi_k^1 = p_k + \Delta t q_k + \frac{\tau^2}{2}\left(\frac{p_{k-1} - 2p_k + p_{k+1}}{h^2} + g_k^0\right),$$

где $\tau = c\Delta t$, и неявную схему, аналогичную схеме Кранка — Николсона, ($\vdots\vdots\vdots$)

$$\frac{\varphi_k^{j+1} - 2\varphi_k^j + \varphi_k^{j-1}}{\tau^2} = \frac{\varphi_{k-1}^{j+1} - 2\varphi_k^{j+1} + \varphi_{k+1}^{j+1}}{2h^2} + \frac{\varphi_{k-1}^{j-1} - 2\varphi_k^{j-1} + \varphi_{k+1}^{j-1}}{2h^2} + g_k^j,$$

$$\varphi_0^{j+1} = 0, \quad \varphi_N^{j+1} = 0, \quad \varphi_k^0 = p_k,$$

$$\varphi_k^1 = \varphi_k^0 + \Delta t q_k + \frac{\tau^2}{2}\left(\frac{p_{k-1} - 2p_k + p_{k+1}}{h^2} + g_k^0\right). \tag{1.5.81}$$

Разложением в ряд Тейлора устанавливаем, что в случае достаточно гладких решений $\varphi(x,t)$ разностные схемы (1.5.80) и (1.5.81) аппроксимируют исходную задачу со вторым порядком как по h, так и по τ.

Решение задач (1.5.80), (1.5.81) будем искать с помощью ряда Фурье по системе собственных функций

$$u_n(k) = \sin(n\pi kh), \quad k, n = 1, 2, \ldots, N-1,$$

аналогично тому, как это было в (1.5.67). Тогда для коэффициентов Фурье приходим к рекуррентным соотношениям:

для явной схемы «крест» (·:·)

$$\frac{\Phi_n^{j+1} - 2\Phi_n^j + \Phi_n^{j-1}}{\tau^2} + \frac{4}{h^2}\sin^2\frac{n\pi h}{2}\Phi_n^j = G_n^j,$$

$$\Phi_n^0 = P_n, \tag{1.5.82}$$

$$\Phi_n^1 = P_n + \Delta t Q_n - \frac{2\tau^2}{h^2}\sin^2\frac{n\pi h}{2}P_n + \frac{\tau^2}{2}G_n^0,$$

где Φ_n, G_n, P_n и Q_n — коэффициенты Фурье элементов φ_k, g_k, p_k и q_k соответственно;

для неявной схемы (:::)

$$\frac{\Phi_n^{j+1} - 2\Phi_n^j + \Phi_n^{j-1}}{\tau^2} + \frac{4}{h^2}\sin^2\frac{n\pi h}{2}\frac{\Phi_n^{j+1} + \Phi_n^{j-1}}{2} = G_n^j,$$

$$\Phi_n^0 = P_n, \tag{1.5.83}$$

$$\Phi_n^1 = P_n + \Delta t Q_n - \frac{2\tau^2}{h^2}\sin^2\frac{n\pi h}{2}P_n + \frac{\tau^2}{2}G_n^0.$$

С целью установления критерия счетной устойчивости для задач с гладкими входными данными изучим поведение линейно независимых решений однородных задач (1.5.82) и (1.5.83) в зависимости от индекса i. Решение будем искать в виде

$$\Phi_n^j = A_n\eta_n^j, \tag{1.5.84}$$

где A_n и η_n — константы, причем здесь η_n возводится в степень с показателем j. Полагая в (1.5.82) и (1.5.83) $G_n^j = 0$ и заменяя в них Φ_n^j по формулам (1.5.84), приходим к следующим выражениям для η_n:

в случае схемы «крест» (·:·)

$$\eta_n^2 - 2(1 - \mu_n^2)\eta_n + 1 = 0; \tag{1.5.85}$$

в случае неявной схемы (:::)

$$\eta_n^2 - \frac{2}{1 + \mu_n^2}\eta_n + 1 = 0, \tag{1.5.86}$$

где $\mu_n^2 = 2\dfrac{\tau^2}{h^2}\sin^2\dfrac{n\pi h}{2}$. Решая квадратные уравнения (1.5.85) и (1.5.86), получим соответственно

$$\eta_n = 1 - \mu_n^2 \pm \sqrt{(1-\mu_n^2)^2 - 1}, \tag{1.5.87}$$

$$\eta_n = \frac{1}{1+\mu_n^2} \pm \sqrt{\left(\frac{1}{1+\mu_n^2}\right)^2 - 1}. \tag{1.5.88}$$

Рассмотрим сначала решение для схемы «крест». Легко видеть, что если

$$\mu_n^2 = 2\frac{\tau^2}{h^2}\sin^2\frac{n\pi h}{2} < 2, \tag{1.5.89}$$

то

$$|\eta_{n_i}| = 1, \quad |\eta_{n_1} - \eta_{n_2}| > C\tau, \tag{1.5.90}$$

и, следовательно, разностная схема будет устойчивой. Необходимо, чтобы условие (1.5.89) выполнялось для всех значений $n = 1, 2, \ldots, N-1$. Это условие, очевидно, выполняется для любых n, если τ и h связаны зависимостью

$$\frac{\tau^2}{h^2} < \frac{1}{\sin^2\dfrac{\pi n h}{2}}, \tag{1.5.91}$$

или, более просто, τ и h связаны условием Куранта

$$\frac{\tau}{h} < 1. \tag{1.5.92}$$

Нетрудно установить, что в случае неявной схемы (::) при любых n и $\tau > 0$ имеет место равенство

$$|\eta_n| = 1. \tag{1.5.93}$$

Это значит, что данная разностная схема абсолютно устойчива.

Аналогичным образом может быть рассмотрена задача о малом колебании мембраны. В этом случае мы приходим к уравнению

$$\frac{1}{c^2}\frac{\partial^2\varphi}{\partial t^2} = \frac{\partial^2\varphi}{\partial x^2} + \frac{\partial^2\varphi}{\partial y^2} + f(x,y,t) \quad \text{в} \quad D \times D_t \tag{1.5.94}$$

при краевом условии

$$\varphi = g \quad \text{на} \quad \partial D \times D_t \tag{1.5.95}$$

и начальных данных

$$\varphi = p, \quad \frac{\partial \varphi}{\partial t} = q \quad \text{в} \quad D \quad \text{при} \quad t = 0.$$

Решение этой задачи в квадрате D при достаточно гладких входных данных находится с помощью метода конечных разностей, как и в случае уравнения малых колебаний струны.

1.5.7. Уравнение движения

К числу основных уравнений математической физики следует отнести уравнение переноса субстанции вдоль траекторий частиц, которое можно представить в виде

$$\frac{d\varphi}{dt} = 0,$$

где

$$\frac{d}{dt} = \frac{\partial}{\partial t} + u\frac{\partial}{\partial x} + v\frac{\partial}{\partial y} + w\frac{\partial}{\partial z}$$

— полная производная от функции $\varphi(x, y, z, t)$ по времени, а u, v и w — компоненты вектора скорости $u = \overrightarrow{ui} + \overrightarrow{vj} + \overrightarrow{wk}$, причем

$$u = \frac{dx}{dt}, \quad u = \frac{dy}{dt}, \quad u = \frac{dz}{dt}.$$

Рассмотренное уравнение решается при дополнительных условиях, простейшим из которых для неограниченной среды будет следующее:

$$\varphi(x, y, z, 0) = f(x, y, z).$$

Аналогичная задача возникает в качестве элемента общего алгоритма при численном решении уравнений гидродинамики, теории переноса излучения и многих других. Учитывая это обстоятельство, проведем подробное обсуждение возможных путей численного решения задач такого вида.

При решении задач гидродинамики, гидротермодинамики, прогноза погоды, динамики океана и других приходится иметь дело с уравнениями переноса субстанций вдоль траекторий. Простейшим

уравнением такого вида является уравнение

$$\begin{gathered}\frac{\partial\varphi}{\partial t}+u\frac{\partial\varphi}{\partial x}=0 \quad \text{в} \quad D\times D_t,\\ \varphi=f(x) \quad \text{при} \quad t=0,\end{gathered} \tag{1.5.96}$$

где u — заданная скорость, а $f(x)$ — начальное распределение φ. Областью D является вся вещественная ось x, а $D_t=\{0\le t\le T\}$. Предполагаем, что $\varphi(x,t)$ и $f(x)$ являются периодическими функциями по x с периодом 2π. Если $u=\mathrm{const}$, то задача (1.5.96) имеет очевидное решение

$$\varphi(x,t)=f(x-ut) \tag{1.5.97}$$

при условии, что $f(x)$ есть дифференцируемая функция. Решение (1.5.97) описывает процесс распространения начального возмущения вдоль характеристик

$$x-ut=\mathrm{const}\,.$$

Это значит, что $\varphi(x,t)=\mathrm{const}$ на любой прямой $x-ut=const$.

Итак, задача (1.5.96) при $u>0$ определяет процесс распространения возмущения в сторону возрастающих значений x. Эти хорошо известные положения следует иметь в виду при конструировании разностных аналогов задачи (1.5.96). Если скорость $u=u(x,t)$ — переменная, то нахождение решения задачи (1.5.96) в аналитическом виде уже затруднительно. Именно в этих случаях оказывается необходимым применение численных методов, основанных на разностных аппроксимациях.

Рассмотрим простейшие разностные схемы с $u=\mathrm{const}$. Ради определенности будем полагать $u>0$. Тогда имеем явную схему

$$\frac{\varphi_k^{j+1}-\varphi_k^j}{\tau}+u\frac{\varphi_k^j-\varphi_{k-1}^j}{\Delta x}=0 \tag{1.5.98}$$

и неявную

$$\frac{\varphi_k^{j+1}-\varphi_k^j}{\tau}+u\frac{\varphi_k^{j+1}-\varphi_{k-1}^{j+1}}{\Delta x}=0. \tag{1.5.99}$$

Обе схемы имеют первый порядок аппроксимации по Δx и τ. В самом деле, предположим, что начальное значение $f(x)$ и решение $\varphi(x,t)$ — достаточно гладкие функции. Решение уравнения (1.5.96)

разложим в ряд Тейлора в окрестности точки $x = x_k$, $t = t_j$:

$$\varphi(x,t) = (\varphi)_k^j + (\varphi_t)_k^j (t - t_j) + (\varphi_x)_k^j (x - x_k) + \dots . \tag{1.5.100}$$

Подставляя ряд (1.5.100) в (1.5.98), получим, в частности,

$$\frac{\partial \varphi}{\partial t} + u\frac{\partial \varphi}{\partial x} = \frac{u\Delta x}{2}\frac{\partial^2 \varphi}{\partial x^2} - \frac{\tau}{2}\frac{\partial^2 \varphi}{\partial t^2} \quad \text{при} \quad x = x_k, \quad t = t_j. \tag{1.5.101}$$

Отброшенные члены в (1.5.101) имеют более высокий порядок малости. Из уравнения (1.5.96) следует, что

$$\frac{\partial^2 \varphi}{\partial t^2} = u^2 \frac{\partial^2 \varphi}{\partial x^2}. \tag{1.5.102}$$

Тогда выражение (1.5.101) примет вид

$$\frac{\partial \varphi}{\partial t} + u\frac{\partial \varphi}{\partial x} = \frac{u\Delta x - \tau u^2}{2}\frac{\partial^2 \varphi}{\partial x^2} \quad \text{при} \quad x = x_k,\ t = t_j. \tag{1.5.103}$$

Анализ соотношения (1.5.103) показывает, что при

$$\frac{u\tau}{\Delta x} < 1$$

соотношение (1.5.103) можно толковать как уравнение теплопроводности с областью определения решения $x_{k-1} \le x \le x_k,\ t_j \le t \le t_{j+1}$. Если предположить, что отброшенные члены в (1.5.103) малы, то в результате мы приходим к уравнению

$$\frac{\partial \varphi}{\partial t} + u\frac{\partial \varphi}{\partial x} = \mu \frac{\partial^2 \varphi}{\partial x^2},$$

где

$$\mu = \frac{u\Delta x - \tau u^2}{2}$$

является так называемым *коэффициентом искусственной, или «счетной», вязкости*. Заметим, кстати, что если

$$\frac{u\tau}{\Delta x} = 1,$$

то μ и все отброшенные слагаемые будут равны нулю; явная схема (1.5.98) оказывается схемой бесконечного порядка аппроксимации по Δx и τ.

Особо следует отметить случай, когда

$$\frac{u\tau}{\Delta x} > 1.$$

В этом случае мы приходим к уравнению

$$\frac{\partial \varphi}{\partial t} + u\frac{\partial \varphi}{\partial x} = -|\mu|\frac{\partial^2 \varphi}{\partial x^2}. \tag{1.5.104}$$

Легко видеть, что уравнение (1.5.104) при начальном условии

$$\varphi = \varphi^0(x) \quad \text{при} \quad t = 0 \tag{1.5.105}$$

приводит к задаче, некорректно поставленной по Адамару. Решение этой задачи неустойчиво по отношению к малым возмущениям начальных данных. При построении разностных уравнений для задачи вида (1.5.96) необходимо всегда учитывать условие корректности задачи

$$\mu = \frac{u\tau}{\Delta x} \leq 1.$$

Исследуем теперь проблему счетной устойчивости схемы (1.5.98). С этой целью сначала рассмотрим спектральную задачу

$$(A^h\omega)_k \equiv \frac{\omega_k - \omega_{k-1}}{\Delta x} = \lambda\omega_k \tag{1.5.106}$$

на бесконечном сеточном интервале $D_h = (-\infty < x_k < \infty)$. Решение уравнения (1.5.106), ограниченное в D_h, имеет вид

$$\omega_k = \mathrm{e}^{ikp\Delta x}, \tag{1.5.107}$$

где p — произвольное целое число. Подставляя (1.5.107) в (1.5.106), приходим к выражению для собственного числа

$$\lambda_p = \frac{u}{\Delta x}\left(2\sin^2\frac{p\Delta x}{2} + i\sin(p\Delta x)\right). \tag{1.5.108}$$

Уравнение (1.5.98) запишем в операторной форме:

$$\frac{\varphi^{j+1} - \varphi^j}{\tau} + A^h\varphi^j = 0. \tag{1.5.109}$$

Решение уравнения (1.5.109) будем искать в виде

$$\varphi^j = \sum_{p=-\infty}^{\infty} \varphi_k^j \, \mathrm{e}^{ikp\Delta x}, \tag{1.5.110}$$

где φ_p^j — коэффициент Фурье функции φ. Для коэффициентов φ_p^j получаем уравнение

$$\frac{\varphi_p^{j+1} - \varphi_p^j}{\tau} + \lambda_p \varphi_p^j = 0.$$

Отсюда

$$\varphi_p^{j+1} = T_p \varphi_p^j,$$

где $T_p = 1 - \tau\lambda_p$ — *множитель перехода* для коэффициентов Фурье.

Найдем условие, при котором все компоненты φ_p^j не возрастают по модулю. Таким условием будет следующее:

$$|1 - \tau\lambda_p| \le 1. \tag{1.5.111}$$

Это неравенство имеет место, если

$$\frac{u\tau}{\Delta x} \le 1.$$

В самом деле,

$$\begin{aligned} |1 - \tau\lambda_p|^2 &= \left(1 - \frac{2u\tau}{\Delta x} \sin^2 \frac{p\Delta x}{2}\right)^2 + \left(\frac{u\tau}{\Delta x}\right)^2 \sin^2(p\Delta x) = \\ &= 1 - 4\frac{u\tau}{\Delta x} \sin^2 \frac{p\Delta x}{2} + \left(\frac{u\tau}{\Delta x}\right)^2 \left(4\sin^4 \frac{p\Delta x}{2} + \sin^2(p\Delta x)\right) = \\ &= 1 - 4\frac{u\tau}{\Delta x} \sin^2 \frac{p\Delta x}{2} + 4\left(\frac{u\tau}{\Delta x}\right)^2 \sin^2 \frac{p\Delta x}{2} \left(\sin^2 \frac{p\Delta x}{2} + \cos^2 \frac{p\Delta x}{2}\right) = \\ &= 1 - 4\sin^2 \frac{p\Delta x}{2} \left(\frac{u\tau}{\Delta x}\right) \left(1 - \frac{u\tau}{\Delta x}\right) \ge 0. \end{aligned} \tag{1.5.112}$$

Из неравенства (1.5.112) сразу следует (1.5.111).

Таким образом, мы приходим к условию счетной устойчивости разностной схемы. Легко видеть, что в рассматриваемом случае установленный критерий устойчивости совпадает с условием корректности уравнения (1.5.103).

Переходим теперь к обсуждению неявной разностной схемы (1.5.99). Методом, изложенным выше, несложно показать, что схема (1.5.99) также имеет первый порядок аппроксимации по Δx и τ. Применяя формулу Тейлора, при $x = x_k$ и $t = t_j$ приходим к «асимптотическому уравнению»

$$\frac{\partial \varphi}{\partial t} + u\frac{\partial \varphi}{\partial x} = \mu\frac{\partial^2 \varphi}{\partial x^2}, \tag{1.5.113}$$

где

$$\mu = \frac{u\Delta x + \tau u^2}{2}.$$

Уже здесь можно констатировать принципиальное различие соотношений (1.5.103) и (1.5.113). В последнем уравнении коэффициент счетной вязкости всегда положителен. Следовательно, уравнение (1.5.113) при соответствующих достаточно гладких начальных данных всегда корректно. Нетрудно показать, что разностное уравнение (1.5.99) устойчиво при любом соотношении шагов, т. е. абсолютно устойчиво, поскольку множитель перехода для каждого коэффициента Фурье равен

$$T_p = \frac{1}{1 + \tau\lambda_p}.$$

Откуда следует, что

$$|T_p| \leq 1.$$

Из интересных и весьма употребительных разностных схем, кроме (1.5.98) и (1.5.99), можно привести следующие:

$$\frac{\varphi_k^{j+1} - \varphi_k^j}{\tau} + u\frac{\varphi_{k+1}^{j+1} - \varphi_{k-1}^{j+1}}{2\Delta x} = 0, \tag{1.5.114}$$

$$\frac{\varphi_k^{j+1} - \varphi_k^j}{\tau} + u\frac{\varphi_{k+1}^{j+1/2} - \varphi_{k-1}^{j+1/2}}{2\Delta x} = 0, \tag{1.5.115}$$

где

$$\varphi_k^{j+1/2} = \frac{1}{2}(\varphi_k^{j+1} + \varphi_k^j).$$

Нетрудно установить, что схема (1.5.114) является схемой первого порядка аппроксимации по τ и второго по Δx. Соответствующее этой схеме дифференциальное уравнение будет иметь вид (1.5.113), где

$$\mu = \frac{u^2\tau}{2}.$$

Устойчивость определяется операторами перехода для коэффициентов Фурье. Приходим к спектральной задаче

$$(A^h\omega)_k \equiv \frac{\omega_{k+1} - \omega_{k-1}}{2\Delta x} = \lambda\omega_k. \tag{1.5.116}$$

Решение уравнения (1.5.116) будем искать в виде (1.5.107). В результате для λ_p получим следующее выражение:

$$\lambda_p = i\frac{u}{\Delta x}\sin(p\Delta x). \tag{1.5.117}$$

Из (1.5.114) непосредственно следуют уравнения для коэффициентов Фурье

$$\frac{\varphi_p^{j+1} - \varphi_p^j}{\tau} + \lambda_p\varphi_p^{j+1} = 0,$$

или

$$\varphi_p^{j+1} = T_p\varphi^j, \tag{1.5.118}$$

где

$$T_p = \frac{1}{1 + \tau\lambda_p}.$$

С учетом (1.5.117)

$$T_p = \frac{1}{1 + i\dfrac{u\tau}{\Delta x}\sin(p\Delta x)}$$

и, следовательно,

$$|T_p| = \frac{1}{\sqrt{1 + \left(\dfrac{u\tau}{\Delta x}\right)^2\sin^2(p\Delta x)}} \le 1.$$

Отсюда следует абсолютная устойчивость схемы (1.5.114).

Наиболее интересной в приложениях является схема Кранка — Николсона (1.5.115). Нетрудно убедиться, что эта схема имеет второй порядок аппроксимации по τ и Δx и не диссипативна. Это означает, что в дифференциальном уравнении (1.5.113) $\mu = 0$, а отброшенные члены имеют порядок $\tau\Delta x$, τ^2 и Δx^2. Что касается счетной устойчивости, то в данном случае

$$T_p = \frac{1 - i\dfrac{u\tau}{2\Delta x}\sin(p\Delta x)}{1 + i\dfrac{u\tau}{2\Delta x}\sin(p\Delta x)}.$$

Отсюда следует, что

$$|T_p| = 1,$$

таким образом, эта схема является абсолютно устойчивой. Следует отметить, что если в схеме (1.5.98) разностное выражение для $u\partial\varphi/\partial x$ в форме

$$u\frac{\varphi_k^j - \varphi_{k-1}^j}{\Delta x}$$

заменить на

$$u\frac{\varphi_{k+1}^j - \varphi_k^j}{\Delta x},$$

то полученная разностная схема при $u > 0$ окажется неустойчивой при любом соотношении шагов.

В заключение рассмотрим еще один интересный метод численного решения задачи (1.5.96) на основе так называемой схемы «бегущего счета». Эта схема имеет вид

$$\frac{\varphi_k^{j+1} - \varphi_k^j}{\tau} + \frac{u}{\Delta x}\left[\left(\frac{\varphi_k^{j+1} + \varphi_k^j}{2}\right) - \left(\frac{\varphi_{k-1}^{j+1} + \varphi_{k-1}^j}{2}\right)\right] = 0. \qquad (1.5.119)$$

Нетрудно показать, что эта схема второго порядка аппроксимации по τ и первого по x. Она реализуется рекуррентным соотношением

$$\varphi_k^{j+1} = \frac{1 - \dfrac{u\tau}{2\Delta x}}{1 + \dfrac{u\tau}{2\Delta x}}\varphi_k^j + \frac{\dfrac{u\tau}{2\Delta x}}{1 + \dfrac{u\tau}{2\Delta x}}(\varphi_{k-1}^{j+1} + \varphi_{k-1}^j). \qquad (1.5.120)$$

На основе анализа устойчивости по Нейману с помощью метода Фурье нетрудно доказать, что схема (1.5.119) абсолютно устойчива.

Аналогичным образом можно построить схему «бегущего счета» для многомерной задачи движения и доказать ее устойчивость в случае уравнения с постоянными коэффициентами.

Выше всюду предполагалось, что u постоянна и положительна. Если u отрицательна, то заменой x на $-x$ приходим к уравнению (1.5.96). Однако особый интерес для приложений имеет случай, когда $u = u(x,t)$. Уже самый простой анализ показывает, что в этом случае, даже при использовании неявных диссипативных разностных схем, возможно нарушение счетной устойчивости. Особенно это проявляется в нелинейных задачах. Суть дела состоит в следующем. Если разложить решение разностной задачи и коэффициент $u(x_k, t_j)$ в

ряд Фурье, то произведения рядов Фурье приведут к гармоникам как к более длинным, чем взаимодействующие, так и к более коротким.

В результате такого процесса в ряде случаев может произойти перекачка «энергии» из длинных волн в наиболее короткие, и процесс вычислений окажется неустойчивым, несмотря на то, что данная разностная схема с постоянным коэффициентом будет счетно-устойчивой. Обычно такую неустойчивость называют *нелинейной*. Она также иногда появляется и при решении линейных задач с переменными коэффициентами. Поэтому построение разностных схем для нелинейных уравнений или уравнений с переменными коэффициентами, устойчивых в отношении к любым возмущениям, является чрезвычайно актуальной задачей. В большинстве случаев подавление счетной неустойчивости возможно с помощью диссипативных разностных схем, отвечающих определенному выбору коэффициента счетной вязкости μ. Однако такие схемы, как правило, оказываются схемами первого порядка аппроксимации либо по τ, либо по Δx, либо и по τ, и по Δx.

Особый интерес в приложениях имеют уравнения вида

$$\frac{\partial \varphi}{\partial t} + \frac{\partial (u\varphi)}{\partial x} = 0, \tag{1.5.121}$$

где $u = u(x, t)$.

Разностные схемы для уравнений такого типа, абсолютно устойчивые и имеющие первый порядок аппроксимации на некоторых классах коэффициентов, будут изучены при рассмотрении многомерных уравнений вида (1.5.121).

Глава 2.

Методы построения разностных схем для дифференциальных уравнений

Известны различные подходы к конструированию разностных уравнений для задач математической физики. Особенно полно этот вопрос изучен для уравнений с коэффициентами, обладающими (вместе с решением) достаточной гладкостью. В этом случае можно строить разностные схемы с высокой степенью аппроксимации. Интерес к таким схемам непрерывно возрастает, поскольку темп формирования новых постановок сложных и трудоемких задач науки и техники опережает темп развития средств вычислительной техники. Поэтому в ряде случаев представляется целесообразным получать приближенное решение с заданной точностью не за счет формального увеличения размерности подпространств (например, уменьшения шага сетки), а путем построения более точных аппроксимаций исходной задачи на основе априорной информации о гладкости решения[1)]. Такая точка зрения оказалась весьма плодотворной и привела исследователей к удобным и достаточно универсальным методам построения разностных уравнений на основе вариационных методов Ритца, Галеркина и метода наименьших квадратов.

Однако следует подчеркнуть, что класс задач с гладкими решениями довольно узок, и поэтому основное внимание должно быть

[1)] См. также гл. 6.

уделено методам построения разностных уравнений для задач с разрывными коэффициентами. Такие задачи возникают, например, при изучении диффузии субстанции, теплопроводности, гидродинамики и т. д.

В силу отмеченного обстоятельства мы пожертвуем возможностью описания ряда оригинальных и весьма общих результатов по построению разностных уравнений высокого порядка точности ради идеи создания общего представления о путях конструирования разностных аналогов уравнений, решения которых могут не обладать высокой гладкостью. Естественно, что все подходы, которые мы будем рассматривать, автоматически применимы для численного решения задач с гладкими данными и решениями.

Для того чтобы дать необходимое представление о путях развития научных идей в области построения разностных уравнений, мы подробно рассмотрим сначала краевые задачи для обыкновенных дифференциальных уравнений и затем изложим более или менее общие подходы к решению двумерных и многомерных задач математической физики. Мы надеемся, что ссылки в главе 12 на оригинальную литературу помогут читателям более глубоко и всесторонне познакомиться с вопросами теории и алгоритмами.

2.1. Вариационные методы в математической физике

В данном разделе будут рассмотрены некоторые из вариационных методов приближенного решения уравнений математической физики. Близость их постановок некоторым задачам вариационного исчисления проиллюстрирована на примере нескольких простых задач. Это позволит в дальнейшем полнее раскрыть существо описываемых вариационных методов.

2.1.1. Некоторые задачи вариационного исчисления

Рассмотрим простейший функционал

$$J(u) = \int_{x_0}^{x_1} \pi(x, u, u')\, dx, \tag{2.1.1}$$

где $\pi(x, y, z)$ — заданная функция, непрерывная вместе со своими производными до второго порядка включительно относительно переменных x, y, z в некоторой области трехмерного евклидового пространства.

Предположим, что функция $u(x)$ непрерывна, имеет непрерывную производную $u'(x)$ на (x_0, x_1) и на концах отрезка $[x_0, x_1]$ принимает заданные значения

$$u(x_0) = u_0, \quad u(x_1) = u_1. \tag{2.1.2}$$

Определим ε-окрестность функции $u = u(x)$ как семейство функций $u_1(x)$, удовлетворяющих на всем отрезке $[x_0, x_1]$ неравенству

$$|u_1(x) - u(x)| \leq \varepsilon. \tag{2.1.3}$$

Сформулируем теперь следующую задачу вариационного исчисления: среди функций, лежащих в ε-окрестности, имеющих непрерывную производную и удовлетворяющих условию (2.1.2), найти функцию, доставляющую экстремум функционалу $J(u)$ (задача с фиксированными концами кривых $u = u(x)$).

Найдем необходимые условия, которым должна подчиняться функция $u(x)$ для того, чтобы она сообщала функционалу $J(u)$ экстремальное значение в ε-окрестности. С этой целью рассмотрим функцию $\eta(x)$, удовлетворяющую условиям

$$\eta(x_0) = \eta(x_1) = 0. \tag{2.1.4}$$

Построим, далее, новую функцию $u_\alpha(x) = u(x) + \alpha\eta(x)$, где α — малый параметр (в силу чего можно предположить, что u_α также принадлежит ε-окрестности). Подставив эту функцию в функционал J,

получим

$$J(u) = \int\limits_{x_0}^{x_1} \pi(x, u(x) + \alpha\eta(x), u'(x) + \alpha\eta'(x))\, dx.$$

Будем рассматривать $J(u_\alpha)$ как функцию от параметра α: $J(u_\alpha) \equiv \Phi(\alpha)$. Назовем производную функции $\Phi(\alpha)$ в точке $\alpha = 0$ первой вариацией функционала J и обозначим ее символом δJ:

$$\delta J(u) = \left.\frac{d\Phi}{d\alpha}\right|_{\alpha=0}.$$

Вторую вариацию $\delta^2 J$ функционала J определим как вторую производную функции $\Phi(\alpha)$ в точке $\alpha = 0$:

$$\delta^2 J(u) = \left.\frac{d^2\Phi}{d\alpha^2}\right|_{\alpha=0}.$$

Используя вид J, найдем выражения для δJ и $\delta^2 J$:

$$\delta J = \int\limits_{x_o}^{x_1} (\pi_u \eta + \pi_{u'} \eta')\, dx, \tag{2.1.5}$$

$$\delta^2 J = \int\limits_{x_o}^{x_1} (\pi_{u'u'} \eta'^2 + 2\pi_{uu'} \eta\eta' + \pi_{uu} \eta^2)\, dx \tag{2.1.6}$$

(здесь использованы обозначения $\pi_u \equiv \dfrac{\partial \pi}{\partial u},\ \pi_{uv} \equiv \dfrac{\partial^2 \pi}{\partial u \partial v}, u' = \dfrac{\partial u}{\partial x}\Big)$. Как известно, необходимым условием экстремума $\Phi(\alpha)$ при $\alpha = 0$ является равенство $\Phi'(0) = 0$, т. е.

$$\delta J(u) = \int\limits_{x_o}^{x_1} (\eta \pi_u + \eta' \pi_{u'})\, dx = 0.$$

Выполняя в последнем выражении интегрирование по частям с учетом условий (2.1.4), получим

$$\delta J(u) = \int\limits_{x_o}^{x_1} \eta(x) \left(\pi_u - \frac{d}{dx}\pi_{u'}\right)\, dx. \tag{2.1.7}$$

В силу произвольности функции $\eta(x)$ приходим к выводу, что кривая $u(x)$, удовлетворяющая условиям (2.1.2) и доставляющая экстремум функционалу (2.1.1), должна удовлетворять дифференциальному уравнению

$$\pi_u - \frac{d}{dx}\pi_{u'} = 0, \tag{2.1.8}$$

которое обычно называют *уравнением Эйлера*.

Отметим, что если $u(x)$ доставляет функционалу J минимум (максимум), то, как известно, в этом случае $\Phi''(0) = \delta^2 J \geq 0$ $(\delta^2 J \leq 0)$.

В качестве иллюстрации изложенного выше, рассмотрим пример, в котором примем $u_0 = u_1 = 0$, а

$$\pi = \left(\frac{du}{dx}\right)^2 + ku^2 - 2fu, \tag{2.1.9}$$

где k, f — достаточно гладкие функции и $k > 0$. Тогда в рассмотренной выше вариационной задаче уравнение Эйлера (2.1.8) имеет вид

$$-\frac{d^2u}{dx^2} + ku = f(x) \tag{2.1.10}$$

(строго говоря, устанавливая необходимое условие экстремума в случае (2.1.9), мы здесь должны требовать, чтобы $u(x)$ обладала непрерывными вторыми производными).

Итак, если функция из области определения функционала

$$\delta J(u) = \int_{x_o}^{x_1} \left(\left(\frac{du}{dx}\right)^2 + ku^2 - 2fu \right) dx, \tag{2.1.11}$$

удовлетворяющая условиям $u(x_0) = u(x_1) = 0$, сообщает экстремум функционалу (2.1.11), то она удовлетворяет условию (2.1.10), т. е. является решением первой краевой задачи вида

$$-\frac{d^2u}{dx^2} + ku = f(x), \tag{2.1.12}$$

$$u(0) = u(1) = 0. \tag{2.1.13}$$

Справедливо также и обратное утверждение (доказательство которого приводится ниже для уравнения более общего вида), а именно: если $u(x)$ является решением задачи (2.1.12), (2.1.13), то эта

функция сообщает экстремум функционалу (2.1.11) на соответствующей области определения.

Рассмотрим теперь еще одну вариационную задачу для функционала (2.1.1): среди кривых $u = u(x)$, концы которых лежат на заданных вертикалях $x = x_0$, $x = x_1$, найти $u(x)$, которая дает экстремум функционалу (2.1.1) (задача со свободными концами). Отметим, что здесь на концы кривых никаких других условий не накладывается. Однако, несмотря на это, оказывается, что если $u(x)$ сообщает экстремум функционалу $J(u)$, то при $x = x_0$ и $x = x_1$ она должна удовлетворять некоторым предельным условиям, которые непосредственно получаются из условия экстремума (2.1.1). Покажем это.

Пусть некоторая кривая $u(x)$ дает экстремум $J(u)$ по сравнению со всеми близкими кривыми $u_\alpha(x) = u(x) + \alpha\eta(x)$ со свободными концами (здесь, в отличие от задачи с закрепленными концами, $\eta(x)$ не обязательно обращается в нуль в точках x_0 и x_1). Необходимое условие экстремума снова приводит к соотношению

$$\delta J(u) = \int\limits_{x_o}^{x_1} (\eta\pi_u + \eta'\pi_{u'})\, dx = 0. \tag{2.1.14}$$

Выполняя интегрирование по частям, получаем

$$\int\limits_{x_o}^{x_1} \eta(x) \left(\pi_u - \frac{d}{dx}\pi_{u'}\right)\, dx + \pi_{u'}\,\eta|_{x=x_1} - \pi_{u'}\,\eta|_{x=x_0} = 0. \tag{2.1.15}$$

В силу произвольности $\eta(x)$ функция $u(x)$ удовлетворяет уравнению Эйлера

$$\pi_u - \frac{d}{dx}\pi_{u'} = 0, \tag{2.1.16}$$

а также предельным условиям

$$\pi_{u'}|_{x=x_1} = 0, \quad \pi_{u'}|_{x=x_0} = 0. \tag{2.1.17}$$

Условия типа (2.1.17), являющиеся одними из необходимых условий экстремума, часто называют естественными граничными условиями (подробнее о них при рассмотрении метода Ритца).

Снова рассмотрим иллюстрирующий пример. Пусть в задаче со свободными концами $\pi(x, u, u')$ имеет вид (2.1.9). Тогда в рассматри-

ваемой задаче уравнения (2.1.16), (2.1.17) принимают вид

$$-\frac{d^2u}{dx^2} + ku = f(x), \tag{2.1.18}$$

$$\frac{du}{dx}(x_1) = \frac{du}{dx}(x_0) = 0, \tag{2.1.19}$$

т. е. в этом случае функция $u(x)$, которая сообщает экстремум функционалу, является решением уже второй краевой задачи вида (2.1.18), (2.1.19)[2)]. Справедливо также и обратное утверждение: если $u(x)$ — решение задачи (2.1.18), (2.1.19), то кривая $u(x)$ сообщает экстремальное значение функционалу вида (2.1.11) в задаче со свободными концами.

Итак, мы рассмотрели выше простейший случай одной функции u и одного независимого переменного x. Аналогичным образом могут быть рассмотрены более общие задачи. Пусть, например,

$$J = \int\limits_D \int \pi(x, y, u, u_x, u_y)\, dx\, dy,$$

где функция π и граница выпуклой ограниченной области D обладают необходимой гладкостью. Поставим задачу: найти функцию $u(x, y)$ непрерывную вместе со своими частными производными до второго порядка включительно, имеющую заданное значение на границе области и доставляющую экстремум функционалу J. Тогда аналогично предыдущему приходим к уравнению Эйлера следующего вида:

$$\pi_u - \frac{\partial}{\partial x}\pi_{u_x} - \frac{\partial}{\partial y}\pi_{u_y} = 0.$$

Распространение результатов на случай n переменных очевидно.

Итак, мы приходим к возможности одни и те же задачи математической физики толковать либо с позиции задач для дифференциальных уравнений (уравнений Эйлера), либо с позиции задач вариационного исчисления об отыскании функций, доставляющих экстремум некоторым функционалам. В последнем случае функции (при наличии необходимой гладкости) будут решениями соответствую-

[2)] Отсюда также видно, что граничные условия (2.1.19), в отличие от условий (2.1.13), будут естественными.

щих уравнений Эйлера. В рассмотренных выше примерах задачи для уравнения Эйлера (т. е. задачи вида (2.1.12) — (2.1.13) и (2.1.18), (2.1.19)) можно записать в операторной форме

$$Lu = f, \quad u \in \Phi(L) \tag{2.1.20}$$

(где $\Phi(L)$ — область определения оператора L). Выше уже отмечалась эквивалентность задачи (2.1.20), соответствующей вариационной задаче

$$J(u) = \min_{v \in \Phi(L)} J(v), \tag{2.1.21}$$

где

$$J(v) = (Lu, u) - 2(f, u) =$$
$$= \int_{x_0}^{x_1} \left(-\frac{d^2u}{dx^2} + ku - 2f \right) u \, dx = \int_{x_0}^{x_1} \left(\left(\frac{du}{dx} \right)^2 + ku^2 - 2fu \right) \, dx.$$

Докажем теперь эту эквивалентность задач (2.1.20) и (2.1.21) для абстрактного оператора L.

Итак, пусть рассматривается уравнение

$$Lu = f, \tag{2.1.22}$$

где L — линейный положительный симметричный оператор с областью определения $\Phi(L)$, являющейся всюду плотным множеством в гильбертовом пространстве H со скалярным произведением $(\cdot, \cdot)$ и областью значений в пространстве H, $u \in \Phi(L)$ и f — некоторый элемент из H. Тогда имеет место следующее утверждение: если решение задачи (2.1.20) существует, то оно доставляет минимум функционалу

$$J(u) = (Lu, u) - 2(u, f).$$

Докажем его. Предположим, что элемент u_0 является решением уравнения (2.1.22), т. е.

$$Lu_0 = f.$$

Пусть η — произвольный ненулевой элемент из $\Phi(L)$ и α — произвольное вещественное число. Определим элемент v_α соотношением

$$v_\alpha = u_0 + \alpha\eta.$$

Тогда

$$J(v_\alpha) = (L(u_0 + \alpha\eta), u_0 + \alpha\eta) - 2(u_0 + \alpha\eta, f).$$

Так как L — самосопряженный оператор, то

$$J(v_\alpha) = J(u_0) + 2\alpha(Lu_0 - f, \eta) + \alpha^2(L\eta, \eta).$$

Отсюда получаем, что

$$J(v_\alpha) = J(u_0) + \alpha^2(L\eta, \eta).$$

Из этого соотношения в силу положительности оператора L следует неравенство

$$J(v_\alpha) > J(u_0) \tag{2.1.23}$$

для любого $\alpha \neq 0$. Это значит, что минимум функционала $J(v)$ достигается на решении $v_\alpha = u_0$.

Имеет место и обратное утверждение. А именно: элемент u_0 гильбертова пространства H, доставляющий минимум функционалу J и принадлежащий $\Phi(L)$, является решением операторного уравнения $Lu = f$.

В самом деле, пусть $u_0 \in \Phi(L)$ — элемент, на котором достигается минимум функционала $J(u)$, η — произвольный элемент из $\Phi(L)$. Известно, что для любых $u, v \in \Phi(L)$ элемент $w = \alpha u + \beta v$ (α и β — константы) также будет элементом из $\Phi(L)$. Поэтому $v_\alpha = u_0 + \alpha\eta \in \Phi(L)$. В силу предположения о том, что на элементе u_0 функционал J достигает минимума, имеем

$$J(u_0 + \alpha\eta) \geq J(u_0). \tag{2.1.24}$$

Число α будем считать вещественным. Соотношение (2.1.24) в предположении симметричности L приводит к неравенству

$$2\alpha(Lu_0 - f, \eta) + \alpha^2(L\eta, \eta) \geq 0.$$

Это возможно только в том случае, если

$$(Lu_0 - f, \eta) = 0. \tag{2.1.25}$$

Таким образом, элемент $Lu_0 - f$ ортогонален ко всем элементам множества $\Phi(L)$. Поэтому имеет место равенство

$$Lu_0 - f = 0,$$

что и завершает доказательство сформулированного утверждения.

В дальнейшем мы приведем ряд дополнительных сведений по вариационным постановкам задач математической физики и опишем некоторые из основных методов их решения. Рассматриваемые вопросы часто будем иллюстрировать на примере эллиптического дифференциального уравнения

$$Lu \equiv -\sum_{i,j=1}^{2} \frac{\partial}{\partial x_i} A_{ij}(x) \frac{\partial u}{\partial x_j} + \sum_{i=1}^{2} B_i(x) \frac{\partial u}{\partial x_i} + q(x)u = f(x), \tag{2.1.26}$$

$$x = (x_1, x_2) \in D,$$

заданного в ограниченной области D с краевыми условиями вида

$$u = 0, \quad x \in \partial D \tag{2.1.27}$$

(первая краевая задача), либо

$$\frac{\partial u}{\partial N} \equiv \sum_{j,k=1}^{2} A_{jk}(x) \frac{\partial u}{\partial x_k} \cos(\nu, x_j) = 0, \tag{2.1.28}$$

где ν — внешняя нормаль к ∂D (вторая краевая задача).

Будем предполагать, что оператор

$$L_0 = -\sum_{i,j=1}^{2} \frac{\partial}{\partial x_i} A_{ij}(x) \frac{\partial}{\partial x_j} \tag{2.1.29}$$

действует в гильбертовом пространстве $F = L_2(D)$, его областью определения является множество $\Phi(L_0)$, состоящее из таких функций $u \in L_2(D)$, что выполняется либо условие (2.1.27), либо условие (2.1.28) и $L_0 u \in L_2(D)$; кроме того, L_0 самосопряжен по Лагранжу и не

вырождается, т. е. для любого ненулевого вектора $\xi = (\xi_1, \xi_2)$ выполняется неравенство

$$\inf_{x \in D} \sum_{i,j=1}^{2} A_{ij} \xi_i \xi_j \geq \mu_0 \sum_{i=1}^{2} \xi_i^2 \tag{2.1.30}$$

с некоторой положительной константой μ_0. Далее будем считать (если не оговорено специально), что функция $q(x)$ ограничена и положительна в области D, а также что решения задач (2.1.26), (2.1.27) и (2.1.26), (2.1.28) существуют. В дальнейшем у читателя могут возникнуть вопросы о том, какова гладкость входных данных, является ли решение классическим, обобщенным и т. д.

Будем, во-первых, предполагать, что решения задач удовлетворяют почти всюду уравнению (2.1.26) и принадлежат пространству Соболева W_2^1, которое состоит из функций пространства $L_2(D)$, имеющих в D суммируемые с квадратом обобщенные производные первого порядка. Норма в W_2^1 определяется соотношением (см. § **??**)

$$\|u\|_{W_2^1} = \left\{ \int\limits_D u^2 \, dD + \int\limits_D \left[\left(\frac{\partial u}{\partial x_1} \right)^2 + \left(\frac{\partial u}{\partial x_2} \right)^2 \right] \, dD \right\}^{1/2} < \infty.$$

Решение первой краевой задачи (2.1.26), (2.1.27), кроме того, предполагается принадлежащим $\overset{\circ}{W}{}_2^1$ — подпространству из W_2^1, состоящему из функций из W_2^1, обращающихся в нуль на границе ∂D.

Будем считать, что входные данные задачи, такие как гладкость коэффициентов и границы области, обеспечивают принадлежность решений указанным пространствам.

Во-вторых, будем считать, если это потребуется при рассмотрении конкретных вопросов, что все нужные нам дополнительные требования о большой гладкости решений или коэффициентов и правой части уравнения также выполняются.

Эти два условия позволяют нам сосредоточить внимание на основной цели главы — изучение принципов построения сеточных аналогов дифференциальных уравнений с частными производными.

2.1.2. Метод Ритца

Одним из известных вариационных методов решения задач математической физики является метод Ритца. Опишем его примени-

тельно к решению операторного уравнения

$$Lu = f \tag{2.1.31}$$

в гильбертовом пространстве F со скалярным произведением (u, v) при условии, что $f \in F$ и оператор L с областью определения $\Phi(L)$, плотной в F, симметричен и положительно определен. Следуя сформулированному в 2.1.1 утверждению, делаем вывод, что задача нахождения решения уравнения (2.1.31) равносильна нахождению элемента $u \in \Phi(L)$, реализующего минимум функционала

$$J(u) = (Lu, u) - 2(f, u). \tag{2.1.32}$$

Отметим, однако, что упомянутое утверждение об эквивалентности задач ничего не говорит о существовании такой $u \in \Phi(L)$, которая была бы решением (2.1.31), а значит, и минимизировала $J(u)$. Поэтому видоизменим постановку вариационной задачи о минимизации $J(u)$ так, чтобы уже можно было гарантировать существование его решения.

Для этого введем в $\Phi(L)$ новое скалярное произведение, определив его соотношением

$$(\varphi, \psi)_L = (L\varphi, \psi), \quad \varphi, \psi \in \Phi(L), \tag{2.1.33}$$

и соответствующую норму

$$\|\varphi\|_L = (\varphi, \varphi)_L^{1/2}.$$

Пополняя $\Phi(L)$ по введенной норме, мы приходим к гильбертову пространству F_L, которое называется *энергетическим* пространством, порождаемым оператором L. Каждая функция из F_L принадлежит пространству F, однако в результате пополнения в F_L могут появиться элементы, не входящие в $\Phi(L)$ (поэтому представление скалярного произведения $(\varphi, \psi)_L$ при произвольных $\varphi, \psi \in F_L$ в виде (2.1.33) уже не имеет места!).

Так как в (2.1.32) предполагалось, что $u \in \Phi(L)$, то, используя (2.1.33), представим $J(u)$ в виде

$$J(u) = (u, u)_L - 2(f, u). \tag{2.1.34}$$

Последня форма записи $J(u)$ позволяет рассматривать $J(u)$ не только в области определения оператора L, но и на всех элементах энергетического пространства F_L. Поэтому расширим функционал (2.1.34) (оставив при этом за ним прежнее обозначение $J(u)$) на все пространство F_L и будем искать его минимум на этом пространстве. Легко показать, что в такой постановке вариационная задача всегда имеет единственное решение. Действительно, так как оператор L предполагается положительно определенным, т. е. $(Lu, u) = (u, u)_L \geq \gamma^2 \|u\|^2$ $(u \in \Phi(L);\ \gamma = \text{const} > 0)$, то в результате пополнения $\Phi(L)$ и получения F_L соотношение определенности $(u, u)_L \geq \gamma^2 \|u\|^2$ останется справедливым для любого элемента $u \in F_L$. Рассматривая теперь функционал (u, f), заметим, что он ограничен в F_L:

$$|(u, f)| \leq \|u\| \cdot \|f\| \leq \frac{1}{\gamma} \|u\|_L \cdot \|f\| \equiv C \|u\|_L.$$

Следовательно, по известной теореме Рисса существует такой элемент $u_0 \in F_L$, что для любого $u \in F_L$ справедливо тождество

$$(u, f) = (u, u_0)_L.$$

Тогда функционал $J(u)$ можно представить в виде

$$J(u) = (u, u)_L - 2(f, u) = (u, u)_L - 2(u, u_0)_L = \|u - u_0\|_L^2 - \|u_0\|_L^2,$$

$$u \in F_L. \tag{2.1.35}$$

Из последнего выражения делаем заключение, что в пространстве F_L функционал $J(u)$ достигает минимума при $u = u_0$. Как уже отмечалось, элемент u_0 единственен и принадлежит F_L. Назовем его *обобщенным* решением уравнения $Lu = f$. Может оказаться, что $u_0 \in \Phi(L)$. В этом случае u_0 будет также классическим решением рассматриваемой задачи, т. е. будет удовлетворять (2.1.31).

Итак, мы свели исходную задачу к задаче минимизации функционала (2.1.34) в энергетическом пространстве F_L. Рассмотрим теперь метод Ритца для приближенного решения последней вариационной задачи.

Введем последовательность конечномерных пространств $F_h \subseteq F_L$, которые определяются такой бесконечной последовательностью параметров $h_1, h_2, \ldots$, что $h_k \to 0$ при $k \to \infty$. Будем говорить, что по-

следовательность $\{F_h\}$ полна в F_L, если для любых $u \in F_L$ и $\varepsilon > 0$ существует такое$\hat{h} = \hat{h}(u, \varepsilon)$, что

$$\inf_{w \in F_h} \|u - w\|_L < \varepsilon \tag{2.1.36}$$

для всех $h < \hat{h}$. Иначе говоря, полнота последовательности подпространств $\{F_h\}$ означает, что всякий элемент $u \in F_L$ может быть с любой степенью точности аппроксимирован элементами пространств F_h.

В предложенной постановке метод Ритца формулируется следующим образом: требуется найти элемент $u^h \in F_h$, минимизирующий $J(u)$ в пространстве F_h.

Справедливо следующее утверждение: при сделанных выше предположениях последовательность $\{u^h\}$ приближений по Ритцу сходится в F_L к решению (обобщенному) задачи u_0. Действительно, так как каждое u^h сообщает на F_h минимум функционалу $J(u)$, то с учетом соотношения (2.1.35) при произвольном $w \in F_h$ имеем

$$\|u_0 - u^h\|_L^2 = J(u_h) - J(u_0) \leq J(w) - J(u_0) = \|u_0 - w\|_L^2.$$

Так как $w \in F_h$ произвольно, то, принимая во внимание (2.1.36), получаем

$$\|u_0 - u^h\|_L \leq \inf_{w \in F_h} \|u_0 - w\|_L \underset{h \to 0}{\to} 0. \tag{2.1.37}$$

В случае, когда базис пространства F_h известен и состоит из функций $\{\varphi_i^h\}_{i=1}^{N_h}$, задача нахождения $u^h \in F_h$ эквивалентна нахождению коэффициентов $\{\alpha_i\}$ разложения

$$u^h = \sum_{i=1}^{N_h} \alpha_i \varphi_i^h \tag{2.1.38}$$

из условия минимума функционала J. Как обычно, подставляя разложение (2.1.38) в функционал J и приравнивая к нулю производные $\partial J(u^h)/\partial \alpha_i$ $(i = 1, 2, \ldots, N_h)$, приходим к системе линейных алгебраических уравнений

$$A\alpha = g, \tag{2.1.39}$$

где α и g есть N_h-мерные векторы, причем

$$g_i = (f, \varphi_i^h), \tag{2.1.40}$$

а $A = (a_{ij})$ — матрица Грама системы векторов $\{\varphi_i^h\}$ в скалярном произведении пространства F_L, т. е.

$$a_{ek} = (\varphi_e^h, \varphi_k^h)_L, \quad 1 \le e, k \le N_h. \tag{2.1.41}$$

Если базисные функции φ_i^h, кроме того, принадлежат $\Phi(L)$, то в этом случае a_{ij} можно представить также в виде

$$a_{ij} = (L\varphi_i, \varphi_j).$$

Так как $a_{ij} = (\varphi_i^h, \varphi_j^h)_L = (\varphi_j^h, \varphi_i^h)_L = a_{ij}$, то матрица A симметрична, а в силу неравенства

$$(A\xi, \xi)_2 \equiv \sum_{i,j=1}^{N_h} a_{ij}\xi_i\xi_j = \left(\sum_{i=1}^{N_h} \xi_i\varphi_i^h, \sum_{j=1}^{N_h} \xi_j\varphi_j^h\right)_L \ge \gamma^2 \left\|\sum_{i=1}^{N_h} \xi_i\varphi_i^h\right\|^2 > 0 \tag{2.1.42}$$

при $\xi \neq 0$—положительно определена.

Рассмотрим один из вопросов, важных для практического использования метода Ритца, — проблему выделения главных и естественных граничных условий.

При определении области определения оператора L, т. е. множества $\Phi(L)$, мы часто накладываем на $u \in \Phi(L)$ те или иные граничные условия. Оказывается, что при построении энергетического пространства F_L в результате пополнения $\Phi(L)$ в метрике $\|\cdot\|_L$ в F_L могут появиться элементы, удовлетворяющие не всем граничным условиям, которым удовлетворяли функции из $\Phi(L)$. Граничные условия, которым обязательно удовлетворяют функции из области определения оператора L и необязательно функции из энергетического пространства F_L, называют *естественными* для оператора L. Граничные условия, которым обязательно удовлетворяют функции из энергетического пространства, называются *главными*.

Практическая важность умения различать эти условия состоит в следующем. Так как в методе Ритца базисные функции $\{\varphi_i\}$ достаточно брать лишь из энергетического пространства (и не обязательно из $\Phi(L)$), то не обязательно подчинять их естественным краевым

условиям. Это обстоятельство в значительной степени облегчает выбор φ_i^h при решении многих практически важных задач, особенно в случае многомерной области D с границей сложной формы. (Отметим, что в случае главных краевых условий проблема построения φ_i^h так, чтобы они удовлетворяли этим условиям, остается.)

Укажем подход, который позволяет при рассмотрении каждой конкретной задачи установить, является краевое условие естественным или нет. Пусть рассматривается задача о минимизации функционала $J(u)$. Предположим, что существует u_0, реализующая минимум $J(u)$ в классе функций, данному условию, вообще говоря, не удовлетворяющих. Используя средства вариационного исчисления, можно найти необходимые условия, которым должна удовлетворять функция u_0. Если к ним относится и рассматриваемое граничное условие, то оно — естественное. Так, например, в 2.1.1 мы таким способом показали, что при рассмотрении задачи (2.1.18), (2.1.19) условия Неймана $\dfrac{du}{dx}(x_0) = \dfrac{du}{dx}(x_1) = 0$ являются естественными. Следовательно, при построении приближенного решения по Ритцу задачи со свободными концами можно выбирать в качестве базиса $\{\varphi_i^h\}$ функции, этим условиям не удовлетворяющие. В то же время, если решается задача (2.1.12), (2.1.13), то φ_i^h обязаны удовлетворять условиям (2.1.13), т. е. граничные условия являются главными.

Наконец, отметим простой метод, который позволяет отличить естественные граничные условия от главных и который применим для ряда краевых задач. Пусть в (2.1.31) оператор L является дифференциальным порядка $2m$ положительным на множестве функций $\Phi(L)$, удовлетворяющих некоторым однородным граничным условиям вида $N_k u = 0$. Тогда такое краевое условие будет естественным, если $N_k u$ содержит производные от u порядка m и выше, и главным, если $N_k u$ не содержит производных от u порядка выше $m-1$.

Рассмотрим теперь метод Ритца для задач (2.1.26), (2.1.27) и (2.1.26), (2.1.28) при дополнительном предположении, что B_i $(i = 1, 2)$ тождественно равны нулю в D. Нетрудно показать, что для всех функций φ и ψ из области определения $\Phi(L)$ оператора L выполняется соотношение

$$(L\varphi, \psi) \equiv \int\limits_D \psi L\varphi \, dD = \int\limits_D \left\{ \sum_{i,j=1}^{2} A_{ij}(x) \frac{\partial \varphi}{\partial x_i} \frac{\partial \psi}{\partial x_j} + q(x)\varphi\psi \right\} dD.$$

Поставим теперь в соответствие каждой из исходных задач соответствующую задачу на нахождение элемента из F_L, на котором функционал

$$J(u) = \int\limits_D \left\{ \sum_{i,j=1}^{2} A_{ij}(x) \frac{\partial u}{\partial x_i} \frac{\partial u}{\partial x_j} + q(x)u^2 - 2uf \right\} dD$$

достигает минимального значения. Минимизируемый функционал в обеих рассматриваемых задачах имеет один и тот же вид. Однако энергетические пространства, соответствующие операторам этих задач, различны. Так, при рассмотрении первой краевой задачи скалярное произведение и норма в F_L есть

$$(\varphi, \psi) = \int\limits_D \left\{ \sum_{i,j=1}^{2} A_{ij}(x) \frac{\partial \varphi}{\partial x_i} \frac{\partial \psi}{\partial x_j} + q(x)\varphi\psi \right\} dD,$$

$$\|\varphi\|_L = (\varphi, \varphi)_L^{1/2} < \infty,$$

но элементы из F_L здесь удовлетворяют однородному граничному условию Дирихле, которое таким образом оказывается главным. Учитывая (2.1.30), получаем, что F_L в данной задаче является подпространством из $\overset{\circ}{W}{}_2^1$ (или совпадает с $\overset{\circ}{W}{}_2^1$, если коэффициенты $A_{ij}(x)$ являются ограниченными функциями).

Если же рассматривается вторая краевая задача, то $(\varphi, \psi)_L$ имеет прежний вид, но краевое условие Неймана оказывается уже естественным. Поэтому получаемое энергетическое пространство оказывается здесь уже подпространством из W_2^1, и если решать эту задачу методом Ритца, то базисные функции можно брать из W_2^1, причем совсем не обязательно, чтобы они удовлетворяли (2.1.28).

2.1.3. Метод Галеркина

Основным недостатком метода Ритца является то обстоятельство, что он применим только для уравнений с симметричными и положительно определенными операторами. От этого недостатка свободен другой вариационный метод, называемый *методом Галеркина* (иногда его называют *проекционным методом Бубнова — Галеркина*). Опишем этот метод применительно к решению операторного

уравнения

$$Lu = f \tag{2.1.43}$$

в гильбертовом пространстве F, где $f \in F$ и область определения $\Phi(L)$ оператора L плотна в F.

Предположим, что $L = L_0 + K$, где L_0 — симметричный и положительно определенный оператор. Пусть $\Phi(L_0) \subseteq \Phi(K)$ и L_0^{-1} вполне непрерывен в F. Введем энергетическое пространство F_{L_0}, соответствующее оператору L_0 со скалярным произведением $(u, v)_{L_0}$ и нормой $\|u\|_{L_0} = (u, u)_{L_0}^{1/2}$. Если умножить (2.1.43) скалярно в F на произвольную функцию $v \in \Phi(L_0)$, то приходим к тождеству

$$(L_0 u, v) + (Ku, v) = (f, v). \tag{2.1.44}$$

Так как $(L_0 u, v) = (u, v)_{L_0}$, то соотношение (2.1.44) принимает вид

$$(u, v)_{L_0} + (Ku, v) = (f, v), \tag{2.1.45}$$

который допускает обобщенную постановку задачи для уравнения (2.1.43). Назовем *обобщенным* решением уравнения (2.1.43) функцию $u_0 \in F_{L_0}$, удовлетворяющую соотношению (2.1.45) при любых $v \in F_{L_0}$. Предположим существование такого обобщенного решения u_0. Если при этом окажется, что $u_0 \in \Phi(L_0)$, то в силу соотношения $(u, v)_{L_0} = (L_0 u, v)$ получаем равенство

$$(L_0 u_0 + K u_0 - f, v) = 0.$$

А так как F_{L_0} плотно в F, то делаем заключение, что u_0 удовлетворяет исходному уравнению (2.1.43).

Как и в предыдущем пункте, введем последовательность конечномерных подпространств $F_h \subset F_{L_0}$ $(h \geq h_1, h_2, \ldots)$ с базисами $\{\varphi_i^h\}_{i=1}^{N_j}$. Тогда приближение по Галеркину находится в виде

$$u^h = \sum_{i=1}^{N_h} \alpha_i \varphi_i^h, \tag{2.1.46}$$

причем коэффициенты α_i выбираются так, чтобы u^h удовлетворяла (2.1.45) при любых $v \in F_h$. Так как $v \in F_h$ можно представить в виде

$$v = \sum_{i=1}^{N_h} b_i \varphi_i^h,$$

то u^h находится из системы уравнений

$$(u^h, \varphi_s^h)_{L_0} + (Ku^h, \varphi_s^h) = (f, \varphi_s^h), \tag{2.1.47}$$

$$s = 1, 2, \ldots, N_h,$$

которую можно записать также в форме

$$A\alpha = g, \tag{2.1.48}$$

где

$$a_{ij} = (\varphi_i^h, \varphi_j^h)_{L_0} + (K\varphi_j^h, \varphi_i^h), \quad g_i = (f, \varphi_i^h), \quad i, j = 1, 2, \ldots, N_h.$$

После вычисления коэффициентов $\{\alpha_i\}$ приближенное решение строится по формуле (2.1.46). Относительно сходимости u^h справедливо следующее утверждение: если уравнение (2.1.43) имеет не более одного обобщенного решения, последовательность $\{F_h\}$ полна в F_{L_0} (в смысле определения п. 2.1.2) и оператор $L_0^{-1}K$ вполне непрерывен в F_{L_0}, то последовательные приближения u^h, получаемые методом Галеркина, сходятся в F_{L_0} к точному (обобщенному) решению уравнения (2.1.43). Отметим, что оператор $L_0^{-1}K$ будет вполне непрерывен F_{L_0}, если K ограничен в F, а L_0^{-1} вполне непрерывен в F.

В приведенном здесь алгоритме метода Галеркина базисные функции φ_i^h можно снова выбирать не обязательно удовлетворяющими естественным граничным условиям.

Пусть уравнение (2.1.26) рассматривается с краевыми условиями (2.1.27) (здесь уже функции $\{B_i(x)\}_{i=1}^2$ не предполагаются тождественно равными нулю, как это было в методе Ритца). В этом случае, как мы уже знаем, в качестве $\{\varphi_i^h\}$ можно взять систему линейно независимых функций, принадлежащих $\overset{\circ}{W}{}_2^1$ и отвечающих требова-

нию полноты F_h в F_{L_0}. Тогда система (2.1.47) имеет вид

$$\int\limits_D \left\{ \sum_{i,j=1}^{2} A_{ij}(x) \frac{\partial u^h}{\partial x_i} \frac{\partial \varphi_s^h}{\partial x_j} + \sum_{i=1}^{2} B_i(x) \frac{\partial u^h}{\partial x_i} \varphi_s^h + q(x) u^h \varphi_s^h - f \varphi_s^h \right\} dD = 0,$$

а элементы матрицы A в (2.1.48) есть

$$a_{ke} = \int\limits_D \left\{ \sum_{i,j=1}^{2} A_{ij}(x) \frac{\partial \varphi_e^h}{\partial x_i} \frac{\partial \varphi_k^h}{\partial x_j} + \sum_{i=1}^{2} B_i(x) \frac{\partial \varphi_e^h}{\partial x_i} \varphi_k^h + q(x) \varphi_e^h \varphi_k^h \right\} dD.$$

Можно показать, что при рассмотрении этой задачи условия сходимости приближений по Галеркину выполнены, если $F = L_2(D)$.

Заметим, что если в (2.1.47) принять $K = 0$, то метод Галеркина приводит к той же самой системе (2.1.48), что и метод Ритца, т. е. эти два метода для положительно определенных симметричных операторов совпадают.

Рассмотрим теперь метод, который в определенном смысле можно считать одной из модификаций метода Галеркина. В этом методе оператор L_0, вообще говоря, не является симметричным и положительно определенным. Пусть существует ограниченный оператор L_0^{-1}, определенный на всем F. Тогда уравнение (2.1.43) эквивалентно следующему:

$$u + L_0^{-1} K u = f', \quad f' = L_0^{-1} f. \tag{2.1.49}$$

Обозначим через F_1 гильбертово пространство со скалярным произведением $(u, v)_1 = (L_0 u, L_0 v)$ и нормой $\|u\|_1 = \|L_0 u\|$. Метод Галеркина для уравнения (2.1.49) можно сформулировать следующим образом. Пусть F_h — конечномерные подпространства из F_1 с базисами $\{\varphi_i^h\}_{i=1}^{N_h}$. Приближенное решение ищется в виде

$$u^h = \sum_{i=1}^{N_h} \alpha_i \varphi_i^h,$$

где неизвестные $\{\alpha_i\}_{i=1}^{N_h}$ определяются из системы линейных уравнений

$$(u^h, \varphi_i^h)_1 + (L_0^{-1} K u^h, \varphi_i^h)_1 = (f', \varphi_i^h)_1, \quad i = 1, 2, \ldots, N_h. \tag{2.1.50}$$

Систему (2.1.50) можно записать в эквивалентной форме

$$(L_0 u^h, L_0 \varphi_i^h) + (K u^h, L_0 \varphi_i^h) = (f, L_0 \varphi_i^h)_1, \quad i = 1, 2, \ldots, N_h. \tag{2.1.51}$$

Формулы (2.1.51) описывают известный вариационный метод — метод моментов.

Ранее было отмечено, что если уравнение $Lu = f$ имеет единственное решение, последовательность $\{F_h\}$ полна в F_1 и оператор $L_0^{-1}K$ вполне непрерывен в F_1, то последовательность сходится к точному решению как в пространстве F, так и в F_1.

Одним из сложных вопросов в рассматриваемом методе (2.1.50) или (2.1.51) является вопрос о выборе базисных функций. Если априори задать функции $\{\varphi_i^h\}_{i=1}^{N_h}$ с известными аппроксимирующими свойствами, то часто трудно исследовать свойства системы $\{\psi_i^h\}_{i=1}^{N_h}$, где $\psi_i^h = L_0\varphi_i^h$. Это в свою очередь затрудняет изучение таких вопросов, как оценка скорости сходимости, учет особенности решения и специфики задачи.

Рассмотрим один из алгоритмов построения базисных функций в данном методе.

Пусть область значений оператора K и функция f принадлежат некоторому подпространству $F(K, f) \subset F$. Зададим в $F(K, f)$ исходную систему координатных функций $\{\psi_i^h\}_{i=1}^{N_h}$ с финитными носителями порядка h так, чтобы последовательность $\{\psi_i^h\}_{i=1}^{N_h}$ $(h = h_1, h_2, \ldots)$ была полна в $F(K, f)$. Построим функции $\{\varphi_i^h\}_{i=1}^{N_h}$, где $\varphi_i^h = L_0^{-1}\psi_i^h$. Эти функции линейно независимы при каждом h. Примем их за базисные при решении уравнения (2.1.49) при помощи метода Галеркина.

Отметим некоторые свойства рассмотренного алгоритма построения базисных функций. По построению функции $\{\varphi_i^h\}_{i=1}^{N_h}$ обладают особенностями решения u, свойственными оператору L_0, а за счет специального выбора системы $\{\psi_i^h\}_{i=1}^{N_h}$ можно учесть те или иные особенности функции $\omega = f - Ku$, которые часто априори известны.

В некоторых случаях может оказаться, что ω обладает лучшими дифференциальными свойствами по сравнению с самим решением уравнения. Тогда можно попытаться при помощи малого числа исходных базисных функций добиться эффективной аппроксимации ω и надеяться на достаточно быструю сходимость u^h к u.

Если решение уравнения (2.1.49) зависит от переменных x_i $(i = 1, 2, \ldots, n)$, а $F(K, f)$ состоит из функций, зависящих лишь от x_i $(i = 1, 2, \ldots, m < n)$, то достаточно ввести координатные функции $\{\psi_i^h\}_{i=1}^{N_h}$, зависящие лишь от x_i $(i = 1, 2, \ldots, m < n)$, и с их помощью аппроксимировать $\omega \subset F(K, f)$. Само же по себе решение u будет прибли-

жаться посредством u^h по всем переменным. Это обстоятельство на практике приводит к значительному уменьшению количества координатных функций, а следовательно, и порядка решаемой системы (2.1.50), что особенно важно при решении многомерных задач математической физики.

Если переписать (2.1.50) в эквивалентном виде

$$\sum_{j=1}^{N_h} \alpha_j(\psi_j^h, \psi_i^h) = -\sum_{j=1}^{N_h} \alpha_j(K\varphi_j^h, \psi_i^h) + (f, \psi_i^h), \tag{2.1.52}$$

$$j = 1, 2, \ldots, N_h,$$

то легко заметить, что в силу финитности ψ_j^h $(j = 1, 2, \ldots, N_h)$ в левой части уравнения (2.1.52) возникает ленточная (либо разреженная) матрица, что во многих случаях облегчает решение системы при помощи итерационных методов. В силу финитности ψ_j^h упрощается также вычисление $\{\varphi_i^h\}$, элементов матриц и значений (f, ψ_j^h).

В силу сказанного выше можно предположить, что метод Галеркина при использовании специальных координатных функций $\{\varphi_i^h\}$ может оказаться достаточно эффективным при решении некоторых краевых задач, в которых можно достаточно быстро строить L_0^{-1} (например, когда оператор L_0 обратим в явном виде: дифференциальный оператор в уравнении переноса, оператор Лапласа в квадрате, круге и т. д.).

2.1.4. Метод наименьших квадратов

Метод наименьших квадратов получил широкое распространение при решении краевых задач математической физики. Для случая операторного уравнения

$$Lu = f \tag{2.1.53}$$

в гильбертовом пространстве F этот метод имеет следующую схему.

Пусть F_h — конечномерные подпространства из F с базисами $\varphi_1^h, \ \ldots \ \varphi_{N_h}^h$, причем $F_h \subseteq \Phi(L)$ — области определения L (заметим, что можно рассматривать не сам оператор L, а его расширение $\widehat{L}$, такое, чтобы $\Phi(\widehat{L})$ было полным пространством).

Тогда приближенные решения (2.1.38) строятся при помощи метода наименьших квадратов, исходя из равенств

$$\frac{\partial}{\partial \alpha_i}\|Lu^h - f\|^2 = 0, \quad i = 1, 2, \ldots, N_h. \tag{2.1.54}$$

При этом возникает система линейных уравнений (2.1.39) с матрицей $A = (a_{ij})$ и вектором $g = (g_i)$, где

$$a_{ij} = (L\varphi_i^h, L\varphi_j^h), \quad g_i = (f, L\varphi_i^h), \quad 1 \le i, j \le N_h. \tag{2.1.55}$$

Если оператор L имеет обратный L^{-1}, то матрица A симметрична и положительно определена.

Сформулируем достаточные условия сходимости метода наименьших квадратов.

Последовательные приближения u^h метода наименьших квадратов сходятся в F к точному решению u уравнения (2.1.53), если оно однозначно разрешимо, последовательность пространства LF_h полна в $\Phi(L)$, а оператор L^{-1} существует и ограничен.

Поясним смысл второго требования. Полнота пространств LF_h, как и в 2.1.2, означает[3)], что для любых $u \subset \Phi(L)$ и $\varepsilon > 0$ найдется такое $\hat{h} = \hat{h}(u, \varepsilon) > 0$, что

$$\inf_{\omega \subset F_h} \|Lu - L\omega\| < \varepsilon \tag{2.1.56}$$

для всех F_h с $h < \hat{h}$. Далее, очевидно, что при единственности решения u задачи (2.1.43) введенные требования будут обеспечивать как

$$Lu^h \underset{h\to 0}{\to} Lu,$$

так и

$$u^h \underset{h\to 0}{\to} u.$$

Более сложным, чем для двух предыдущих методов, является вопрос об удовлетворении предельного решения граничным условиям, когда метод наименьших квадратов применяется для решения краевых задач математической физики. Опишем кратко два возможных подхода к решению этого вопроса.

[3)] Заметим, что выражение LF_h имеет смысл, так как по предположению $F_h \subset \Phi(L)$.

Первый, наиболее очевидный путь — требовать от функций пространств F_h точного удовлетворения граничным условиям. Уже в случае смешанной краевой задачи для эллиптического уравнения (2.1.26) этот подход оказывается весьма сложным в практической реализации.

Второй возможный путь — использовать «весовой метод» для постановки дополнительной вариационной задачи. Идея подхода заключается в следующем. Дифференциальному уравнению с частными производными порядка $2m$

$$Lu = f \quad \text{в} \quad D \tag{2.1.57}$$

с краевыми условиями

$$L_i u = f_i \quad \text{на} \quad \partial D, \quad i = 1, 2, \ldots, m, \tag{2.1.58}$$

ставится в соответствие функционал

$$J_h(u) = \|Lu - f\|^2 + \sum_{i=1}^{m} c_i(h) \|L_i u - f_i\|^2, \tag{2.1.59}$$

где $\{c_i(h)\}_{i=1}^m$ — положительные функции параметра h, характеризующего последовательность подпространств F_h. Для разностного аналога самосопряженной задачи (2.1.57), (2.1.58) на гладких решениях берем

$$c_i(h) = h^{-2(2m - m_i - 1/2)},$$

где m_i — порядок старшей производной в операторе L_i. Теперь приближения u^h методом наименьших квадратов ищутся как решения вариационных задач

$$\inf_{u \subset F_h} J_h(u) = J_h(u^h).$$

Функции u^h сходятся к u при $h \to 0$, причем асимптотически удовлетворяется как само уравнение (2.1.57), так и краевые условия (2.1.58). При этом функции подпространств F_h не обязательно удовлетворяют краевым условиям.

2.2. Построение базисных функций для решения одномерных задач

В предыдущем разделе были рассмотрены некоторые методы для приближенного решения задач — методы Ритца, Галеркина, наименьших квадратов. При их формулировке приближенное решение задачи $u^h = \sum\limits_{i=1}^{N_h} \alpha_i \varphi_i^h$ принадлежало некоторому подпространству F_h, базисом в котором являются функции $\varphi_1^h,\ \varphi_2^h,\ \ldots,\ \varphi_{N_h}^h$. При этом вид самих функций $\{\varphi_i^h\}$ не конкретизировался. Построение этих функций и исследование их аппроксимирующих свойств будут осуществлены в этом и следующем разделах.

Основное внимание будет уделяться базисным функциям с конечным (финитным) носителем, т. е. таким, каждая из которых только в сравнительно небольшой (порядка шага сетки) окрестности отлична от нуля, а вне ее тождественна равно нулю.

Оказалось, что решение искомой задачи зачастую удобно искать в виде линейной комбинации функций с конечным носителем при неизвестных коэффициентах, которые выбираются на основе минимума того или иного функционала, связанного с вариационным принципом. Эта методология была применена к различным классам задач и привела к весьма эффективному алгоритму построения разностных систем, который мы постараемся проиллюстрировать в дальнейшем.

2.2.1. Кусочно-постоянные финитные функции

Рассмотрим наиболее простые функции с конечным носителем — кусочно-постоянные (ступенчатые) функции. Пусть $D = (a, b) \subset R$. Введем на $\overline{D} = [a, b]$ сетку $a = x_0 < x_1 < \ldots < x_{N-1} < x_N = b,\ h_i = x_i - x_{i-1},\ h = \max\ h_i\ (i = 1, 2, \ldots, N)$, разбив тем самым $[a, b]$ на $N \equiv N_h$ подобластей $D_i = (x_{i-1}, x_i)\ (i = 1, 2, \ldots, N)$ (*конечных элементов*). Зададим на каждом (x_{i-1}, x_i) характеристическую функцию

$$\varphi_i^h(x) = \begin{cases} 1, & x \in (x_{i-1}, x_i], \\ 0, & x \notin (x_{i-1}, x_i]. \end{cases}$$

Набор таких функций $\{\varphi_i^h\}$ примем в качестве базисных при решении соответствующей задачи (например, интегрального уравнения типа Фредгольма, рассматриваемого в $L_2(D)$). Линейную оболочку функций $\varphi_i^h(x)$ $(i = 1, 2, \ldots, N)$ обозначим через F_h.

Рассмотрим некоторые свойства этих базисных функций $\varphi_1^h,\ \varphi_2^h,\ \ldots,\ \varphi_{N_h}^h$. Прежде всего отметим, что все они линейно независимы, причем $(\varphi_i^h, \varphi_k^h)_{L_2(D)} = 0$ при $i \neq k,\ (\varphi_i^h, \varphi_i^h)_{L_2(D)} = h_i$. Множество F_h принадлежит любому из пространств $L_p(D)$ $(p = 1, 2, \ldots)$. Аппроксимирующие свойства данных функций даются следующим утверждением: для любой функции $u(x) \in W_p^1(D)$ существует такая линейная комбинация $u_I(x) \in F_h$, что

$$\inf_{v \in F_h} \|u - v\|_{L_p(D)} \leq \|u - u_I\|_{L_p(D)} \leq h\|u\|_{W_p^1(D)}, \tag{2.2.1}$$

где нормы в $L_p(D)$, $W_p^1(D)$ задаются выражениями

$$\|u\|_{W_p^1(D)} = \|u\|_{L_p(D)} + \left\|\frac{du}{dx}\right\|_{L_p(D)},$$

$$\|u\|_{L_p(D)} = \left(\int_D |u(x)|^p\right)^{1/p}, \quad 1 \leq p < \infty,$$

$$\|u\|_{L_\infty(D)} = \operatorname*{supvrai}_{x \in D} |u(x)|.$$

Для доказательства данного утверждения выберем в качестве u_I функцию вида

$$u_I(x) = \sum_{i=1}^{N} u_i \varphi_i^h(x),\ u_i = \frac{1}{h_i} \int_{x_{i-1}}^{x_i} u(t)dt.$$

Тогда

$$\|u-u_I\|_{L_p(D)} = \left(\int\limits_a^b |u-u_I|^p dx\right)^{1/p} = \left(\sum_{i=1}^N \int\limits_{x_{i-1}}^{x_i} |u-u_I|^p dx\right)^{1/p} =$$

$$= \left(\sum_{i=1}^N \int\limits_{x_{i-1}}^{x_i} \left|\frac{1}{h_i}\int\limits_{x_{i-1}}^{x_i} dt \int\limits_t^x \frac{du}{d\eta}(\eta)\, d\eta\right|^p dx\right)^{1/p} \leq$$

$$\leq \left(\sum_{i=1}^N h_i \left(\int\limits_{x_{i-1}}^{x_i} \left|\frac{du}{d\eta}\right| d\eta\right)^p\right)^{1/p}.$$

Используя *неравенство Гельдера*

$$\left|\int\limits_D u(x)v(x)\, dx\right| \leq \left(\int\limits_D |u(x)|^p\, dx\right)^{1/p} \left(\int\limits_D |v(x)|^q\, dx\right)^{1/q},$$

$$q \geq 1,\ \frac{1}{p}+\frac{1}{q}=1,$$

получим

$$\int\limits_{x_{i-1}}^{x_i} \left|\frac{du}{d\eta}\right| d\eta \leq (x_i - x_{i-1})^{1/q} \left(\int\limits_{x_{i-1}}^{x_i} \left|\frac{du}{d\eta}\right|^p d\eta\right)^{1/p} \leq$$

$$\leq h_i^{1/q} \left(\int\limits_{x_{i-1}}^{x_i} \left|\frac{du}{d\eta}\right|^p d\eta\right)^{1/p}.$$

Следовательно,

$$\|u-u_I\|_{L_p} \leq \left(\sum_{i=1}^N h_i^{1+\frac{p}{q}} \int\limits_{x_{i-1}}^{x_i} \left|\frac{du}{dx}\right|^p dx\right)^{1/p} \leq h\left(\sum_{i=1}^N \int\limits_{x_{i-1}}^{x_i} \left|\frac{du}{dx}\right|^p dx\right)^{1/p} =$$

$$= h\left\|\frac{du}{dx}\right\|_{L_p(D)} \leq h\|u\|_{W_p^1(D)},$$

что и требовалось доказать.

Из данного утверждения следует, что подпоследовательность пространств $\{F_h\}$ полна в $L_p(D),\ 1 \leq p < \infty$.

2.2.2. Кусочно-линейные базисные функции

Рассмотрим одни из наиболее распространенных финитных функций, нашедших широкое применение при построении разностных схем, а именно кусочно-линейные финитные функции (которые также называют *«функциями-крышками», «функциями-домиками»*).

Пусть функции $u(x)$, для которых мы хотим построить подходящие аппроксимации, определены на конечной области $D = (a, b)$. Введем на $[a, b] = \overline{D}$ сетку $a = x_0 < x_1 < \ldots < x_{N-1} < x_N = b,\ h_i = x_i - x_{i-1},\ h = \max\ h_i\ (i = 1, 2, \ldots, N)$ и поставим в соответствие каждому узлу сетки функцию

$$\varphi_i^h(x) = \begin{cases} \dfrac{x - x_{i-1}}{h_i}, & x \in (x_{i-1}, x_i), \\ \dfrac{x_{i+1} - x}{h_{i+1}}, & x \in (x_i, x_{i+1}), \\ 0, & x \notin (x_{i-1}, x_{i+1}),\ i = 1, 2, \ldots, N-1, \end{cases} \tag{2.2.2}$$

$$\varphi_0^h(x) = \begin{cases} \dfrac{x_1 - x}{h_1}, & x \in (x_0, x_1), \\ 0, & x \notin (x_0, x_1), \end{cases} \quad \varphi_N^h(x) = \begin{cases} \dfrac{x - x_{N-1}}{h_N}, & x \in (x_{N-1}, x_N), \\ 0, & x \notin (x_{N-1}, x_N). \end{cases}$$

Очевидно, что эти функции линейно независимы и каждая из них отлична от нуля лишь в интервале длиной порядка $2h$. Линейную оболочку $\{\varphi_i^h\}$ обозначим через F_h. Функции из F_h являются непрерывными кусочно-линейными функциями, обладающими суммируемой с любой конечной степенью первой производной. Таким образом, $F_h \subset C(D)$, $F_h \subset W_2^1(D)$, а множество F_h можно обозначить соответственно через $C_h(D)$ или $W_2^{1,h}(D)$ в зависимости от того, в каком пространстве исследуется проблема аппроксимации с помощью $\{\varphi_i^h(x)\}$.

Если взять $v^h = \sum\limits_{i=1}^{N} \alpha_i \varphi_i^h(x) \in F_h$, то, как нетрудно заметить, $v^h(x_i) = \alpha_i$, а, стало быть, коэффициенты в этой линейной комбинации несут явный смысл: коэффициент α_i при $\varphi_i^h(x)$ равен значению функции $v^h(x)$ в точке x_i. Отметим, что функции $\{\varphi_i^h\}$ почти ортогональны, т. е. только для соседних функций скалярное произведение (например, в

$L_2(D)$) отлично от нуля:

$$\int_a^b \varphi_i^h(x)\varphi_j^h(x)\,dx = \begin{cases} 0 & \text{при} \quad |i-j|>1, \\ \neq 0 & \text{при} \quad |i-j|\le 1. \end{cases}$$

Это свойство является причиной того, что система уравнений в методах Ритца и Галеркина (в применении к дифференциальным уравнениям) при использовании функций (2.2.2) в качестве базисных имеет разреженную матрицу.

Изучим аппроксимирующие свойства функций (2.2.2), или, что одно и то же, множеств F_h-линейных оболочек $\{\varphi_i^h\}_{i=1}^N$ $(N=1,2,3,\ldots)$.

Справедливо следующее утверждение: если $u(x)\in W_2^2(D)$, то существует такая функция $u_1\in F_h\equiv W_2^{1,h}(D)$, что

$$\begin{aligned} \|u-u_I\|_{L_2(D)} &\le h^2\left\|\frac{d^2u}{dx^2}\right\|_{L_2(D)} \le h^2\|u\|_{W_2^2(D)}, \\ \|u-u_I\|_{W_2^1(D)} &\le ch\left\|\frac{d^2u}{dx^2}\right\|_{L_2(D)} \le ch\|u\|_{W_2^2(D)}, \end{aligned} \tag{2.2.3}$$

где постоянная c не зависит от h и от $u(x)$. Докажем это утверждение. Так как функцию $u(x)\in W_2^1(D)$ (а тем более $u(x)\in W_2^2(D)$) в одномерном случае можно отождествить с непрерывной функцией (т. е. «$u(x)$ является непрерывной»), то $u(x)$ имеет конечное значение $u(x_i)$ в любом узле x_i $(i=0,1,2,\ldots,N)$. Поэтому можно рассмотреть линейную комбинацию

$$u_I(x)=\sum_{i=0}^{N} u(x_i)\varphi_i^h(x).$$

Оценим разность $u-u_I$ в произвольной точке $x\in(x_{i-1},x_i)$. Для этого запишем следующее тождество при $x\in(x_{i-1},x_i)$:

$$u(x)-u_I(x)=\frac{1}{h_i}\int_{x_{i-1}}^{x} dx'\int_{x_{i-1}}^{x_i} dx''\int_{x''}^{x'}\frac{d^2u(t)}{dt^2}\,dt. \tag{2.2.4}$$

Применим к (2.2.4) неравенство Коши — Буняковского, расширяя при этом пределы интегрирования:

$$|u(x)-u_I(x)| \le \frac{1}{h_i}\int\limits_{x_{i-1}}^{x_i} dx' \int\limits_{x_{i-1}}^{x_i} dx'' \int\limits_{x_{i1}}^{x_i} \left|\frac{d^2u}{dt^2}(t)\right| \, dt \le$$

$$\le h_i^{3/2} \left(\int\limits_{x_{i-1}}^{x_i} \left|\frac{d^2u}{dt^2}\right|^2 \, dt \right)^{1/2}, \quad x \in (x_{i-1}, x_i).$$

Следовательно,

$$\int\limits_{x_{i-1}}^{x_i} |u(x)-u_I(x)|^2 \, dx \le h_i^4 \int\limits_{x_{i-1}}^{x_i} \left|\frac{d^2u}{dx^2}\right|^2 \, dx.$$

Суммируя последнее неравенство по $i = 1, 2, \ldots, N$, а также оценивая h_i через h, получим первую из оценок (2.2.3).

Если теперь сначала продифференцировать (2.2.4), а затем провести те же самые рассуждения и учесть оценку для $u - u_I$ в норме $\|\cdot\|_{L_2}$, то придем ко второй оценке из (2.2.3).

Аналогично (исходя из (2.2.4)) доказывается следующий результат: если $u(x) \in C^{(2)}(\overline{D})$, то

$$\|u-u_I\|_{C(\overline{D})} \le h^2 \|u\|_{C^{(2)}(\overline{D})}. \tag{2.2.5}$$

Из приведенных выше результатов следует, что

$$\begin{gathered} \inf_{v \in W_2^{1,h}(D)} \|u-v\|_{W_2^k(D)} \le Ch^{2-k}\|u\|_{W_2^2(D)}, \quad k = 0, 1, \ldots, \\ \inf_{v \in C_h)} \|u-v\|_{C(\overline{D})} \le h^2 \|u\|_{C^{(2)}(\overline{D})} \end{gathered} \tag{2.2.6}$$

(где $W_2^0(D) \equiv L_2(D)$) и что последовательность подпространств F_h полна в каждом из пространств $L_2(D)$, $W_2^1(D)$, $C(D)$.

Отметим, что все сформулированные утверждения остаются справедливыми, если вместо функций (2.2.2) нормированные кусочно-

линейные функции

$$\varphi_i^h(x) = \frac{1}{\sqrt{h_i}} \begin{cases} \dfrac{x - x_{i-1}}{h_i}, & x \in (x_{i-1}, x_i), \\ \dfrac{x_{i+1} - x}{h_{i+1}}, & x \in (x_i, x_{i+1}), \\ 0, & x \notin (x_{i-1}, x_{i+1}), \end{cases} \tag{2.2.7}$$

$$i = 0, 1, 2, \ldots, N, \quad h_0 \equiv h_1.$$

В данном случае для доказательства всех утверждений в качестве u_I достаточно принять функцию $u_I(x) = \sum\limits_{i=0}^{N} \sqrt{h_i} u(x_i) \varphi_i^h(x)$. Использование нормированных кусочно-линейных функций в ряде случаев может оказаться предпочтительным. Так, при дополнительном условии $h \le C \min\limits_i h_i$ — условии квазиравномерности сетки — применение их в качестве базисных приводит к системам уравнений, число обусловленности которых оказывается меньшим по сравнению со случаем, когда используются функции (2.2.2).

Расмотренные в данном пункте кусочно-линейные функции широко используются при численном решении одномерного уравнения диффузии

$$Lu \equiv -\frac{d}{dx} p(x) \frac{du}{dx} + r(x) \frac{du}{dx} + q(x) u(x) = f(x), \tag{2.2.8}$$

$$a < x < b,$$

с граничными условиями

$$u(a) = u(b) = 0 \tag{2.2.9}$$

или условиями вида

$$\frac{du}{dx}(a) = \frac{du}{dx}(b) = 0 \tag{2.2.10}$$

(а также с другими условиями при $x = a$, $x = b$). Здесь $p = p(x) \ge p_0 = = \text{const} > 0$ — коэффициент диффузии $q = q(x) \ge 0$ — коэффициент поглощения частиц $f = f(x)$ — источники диффундирующей субстанции. Считаем, что p, q, r, f — кусочно-непрерывные функции на $[a, b]$ с возможными точками разрыва первого рода. Как мы знаем из 2.1, задача (2.2.8), (2.2.10) является задачей с естественными граничными условиями. Поэтому при ее численном решении с помощью мето-

да Галеркина можно воспользоваться базисными функциями (2.2.2) или (2.2.7), а в качестве F_h принять $W_2^{1,h}(D)$ — линейную оболочку $\{\varphi_i^h\}_{i=0}^N$. Если решается задача (2.2.8), (2.2.9), то, поскольку она является задачей с главными краевыми условиями, нам необходимо выбрать F_h таким, чтобы любая функция $v^h = \sum\limits_{i=0}^{N} \alpha_i \varphi_i^h(x) \in F_h$ удовлетворяла условиям (2.2.9). В одномерном случае добиться этого просто. Так, условия $v^h(a) = v^h(b) = 0$ приводятся к соотношениям $\alpha_0 \equiv 0$, $\alpha_N \equiv 0$. И мы заключаем, что в качестве подпространства F_h при расмотрении задачи (2.2.8), (2.2.9) необходимо брать всевозможные линейные комбинации вида $v^h = \sum\limits_{i=1}^{N-1} \alpha_i \varphi_i^h(x)$. Базисом в таком подпространстве F_h являются функции $\{\varphi_i^h\}_{i=1}^{N-1}$, каждая из которых удовлетворяет условиям (2.2.9). Замечаем, что это подпространство принадлежит $\overset{\circ}{W}{}_2^1(a,b) = \{u(x) : u \in W_2^1(a,b);\ u(a) = u(b) = 0\}$. Поэтому его можно также обозначить через $\overset{\circ}{W}{}_2^{1,h}(D)$. Таким образом, при численном решении (2.2.8), (2.2.9) надо использовать подпространство $F_h \equiv \overset{\circ}{W}{}_2^{1,h}(D) \subset \overset{\circ}{W}{}_2^1(D)$, определяемое как линейная оболочка кусочно-линейной функции $\varphi_1^h(x),\ \varphi_2^h(x),\ \ldots\ \varphi_{N-1}^h(x)$. Нетрудно заметить, что для функции $u(x) \in \overset{\circ}{W}{}_2^1(D) \cap W_2^2(D)$ оценки погрешности аппроксимации ее с помощью функций из $\overset{\circ}{W}{}_2^{1,h}(D)$ останутся справедливыми (сохраняются также и доказательства этих оценок).

2.2.3. Общий подход к построению подпространств кусочно-полиномиальных функций

В предыдущих пунктах мы ввели кусочно-постоянные и кусочно-линейные базисные функции. Однако при численном решении, например задачи (2.2.8), (2.2.9), весьма эффективным может оказаться использование подпространств F_h, являющихся линейными оболочками кусочно-полиномиальных функций высокого порядка. Рассмотрим один из способов построения таких подпространств.

Введем базисные функции $\{\varphi_i(x)\}_{i=1}^m,\ m = (p+1)/2$, где p — нечетное положительное число: $\varphi_i(x) = 0$, если $x \notin [-1,1]$, на каждом из отрезков $[-1,0]$, $[0,1]$ функция $\varphi_i(x)$ есть многочлен степени p, причем в точках $x = -1$ и $x = 1$ функция $\varphi_i(x)$ и все ее производные до $(m-1)$-го порядка включительно равны нулю, а в точке $x = 0$ единственной

ненулевой производной является

$$\left.\frac{d^{i-1}\varphi_i(x)}{dx^{i-1}}\right|_{x=0} = 1, \quad 1 \le i \le m. \tag{2.2.11}$$

Построим теперь на $[a,b]$ равномерную сетку с шагом h и узлами $a = x_0 < x_1 = a + h < \ldots < a + Nh = b$, где $h = (b-a)/N$, N — целое положительное число. Образуем множество функций вида

$$u^h(x) = \sum_{i=1}^{m}\sum_{j=0}^{N} u_{ij}^h \varphi_{ij}^h(x), \tag{2.2.12}$$

где

$$\varphi_{ij}^h(x) = \varphi_i\left(\frac{x-a}{h} - j\right), \quad 0 \le j \le N,\ 1 \le i \le m. \tag{2.2.13}$$

Обозначим его через $F_h \equiv H^{(m)_h}$. Базисом в $H^{(m)_h}$ является система функций $\{\varphi_{ij}^h\}_{1,m}^{j=\overline{0,N}}$.

Отметим, что из формулы (2.2.11) следует полезное свойство интерполирования при помощи функций из $H_h^{(m)}$: в точках сетки значения $u^h(x)$ и ее первых $m-1$ производных задаются коэффициентами

$$u_{ij}^h = \frac{d^{i-1}u^h(a+jh)}{dx^{i-1}}. \tag{2.2.14}$$

Рассмотрим некоторые случаи введенных подпространств.

$p = 1$ $(m = 1)$. Здесь имеется единственная базисная функция $\varphi_1(x)$, являющаяся на $[-1,0]$ и $[0,1]$ полиномом первой степени

$$\varphi_1(x) = \begin{cases} a + bx, & -1 \le x \le 0, \\ c + dx, & 0 \le x \le 1 \end{cases} \tag{2.2.15}$$

и определяемая из условий

$$\begin{aligned} a + bx|_{x=-1} = 0, \quad c + dx|_{x=1} = 0, \\ a + bx|_{x=0} = 1, \quad c + dx|_{x=0} = 1. \end{aligned} \tag{2.2.16}$$

Отсюда получаем, что $a = b = 1$, $c = -d = 1$, т. е. $\varphi_1(x)$ есть обычная «*функция-домик*» (см. рис. 2.1), а $H_h^{(1)}$ — пространство кусочно-линейных функций, рассмотренных нами в 2.2.2.

$p = 3$ $(m = 2)$. В этом случае будем иметь две базисные функции $\varphi_1(x)$ и $\varphi_2(x)$. Построим сначала $\varphi_1(x)$. Согласно изложенному выше,

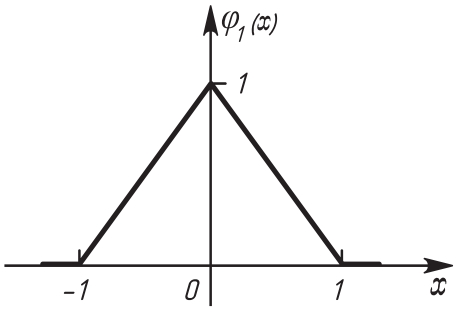


Рис. 2.1.

на отрезке $[0,1]$ функция $\varphi_1(x)$ имеет следующий вид:

$$\varphi_1(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3. \tag{2.2.17}$$

Неизвестные $a_0,\ a_1, \ldots, a_3$ определяются из условий

$$\begin{aligned} &\varphi_1(0) = 1, \quad \varphi_1(0) = 0, \\ &\varphi_1'(0) = 0, \quad \varphi_1'(1) = 0. \end{aligned} \tag{2.2.18}$$

Откуда вытекает

$$a_0 = 1,\ a_1 = 0,\ a_2 = -3,\ a_3 = 2. \tag{2.2.19}$$

Следовательно

$$\varphi_1(x) = \begin{cases} 0 \quad \text{при} \quad |x| \geq 1, \\ 1 - 3x^2 + 2x^3 = (1-x^2)(1+2x), \quad 0 \leq x \leq 1, \\ \varphi_1(-x) \quad \text{при} \quad -1 \leq x \leq 0. \end{cases} \tag{2.2.20}$$

Аналогично осуществляется построение $\varphi_2(x)$, которое на $[0,1]$ записывается в виде $\varphi_2(x) = b_0 + b_1 x + b_2 x^2 + b_3 x^3$. Коэффициенты $b_i,\ 0 \leq i \leq 3$, ищутся из условий

$$\begin{aligned} &\varphi_2(0) = 0, \quad \varphi_2(1) = 0, \\ &\varphi_2'(0) = 1, \quad \varphi_2'(1) = 0. \end{aligned} \tag{2.2.21}$$

Определив b_i, получим

$$\varphi_2(x) = \begin{cases} \quad \text{при} \quad |x| \geq 1, \\ (1-x)^2 x, \quad 0 \leq x \leq 1, \\ -\varphi_2(-x) \quad \text{при} \quad -1 \leq x \leq 0 \end{cases} \tag{2.2.22}$$

(см. рис. 2.2, 2.3). Функции из $H_h^{(2)}$ здесь имеют вид

$$u^h(x) = \sum_{j=0}^{N} (a_j \varphi_{1,j}^h(x) + b_j \varphi_{2,j}^h(x)). \tag{2.2.23}$$

При помощи таких функций легко можно определить квазиинтерполяцию u_1^h, например, решения $u(x)$ задачи (2.2.8), (2.2.9), положив

$$u_1^h(x) = \sum_{j=0}^{N} (u(a+jh)\varphi_{1,j}^h(x) + u'(a+jh)\varphi_{2,j}^h(x)). \tag{2.2.24}$$

Это еще раз дает нам представление о физической интерпретации значений коэффициентов при функциях $\varphi_{1,j}^h(x)$ и $\varphi_{2,j}^h(x)$ в приближенном решении $u^h(x)$, построенном при помощи базиса $\{\varphi_{i,j}^h\}_{i=1,2}^{j=\overline{0,N}}$ методом Галеркина.

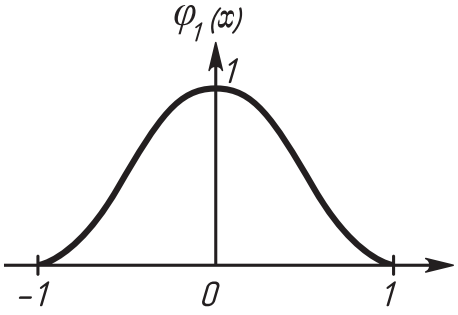


Рис. 2.2.

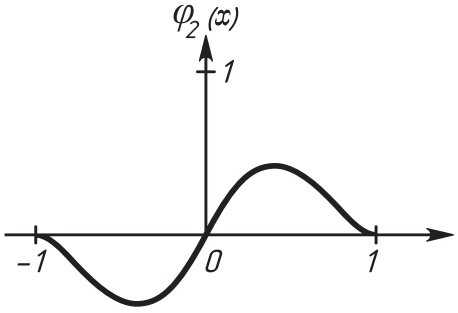


Рис. 2.3.

Опишем еще один способ построения пространства F_h. Для этого, как и ранее, возьмем отрезок $[a,b]$ вещественной оси и построим на нем сетку с узлами $a = x_0 < x_1 < \ldots < x_N < x_{N+1} = b$, где $x_k = a + kh$ $(k = 0, 1, \ldots, N+1)$ и $h = (b-a)/(N+1)$. Обозначим через $M_N^m(a,b)$ множество функций $g(x)$, удовлетворяющих следующим требованиям: во-первых, на каждом из отрезков $[x_k, x_{k+1}]$ функция $g(x)$ является многочленом степени m; во-вторых, для любых $0 \le k \le N$ и $0 \le j \le m$ выполняются равенства

$$g(x_{k,j}) = d_{k,j},$$

где $x_{k,j} = x_k + \dfrac{h}{m} j$ и $d_{k,j}$ — заданные числа; наконец, $g(a) = g(b) = 0$, т. е. всегда $d_{0,0} = d_{N,m} = 0$. Отсюда следует, что функция $g \in M_N^m(a,b)$ является кусочно-полиномиальной функцией пространства $\overset{\circ}{W}{}_2^1$, т. е. $g(x)$ — непрерывная функция с возможными разрывами первых про-

изводных в точках $\{x_k\}_{k=1}^N$. Рассмотрим, как можно явно построить функцию $g(x)$ через значения $\{d_{k,j}\}$.

Выберем произвольное значение $0 \le k \le N$ и построим функцию $g(x)$ на отрезке $[x_k, x_{k+1}]$ (обозначим ее $g_k(x)$). Как известно из теории аппроксимации, многочлен степени m, проходящий через $m+1$ точку $d_{k,0}, \ldots, d_{k,m}$, существует и единственен. Этим многочленом будет интерполяционный полином Лагранжа, который вычисляется по формуле

$$g_k(x) = \sum_{i=0}^{m} d_{k,i} \prod_{l=0,\, l\ne i}^{m} \frac{x_{k,l} - x}{x_{k,l} - x_{k,i}}. \tag{2.2.25}$$

Для случая $m = 1$ получаем обычную линейную функцию

$$g_k(x) = d_k \frac{x_{k+1} - x}{x_{k+1} - x_k} + d_{k+1} \frac{x_k - x}{x_k - x_{k+1}}. \tag{2.2.26}$$

Отсюда, в частности, следует, что пространство $M_N^1(a,b)$ совпадает с пространством $H_h^{(1)}(a,b)$.

Таким образом, для пространства $F = \overset{\circ}{W}{}_2^1(a,b)$ мы построили два вида пространств F_h (определяемых параметром h), последовательность которых полна в $\overset{\circ}{W}{}_2^1(a,b)$.

Естественно, что при наличии соответствующей гладкости у функции $u(x)$ аппроксимация ее с помощью кусочно-полиномиальных функций высокой степени будет более точной.

Так, например, если $u(x) \in W_2^4(D) \cap \overset{\circ}{W}{}_2^1(D)$, где $D = (a,b)$, то существует функция u_I вида (2.2.23), для которой справедлива оценка[4)]

$$\|u - u_I\|_{W_2^S(D)} \le c h^{4-s} \|u\|_{W_2^4(D)}, \quad 0 \le S \le 3, \tag{2.2.27}$$

с постоянной c, не зависящей от h и $u(x)$.

[4)] Г. Стрэнг, Дж. Фикс [3].

2.2.4. Построение базиса на основе тригонометрических функций и использование его в вариационных задачах

Для уяснения основных принципов метода рассмотрим конкретную задачу (2.2.8), (2.2.29) при $r \equiv 0$, $a = 0$, $b = 1$:

$$Lu \equiv -\frac{d}{dx}p(x)\frac{d\varphi}{dx} + q(x)\varphi = f(x), \tag{2.2.28}$$

$$\varphi(0) = \varphi(1) = 0, \tag{2.2.29}$$

которая сводится к минимизации в $\overset{\circ}{W}{}_2^1\,(0,1)$ функционала

$$I(v) = \int_0^1 \left[p\left(\frac{dv}{dx}\right)^2 + qv^2 - 2fv \right] dx, \tag{2.2.30}$$

т. е.

$$I(\varphi) = \inf_{v \in \overset{\circ}{W}{}_2^1(0,1)} I(v). \tag{2.2.31}$$

Предположим, что параметры задачи $p,\ q,\ f$ могут иметь разрывы первого рода в точках $\{y_i\}_{i=1}^m$, а на каждом из промежутков $y_i < x < y_{i+1}$ имеют достаточно высокую гладкость. Пусть гладкость параметров задачи обеспечивает наличие у решения φ производных ν-го порядка, непрерывных на упомянутых интервалах. Класс таких функций обозначим через $Q^{(\nu)}[0,1]$ (таким образом, предполагается, что $\varphi \in Q^{(\nu)}[0,1]$).

При достаточно большом ν для построения базисной системы $\{\omega_k\}_{k=1}^\infty$ естественно, например, использовать хорошие аппроксимационные качества тригонометрических полиномов, периоды которых больше соответствующих промежутков гладкости. Изложим более подробно это утверждение о тригонометрических полиномах.

Если на промежутке $0 \le x \le 1$ задана функция $f(x) \in C^{(\nu)}[0,1]$, то ее всегда можно (неоднозначно) доопределить в остальных точках действительной оси так, что расширенная функция $\widetilde{f}(x)$ будет обладать следующими важными свойствами:

$$\widetilde{f}(x+T) = \widetilde{f}(x), \quad -\infty < x < \infty,$$

$$\widetilde{f} \in C^{(\nu)}[x_0, x_0+T] \quad \text{для любого} \quad x_0 \in (-\infty, \infty),$$

где $T > 1$. Скорость сходимости ряда Фурье функции $\widetilde{f}(x)$ (как в пространстве $C[0,1]$, так и в пространстве $L_2[0,1]$) допускает оценку

$$\|f - \tau_N\| \le \text{const}\,\frac{\ln\ N}{N^\nu}, \tag{2.2.32}$$

где

$$\tau_N = \sum_{k=0}^{N}\left(a_k \cos\frac{2\pi kx}{T} + b_k \sin\frac{2\pi kx}{T}\right).$$

Заметим, что, не ухудшая оценки (2.2.32), можно так изменить коэффициенты a_k и b_k, что дополнительно будем иметь $\tau_N(0) = f(0)$, $\tau_N(1) = f(1)$.

Возвращаясь к задаче (2.2.28), (2.2.29) и учитывая только что сформулированное утверждение, будем аппроксимировать решение задачи на каждом из промежутков гладкости $y_i \le x \le y_{i+1}$ тригонометрическим полиномом

$$\varphi_{N_i}(x) = \sum_{k=0}^{2N_i} C_k^i T_k^i(x), \quad i = 0, 1, \ldots, m,$$

где, например,

$$T_{2l-1}^i(x) = \sin\frac{l\pi(x - y_i)}{t_i},\ T_{2l}^i(x) = \cos\frac{l\pi(x - y_i)}{t_i}, \tag{2.2.33}$$

$$t_i = y_{i+1} - y_i.$$

Рассмотрим систему функций

$$\omega_k^i(x) = \begin{cases} T_k^i(x), & \text{если} \quad x \in [y_i, y_{i+1}], \\ 0 & \text{вне этого промежутка} \end{cases} \tag{2.2.34}$$

$(i = 0, 1, \ldots, m;\ k = 0, 1, \ldots)$. Примем ее за последовательность базисных элементов (разрывных при четном k) и будем искать приближенное решение задачи (2.2.31) в виде

$$\varphi(x) = \sum_{i=1}^{m}\sum_{k=0}^{2N_i} C_k^i \omega_k^i(x), \quad 0 \le x \le 1. \tag{2.2.35}$$

В силу разрывности базисных функций (2.2.34) их линейные комбинации (2.2.35) не обеспечивают непрерывности φ и $p\ d\varphi/dx$ в точках $\{y_i\}_{i=1}^m$ и не удовлетворяют краевым условиям[5)].

Вариационный метод, основанный на минимизации функционала (2.2.30), требует выполнения главных условий

$$\varphi(0) = \varphi(1) = 0, \quad \varphi(x_i - 0) = \varphi(x_i + 0),$$

$$i = 1, 2, \ldots, m,$$

и не требует выполнения естественных

$$p\frac{d\varphi}{dx}\bigg|_{x_i-0} = p\frac{d\varphi}{dx}\bigg|_{x_i+0},$$

$$i = 1, 2, \ldots, m;$$

последние будут выполнены автоматически на элементе $\varphi^0 \in \Phi$, решающем вариационную задачу. Таким образом, от функции (2.2.35) следует потребовать выполнения главных условий

$$\varphi_{N_0}(0) = \varphi_{N_m}(1) = 0, \quad \varphi_{N_{i-1}(x_i)} = \varphi_{N_i}(x_i), \tag{2.2.36}$$

$$i = 1, 2, \ldots, m.$$

Покажем, что отклонение приближенного решения $\varphi(x)$ вариационной задачи от точного $\varphi^0(x)$ оценивается следующим образом:

$$\|\varphi - \varphi^0\|_{L_2[0,1]} \le \text{const}\,\frac{\ln N}{N^{\nu-1}}, \quad N = \min_i N_i. \tag{2.2.37}$$

Рассмотрим точное решение $\varphi^0(x)$ только в пределах промежутка гладкости $y_i \le x \le y_{i+1}$, а вне его доопределим $\varphi^0(x)$ до функции

[5)] Краевым условиям легко удовлетворить, если на промежутках $[0, x_1]$ и $[x_m, 1]$ функции $\omega_k^0(x)$ и $\omega_k^m(x)$ определить следующим образом:

$$\omega_k^0(x) = \sin\frac{k\pi x}{2x_1}, \quad \omega_k^m(x) = \sin\frac{k\pi(1-x)}{2t_m}.$$

$\psi^i(x)$ следующим образом (см. выше доопределение $f(x)$ до $\widetilde{f}(x)$):

$$\psi^i(x) \equiv \varphi^0(x) \quad \text{при} \quad y_1 \le x \le y_{i+1},$$
$$\psi^i \in C^{(\nu)}[x_0, x_0 + T_i] \quad \text{для любого} \quad x_0 \in (-\infty, \infty),$$
$$\psi^i(x + T_i) = \psi^i(x), \quad -\infty < x < \infty,$$

где $T_i = 2(y_{i+1} - y_i)$.

Каждую из функций $\varphi^i(x)$ $(i = 0, 1, \ldots, m)$ аппроксимируем конечным отрезком ее ряда Фурье

$$\sum_{k=0}^{2N_i} a_k^i T_k^i(x).$$

С учетом того, что ряд Фурье функции $\psi^i(x)$ при $\nu \ge 2$ допускает почленное дифференцирование, можем записать следующие оценки:

$$\|R_i\| \le \text{const}\, \frac{\ln N_i}{N_i^\nu}, \quad \left\| \frac{dR_i}{dx} \right\| \le \text{const}\, \frac{\ln N_i}{N_i^{\nu-1}},$$

где

$$R_i(x) \equiv \psi^i(x) - \sum_{k=0}^{2N_i} a_k^i T_k^i(x),$$

а под $\|\cdot\|$ можно понимать как $\|\cdot\|_C$, так и $\|\cdot\|_{L_2}$. Не ухудшая выписанных оценок, можно «подправить» коэффициеты a_k^i так, чтобы $R_i(y_i) = R(y_{i+1}) = 0$. Полагая, что такая корректировка осуществлена, построим непрерывную функцию, удовлетворяющую краевым условиям (2.2.29):

$$\psi(x) = \sum_{i=0}^{m} \sum_{k=0}^{2N_i} a_k^i \omega_k^i(x), \quad 0 \le x \le 1.$$

В силу положительной определенности оператора L задачи (2.2.28), (2.2.29) для всякой функции $u(x) \in \Phi(L)$ имеем

$$(Lu, u) \ge \gamma \|u\|_{L_2}^2, \quad \gamma > 0.$$

Если обозначить через V конечномерное пространство непрерывных функций вида (2.2.35), удовлетворяющих нулевым краевым

условиям, то получим, что

$$\|\varphi^0-\varphi\|_{L_2}^2 \le \frac{1}{\gamma}(L(\varphi^0-\varphi),\varphi^0-\varphi) = \frac{1}{\gamma}\min_{u\in V}(L(\varphi^0-u),\varphi^0-u) \le$$
$$\le \frac{1}{\gamma}(L(\varphi^0-\psi),\varphi^0-\psi) = \frac{1}{\gamma}\int\limits_0^1 \left[p(x)\left(\frac{dR}{dx}\right)^2 + q(x)R^2(x)\right]dx,$$

где $R(x) \equiv \varphi^0(x) - \psi(x)$.

Приведенные выше оценки для R_i и dR_i/dx позволяют оценить $R(x)$ и ее производную

$$\|R\| \le \text{const}\,\frac{\ln N}{N^\nu}, \quad \left\|\frac{dR}{dx}\right\| \le \text{const}\,\frac{\ln N}{N^{\nu-1}}, \quad N = \min_i N_i.$$

Из этих неравенств и предыдущего соотношения получаем оценку (2.2.37).

Опишем теперь кратко алгоритм численной реализации.

Минимизация функционала (2.2.30) на функциях вида (2.2.35) при дополнительных требованиях (2.2.36) с использованием метода неопределенных множителей Лагранжа приводит к следующей системе уравнений:

$$\sum_{k=0}^{2N_i}\alpha_k^i(L\omega_k^i,\omega_j^i) + \sum_{s=i}^{i+1}\lambda_s\beta_{js}^i = (f,\omega_j^i),$$

$$i = 0,1,\ldots,m; \quad j = 0,1,\ldots,2N_i, \tag{2.2.38}$$

$$\sum_{i=s-1}^{s}\sum_{k=0}^{2N_i}\alpha_k^i\beta_{ks}^i = 0, \quad s = 0,1,\ldots,m+1,$$

где

$$\beta_{ks}^i = \begin{cases} \omega_k^{s-1}(x_s), & i = s-1, \\ -\omega_k^s(x_s), & i = s,\ \omega_k^{-1}(x) \equiv \omega_k^{m+1}(x) \equiv 0, \\ 0, & i = s,\ s-1. \end{cases} \tag{2.2.39}$$

Система (2.2.38) имеет блочную структуру:

$$\begin{Vmatrix} \hat{A} & B^T \\ B & 0 \end{Vmatrix} \begin{Vmatrix} X \\ \Lambda \end{Vmatrix} = \begin{Vmatrix} F \\ 0 \end{Vmatrix}, \tag{2.2.40}$$

где X — вектор, объединяющий множество всех α_k^i; Λ — вектор, компоненты которого есть множители Лагранжа; T — символ транспонирования.

Из (2.2.40) имеем

$$\hat{A}X + B^T\Lambda = F, \quad BX = 0, \tag{2.2.41}$$

откуда после исключения X формально получаем

$$(B\hat{A}^{-1}B^T)\Lambda = B\hat{A}^{-1}F. \tag{2.2.42}$$

Из положительной определенности матрицы $\hat{A}$ следуют существование $\hat{A}^{-1}$ и положительная определенность $B\hat{A}^{-1}B^T$. Действительно,

$$(B\hat{A}^{-1}B^TW, W) = (\hat{A}^{-1}B^TW, B^TW) = (\hat{A}V, V) > 0,$$

где $V = \hat{A}^{-1}B^TW \neq 0$ при $W \neq 0$ (обращение W в нуль при $B^TW = 0$ элементарно следует из (2.2.39) и (2.2.33)).

Из блочно-диагональной структуры матрицы $\hat{A}$ легко находим

$$B\hat{A}^{-1}B^T = \sum_{i=0}^{m} B_iA_i^{-1}B_i^T, \quad BA^{-1}F = \sum_{i=0}^{m} B_iA_i^{-1}F_i, \tag{2.2.43}$$

где A_i и B_i — матрицы с элементами $(A\omega_k^i, \omega_j^i)$, (β_{ks}^i) соответственно, а F_i — векторы с компонентами $(f, \omega_0^i), \ \ldots, \ (f, \omega_{2N_i}^i)$.

Как следует из (2.2.43), обращение матрицы $\hat{A}$ сводится к обращению каждого из блоков этой матрицы, т. е., как правило, к обращению матриц невысокого порядка. Если, далее, число m невелико, то порядок системы (2.2.42) мал, а эта система с положительно определенной симметричной матрицей решается без труда.

Наконец, из первого уравнения в (2.2.41) находим

$$X_i = A_i^{-1}(F_i - B_i^T\Lambda), \tag{2.2.44}$$

где

$$X_i = (\alpha_0^i, \ \ldots, \ \alpha_{2N_i}^i), \quad i = 0, 1, \ldots, m.$$

Проиллюстрируем рассмотренный подход на одной из типичных диффузионных задач теории переноса частиц:

$$\frac{1}{r}\frac{d}{dr}rp\frac{d\varphi}{dr} + q\varphi = f,$$

$$\left.\frac{d\varphi}{dr}\right|_{r=0} = \left.\frac{d\varphi}{dr}\right|_{r=R} = 0,$$

$p(r)$, $q(r)$ и $f(r)$ — кусочно-постоянные неотрицательные функции.

Рассмотрим конкретный пример.

Таблица 2.1.

r	$p(r)$	$q(r)$	$f(r)$
$0 \le r < 12,7$	$1,333$	$0,2$	0
$12,7 \le r < 13$	$0,3115$	$0,15$	0
$13 \le r \le 15$	$0,1282$	$0,015$	1

Таблица 2.2.

n_1	n_2	n_3	n	E
2	3	3	8	$0,138$
3	5	5	13	$0,141 \cdot 10^{-1}$
5	7	9	21	$0,501 \cdot 10^{-3}$
6	9	11	26	$0,782 \cdot 10^{-4}$

В таблице 2.2 приводится относительная погрешность приближенного решения в форме

$$E = \|\varphi - \varphi^0\| / \|\varphi^0\| \tag{2.2.45}$$

в зависимости от числа базисных функций[6)], причем норма в (2.2.45) вычислена в равномерной метрике, т. е. $\|\cdot\|_C$.

Здесь $n = \sum\limits_i n_i$, n_i — число базисных функций в i-м промежутке гладкости.

[6)] Из таблицы 2.2 видно быстрое убывание E по мере увеличения числа базисных функций, что свидетельствует о больших возможностях рассматриваемого подхода в реализации вариационно-разностных методов решения задач математической физики.

2.3. Построение базисных функций для решения многомерных задач

Рассмотрим проблему построения базисов в многомерном случае. Поскольку определение кусочно-постоянных базисных функций здесь очевидно и аналогично одномерному случаю, то дальнейшее изложение мы начинаем с изучения кусочно-линейных финитных функций.

2.3.1. Кусочно-линейные функции на прямоугольнике

Пусть в прямоугольнике $D = \{0 < x < a,\ 0 < y < b\}$ ставится задача аппроксимации заданной функции $u(x, y) \in \overset{\circ}{W}{}_2^2(D)$ с помощью кусочно-линейных функций. Разобьем D на подобласти D_{ij} прямыми $x_i = ih_x$, $y_j = jh_y$, $h_x = a/N_x$, $h_y = b/N_y$ (N_x, N_y — положительные целые числа), а затем разделим каждый из прямоугольников D_{ij} диагонально, как это сделано на рис. 2.4 (т. е. осуществим триангуляцию области D). Каждому узлу (x_i, y_j) $(i = 0, 1, \ldots, N_x,\ \ j = 0, 1, \ldots, N_y)$ поставим в соответствие функцию $\varphi_{ij}(x, y)$, равную единице в данном узле и нулю во всех остальных и линейную в каждом треугольнике. Каждую из эти функций $\varphi_{ij}(x, y)$ для введенной сетки можно выразить через «стандартную» функцию $\varphi(s, t)$ вида

$$\varphi(s,t) = \begin{cases} 1-s, & 0 \le s \le 1, & 0 \le t \le s, \\ 1-t, & 0 \le s \le 1, & s \le t \le 1, \\ 1+s-t, & -1 \le s \le 0, & 0 \le t \le s+1, \\ 1+s, & -1 \le s \le 0, & s \le t \le 0, \\ 1+t, & -1 \le s \le 0, & -1 \le t \le s, \\ 1-s+t, & 0 \le s \le 1, & s-1 \le t \le 0. \end{cases} \tag{2.3.1}$$

Носитель этой функции изображен на рис. 2.5, а общий ее вид — на рис. 2.6. Теперь $\varphi_{ij}(x, y)$ можно представить в виде

$$\varphi_{ij}(x, y) = \varphi(x/h_x - i,\ y/h_y - j); \tag{2.3.2}$$

эти функции часто называют *функциями Куранта*.

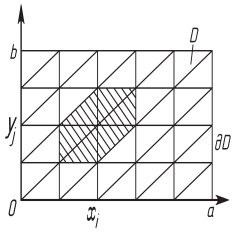


Рис. 2.4.

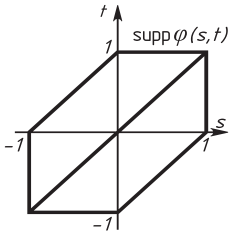


Рис. 2.5.

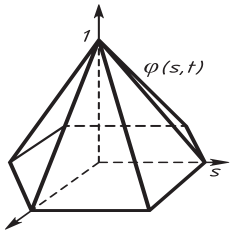


Рис. 2.6.

Введем для заданной функции $u(x,y) \in W_2^2(D) \subset C(D)$ набор чисел

$$u_{ij} = u(x_i, y_j), \quad i = 0, 1, \ldots, N_x, \quad j = 0, 1, \ldots, N_y, \tag{2.3.3}$$

,а также построим линейную комбинацию

$$u_I(x,y) = \sum_{i=0}^{N_x} \sum_{j=0}^{N_y} u_{ij} \varphi_{ij}(x,y), \tag{2.3.4}$$

которая называется *кусочно-линейным восполнением* функции $u(x,y)$. Очевидно, что $u_I \in C(\overline{D}) \cap W_2^1(D)$. Если же $u \in \overset{\circ}{W}{}_2^1 \cap W_2^2$, то автоматически получаем, что

$$u_I(x,y) = \sum_{i=1}^{N_x - 1} \sum_{j=1}^{N_y - 1} u_{ij} \varphi_{ij}(x,y) \tag{2.3.5}$$

(суммирование ведется по узлам, лежащим строго внутри D), т. е. $u_I \in \overset{\circ}{W}{}_2^1(D)$. Для $u_I(x,y)$ справедливо утверждение: если $u(x,y) \in W_2^2(D)$, то (Г. И. Марчук, В. И. Агошков [5])

$$\|u - u_I\|_{L_2(D)} \le ch^2 \|u\|_{W_2^2(D)}, \tag{2.3.6}$$

$$\|u - u_I\|_{W_2^1(D)} \le ch \|u\|_{W_2^2(D)}, \tag{2.3.7}$$

где $h = \max(h_x, h_y)$, а постоянная c не зависит от h и $u(x,y)$.

Рассмотрим пространства $W_2^{1,h}$ и $\overset{\circ}{W}{}_2^{1,h}$, которые часто используются, например, при решении эллиптических задач второго порядка. Обозначим через $W_2^{1,h}$ множество функций вида

$$u^h = \sum_{i=0}^{N_x}\sum_{j=0}^{N_y} a_{ij}\varphi_{ij}(x,y) \quad \text{или} \quad u^h = \sum_{i,j} a_{ij}\varphi_{ij}(x,y),$$

$$(x_i, y_j) \in \overline{D} = D + \partial D. \tag{2.3.8}$$

Очевидно, что $W_2^{1,h}$ образует подпространство пространства $W_2^1(D)$. (Отметим, что вторая форма записи удобна, когда D — некоторая многоугольная область, не обязательно прямоугольник.)

Если рассматриваются функции $u(x) \in \overset{\circ}{W}{}_2^1 \subset W_2^1$, то соответствующее подпространство $\overset{\circ}{W}{}_2^{1,h} \subset \overset{\circ}{W}{}_2^1$ состоит из линейных комбинаций вида

$$u^h = \sum_{i=1}^{N_x-1}\sum_{j=1}^{N_y-1} a_{ij}\varphi_{ij}(x,y) \quad \text{или} \quad u^h = \sum_{i,j} a_{ij}\varphi_{ij}(x,y), \tag{2.3.9}$$

$$(x_i, y_j) \in D.$$

Поскольку функции вида (2.3.4) принадлежат $W_2^{1,h}$, то соотношения (2.3.6), (2.3.7) характеризуют аппроксимирующие свойства каждого из пространств $W_2^{1,h}$, $\overset{\circ}{W}{}_2^{1,h}$. Из данных соотношений получаем, что для любой функции $u(x,y) \in W_2^2(D)$ существует такая функция $u_I \in F_h \equiv W_2^{1,h}$ (или $F_h \equiv \overset{\circ}{W}{}_2^{1,h}$, если $u(x,y) \in W_2^2(D) \cap \overset{\circ}{W}{}_2^1(D)$), что

$$\inf_{v \in F_h} \|u - v\|_{W_2^k(D)} \le \|u - u_I\|_{W_2^k(D)} \le ch^{2-k}\|u\|_{W_2^2(D)}, \tag{2.3.10}$$

$$k = 0, 1.$$

Все сформулированные утверждения остаются справедливыми и для нормированных функций Куранта

$$\varphi_{ij}(x,y) = \frac{1}{\sqrt{h_x h_y}}\varphi\left(\frac{x}{h_x} - i, \frac{y}{h_y} - j\right), \tag{2.3.11}$$

$$i = 0, 1, \ldots, N_x, \quad j = 0, 1, \ldots, N_y,$$

использование которых при численном решении задач может оказаться предпочтительным.

2.3.2. Кусочно-линейные базисные функции на многоугольной области

В этом пункте мы рассмотрим аппроксимацию функций, определенных на многоугольной области D из $\mathbf{R}^2$ с помощью кусочно-линейных базисных функций. При этом D будет разбиваться на треугольники (не обязательно прямоугольные).

Пусть задана многоугольная область $D \subset \mathbf{R}^2$. Введем разбиение ее на треугольники T_i так, чтобы: 1) каждая пара треугольников имела либо одну общую вершину, либо одну общую сторону, либо они не пересекались; 2) объединение треугольников составляло D. Множество вершин (точек разбиения) обозначим через $(P_0,\ P_1,\ \ldots\ , P_N)$, где $P_i = (x_i, y_i)$ (рис. 2.7).

Определим набор кусочно-линейных функций $\{\varphi_i(x, y)\}$. Для этого каждой точке разбиения P_i поставим в соответствие функцию $\varphi_i(x, y)$, которая в точке P_i равна единице, а в остальных P_j равна нулю и линейна на каждом треугольнике. Вид функции $\varphi_i(x, y)$ приведен на рис. 2.8.

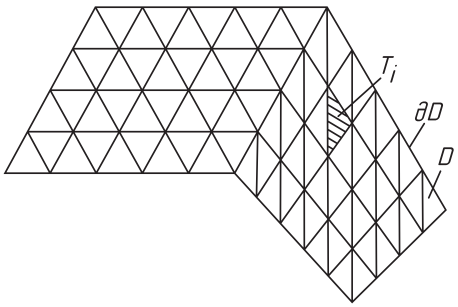


Рис. 2.7.

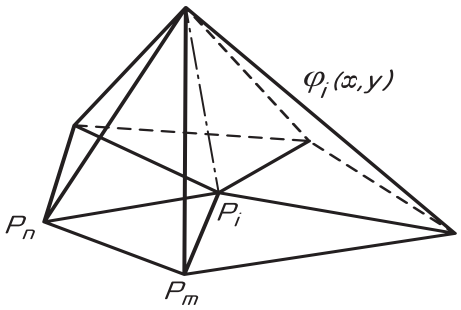


Рис. 2.8.

Чтобы задать $\varphi_i(x, y)$ аналитически, достаточно для каждого треугольника, входящего в носитель $\varphi_i(x, y)$, составить уравнение плоскости, проходящей через единицу в P_i, а в остальных двух вершинах его — через нуль. Например, на треугольнике с вершинами $(P_i,\ P_n,\ P_m)$ функция $\varphi_i(x, y)$ имеет вид

$$\varphi_i(x,y) = \left(1 - \frac{y_i - y_m}{y_n - y_m} - \frac{x_i - x_n}{x_m - x_n}\right)^{-1} \left(1 - \frac{y - y_m}{y_n - y_m} - \frac{x - x_n}{x_m - x_n}\right). \qquad (2.3.12)$$

Используя представление (2.3.12), можно определить вид $\varphi_i(x, y)$ на любом другом треугольнике из носителя $\varphi_i(x, y)$. В силу произволь-

ности индекса i считаем все базисные функции заданными аналитически.

Здесь следует отметить, что построение базиса во многом зависит от разбиения D на T_i и от порядка нумерации узлов сетки. После того как расположение узлов выбрано и их координаты хранятся, например в памяти ЭВМ, легко можно определить, какие вершины являются соседними с вершиной P_i, а это, в свою очередь, позволяет по формуле (2.3.12) построить функцию $\varphi_i(x, y)$. Таким образом, задача построения базисных функций даже при произвольной нумерации узлов сетки решается сравнительно просто.

Обозначим через $W_2^{1,h}$ множество функций вида

$$u_N(x, y) = \sum_{i=0}^{N} a_i \varphi_i(x, y), \tag{2.3.13}$$

где a_i $(i = 0, 1, \ldots, N)$ — всевозможные наборы чисел. Очевидно, что $W_2^{1,h} \subset C(D) \cap W_2^1(D)$.

Пусть в дальнейшем через $\theta_0 > 0$ обозначен минимальный из углов во всех трейгольниках $T_i \subset D$, h — максимальная из сторон треугольников. Аппроксимирующие свойства подпространства $W_2^{1,h}(D)$ характеризуются следующими утверждениями: если $u(x) \in W_2^2(D)$, то существует функция $u_I \in W_2^{1,h}(D)$, для которой справедливы оценки[7)]

$$\|u - u_I\|_{L_2(D)} \leq ch^2 \|u\|_{W_2^2(D)}, \tag{2.3.14}$$

$$\|u - u_I\|_{W_2^1(D)} \leq \frac{ch}{\sin\theta_0} \|u\|_{W_2^2(D)}, \tag{2.3.15}$$

если при этом $u(x, y) \in C^{(2)}(D)$, то также имеет место соотношение

$$\|u - u_I\|_{C(D)} \leq \frac{ch^2}{\sin\theta_0} \|u\|_{C^{(2)}(D)}. \tag{2.3.16}$$

(Здесь постоянная c не зависит от h, θ_0, $u(x, y)$.) Отсюда делаем вывод, что нужно стремиться триангуляцию области проводить так, чтобы среди треугольников не было вырождающихся (т. е. чтобы θ_0 не был близок к нулю) и чтобы углы в них были одного порядка.

[7)] Г. И. Марчук, В. И. Агошков [5].

2.3.3. Билинейные базисные функции

В данном разделе будет показано, как с помощью одномерных «функций-крышек» (см. 2.2.2) можно строить базисные функции в случае двух независимых переменных.

Пусть D — прямоугольник из $\mathbf{R}^2$. Введем на D сетку

$$A_0 = x_0 < x_1 < \ldots < x_{N_x} = A_1, \quad B_0 = y_0 < y_1 < \ldots < y_{N_y} = B_1,$$

$$\Delta y_k = y_k - y_{k-1}, \quad \Delta x_i = x_i - x_{i-1},$$

$$\Delta x = \max_i \Delta x_i, \quad \Delta y = \max_k \Delta y_k, \quad h = \max(\Delta x, \Delta y).$$

Узлам сетки, принадлежащим $D = D \cup \partial D$, поставим в соответствие функции

$$\varphi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{\Delta x_i}, & x \in (x_{i-1}, x_i), \\ \dfrac{x_{i+1} - x}{\Delta x_{i+1}}, & x \in (x_i, x_{i+1}), \\ 0, & x \notin (x_{i-1}, x_{i+1}); \end{cases}$$

$$\varphi_j(y) = \begin{cases} \dfrac{y - y_{j-1}}{\Delta y_j}, & y \in (y_{j-1}, y_j), \\ \dfrac{y_{j+1} - y}{\Delta y_{j+1}}, & y \in (y_j, y_{j+1}), \\ 0, & y \notin (y_{j-1}, y_{j+1}); \end{cases} \tag{2.3.17}$$

$$\varphi_{ij}(x, y) = \varphi_i(x)\varphi_j(y).$$

Рассмотрим всевозможные линейные комбинации функций $\varphi_{ij}(x, y)$:

$$u^h = \sum_{i,j} a_{ij}\varphi_{ij}(x, y), \quad (x_i, y_j) \in \overline{D}, \tag{2.3.18}$$

которые, как легко заметить, образуют множество $F_h \equiv W_2^{1,h} \subset C(D) \cap$ $\cap W_2^1(D)$. Функции $\varphi_{ij}(x, y)$ назовем *билинейными базисными функциями*. Для функций из $W_2^{1,h}$ справедливы следующие утверждения: если $u(x, y) \in C^{(2)}(D)$, то существует такая $u_I \in W_2^{1,h}$, что (Г. И. Марчук, В. И.

Агошков [5])

$$\|u - u_I\|_{L_2(D)} \le ch^2 \|u\|_{C^{(2)}(D)}, \tag{2.3.19}$$

$$\|u - u_I\|_{W_2^1(D)} \le ch \|u\|_{C^{(2)}(D)}, \tag{2.3.20}$$

где постоянная c не зависит от h и $u(x, y)$. Эти утверждения остаются справедливыми для случая, когда $u(x, y) \in C^{(2)}(D) \cap \overset{\circ}{W}{}_2^1(D)$ (т. е. $u(x, y)$ удовлетворяет граничному условию $u = 0$ на ∂D), а вместо $W_2^{1,h}$ берется его подпространство $\overset{\circ}{W}{}_2^{1,h}$:

$$\begin{gathered} \overset{\circ}{W}{}_2^{1,h} = \{u^h : u^h \in W_2^{1,h}, \\ u^h = \sum_{i,j} a_{ij}\varphi_{ij},\ \|u^h\|_{\overset{\circ}{W}{}_2^{1,h}} = \|u\|_{W_2^1},\ (x_i, y_j) \in D\}, \end{gathered} \tag{2.3.21}$$

т. е. суммирование ведется лишь по тем индексам, которые соответствуют строго внутренним узлам сетки, в результате чего получаем функции u^h, удовлетворяющие условию $u^h = 0$ на ∂D.

Пусть теперь область D является объединением r прямоугольников $\{D_i\}_{i=1}^r$ со сторонами, параллельными осям координат, и $\widetilde{D}$ — наименьший по площади прямоугольник со сторонами, параллельными координатным осям, содержащий область D (рис. 2.9): $\widetilde{D} = \{(x, y) :\ a \le x \le b,\ c \le y \le d\}$. Построим на отрезках $[a, b]$ и $[c, d]$ сетки $a = x_0 < x_1 < \ldots < x_{N+1} = b$ и $c = y_0 < y_1 < \ldots < y_{M+1} = d$ таким образом, чтобы любая из сторон, составляющих область D прямоугольников, обязательно принадлежала какой-нибудь из линий

$$x = x_k,\ y = y_l,\ k = 0, 1, \ldots, N + 1;\ l = 0, 1, \ldots, M + 1. \tag{2.3.22}$$

После этого мы определим сеточную область D^h как совокупность точек (x_k, y_l), принадлежащих D $(k = 1, 2, \ldots, N;\ l = 1, 2, \ldots, M)$.

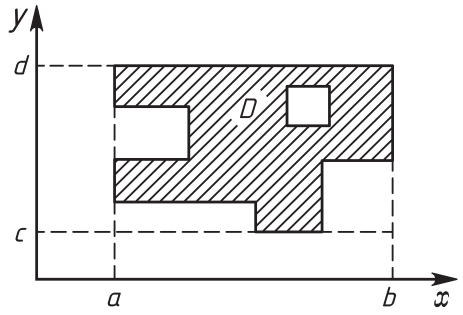


Рис. 2.9.

Перейдем к построению пространства $F_h \subset \overset{\circ}{W}{}_2^1(D)$. На отрезках $[a,b]$ и $[c,d]$ введем соответственно функции

$$\varphi_k(x) = \begin{cases} \dfrac{x - x_{k-1}}{x_k - x_{k-1}}, & \text{если} \quad x \in [x_{k-1}, x_k], \\ \dfrac{x - x_{k+1}}{x_k - x_{k+1}}, & \text{если} \quad x \in [x_k, x_{k+1}], \\ 0 & \text{в остальных точках}, \\ k = 1, 2, \ldots, N, \end{cases} \tag{2.3.23}$$

$$\varphi_l(y) = \begin{cases} \dfrac{y - y_{l-1}}{y_l - y_{l-1}}, & \text{если} \quad y \in [y_{l-1}, y_l], \\ \dfrac{y - y_{l+1}}{y_l - y_{l+1}}, & \text{если} \quad y \in [y_l, y_{l+1}], \\ 0 & \text{в остальных точках}, \\ l = 1, 2, \ldots, M. \end{cases} \tag{2.3.24}$$

Положим

$$\varphi_{k,l}(x,y) = \varphi_k(x)\varphi_l(y), \quad (x_k, y_l) \in D^h \tag{2.3.25}$$

и выберем в качестве F_h линейную оболочку функций $\{\varphi_{k,l}\}$. Так как система $\{\varphi_{k,l}\}$ линейно независима, то она, очевидно, образует базис пространства F_h.

Если рассматривается пространство $W_2^1(D)$, то построение множества $F_h \equiv W_2^{1,h}$ осуществляется аналогичным образом, а состоит

оно из всевозможных линейных комбинаций функций $\{\varphi_{k,l}(x,y)\}$, где $(x_k, y_l) \in D^h \cup \partial D^h$ (здесь $D^h \cup \partial D^h$ есть совокупность точек (x_k, y_l), принадлежащих $D \cup \partial D = \overline{D}$).

Сформулированные выше утверждения и оценки (2.3.19), (2.3.20) здесь также справедливы.

2.3.4. Способы построения подпространств в областях с криволинейной границей

В этом пункте мы рассмотрим простейшие способы построения триангуляций ограниченной двумерной области D^h с гладкой границей S.

Сформулируем основные требования, которым должна удовлетворять сеточная область D^h, являющаяся объединением конечного числа треугольников $\Delta_k \subset D$, не имеющих общих внутренних точек, с кусочно-линейной границей S^h.

1. Между точками S^h и S с помощью нормалей к S устанавливается взаимо-однозначное соответствие, и расстояния между соответствующими токами не превосходят величины $\delta_1 h^2$, где $\delta_1 > 0$ и не зависит от h.

2. Длины сторон и площади треугольников Δ_k, из которых составлена сеточная область, лежат соответственно в пределах $[l_1 h, l_2 h]$ и $[\gamma_1 h^2, \gamma_2 h^2]$, где положительные константы $l_1,\ l_2,\ \gamma_1$ и γ_2 не зависят от h.

Легко видеть, что для области D можно построить различные сеточные области, удовлетворяющие перечисленным выше условиям, и, следовательно, у нас есть возможность наложить дополнительное требование на структуру сеточной области. Прежде чем сформулировать такое дополнительное условие, напомним, что для прямоугольных областей и областей, составленных из прямоугольников в 2.3.1 и 2.3.3, триангуляции строились естественным образом на основе прямоугольных сеток. В качестве дополнительного требования на структуру сеточной области D^h целесообразно выбрать следующее условие.

3. Существует непрерывное взаимно однозначное преобразование сеточной области D^h на область, граница которой состоит из отрезков, параллельных осям координат или образующих с ними угол в

$45°$; это преобразование линейно на каждом треугольнике Δ_k и отображает его в прямоугольный треугольник с катетами, равными h.

Из этого условия, например, следует, что общую вершину могут иметь не более восьми треугольников сеточной области.

Один из наиболее простых способов построения таких триангуляций состоит в следующем. Заключим область D в прямоугольник Π и покроем Π квадратной сеткой с шагом h. Узлы построенной сетки, лежащие вблизи границы S области D, сдвинем к S так, чтобы соединяющая их ломаная S^h хорошо приближала границу (т. е. чтобы выполнялось первое требование для сеточных областей). Затем разобьем все четырехугольники на треугольники так, чтобы выполнялось второе требование. Справедливость третьего условия следует непосредственно из способа построения триангуляции. На рис. 2.10 приводится пример такого построения.

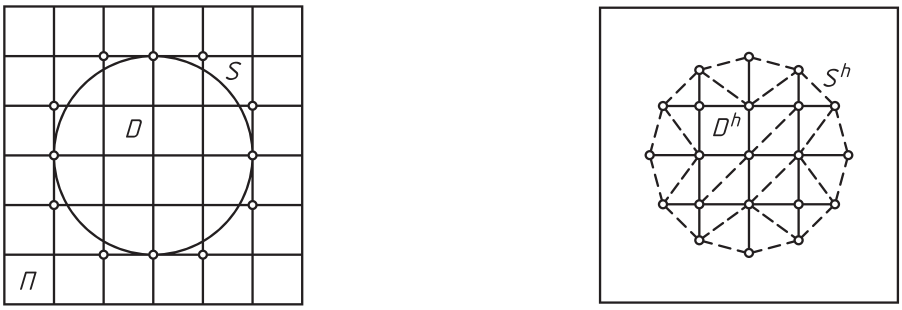


Рис. 2.10.

Нетрудно вычислить конкретные значения постоянных, характеризующих длины сторон и площади треугольников таких триангуляций и не зависящих от конфигурации области D: $l_1 = 0,5$, $l_2 = \sqrt{18/2}$, $\gamma_1 = 0,125$ и $\gamma_2 = 1,125$. Естественно, что постоянная δ_1, характеризующая приближение границы, зависит от кривизны кривой S.

Отметим, что описанный способ триангуляции применим как к односвязным, так и к многосвязным областям и легко реализуем.

Для односвязных областей можно предложить способ построения сеточных областей, удовлетворяющих перечисленным ранее трем условиям и преобразуемых непрерывным кусочно-линейным преобразованием в прямоугольник.

Простоты ради ограничимся случаем, когда область D выпукла. Идея построения сеточной области очень проста. Сначала область D непрерывным преобразованием, имеющим разрывы производных внутри области, переводится в прямоугольник. Полученный прямо-

угольник триангулируется так, чтобы стороны треугольников не пересекали линии разрывов производных использованного отображения. Обратное преобразование дает криволинейную триангуляцию области D, спрямление которой приводит к требуемой сеточной области D^h.

Отображения на прямоугольник можно строить достаточно просто: можно, например, заключить область D в квадрат и вытянуть ее вдоль лучей, выходящих из точки пересечения диагоналей квадрата. Тогда эти диагонали и будут линиями разрыва производной такого преобразования. В действительности отображение на прямоугольник используется только для получения правила, по которому строится сеточная область. Следовательно, рассмотренный способ триангуляции можно назвать методом фиктивного отображения на прямоугольник. На рис. 2.11 представлена такая триангуляция для L-образной области.

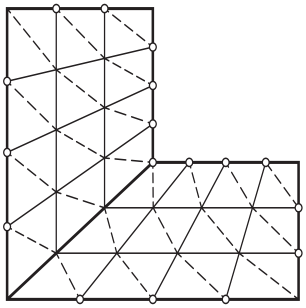


Рис. 2.11.

Теперь проиллюстрируем целесообразность использования таких сеточных областей. Обозначим через F^h множество непрерывных функций, линейных на каждом элементарном треугольнике Δ_k из D^h и равных нулю вне области D^h. Система вариационно-разностных уравнений для краевой задачи

$$\begin{aligned} -\Delta u(x,y) &= f(x,y), \qquad (x,y) \in D, \\ u(x,y) &= 0, \qquad\quad (x,y) \in S, \end{aligned}$$

строится по методу Ритца из условия минимизации квадратичного функционала

$$J(v^h) = \int\limits_{D^h} \left(\left| \frac{\partial v_h}{\partial x} \right|^2 + \left| \frac{\partial v_h}{\partial y} \right|^2 \right) \, dx \, dy - 2 \int\limits_{D^h} f v_h \, dx \, dy$$

в конечномерном пространстве F_h.

Так как триангуляция сеточной области D^h топологически эквивалентна простейшей триангуляции прямоугольной области, то мы можем занумеровать вершины треугольников согласно этому соответствию и определить в пространстве F_h базис $\{\omega_{k,l}(x,y)\}$, как это сделано в 2.3.2. Тогда структура матрицы системы алгебраических линейных уравнений, получаемой из приближенного вариационного уравнения, принципиально не будет отличаться от структуры матрицы системы, возникающей в случае, когда D является прямоугольником.

В заключение отметим, что рассмотренные в этом пункте способы построения сеточных областей можно обобщить и на случай трехмерных областей.

2.3.5. Способы построения подпространств F_h для многомерных задач

Кратко остановимся на возможных путях конструирования подпространств для многомерных задач, когда число независимых переменных больше двух.

Предположим, что в пространстве трех переменных x, y и z задана ограниченная область D с кусочно-линейной границей ∂D. Для этого случая наиболе известный способ построения подпространств $F_h \subset \overset{\circ}{W}{}_2^1(D)$ заключается в следующем. Сначала осуществляется пространственная триангуляция области, т. е. область D покрывается конечным числом треугольных пирамид Δ_k, не имеющих общих внутренних точек так, чтобы $D = \bigcup\limits_{k=1}^{N} \Delta_k$. Если максимальную длину ребер пирамиды Δ_k обозначить h_k, то можно легко построить последовательность пространств F_h кусочно-полиномиальных функций, определяемых параметром $h = \max\limits_{1 \le k \le N} h_k$. Такие пространства определяются аналогично одномерному и двумерному случаям.

Проиллюстрируем это на примере кусочно-линейных аппроксимаций, когда область $D = \{x, y, z : \ 0 < x < 1, \ 0 < y < 1, \ 0 < z < 1\}$ является кубом. Разобьем отрезок $[0, 1]$ на равные отрезки точками $0 = \xi_0 < \xi_1 < \ldots < \xi_{n+1} = 1$, где $\xi_k = \Delta\xi \cdot k$ $(\Delta\xi = 1/(n+1))$. Затем покроем область D одинаковыми элементарными кубами со стороной $\Delta\xi$; для этого достаточно рассечь D плоскостями

$$x = \xi_k, \quad y = \xi_k, \quad z = \xi_k, \quad k = 0, 1, \ldots, n+1.$$

Триангуляцию элементарного кубика со стороной $\Delta\xi$ можно осуществить различными способами. Одна из возможных триангуляций показана на рис. 2.12, где кубик предварительно разбит на две призмы.

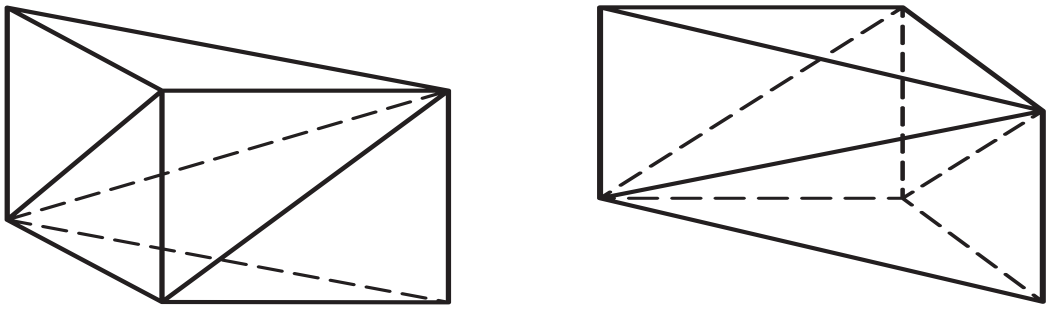


Рис. 2.12.

Таким образом, каждый элементарный кубик можно представить как объединение шести пирамид, причем длина максимального ребра пирамиды равна $h = \sqrt{3}\Delta\xi = \sqrt{3}/(n \ + \ 1)$.

Проделав такую триангуляцию для всех кубиков, получим, что область D будет полностью покрыта $N = 6 \cdot (n+1)^3$ пирамидами.

Пространство F_h конструируется следующим образом. В каждой пирамиде Δ_k строится многочлен

$$g_k(x, y, z) = \sum_{t=0}^{m} \sum_{i_1+i_2+i_3=t} C^k_{i_1,i_2,i_3} x^{i_1} y^{i_2} z^{i_3} \tag{2.3.26}$$

так, чтобы функция

$$g(x, y, z) = \begin{cases} g_1(x, y, z), & \text{если} \quad (x, y, z) \in \Delta_1, \\ \cdots & \\ g_N(x, y, z), & \text{если} \quad (x, y, z) \in \Delta_N, \end{cases} \tag{2.3.27}$$

принадлежала пространству $\overset{\circ}{W}{}_2^1(D)$, т. е. имела суммируемые с квадратом первые производные и обращалась в нуль на границе (чтобы обеспечить выполнение этого условия, достаточно потребовать непрерывности кусочно-полиномиальной функции g и ее обращения в нуль на ∂D).

В случае кусочно-линейных функций

$$g_k(x,y,z) = C^k_{0,0,0} + C^k_{1,0,0}x + C^k_{0,1,0}y + C^k_{0,0,1}z \tag{2.3.28}$$

для непрерывности $g(x,y,z)$ достаточно определить коэффициенты функций g_k через значения g в вершинах соответствующей пирамиды Δ_k.

Описанный подход является достаточно универсальным и может быть применен для областей с более общими границами. Однако даже для простейших областей он весьма сложен алгоритмически. Ниже мы остановимся на еще одном способе конструирования подпространств F_h, который для многомерных областей специального вида представляется нам предпочтительнее первого.

Пусть D — область p-мерного евклидова пространства, которую можно представить в виде объединения конечного числа p-мерных прямоугольных параллелепипедов $\{D_\nu\}$, и пусть $\overline{D} = \{x_i : a_i \le x \le b_i,\ i = 1,2,\ldots,p\}$ — параллелепипед минимального объема, содержащий D (предполагается, что стороны D_ν параллельны координатным гиперплоскостям). Тогда для каждого i $(1 \le i \le p)$ на отрезке $[a_i, b_i]$ оси переменной x_i можно построить сетку

$$a_i = x_{i,0} < x_{i,1} < \ldots < x_{i,N_i} < x_{i,N_i+1} = b_i, \quad i = 1,2,\ldots,p,$$

и определить сетку D^h как совокупность точек $x_k = (x_{1,k_1}, x_{2,k_2}, \ldots, x_{p,k_p})$, принадлежащих области D $(1 \le k_i \le N_i$ для $i = 1,2,\ldots,p)$. Если теперь на каждой одномерной сетке ввести набор одномерных базис-

ных функций

$$\varphi_{i,k_i}(x_i) = \begin{cases} \dfrac{x - x_{i,k_i-1}}{x_{i,k_i} - x_{i,k_i-1}}, & \text{если} \quad x_i \in [x_{i,k_i-1}, x_{i,k_i}], \\ \dfrac{x - x_{i,k_i+1}}{x_{i,k_i} - x_{i,k_i+1}}, & \text{если} \quad x_i \in [x_{i,k_i}, x_{i,k_i+1}], \\ 0 \quad \text{в противном случае,} \\ k_i = 1, 2, \ldots, N_i;\ 1 \le i \le p, \end{cases} \tag{2.3.29}$$

то систему функций

$$\varphi_k(x) = \prod_{i=1}^{p} \varphi_{i,k_i}(x_i), \quad x_k \in D^h, \quad k = (k_1, k_2, \ldots, k_p), \tag{2.3.30}$$

можно выбрать как базис пространства полилинейных функций $F^h \subset \overset{\circ}{\subset} W_2^1(D)$. Для случая двух переменных $(p = 2)$ такой подход достаточно полно рассмотрен в 2.3.3.

2.4. Вариационно-разностные и проекционно-сеточные схемы

В 2.1 мы рассмотрели ряд вариационных и проекционных методов. Известны их хорошие качества: сохранение у матриц возникающих систем свойств симметричности или положительной определенности, если им обладал оператор исходной задачи; получение достаточно хороших приближений к решению задачи при небольшом числе базисных функций и др. Поэтому привлекательным становится конструирование таких алгоритмов приближенного решения задачи, которые, с одной стороны, по форме были бы вариационными, а с другой стороны, чтобы эти алгоритмы приводили к системам уравнений, подобным системам в разностных методах (т. е. незначительное число элементов матриц этих систем были бы ненулевыми). Оказывается, последнее свойство часто достигается путем использования базисных функций с финитным носителем, в результате чего мы будем получать систему уравнений, аппроксимирующую исход-

ную задачу, которую можно называть схемой. И если для получения этих систем привлекаются вариационные методы (Ритца, наименьших квадратов и др.), то они будут *вариационно-разностными схемами*. Если применяются проекционные методы (Галеркина, Галеркина — Петрова, метод моментов и др.), то они относятся к классу проекционно-сеточных схем (который включает в себя также и вариационно-разностные схемы). И в данном разделе мы проиллюстрируем применение вариационных или проекционных методов и финитных базисных функций к построению этих схем для некоторых задач математической физики.

2.4.1. Вариационно-разностная схема для одномерного уравнения диффузии

Рассмотрим задачу об отыскании непрерывной на $D = (a, b)$ функции $u(x)$, удовлетворяющей уравнению

$$-\frac{d}{dx}p(x)\frac{du}{dx} + q(x)u(x) = f(x) \tag{2.4.1}$$

и краевым условиям

$$u(a) = u(b) = 0. \tag{2.4.2}$$

Здесь $f(x) \in L_2(a, b)$; $p(x)$, $q(x)$ — ограниченные функции; $0 < p_0 \le p(x) \le p_1;\ 0 \le q(x) \le q_1$; $p_0,\ p_1,\ q_1$ — постоянные.

Обозначим через L оператор задачи (2.4.1), (2.4.2), определяемый выражением $Lu = -\frac{d}{dx}p\frac{du}{dx} + qu$ и областью определения $D(L)$. Пусть $D(L)$ состоит из непрерывных функций $u(x)$, обладающих производной $du/dx \in L_2(a, b)$, таких, что $Lu \in L_2$, и удовлетворяющих краевым условиям (2.4.2). Теперь задачу (2.4.1), (2.4.2) можно записать в виде операторного уравнения

$$Lu = f, \tag{2.4.3}$$

которое будем рассматривать в гильбертовом пространстве $F = L_2(D)$ со скалярным произведением $(u, v) = (u, v)_L$ и нормой $\|u\| = \|u\|_L$.

Изучим свойства оператора L. Прежде всего отметим, что множество $D(L)$ плотно в L_2 и что оператор L является симметричным:

$$(Lu, v) = \int_a^b \left(-\frac{d}{dx} p \frac{du}{dx} + qu \right) v \, dx =$$

$$= -p \frac{du}{dx} \, v|_{x=a}^{x=b} + \int_a^b \left(p \frac{du}{dx} \frac{dv}{dx} + quv \right) \, dx, \quad u, v, \in D(L), \tag{2.4.4}$$

а поскольку $v(a) = v(b) = 0$, то

$$(Lu, v) = \int_a^b \left(p \frac{du}{dx} \frac{dv}{dx} + quv \right) \, dx = (Lv, u).$$

Оператор L является также положительно определенным, т. е. для него выполнено условие

$$\gamma^2 \|u\|^2 \leq (Lu, u), \quad u \in D(L),$$

где $\gamma > 0$ — постоянная, не зависящая от $u(x)$. Для доказательства этого факта достаточно воспользоваться неравенством Стеклова (см. (1.1.25) из § 1.1).

В силу отмеченных свойств операторов L мы можем ввести энергетическое пространство F_L, соответствующее L. Скалярное произведение и норма в нем будут иметь вид

$$(u, v)_L = \int_a^b \left(p \frac{du}{dx} \frac{dv}{dx} + quv \right) \, dx, \quad \|u\|_L = (u, u)_L^{1/2}. \tag{2.4.5}$$

Учитывая ограничения на p и q и неравенство Стеклова, замечаем, что в нашей задаче F_L совпадает с пространством $\overset{\circ}{W}{}_2^1$ и имеют место соотношения эквивалентности норм

$$c_0 \|u\|_{W_2^1} \leq \|u\|_L \leq c_1 \|u\|_{W_2^1}. \tag{2.4.6}$$

Задача (2.4.1), (2.4.2), согласно теории метода Ритца, сводится к проблеме минимизации функционала

$$F(u) = (u,u)_L - 2(u,f) \tag{2.4.7}$$

в пространстве $F_L = \overset{\circ}{W}{}_2^1$. Данная проблема о минимизации $F(v)$ имеет единственное решение $u \in \overset{\circ}{W}{}_2^1\,(D)$, причем $u(x) \in\ W_2^2(D)$ при $|dp/dx| < \infty$ и $\|u\|_{W_2^2} \le c\|f\|_{L_2}$.

Зададим базисные функции. Поскольку в задаче о минимизации $F(u)$ допускаются функции из $\overset{\circ}{W}{}_2^1$ (т. е. область определения функционала $F(u)$ есть $\overset{\circ}{W}{}_2^1$), то $\overset{\circ}{W}{}_2^1$ принадлежат кусочно-линейные функции, удовлетворяющие условию (2.4.2). Поэтому выберем в качестве базисных кусочно-линейные функции. Для их построения введем на (a,b) сетку

$$a = x_0 < x_1 < \ldots < x_N = b, \quad h_i = x_i - x_{i-1},$$
$$i = 1,2,\ldots,N, \tag{2.4.8}$$

которая удовлетворяет ограничениям

$$c_2 h \le h_i \le c_3 h, \tag{2.4.9}$$

где $c_2,\ c_3 > 0$ — постоянные, не зависящие от h_i и от h. Поставим в соответствие каждому узлу кусочно-линейную функцию

$$\varphi_i(x) = \frac{1}{\sqrt{h}} \begin{cases} \dfrac{x - x_{i-1}}{x_i - x_{i-1}}, & x \in (x_{i-1}, x_i), \\ \dfrac{x_{i+1} - x}{x_{i+1} - x_i}, & x \in (x_i, x_{i+1}), \\ 0, & x \notin (x_{i-1}, x_{i+1}). \end{cases} \tag{2.4.10}$$

Введем линейную комбинацию

$$u_h(x) = \sum_{i=0}^{N} a_i \varphi_i(x)$$

и потребуем, чтобы она удовлетворяла главным краевым условиям задачи, т. е. чтобы $u_h(a) = a_0 = u_h(b) = a_N = 0$. Этим требованиям будет

удовлетворять линейная комбинация вида

$$u_h(x) = \sum_{i=1}^{N-1} a_i \varphi_i(x); \tag{2.4.11}$$

множество таких линейных комбинаций обозначим через $\overset{\circ}{W}{}_2^{1,h} = F_h$. Очевидно, что $\overset{\circ}{W}{}_2^{1,h} \subset \overset{\circ}{W}{}_2^1 = F_L$.

Согласно теории метода Ритца в энергетических пространствах, за приближенное решение задачи u_h можно принять функцию вида (2.4.11), минимизирующую $F(v)$ на подпространстве $\overset{\circ}{W}{}_2^{1,h}$. Коэффициенты a_i этой функции находятся из условий $\partial F(u_h)/\partial a_i = 0$ $(i = 1, 2, \ldots, N-1)$, которые приводят к системе уравнений

$$\hat{A}a = f, \quad a = (a_1, a_2, \ldots, a_{N-1})^T, \tag{2.4.12}$$

где элементы матрицы $\hat{A} = (A_{ij})$ и составляющие вектора $f = (f_1, f_2, \ldots, f_{N-1})^T$ имеют соответственно вид

$$A_{i,j} = (\varphi_i, \varphi_j)_L = \int\limits_a^b \left(p \frac{d\varphi_i}{dx} \frac{d\varphi_j}{dx} + q\varphi_i\varphi_j \right) dx =$$

$$= \int\limits_{D_{ij}} \left(p \frac{d\varphi_i}{dx} \frac{d\varphi_j}{dx} + q\varphi_i\varphi_j \right) dx, \tag{2.4.13}$$

где $D_{i,j} = (a,b) \cap \operatorname{supp}\varphi_i \cap \operatorname{supp}\varphi_j$, $f_i = (f, \varphi_i) = \int\limits_a^b f\varphi_i \, dx = \int\limits_{D_i} f\varphi_i \, dx$, $D_i = (a,b) \cap \operatorname{supp}\varphi_i = (x_{i-1}, x_{i+1})$. Если вычислить A_{ij} и f_i, представляющие собой интегралы от известных функций, то матрица $\hat{A}$ и вектор f будут определены.

Поскольку в методе Ритца матрица системы (2.4.12) сохраняет такие свойства оператора A, как симметричность и положительную определенность, то можно гарантировать, что (2.4.12) имеет единственное решение $a = (a_1, a_2, \ldots, a_{N-1})^T$, которое однозначно определяет приближенное решение $u_h(x)$ задачи по формуле (2.4.11). Отметим, что так как $A_{i,j} = 0$ при $|i-j| > 1$ (носители функций φ_i и φ_j в этом случае не пересекаются), то матрица $\hat{A}$ оказывается трехдиагональной. Отсюда заключаем, что решить (2.4.12) можно с помощью метода прогонки (метода факторизации), который особенно эффективен

для таких систем. Ряд других сведений о структуре матрицы $\hat{A}$ и ее свойствах, которые позволяют рассматривать (2.4.12) как некоторую систему разностных уравнений, следуют из приводимых в конце параграфа примеров.

Исследуем вопрос о сходимости приближенного решения u_h к точному решению $u(x)$ задачи (2.4.1), (2.4.2). Как следует из 2.1.2, для приближенного решения по методу Ритца справедливо неравенство $\|u-u_h\|_L \le \|u-v_h\|_L$, где $v_h = \sum\limits_{i=1}^{N-1} b_i\varphi_i$ — произвольная функция из $H_A^{(N)} = \overset{\circ}{W}{}_2^{1,h}$. Учитывая (2.4.6), получаем

$$\begin{gathered}\|u-u_h\|_{W_2^1} \le c\|u-v_h\|_{W_2^1},\\ \|u-u_h\|_{W_2^1} \le c\inf_{v_h}\|u-v_h\|_{W_2^1}, \quad v_h \in \overset{\circ}{W}{}_2^{1,h}.\end{gathered} \tag{2.4.14}$$

В силу результатов об аппроксимации с помощью кусочно-линейных базисных функций $\{\varphi_i\}$ из гладкости точного решения $u(x)$ и из (2.4.14) следует не только сам факт сходимости u_h к u при $h \to 0$, но и оценка скорости сходимости: если $|dp/dx| < \infty$, то

$$\|u-u_h\|_{W_2^1} \le ch\|u\|_{W_2^2} \le ch\|f\|. \tag{2.4.15}$$

Дополнительный анализ позволяет также получить оценку погрешности в $L_2(D)$ (Г. И. Марчук, В. И. Агошков [5]):

$$\|u-u_h\|_{L_2(D)} \le ch^2\|f\|. \tag{2.4.16}$$

Рассмотрим несколько примеров, иллюстрирующих алгоритм решения задачи (2.4.1), (2.4.2).

Пример 1. Пусть $p(x)$, $q(x)$ — кусочно-постоянные функции с конечным числом разрывов, совпадающих с узлами сетки, и $p_{i-1/2} = p(x)$ при $x \in (x_{i-1}, x_i)$ $(i = 1, 2, \ldots, N)$, $q_{i-1/2} = q(x)$ при $x \in (x_{i-1}, x_i)$ $(i = 1, 2, \ldots, N)$. Найдем явный вид матрицы A в (2.4.12). Так как $A_{ij} = 0$ при $|i-j| > 1$, то вычислить необходимо лишь элементы $A_{j-1,j}$, A_{jj}, $A_{j+1,j}$. Выпишем элемент A_{ij} и рассмотрим его составляющие:

$$A_{ij} = (\varphi_i, \varphi_j)_L = \left(p\frac{d\varphi_i}{dx}, \frac{d\varphi_j}{dx}\right) + (q\varphi_i, \varphi_j), \quad i = j-1, j, j+1,$$

$$\left(p\frac{d\varphi_{j-1}}{dx},\frac{d\varphi_j}{dx}\right)=\int\limits_{x_{j-1}}^{x_j}p\frac{d\varphi_{j-1}}{dx}\frac{d\varphi_j}{dx}\,dx=-p_{j-1/2}\frac{1}{hh_j},$$

$$\left(p\frac{d\varphi_j}{dx},\frac{d\varphi_j}{dx}\right)=\int\limits_{x_{j-1}}^{x_j}p\left(\frac{d\varphi_j}{dx}\right)^2\,dx+\int\limits_{x_j}^{x_{j+1}}p\left(\frac{d\varphi_j}{dx}\right)^2\,dx=$$

$$=p_{j-1/2}\frac{1}{hh_j}+p_{j+1/2}\frac{1}{hh_{j+1}},$$

$$\left(p\frac{d\varphi_{j+1}}{dx},\frac{d\varphi_j}{dx}\right)=\int\limits_{x_j}^{x_{j+1}}p\frac{d\varphi_{j+1}}{dx}\frac{d\varphi_j}{dx}\,dx=-p_{j+1/2}\frac{1}{hh_{j+1}},$$

$$(q\varphi_{j-1},\varphi_j)=\int\limits_{x_{j-1}}^{x_j}q\varphi_{j-1}\varphi_j\,dx=q_{j-1/2}\int\limits_{x_{j-1}}^{x_j}\frac{(x_j-x)(x-x_{j-1})}{h_j^2h}\,dx=q_{j-1/2}\frac{h_j}{6h},$$

$$(q\varphi_j,\varphi_j)=q_{j-1/2}\frac{2h_j}{6h}+q_{j+1/2}\frac{2h_{j+1}}{6h},$$

$$(q\varphi_{j+1},\varphi_j)=q_{j+1/2}\frac{h_{j+1}}{6h},\quad j=1,2,\ldots,N-1.$$

В итоге получаем матрицу

$$\hat{A}=\begin{bmatrix}\frac{p_{1/2}}{hh_1}+\frac{p_{3/2}}{hh_2} & -\frac{p_{3/2}}{hh_2} & & & \\ -\frac{p_{3/2}}{hh_2} & \ddots & & & \\ & & \ddots & -\frac{p_{N-3/2}}{hh_{N-1}} & \\ & & -\frac{p_{N-3/2}}{hh_{N-1}} & -\frac{p_{N-3/2}}{hh_{N-1}}+\frac{p_{N-1/2}}{hh_N} & \end{bmatrix}+$$

$$+$$

$$+\frac{1}{6h}\begin{bmatrix} 2q_{1/2}h_1+2q_{3/2}h_2 & q_{3/2}h_2 & & & \\ q_{3/2}h_2 & \ddots & & & \\ & & \ddots & & \\ & & & & q_{N-3/2}h_{N-1} \\ & & & q_{N-3/2}h_{N-1} & 2q_{N-3/2}h_{N-1}+2q_{N-1/2}h_N \end{bmatrix}.$$

Пример 2. Пусть $q=0$, $p=\mathrm{const}$ и пусть $h_i=h$ $(i=1,2,\ldots,N)$. Требуется найти собственные числа матрицы из примера 1. При сделанных ограничениях матрица $\hat{A}$ совпадает с матрицей, возникающей при решении задачи (2.4.1), (2.4.2) разностным методом с применением для аппроксимации d^2u/dx^2 трехточечного соотношения

$$\left.\frac{d^2u}{dx^2}\right|_{x=x_i} \approx \frac{u(x_{i-1})-2u(x_i)+u(x_{i+1})}{h^2},$$

т. е. имеет вид

$$\hat{A}=\begin{bmatrix} \frac{2}{h^2} & -\frac{1}{h^2} & & \\ -\frac{1}{h^2} & \ddots & & \\ & & \ddots & -\frac{1}{h^2} \\ & & -\frac{1}{h^2} & \frac{2}{h^2} \end{bmatrix}.$$

Собственные числа λ_k и соответствующие им собственные векторы $u^{(k)}=(u_1^{(k)},u_2^{(k)},\ldots,u_{N-1}^{(k)})^T$ такой матрицы хорошо известны; они имеют вид

$$\lambda_k=p\frac{4}{h^2}\sin^2\frac{k\pi h}{2}, \quad u_j^{(k)}=\sin(jk\pi h),$$
$$k=1,2,\ldots,N-1, \quad j=1,2,\ldots,N-1.$$

2.4.2. Вариационно-разностная схема для эллиптического уравнения

Рассмотрим двумерную эллиптическую задачу вида

$$-\sum_{i,j=1}^{2}\frac{\partial}{\partial x_i}A_{ij}(x)\frac{\partial u}{\partial x_j}=f \quad \text{в} \quad D, \tag{2.4.17}$$

$$u=0 \quad \text{на} \quad \partial D, \tag{2.4.18}$$

где $A_{ij}(x)=A_{ji}(x)$ есть ограниченные функции, причем для любого вектора $\xi=(\xi_1,\xi_2)$ выполняется неравенство

$$\mu_0\sum_{i=1}^{2}\xi_i^2\le\inf_{x\in D}\sum_{i,j=1}^{2}A_{ij}(x)\xi_i\xi_j\le\sup_{x\in D}\sum_{i,j=1}^{2}A_{ij}(x)\xi_i\xi_j\le\mu_1\sum_{i=1}^{2}\xi_i^2 \tag{2.4.19}$$

с некоторыми положительными константами $\mu_0\le\mu_1$; граница ∂D области D является кусочно-линейной. Так же как и в одномерной задаче из 2.4.1, можно показать, что оператор задач (2.4.17), (2.4.18) является симметричным и положительно определенным. Далее, эта задача, как видно из 2.1, эквивалентна нахождению функции, минимизирующей в пространстве $\overset{\circ}{W}{}_2^1(D)$ квадратичный функционал

$$J(u)=\int\limits_D\left(\sum_{i,j=1}^{2}A_{ij}(x)\frac{\partial u}{\partial x_i}\frac{\partial u}{\partial x_j}\right)\,dD-2\int\limits_D uf\,dD. \tag{2.4.20}$$

Применим для приближенного нахождения решения последней задачи метод Ритца с подпространствами F_h, состоящими из кусочно-линейных функций v^h, удовлетворяющих условию (2.4.18). Способы построения этих подпространств сформулированы в 2.3. Поэтому здесь для иллюстрации построения схемы и для упрощения изложения предположим, что $D=\{(x_1,x_2):0<x_1,x_2<1\}$ является единственным квадратом. Покроем D обычной равномерной сеткой с шагом $h=\dfrac{1}{N+1}$ (N — целое положительное число) и триангулируем D, как это показано на рис. 2.13. Каждому внутреннему узлу сетки поставим в соответствие кусочно-линейную базисную функцию (см. § 2.3) и обозначим систему базисных функций через $\{\varphi_{k,l}(x)\}_{k,l=1}^{N}$.

Для построения приближенного решения $u^h(x)$ задачи (2.4.17), (2.4.18) воспользуемся методом Ритца с применением базиса $\{\varphi_{k,l}(x)\}_{k,l=1}^{N}$.

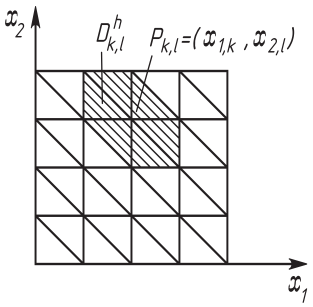


Рис. 2.13.

В результате приходим к системе линейных уравнений

$$A\alpha = g, \tag{2.4.21}$$

где $\alpha = (\alpha_1, \alpha_2 \ldots, \alpha_{N^2})^T$ — вектор, составленный из коэффициентов $\{\alpha_{N(k-1)+l} = \alpha_{k,l}\}_{k,l=1}^N$ разложения

$$u^h(x) = \sum_{k,l=1}^{N} \alpha_{k,l}\varphi_{k,l}(x), \tag{2.4.22}$$

$g = (g_1, g_2 \ldots, g_{N^2})^T$ — вектор с компонентами

$$g_{N(k-1)+l} = g_{k,l} = \int\limits_{D_{k,l}} f\varphi_{k,l}(x)\, dD, \quad k, l = 1, 2, \ldots, N, \tag{2.4.23}$$

а элементы матрицы A вычисляются по формулам

$$a_{N(k-1)+l,N(i-1)+j} = \int\limits_{D} \sum_{s,t=1}^{2} A_{st}(x) \frac{\partial\varphi_{k,l}}{\partial x_s} \frac{\partial\varphi_{i,j}}{\partial x_t}\, dD, \tag{2.4.24}$$

$$k, l, i, j = 1, 2, \ldots, N.$$

Введем обозначение $a^{i,j}_{k,l} = a_{N(k-1)+l,N(i-1)+j}$.

Учитывая вид функций $\{\varphi_{k,l}(x)\}_{k,l=1}^N$ (см. рис. 2.13), нетрудно показать, что

$$a^{i,j}_{k,l} = 0,$$

если выполнено хотя бы одно из двух неравенств

$$|i-k| > 1,\ |j-l| > 1, \quad k, l, i, j = 1, 2, \ldots, N.$$

Отсюда сразу следует, что матрица A является блочно-трехдиагональной вида

$$A = \left\| \begin{array}{cccccc} A_{11} & A_{12} & 0 & \dots & 0 & 0 \\ A_{21} & A_{22} & A_{23} & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & A_{N,N-1} & A_N N \end{array} \right\|,$$

где $A_{kk} = A^*_{kk}$, $A_{k,k+1} = A^*_{k+1,k}$ $(k = 1, 2, \dots, N)$ и каждая из матриц $A_{k,l}$ является трехдиагональной матрицей порядка N. Более точный анализ показывает, что матрицы $\{A_{k,k-1}\}_{k=2}^N$ — двухдиагональные:

$$A_{k,k-1} = \left\| \begin{array}{cccccc} a_{k,1}^{k-1,1} & a_{k,1}^{k-1,2} & 0 & \dots & 0 & 0 \\ 0 & a_{k,2}^{k-1,2} & a_{k,2}^{k-1,3} & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & a_{k,N}^{k-1,N} \end{array} \right\|. \qquad (2.4.25)$$

Вычислим элементы $(a_{k,l}^{i,j})$ матрицы A для частного случая задачи (2.4.17), (2.4.18):

$$-\frac{\partial}{\partial x} p(x,y) \frac{\partial u}{\partial x} - \frac{\partial}{\partial y} q(x,y) \frac{\partial u}{\partial y} = f \quad \text{в} \quad D, \qquad (2.4.26)$$

$$u = 0 \quad \text{на} \quad \partial D. \qquad (2.4.27)$$

Можно представить $D_{k,l}^h$ в виде объединения шести треугольников $\{D_{k,l,m}^h\}_{m=1}^6$, порядок нумерации которых указан на рис. 2.14. Непо-

средственно вычисления показывают, что

$$\varphi_{k,l}(x,y)=\begin{cases} 1-\dfrac{1}{h}(x_k-x)-\dfrac{1}{h}(y_l-y), & \text{если} \quad x,y\in D^h_{k,l,1},\\ 1-\dfrac{1}{h}(x_k-x), & \text{если} \quad x,y\in D^h_{k,l,2},\\ 1+\dfrac{1}{h}(y_l-y), & \text{если} \quad x,y\in D^h_{k,l,3},\\ 1+\dfrac{1}{h}(x_k-x)+\dfrac{1}{h}(y_l-y), & \text{если} \quad x,y\in D^h_{k,l,4},\\ 1+\dfrac{1}{h}(x_k-x), & \text{если} \quad x,y\in D^h_{k,l,5},\\ 1-\dfrac{1}{h}(y_l-y), & \text{если} \quad x,y\in D^h_{k,l,6}. \end{cases} \tag{2.4.28}$$

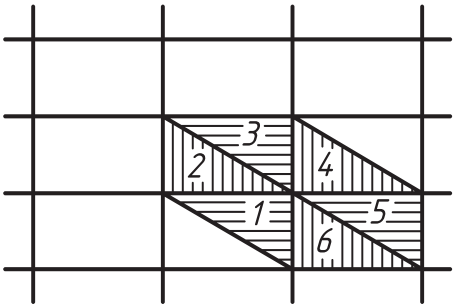


Рис. 2.14.

В силу симметрии A, трехдиагональности матриц $\{A_{k,k}\}_{k=1}^N$ и двухдиагональности матриц $\{A_{k,k-1}\}_{k=2}^N$ и $\{A_{k,k+1}\}_{k=1}^{N-1}$ нам достаточно указать формулы для вычисления элементов

$$a^{k,l}_{k,l},\quad a^{k,l-1}_{k,l},\quad a^{k-1,l}_{k,l},\quad a^{k-1,l+1}_{k,l},\quad 1\le k,l\le N.$$

Эти формулы, согласно (2.4.24) и (2.4.28), имеют вид (для простоты используются обозначения $D_i=D^h_{k,l,i}$)

$$a^{k,l}_{k,l}=\int\limits_{D^h_{k,l}}\left[p(x,y)\left(\frac{\partial\varphi_{k,l}}{\partial x}\right)^2+q(x,y)\left(\frac{\partial\varphi_{k,l}}{\partial y}\right)^2\right]\,dx\,dy=$$

$$= \frac{1}{h^2}\left[\int\limits_{D_1\cup D_2} p(x,y)\,dx\,dy + \int\limits_{D_4\cup D_5} p(x,y)\,dx\,dy + \right.$$

$$\left. + \int\limits_{D_3\cup D_4} q(x,y)\,dx\,dy + \int\limits_{D_1\cup D_6} q(x,y)\,dx\,dy\right],$$

$$a_{k,l}^{k,l-1} = \int\limits_{D_{k,l}^h\cup D_{k,l-1}^h}\left[p(x,y)\frac{\partial\varphi_{k,l}}{\partial x}\frac{\partial\varphi_{k,l-1}}{\partial x} + q(x,y)\frac{\partial\varphi_{k,l}}{\partial y}\frac{\partial\varphi_{k,l-1}}{\partial y}\right]\,dx\,dy =$$

$$= -\frac{1}{h^2}\left[\int\limits_{D_1\cup D_6} q(x,y)\,dx\,dy\right], \tag{2.4.29}$$

$$a_{k,l}^{k-1,l} = -\frac{1}{h^2}\left[\int\limits_{D_1\cup D_5} p(x,y)\,dx\,dy\right],$$

$$a_{k,l}^{k-1,l+1} = 0.$$

Отсюда сразу следует, что матрицы $\{A_{k,k-1}\}_{k=2}^N$ являются диагональными. Кроме того, если ввести вектор u с компонентами $u_{k,l} = \alpha_{k,l}$, где α — вектор системы (2.4.21) метода Ритца, то систему $Au = g$ можно записать в виде

$$(A_1 + A_2)u = g, \tag{2.4.30}$$

где

$$(A_1u)_{k,l} = -\widetilde{P}_{k-1/2,l}u_{k-1,l} + (\widetilde{P}_{k-1/2,l} + \widetilde{P}_{k+1/2,l})u_{k,l} - \widetilde{P}_{k+1/2,l}u_{k+1,l}, \tag{2.4.31}$$

$$(A_2u)_{k,l} = -\widetilde{Q}_{k,l-1/2}u_{k,l-1} + (\widetilde{Q}_{k,l-1/2} + \widetilde{Q}_{k,l+1/2})u_{k,l} - \widetilde{Q}_{k,l+1/2}u_{k,l+1}. \tag{2.4.32}$$

Здесь использованы обозначения

$$\widetilde{P}_{k\pm1/2,l} = \frac{1}{h^2}\int\limits_{D_{k,l}^h\cap D_{k\pm1,l}^h} p(x,y)\,dx\,dy,$$

$$\widetilde{Q}_{k,l\pm1/2} = \frac{1}{h^2}\int\limits_{D_{k,l}^h\cap D_{k,l\pm1}^h} q(x,y)\,dx\,dy.$$

Объединяя (2.4.30)—(2.4.32), легко видеть, что построенная нами по методу Ритца вариационно-разностная схема по структуре расположения ненулевых элементов и их виду практически совпадает

с чисто разностными схемами. В частности, для случая постоянных $p(x,y)$ и $q(x,y)$ вариационно-разностный и разностный аналоги дифференциального оператора полностью совпадают. Отмеченное обстоятельство позволяет применять для решения системы (2.4.30) эффективные итерационные методы, такие как метод расщепления, последовательной верхней релаксации и др.

На основе аппроксимирующих свойств кусочно-линейных базисных функций (аналогично тому, как это сделано в 2.4.1) показывается, что $\|u-u_h\|_{W_2^1(D)} \to 0$ при $h \to 0$, а если $\|u\|_{W_2^2(D)} \le c\|f\|_{L_2(D)}$, то справедливы также оценки

$$\|u-u_h\|_{L_2(D)} \le ch^2\|f\|_{L_2(D)}, \tag{2.4.33}$$

$$\|u-u_h\|_{W_2^1(D)} \le ch\|f\|_{L_2(D)}, \tag{2.4.34}$$

где постоянная c не зависит от h.

2.4.3. Проекционно-сеточная схема для эллиптического уравнения

В предыдущих двух пунктах рассматривались задачи с симметричными положительно определенными операторами. Поэтому для построения схем для данных задач мы применяли метод Ритца с использованием финитных базисных функций. Однако если оператор задачи не является симметричным, то необходимо привлечь другой вариационный или проекционный метод. Ниже на примере конкретной задачи с несимметричным оператором будет проиллюстрировано применение метода Галеркина для построения проекционно-сеточной схемы, аппроксимирующей данную задачу.

Пусть область D является объединением r прямоугольников $\{D_i\}_{i=1}^r$ со сторонами, параллельными осям координат, и $\widetilde{D}$ — наименьший по площади прямоугольник со сторонами, параллельными осям, содержащий область D (рис. 2.9): $\widetilde{D} = \{(x,y) : \ a \le x \le b, \ c \le x \le d\}$. Построим на $\widetilde{D}$ сетку так, как это сделано в 2.3.3, и введем билинейные базисные функции (2.3.25). Определим сеточную область D^h как множество внутренних точек (x_k, y_l) из D. Линейную оболочку билинейных базисных функций $\{\varphi_{k,l}\}$, соответствующих этим узлам, обозначим через $F_h \equiv \overset{\circ}{W}{}_2^{1,h}$. Очевидно, что $F_h \in \overset{\circ}{W}{}_2^1(D)$.

Рассмотрим теперь задачу

$$-\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} + \frac{\partial u}{\partial x} = f \quad \text{в} \quad D, \tag{2.4.35}$$

$$u = 0 \quad \text{на} \quad \partial D, \tag{2.4.36}$$

где $f \in F \equiv L_2(D)$. Обобщенную постановку такой задачи сформулируем следующим образом: требуется отыскать функцию $u(x,y) \in \overset{\circ}{W}{}_2^1(D)$, удовлетворяющую равенству

$$a(u,v) \equiv (u,v)_{L_0} + (Ku,v) = (f,v) \tag{2.4.37}$$

при любой функции $v \in \overset{\circ}{W}{}_2^1(D)$. Здесь $(\cdot,\cdot) \equiv (\cdot,\cdot)_{L_2(D)}$, $Ku = \partial u/\partial x$, а также

$$(u,v)_{L_0} = \left(\frac{\partial u}{\partial x}, \frac{\partial v}{\partial x}\right) + \left(\frac{\partial u}{\partial y}, \frac{\partial v}{\partial y}\right).$$

Можно показать, что (2.4.35), (2.4.36) имеет единственное обобщенное решение $u(x,y) \in \overset{\circ}{W}{}_2^1(D)$, причем $\|u\|_{W_2^1(D)} \le c\|f\|_{L_2(D)}$. Если же, например, область D является прямоугольником, то $\|u\|_{W_2^2(D)} \le c\|f\|_{L_2(D)}$.

Приближенное решение u^h будем искать в виде

$$u^h = \sum_{(x_i,y_j)\in D^h} u_{i,j}\varphi_{ij}(x,y), \tag{2.4.38}$$

где неизвестные коэффициенты $\{u_{ij}\}$ определим методом Галеркина (см. 2.1.3). Тогда соответствующая алгебраическая система будет иметь вид

$$Au = g, \tag{2.4.39}$$

где u и g — векторы с компонентами $\{u_{k,l}\}$ из (2.4.38) и $\{g_{k,l}\}$ ($g_{k,l} = \int\limits_D f\varphi_{k,l}\,dD$) соответственно; элементы $\{a_{k,l}^{i,j}\}$ матрицы A вычисляются по формуле

$$a_{k,l}^{i,j} = \int\limits_D \left(\frac{\partial\varphi_{i,j}}{\partial x}\frac{\partial\varphi_{k,l}}{\partial x} + \frac{\partial\varphi_{i,j}}{\partial y}\frac{\partial\varphi_{k,l}}{\partial y} + \frac{\partial\varphi_{i,j}}{\partial x}\varphi_{k,l}\right) dD =$$

$$= \int\limits_D \left(\varphi_j\varphi_l\frac{d\varphi_i}{dx}\frac{d\varphi_k}{dx} + \varphi_i\varphi_k\frac{d\varphi_j}{dy}\frac{d\varphi_l}{dy} + \varphi_j\frac{d\varphi_i}{dx}\varphi_k\varphi_l\right) dD, \tag{2.4.40}$$

где $\varphi_i = \varphi_i(x)$, $\varphi_k = \varphi_k(x)$, $\varphi_j = \varphi_j(y)$, $\varphi_l = \varphi_l(y)$.

Так же как и в предыдущем пункте, нетрудно видеть, что $a_{k,l}^{i,j} = 0$, если выполнено хотя бы одно из двух неравенств

$$|i-k| > 1, |j-l| > 1.$$

Отсюда сразу следует, что A будет блочной трехдиагональной матрицей вида (2.4.25). Приведем окончательные значения элементов $\{a_{k,l}^{i,j}\}$, предполагая для простоты, что сетка равномерная и ее шаг равен h:

$$a_{k,l}^{k,l} = \frac{1}{h^2}\int\limits_{x_{k-1}}^{x_{k+1}} dx \int\limits_{y_{l-1}}^{y_{l+1}} dy[\varphi_l^2(y) + \varphi_k^2(x)] = \frac{8}{3},$$

$$a_{k,l}^{k,l-1} = \frac{1}{h^2}\int\limits_{x_{k-1}}^{x_{k+1}} dx \int\limits_{y_{l-1}}^{y_l} dy[\varphi_{l-1}(y)\varphi_l(y) - \varphi_k^2(x)] = -\frac{1}{3},$$

$$a_{k,l}^{k,l+1} = -\frac{1}{3},$$

$$a_{k,l}^{k-1,l} = \int\limits_{x_{k-1}}^{x_k} dx \int\limits_{y_{l-1}}^{y_{l+1}} dy \left[-\frac{\varphi_l^2(y)}{h^2} + \frac{\varphi_{k-1}(x)\varphi_k(x)}{h^2} - \frac{1}{h}\varphi_k(x)\varphi_l^2(y)\right] = -\frac{1}{3} - \frac{h}{3},$$

$$a_{k,l}^{k-1,l-1} = \int\limits_{x_{k-1}}^{x_k} dx \int\limits_{y_{l-1}}^{y_l} dy \left[-\frac{1}{h^2}\varphi_l(y)\varphi_{l-1}(y) -\right.$$

$$\left. - \frac{1}{h^2}\varphi_k(x)\varphi_{k-1}(x) - \varphi_k(x)\varphi_l(y)\varphi_{l-1}(y)\right] = -\frac{1}{3} - \frac{h}{12}, \qquad (2.4.41)$$

$$a_{k,l}^{k-1,l+1} = \int\limits_{x_{k-1}}^{x_k} dx \int\limits_{y_l}^{y_{l+1}} dy \left[-\frac{\varphi_{l+1}(y)\varphi_l(y)}{h^2} -\right.$$

$$\left. - \frac{\varphi_{l+1}(x)\varphi_{k+1}(x)}{h^2} - \frac{\varphi_k(x)\varphi_l(y)\varphi_{l+1}(y)}{h}\right] = -\frac{1}{3} - \frac{h}{12},$$

$$a_{k,l}^{k+1,l} = -\frac{1}{3} + \frac{h}{3}, \quad a_{k,l}^{k+1,l-1} = -\frac{1}{3} + \frac{h}{12},$$

$$a_{k,l}^{k+1,l+1} = -\frac{1}{3} + \frac{h}{12}, \quad (x_k, y_l) \in D^h.$$

Таким образом, оказывается, что построенная проекционно-сеточная схема при $f(x,y) = f \equiv \text{const}$ совпадает с несколько необыч-

ной разностной схемой

$$\frac{8}{3}u_{k,l} - \frac{1}{3}u_{k,l+1} - \frac{1}{3}u_{k,l-1} - (\frac{1}{3} + \frac{h}{3})u_{k-1,l} -$$

$$- (\frac{1}{3} + \frac{h}{12})u_{k-1,l-1} - (\frac{1}{3} + \frac{h}{12})u_{k-1,l+1} - (\frac{1}{3} - \frac{h}{3})u_{k+1,l} -$$

$$- (\frac{1}{3} - \frac{h}{12})u_{k+1,l-1} - (\frac{1}{3} - \frac{h}{12})u_{k+1,l+1} = h^2 f. \qquad (2.4.42)$$

Эту разностную схему легко построить с помощью обычных трехточечных аппроксимаций вторых производных, если предварительно заменить дифференциальную часть уравнения (2.4.35) в узлах D^h на приближенную следующим образом:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} - \frac{\partial u}{\partial x}\Big|_{\substack{x=x_k \\ y=y_l}} \approx$$

$$\approx \frac{1}{6}\left[\sum_{l=k-1}^{k+1} \beta_{k-i}\left(\frac{\partial^2 u}{\partial y^2}\right)_{\substack{x=x_k \\ y=y_l}} + \sum_{j=l-1}^{l+1} \beta_{l-j}\left(\frac{\partial^2 u}{\partial x^2} - \frac{\partial u}{\partial x}\right)_{\substack{x=x_k \\ y=y_l}}\right], \qquad (2.4.43)$$

где

$$\beta_{-1} = \beta_1 = 1, \quad \beta_0 = 4 \quad \text{и} \quad (x_k, y_l) \in D^h.$$

Изучим вопрос о сходимости u^h к u при $h = \max(h_x, h_y)$, где $h_x = \max\limits_k(x_k - x_{k-1})$, $h_y = \max\limits_l(y_l - y_{l-1})$. Сделаем это, не пользуясь утверждением о сходимости из 2.1.3. Отмечаем, что в рассматриваемой задаче форма $a(u,v)$ является $\overset{\circ}{W}{}_2^1$-определенной и $\overset{\circ}{W}{}_2^1$-ограниченной, т. е. имеют место соотношения

$$c_0\|u\|^2_{\overset{\circ}{W}{}_2^1} \le a(u,u), \qquad (2.4.44)$$

$$|a(u,v)| \le c_1\|u\|_{\overset{\circ}{W}{}_2^1}\|v\|_{\overset{\circ}{W}{}_2^1} \qquad (2.4.45)$$

при любых $u, v \in \overset{\circ}{W}{}_2^1(D)$. (Здесь постоянные $c_0, c_1 > 0$ не зависят от $u,\ v$.) Если воспользоваться (2.4.44), (2.4.45), то нетрудно доказать сходимость u^h к u. Действительно, запишем равенства

$$a(u, \varphi_{k,l}) = (f, \varphi_{k,l}), \quad a(u^h, \varphi_{k,l}) = (f, \varphi_{k,l}), \quad a(u - u^h, \varphi_{k,l}) = 0$$

при произвольной базисной функции $\varphi_{k,l}$, для которой $(x_k, y_l) \in D^h$. Отсюда легко получаем соотношение

$$a(u-u^h, u-u^h) = a(u-u^h, u-v), \quad \forall v \in \overset{\circ}{W}{}_2^{1,h},$$

из которого с привлечением (2.4.44), (2.4.45) имеем

$$c_0 \|u-u^h\|^2_{\overset{\circ}{W}{}_2^1 D} \le c_1 \|u-u^h\|_{\overset{\circ}{W}{}_2^1(D)} \|u-v\|_{\overset{\circ}{W}{}_2^1(D)}.$$

Следовательно,

$$\|u-u^h\|_{\overset{\circ}{W}{}_2^1(D)} \le \frac{c_1}{c_0} \inf_{v \in \overset{\circ}{W}{}_2^{1,h}} \|u-v\|_{\overset{\circ}{W}{}_2^1(D)} \to 0 \tag{2.4.46}$$

при $h \to 0$. Если же $\|u\|_{C^{(2)}} < \infty$, то, вспоминая аппроксимирующие свойства билинейных базисных функций, получаем также оценку скорости сходимости

$$\|u-u^h\|_{\overset{\circ}{W}{}_2^1(D)} \le ch \tag{2.4.47}$$

(отметим, что в действительности оценки (2.4.47) справедливы и при $u(x,y) \in W_2^2(D)$).

Аналогично изложенному выше можно построить проекционно-сеточную схему и провести ее исследование и для более общей задачи (2.1.26), (2.1.27). Для этого достаточно задать конкретный вид коэффициентов в области D, воспользоваться подходящими финитными базисными функциями (не обязательно билинейными), а затем определить элементы матрицы и правой части системы (2.1.48). Общий вид этих элементов приведен в 2.1.3.

2.4.4. Решение третьей краевой задачи для эллиптического уравнения второго порядка

Ранее было показано, как получить вариационно-разностные или проекционно-сеточные схемы в задачах с главными граничными условиями. Теперь же сделаем это для задачи с естественным граничным условием. Для простоты изложение будет осуществляться в применении к уравнению с оператором Лапласа. Однако легко заметить, что изложение распространяется на задачи для эллиптического уравнения второго порядка более общего вида.

Рассмотрим краевую задачу вида

$$-\Delta u = f(x), \quad x = (x_1, x_2) \in D, \tag{2.4.48}$$

$$\sigma u + \frac{\partial u}{\partial n} = g(x), \quad x \in \partial D. \tag{2.4.49}$$

Предполагается, что $f(x) \in F \equiv L_2(D)$; граница ∂D достаточно гладкая; $g(x) \in L_2(\partial D)$; $\sigma(x)$ — ограниченная функция, определенная на ∂D, причем $0 < \sigma_0 \le \sigma(x) \le \sigma_1$; $\sigma_0,\ \sigma_1$ —постоянные.

Запишем (2.4.48), (2.4.49) в обобщенной формулировке (минуя классическую операторную постановку). Для этого умножим (2.4.48) скалярно в $L_2(D)$ на произвольную функцию $v \in W_2^1(D)$ и выполним интегрирование по частям с учетом краевых условий. В результате приходим к равенству

$$(u, v)_L = (f, v) + \int\limits_{\partial D} gv \, d\omega, \tag{2.4.50}$$

где

$$(u, v)_L = \sum_{i=1}^{2} \left(\frac{\partial u}{\partial x_i}, \frac{\partial v}{\partial x_i} \right) + \int\limits_{\partial D} \sigma uv \, d\omega.$$

Назовем *обобщенным решением* задачи (2.4.48), (2.4.49) функцию $u \in W_2^1(D)$, удовлетворяющую (2.4.50) при произвольной функции $v \in W_2^1(D)$.

Для исследования задачи в обобщенной постановке нам потребуется неравенство

$$\int\limits_D |v|^2 \, dx \le c \left(\sum_{i=1}^{2} \int\limits_D \left| \frac{\partial v}{\partial x_i} \right|^2 + \int\limits_{\partial D} |v|^2 \, d\omega \right), \tag{2.4.51}$$

справедливое для любой функции $v \in W_2^1(D)$.

С помощью (2.4.51) легко показать, что форма $(u, v)_L$ удовлетворяет всем аксиомам скалярного произведения, а величина $\|u\|_L = (u, u)^{1/2}$ может быть принята в качестве нормы в $W_2^1(D)$. Имеет место эквивалентность норм $\|\cdot\|_L$ и $\|\cdot\|_{W_2^1(D)}$. Действительно, так как для любой функции $v \in W_2^1(D)$ выполняется неравенство $\|v\|_{L_2(\partial D)} \le c_1 \|v\|_{W_2^1}$,

то

$$\|v\|_L = \left(\sum_{i=1}^{2}\left\|\frac{\partial v}{\partial x_i}\right\|^2 + \int\limits_{\partial D}\sigma v^2\,d\omega\right)^{1/2} \le \left(\sum_{i=1}^{2}\left\|\frac{\partial v}{\partial x_i}\right\|^2 + \sigma_1 c_1^2\|v\|^2_{W_2^1}\right)^{1/2} \le$$

$$\le \sqrt{1+\sigma_1 c_1^2}\,\|v\|_{W_2^1}. \tag{2.4.52}$$

С другой стороны, в силу (2.4.51) имеем

$$\|v\|_{W_2^1(D)} = \left(\sum_{i=1}^{2}\left\|\frac{\partial v}{\partial x_i}\right\|^2 + \|v\|^2\right)^{1/2} \le \left(\sum_{i=1}^{2}(c+1)\left\|\frac{\partial v}{\partial x_i}\right\|^2 + c\|v\|^2_{L_2(\partial D)}\right)^{1/2} \le$$

$$\le \max\left(\sqrt{c+1}, \sqrt{\frac{c}{\sigma_0}}\right)\left(\sum_{i=1}^{2}\left\|\frac{\partial v}{\partial x_i}\right\|^2 + \int\limits_{\partial D}\sigma v^2\,d\omega\right)^{1/2}, \tag{2.4.53}$$

т. е. имеют место неравенства

$$c_2\|v\|_{W_2^1(D)} \le \|v\|_L \le c_3\|v\|_{W_2^1(D)}, \quad c_2, c_3 > 0.$$

Итак, пусть в $W_2^1(D)$ введены скалярное произведение $(u,v)_L$ и норма $\|u\|_Z = (u,u)_L^{1/2}$. Покажем, что обобщенное решение задачи (2.4.48), (2.4.49) существует и единственно. Для этого рассмотрим функционал, стоящий в правой части (2.4.50). Он является ограниченным в

$$\left|(f,v) + \int\limits_{\partial D} gv\,d\omega\right| \le \|f\|\|v\| + \sigma_0^{-1/2}\|g\|_{L_2(\partial D)}\|v\|_L \le$$

$$\le c(\|f\|^2 + \sigma_0^{-1}\|g\|^2_{L_2(\partial D)})^{1/2}\|v\|_L \le c\|v\|_L.$$

Следовательно, по теореме Рисса о представлении ограниченного функционала в W_2^1 существует такая функция u_0, что

$$(f,v) + \int\limits_{\partial D} gv\,d\omega = (u_0,v)_L,$$

$$\|u_0\|_L \le c(\|f\|^2 + \sigma_0^{-1}\|g\|^2_{L_2(\partial D)})^{1/2}. \tag{2.4.54}$$

Но тогда (2.4.50) принимает вид

$$(u, v)_L = (u_0, v)_L, \quad v \in W_2^1.$$

Отсюда следует, что обобщенное решение задачи существует, единственно и совпадает с функцией u_0. Следовательно, для него справедлива оценка из (2.4.54). Соотношение (2.4.55) есть уравнение Эйлера в обобщенной форме, которому удовлетворяет функция, реализующая минимум функционала $F(u) = \|u\|_L^2 - 2(u, u_0)_L = (u-u_0)_L^2 - \|u_0\|_L^2$ в пространстве W_2^1. Используя (2.4.54), можно записать этот функционал в виде

$$F(u) = \|u\|_L^2 - 2(u, f) - 2\int\limits_{\partial D} gu \, d\omega. \tag{2.4.55}$$

Итак, доказано, что обобщенное решение минимизирует $F(u)$ в W_2^1.

Заметим, что если предположить гладкость функции $g(x)$, то, как известно из теории эллиптических уравнений, обобщенное решение задачи принадлежит $W_2^2(D)$, причем $\|u\|_{W_2^2(D)} \le c$, где постоянная c зависит лишь от области D и от функций f, g.

Сформулируем алгоритм приближенного решения задачи (2.4.48), (2.4.49). Опишем вокруг D многоугольник (см. рис. 2.15), триангулируем его, стремясь при этом, чтобы все элементарные треугольники D_i имели площадь одного порядка, минимальный из углов θ_0 в треугольниках был всегда строго положителен и не зависел от триангуляций, величины $\operatorname{mes}(D_i), \operatorname{mes}(D \cap D_j)$ были также одного порядка при различных i и j. Обозначим через D_h объединение всех треугольников, имеющих ненулевые пересечения с D. Число вершин треугольников в D_h обозначим через N. В качестве базисных функций выберем кусочно-линейные функции $\{\varphi_i(x)\}$ (см. 2.3.2).

Приближенное решение ищем в виде

$$u_h = \sum_{i=1}^{N} a_i \varphi_i(x),$$

где a_i определяются из системы уравнений

$$(u_h, \varphi_i)_L = (f, \varphi_i) + \int\limits_{\partial D} g\varphi_i \, d\omega, \quad i = 1, 2, \ldots, N, \tag{2.4.56}$$

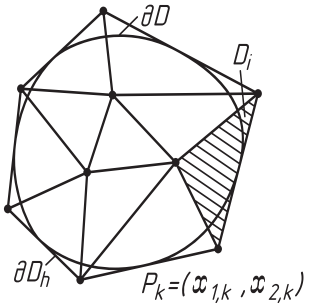


Рис. 2.15.

или

$$\hat{A}a = f, \tag{2.4.57}$$

где $a = (a_1, a_2, \ldots, a_N)^T$, $f = (f_1, f_2, \ldots, f_N)^T$, $\hat{A} = A_{ij}$,

$$A_{ij} = A_{ji} = [\varphi_i, \varphi_j] = \sum_{k=1}^{2} \int\limits_{D_{ij}} \frac{\partial \varphi_i}{\partial x_k} \frac{\partial \varphi_j}{\partial x_k}\, dx_1\, dx_2 + \int\limits_{\partial D} \sigma \varphi_i \varphi_j\, d\omega,$$

$$f_i = \int\limits_{D_i} f\varphi_i\, dx_1\, dx_2 + \int\limits_{\partial D} g\varphi_i\, d\omega, \quad D_i = D \cap \operatorname{supp} \varphi_i,$$

$$D_{ij} = D \cap \operatorname{supp} \varphi_i \cap \operatorname{supp} \varphi_j, \quad i, j = 1, 2, \ldots, N.$$

Система (2.4.57) имеет симметричную положительно определенную матрицу, в силу чего она имеет единственное решение, однозначно определяющее $u_h = \sum\limits_{i=1}^{N} a_i \varphi_i(x)$.

Оценим скорость сходимости u_h к u. Из теории метода Ритца следует, что

$$\|u - u_h\|_L^2 = (u - u_h, u - w_h)_L \le \|u - u_h\|_L \|u - w_h\|_L,$$

где $w_h = \sum\limits_{i=1}^{N} b_i \varphi_i$, b_i — произвольные величины. Отсюда, применяя (2.4.52), (2.4.53), получаем

$$\|u - u_h\|_{W_2^1(D)} \le c_4 \|u - w_h\|_{W_2^1(D)}, \tag{2.4.58}$$

$$\|u - u_h\|_{W_2^1(D)} \le c_4 \inf_{b_i} \left\| u - \sum_{i=1}^{N} b_i \varphi_i \right\|_{W_2^1(D)}.$$

Из 2.3.2 известно, что для произвольной функции $u \in W_2^2(D)$ существует линейная комбинация $u_I = \sum\limits_{i=1}^{N} c_i \varphi_i$, при которой

$$\|u - u_I\|_{W_2^1(D)} \le c_5 \frac{h}{\sin\theta_0} \|u\|_{W_2^2(D)},$$

где h — максимальная длина сторон треугольников, θ_0 — минимальный из углов, $c_5 > 0$ — постоянная, не зависящая от h, θ_0, $u(x)$. Тогда, используя этот результат и предполагая, что решение принадлежит W_2^2, приходим к оценке в метрике W_2^1:

$$\|u - u_h\|_{W_2^1(D)} \le c_4 \inf_{b_i} \left\| u - \sum_{i=1}^{N} b_i \varphi_i \right\|_{W_2^1(D)} \le$$

$$\le c_4 \|u - u_I\|_{W_2^1(D)} \le c \frac{h}{\sin\theta_0} \|u\|_{W_2^2(D)}. \qquad (2.4.59)$$

Можно показать, что справедлива также оценка вида[8)]

$$\|u - u_h\|_{L_2(D)} \le c \left(\frac{h}{\sin\theta_0} \right)^2 \|u\|_{W_2^2(D)}. \qquad (2.4.60)$$

В заключение отметим, что поскольку в рассматриваемой задаче краевое условие естественное и базисные функции могут ему не удовлетворять, то часто в таких задачах используют равномерные сетки. Это упрощает построение базисных функций и весь алгоритм в целом. Однако здесь может нарушиться условие равновеликости $\mathrm{mes}(D_i)$ и $\mathrm{mes}(D \cap D_i)$, что может привести к плохой обусловленности матрицы системы. Поэтому, чтобы добиться выполнения и этого условия, видоизменяют лишь регулярные треугольники (построенные на равномерной сетке), прилегающие к ∂D.

2.4.5. Метод штрафа

Как видно из 2.4.4, в задаче с естественными граничными условиями базисные функции можно выбирать не удовлетворяющими этим условиям. Последнее обстоятельство наводит на мысль: а нельзя ли задачу с главными граничными условиями приближенно свести к задаче с естественными условиями? Оказывается, что в ряде

[8)] Г. И. Марчук, В. И. Агошков [5].

случаев это можно сделать при помощи *метода штрафа*. Сущность данного метода состоит в следующем. Пусть в $D \subset \mathbf{R}^2$ с криволинейной границей ∂D решается задача (2.4.17), (2.4.18), т. е. задача с главными граничными условиями. Осуществим приближенную замену (2.4.17), (2.4.18) третьей краевой задачей (в которой граничное условие уже является естественным):

$$-\sum_{i=1}^{2} \frac{\partial}{\partial x_i} A_{i,j} \frac{\partial u_\varepsilon}{\partial x_j} = f \quad \text{в} \quad D, \tag{2.4.61}$$

$$u_\varepsilon + \varepsilon \frac{\partial u_\varepsilon}{\partial \nu} = 0 \quad \text{на} \quad \partial D, \tag{2.4.62}$$

где ε — малое положительное число, $\dfrac{\partial u_\varepsilon}{\partial \nu} = \sum\limits_{i,j=1}^{2} A_{ij}(x) \cos(n, x_i)$, n — единичный вектор внешней нормали к ∂D. Задача (2.4.61), (2.4.62), как мы уже знаем, сводится к минимизации функционала

$$J(v) = \int\limits_D \left(\sum_{i,j=1}^{2} A_{i,j}(x) \frac{\partial v}{\partial x_i} \frac{\partial v}{\partial x_j} - 2fv \right) dx + \frac{1}{\varepsilon} \int\limits_D v^2 \, d\omega \tag{2.4.63}$$

в пространстве $W_2^1(D)$ (а не в $\overset{\circ}{W}{}_2^1(D)$!). Если теперь методом Ритца построить приближенное решение u_ε^h задачи (2.4.61), (2.4.62), согласно изложенному в предыдущем пункте алгоритму, то u_ε^h можно принять за приближенное решение задачи (2.4.17), (2.4.18). Оценка погрешности $u_\varepsilon - u_\varepsilon^h$ осуществляется с помощью изложенных выше подходов. Если же известна оценка погрешности $u - u_\varepsilon$, то в результате будем иметь оценку для $u - u_\varepsilon^h$. Так, например, для задач (2.4.17), (2.4.18) и (2.4.61), (2.4.62) частного вида

$$-\Delta u = f \quad \text{в} \quad D, \quad u = 0 \quad \text{на} \quad \partial D, \tag{2.4.64}$$

$$-\Delta u_\varepsilon = f \quad \text{в} \quad D, \quad u_\varepsilon + \varepsilon \frac{\partial u_\varepsilon}{\partial n} = 0 \quad \text{на} \quad \partial D \tag{2.4.65}$$

при $\partial D \in C^{(2)}$ имеем (Г. И. Марчук, В. И. Агошков [5])

$$\|u - u_\varepsilon\|_{W_2^1(D)} \le c_0 \varepsilon \|u_\varepsilon\|_{W_2^2(D)} \le c_0 c_1 \varepsilon \|f\|_{L_2(D)}, \tag{2.4.66}$$

где $c_i = \text{const} > 0$. Теперь из (2.4.59) получаем

$$\|u_\varepsilon - u_\varepsilon^h\|_{W_2^1(D)} \le c\frac{h}{\sin\theta_0}\|f\|_{L_2(D)}. \tag{2.4.67}$$

Следовательно, при $\varepsilon = h$ справедлива оценка

$$\|u - u_\varepsilon^h\|_{W_2^1(D)} \le c\frac{h}{\sin\theta_0}\|f\|_{L_2(D)}. \tag{2.4.68}$$

Отметим, что сходимость u_ε^h к u при $\varepsilon \to 0$, $h \to 0$ можно доказать и для эллиптических краевых задач более общего вида.

2.5. Метод интегральных тождеств

В главе 1 на примере ряда задач мы проиллюстрировали применение метода конечных разностей к построению схем, аппроксимирующих исходные задачи. В предыдущих параграфах этой главы для этой цели мы воспользовались вариационными и проекционными методами. В настоящем параграфе в применении к уравнению диффузии будет рассмотрен еще один метод построения конечно-разностных уравнений на основе интегрального тождества, полученного автором, и на основе вариационной формы этого тождества.

2.5.1. Построение разностных уравнений для задач с разрывными коэффициентами на основе интегрального тождества

Рассмотрим уравнение диффузии для одномерных областей. Оно имеет вид

$$-\frac{d}{dx}p\frac{d\varphi}{dx} + q\varphi = f, \tag{2.5.1}$$

где $p = p(x) \ge p_0 > 0$ — коэффициент диффузии, $q = q(x) \ge 0$ — сечение поглощения частиц и $f = f(x)$ — источник диффундирующей субстанции. Будем считать, что p, q и f — кусочно-непрерывные функции с возможными точками разрыва первого рода.

Требуется найти решение уравнения (2.5.1), являющееся непрерывной функцией, обладающее дифференцируемым «потоком»

$$J = p\frac{d\varphi}{dx}$$

и удовлетворяющее граничным условиям

$$\varphi(0) = 0, \quad \varphi(1) = 0. \tag{2.5.2}$$

Рассмотрим на интервале изменения переменной x две системы узловых точек: основную — $\{x_k\}$ и вспомогательную — $\{x_{k+1/2}\}$. Точки этих двух систем взаимно чередуются, т. е. $x_k < x_{k+1/2} < x_{k+1}$. В дальнейшем предполагается, что

$$x_{k+1/2} = \frac{x_{k+1} + x_k}{2}, \quad 0 = x_0 < x_{1/2} < \ldots < x_{n-1/2} < x_n = 1.$$

Проинтегрируем уравнение (2.5.1) по x в пределах $(x_{k-1/2}, x_{k+1/2})$. В результате получим соотношение баланса

$$-J_{k+1/2} + J_{k-1/2} + \int\limits_{x_{k-1/2}}^{x_{k+1/2}} (q\varphi - f)\, dx = 0, \tag{2.5.3}$$

где

$$J_{k\pm1/2} = J(x_{k\pm1/2}).$$

Для нахождения $J_{k\pm1/2}$ поступим следующим образом. Проинтегрируем уравнение (2.5.1) в пределах $(x_{k-1/2}, x)$. Получим соотношение

$$p\frac{d\varphi}{dx} = J_{k-1/2} + \int\limits_{x_{k-1/2}}^{x} (q\varphi - f)\, d\xi. \tag{2.5.4}$$

Выражение (2.5.4) разделим на p и проинтегрируем в пределах (x_{k-1}, x_k). В результате получим

$$\varphi_k - \varphi_{k-1} = J_{k-1/2} \int\limits_{x_{k-1}}^{x_k} \frac{dx}{p} + \int\limits_{x_{k-1}}^{x_k} \frac{dx}{p} \int\limits_{x_{k-1/2}}^{x} (q\varphi - f)\, d\xi. \tag{2.5.5}$$

Разрешая уравнение (2.5.5) относительно $J_{k-1/2}$, приходим к следующему соотношению:

$$J_{k-1/2} = \frac{1}{\int\limits_{x_{k-1}}^{x_k} \frac{dx}{p}} \left[\varphi_k - \varphi_{k-1} - \int\limits_{x_{k-1}}^{x_k} \frac{dx}{p} \int\limits_{x_{k-1/2}}^{x} (q\varphi - f)\, d\xi \right]. \tag{2.5.6}$$

Аналогичное выражение для $J_{k+1/2}$ получаем, заменив в (2.5.6) индекс k на $k+1$. Таким образом, нам удалось потоки $J_{k\pm1/2}$ выразить через известные функции и решение задачи. Соотношение (2.5.6) точное. Подставив полученное выражение (2.5.6) для $J_{k-1/2}$ и соответствующее значение для $J_{k+1/2}$ в равенство (2.5.3), приходим к соотношению

$$\begin{aligned} &-\frac{\varphi_{k+1} - \varphi_k}{\int\limits_{x_k}^{x_{k+1}} \frac{dx}{p}} + \frac{\varphi_k - \varphi_{k-1}}{\int\limits_{x_{k-1}}^{x_k} \frac{dx}{p}} + \int\limits_{x_{k-1/2}}^{x_{k+1/2}} (q\varphi - f)\, dx = \\ &= -\frac{1}{\int\limits_{x_k}^{x_{k+1}} \frac{dx}{p}} \int\limits_{x_k}^{x_{k+1}} \frac{dx}{p} \int\limits_{x_{k+1/2}}^{x} (q\varphi - f)\, d\xi + \\ &+ \frac{1}{\int\limits_{x_{k-1}}^{x_k} \frac{dx}{p}} \int\limits_{x_{k-1}}^{x_k} \frac{dx}{p} \int\limits_{x_{k-1/2}}^{x} (q\varphi - f)\, d\xi. \end{aligned} \tag{2.5.7}$$

Формулу (2.5.7) будем называть *основным тождеством* для получения конечно-разностных уравнений.

Введем в рассмотрение оператор A, который на классе решений Φ уравнения (2.5.1) определяется соотношениями

$$(A\varphi)_k = -\frac{1}{\Delta x_k} \left(\frac{\varphi_{k+1} - \varphi_k}{\int\limits_{x_k}^{x_{k+1}} \frac{dx}{p}} - \frac{\varphi_k - \varphi_{k-1}}{\int\limits_{x_{k-1}}^{x_k} \frac{dx}{p}} - \int\limits_{x_{k-1/2}}^{x_{k+1/2}} q\varphi\, dx - \right.$$

$$- \frac{1}{\int\limits_{x_k}^{x_{k+1}} \frac{dx}{p}} \int\limits_{x_k}^{x_{k+1}} \frac{dx}{p} \int\limits_{x_{k+1/2}}^{x} q\varphi \, d\xi + \frac{1}{\int\limits_{x_{k-1}}^{x_k} \frac{dx}{p}} \int\limits_{x_{k-1}}^{x_k} \frac{dx}{p} \int\limits_{x_{k-1/2}}^{x} q\varphi \, d\xi \Bigg),$$

и вектор f с компонентами[9)]

$$(f)_k = \frac{1}{\Delta x_k} \int\limits_{x_{k-1/2}}^{x_{k+1/2}} f \, dx +$$

$$+ \frac{1}{\Delta x_k} \left(\frac{1}{\int\limits_{x_k}^{x_{k+1}} \frac{dx}{p}} \int\limits_{x_k}^{x_{k+1}} \frac{dx}{p} \int\limits_{x_{k+1/2}}^{x} f \, d\xi - \frac{1}{\int\limits_{x_{k-1}}^{x_k} \frac{dx}{p}} \int\limits_{x_{k-1}}^{x_k} \frac{dx}{p} \int\limits_{x_{k-1/2}}^{x} f \, d\xi \right), \qquad (2.5.8)$$

где

$$\Delta x_k = x_{k+1/2} - x_{k-1/2}, \quad k = 1, 2, \ldots, n-1.$$

В дальнейшем ради простоты будем считать, что класс Φ решений уравнения (2.5.1) состоит из функций φ, имеющих определенную гладкость и удовлетворяющих граничным условиям (2.5.2). Объединяя соотношения (2.5.7) для $k = 1, 2, \ldots, n-1$, получим систему

$$A\varphi = f. \qquad (2.5.9)$$

Рассмотрим теперь различные аппроксимации уравнения (2.5.9). С этой целью введем в пространстве F_h сеточных функций вида

$$\varphi^h = (\varphi_1^h, \ldots, \varphi_{n-1}^h),$$

заданных в точках $x_1, x_2, \ldots, x_{n-1}$, норму

$$\|\varphi^h\|_{F_h}^2 = \sum_{k=1}^{n-1} (\varphi_k^h)^2 \Delta x_k. \qquad (2.5.10)$$

[9)] Не следует смешивать $(f)_k$ с $f(x_k)$.

В качестве приближенной задачи рассмотрим задачу

$$A^h\varphi^h = f^h, \tag{2.5.11}$$

где

$$(A^h\varphi^h)_k = -\frac{1}{\Delta x_k}\left(\frac{\varphi^h_{k+1}-\varphi^h_k}{\int\limits_{x_k}^{x_{k+1}}\frac{dx}{p}} - \frac{\varphi^h_k-\varphi^h_{k-1}}{\int\limits_{x_{k-1}}^{x_k}\frac{dx}{p}} - \varphi^h_k\int\limits_{x_{k-1/2}}^{x_{k+1/2}} q\,dx\right), \tag{2.5.12}$$

$$(f^h)_k = \tfrac{1}{\Delta x_k}\int\limits_{x_{k-1/2}}^{x_{k+1/2}} f\,dx$$

для $k = 1, 2, \ldots, n-1$ и, кроме того, $\varphi^h_0 = \varphi^h_n = 0$. В соответствии с определением аппроксимации, используя неравенство треугольника, получим[10)]

$$\|A^h(\varphi)_h - f^h\|_{F_h} \le \|\xi^h\|_{F_h} + \|\eta^h\|_{F_h} + \|\theta^h_0\|F_h, \tag{2.5.13}$$

где

$$(\xi^h)_k = \frac{1}{\Delta x_k}\left(\int\limits_{x_{k-1/2}}^{x_{k+1/2}} q\varphi\,dx - \varphi^h_k\int\limits_{x_{k-1/2}}^{x_{k+1/2}} q\,dx\right),$$

$$(\eta^h)_k = -\frac{1}{\Delta x_k}\left(\frac{1}{\int\limits_{x_k}^{x_{k+1}}\frac{dx}{p}}\int\limits_{x_k}^{x_{k+1}}\frac{dx}{p}\int\limits_{x_{k+1/2}}^{x} q\varphi\,d\xi - \frac{1}{\int\limits_{x_{k-1}}^{x_k}\frac{dx}{p}}\int\limits_{x_{k-1}}^{x_k}\frac{dx}{p}\int\limits_{x_{k-1/2}}^{x} q\varphi\,d\xi\right),$$

$$(\theta^h)_k = -\frac{1}{\Delta x_k}\left(\frac{1}{\int\limits_{x_k}^{x_{k+1}}\frac{dx}{p}}\int\limits_{x_k}^{x_{k+1}}\frac{dx}{p}\int\limits_{x_{k+1/2}}^{x} f\,d\xi - \frac{1}{\int\limits_{x_{k-1}}^{x_k}\frac{dx}{p}}\int\limits_{x_{k-1}}^{x_k}\frac{dx}{p}\int\limits_{x_{k-1/2}}^{x} f\,d\xi\right).$$

[10)] Здесь и далее для любой непрерывной на $[0,1]$ функции u обозначение $(u)_h$ принято для вектора размерности $n-1$ из F_h с компонентами $u(x_k)$.

Оценим величины $\|\xi^h\|_{F_h}$, $\|\eta^h\|_{F_h}$ и $\|\theta^h\|_{F_h}$. Предположим, что $q,\ f \in$ $\in\ Q^{(2)}(0,1)$ и $p\ \in\ Q^{(3)}(0,1)$, где $Q^{(\varepsilon)}(0,1)$ — пространство кусочно-непрерывных вместе с производными до s-го порядка функций на $(0,1)$ с возможными разрывами первого рода в точках $0 < y_1 < y_2 < \ldots < y_m < 1$. В дальнейшем мы везде будем предполагать, что множество точек $\{y_l\}_{l=1}^m$ принадлежит множеству узлов сетки $\{x_k\}_{k=1}^{n-1}$. Это требование будет необходимо при анализе погрешностей аппроксимации.

Из сделанных предположений следует, что решение φ задачи (2.5.1) будет непрерывной функцией, причем на каждом из отрезков $(y_l, y_{l+1})_{l=1}^{m-1}$ решение будет непрерывно вместе с производными до четвертого порядка включительно, т. е. $\varphi \in Q^{(4)}(0,1)$.

Исследуем теперь поведение компонент векторов ξ^h, η^h и θ^h в предположении $h \ll 1$, где

$$h = \max_{0 \le k \le n-1} |x_{k+1} - x_k|.$$

Используя обычное разложение по формуле Тейлора в окрестностях узлов сетки, нетрудно показать, что модули компонент каждого из этих векторов мажорируются сверху соответствующими компонентами вектора ω^h, где

$$(\omega^h)_k = \begin{cases} Nh, & \text{если } x_k \text{ является одной из точек } y_l, \quad (l = 1, 2, \ldots, m); \\ M(|\Delta x_{k+1/2} - \Delta x_{k-1/2}| + h^2) & \text{в остальных случаях;} \end{cases}$$

$M,\ N$ — некоторые положительные константы. Здесь введено обозначение $\Delta x_{k+1/2} = x_{k+1} - x_k$. Предположим, что на интервале $(0,1)$ имеется точка разрыва коэффициентов $x = x_l$ $(1 \le l \le n)$ и $\Delta x_{k+1/2} = \Delta x_{k-1/2}$ при $k \ne l$. Тогда на основании (2.5.10) имеем

$$\|\omega^h\|_{F_h}^2 = \sum_{k=1, k\ne l}^{n-1} (\omega^h)_k^2 \Delta x_k + (\omega^h)_l^2 \Delta x_l.$$

Предположим, далее, что $\widetilde{h} = \max\{\Delta x_k, 2(1 - x_{n-1/2}), 2x_{1/2}\}$. Учитывая соотношение

$$1 - \widetilde{h} \le \sum_{k=1}^{n-1} \Delta x_k < 1$$

и используя при оценке членов в равенстве для квадрата нормы ω^h приведенные выше локальные оценки для $(\omega^h)_k$, получим оценку

$$\|\omega^h\|^2_{F_h} \le M^2h^4 + N^2h^3;$$

отсюда

$$\|\omega^h\|_{F_h} \le Ch^{3/2}.$$

Таким образом, для норм векторов погрешностей аппроксимации ξ^h, η^h и θ^h справедлива оценка

$$\max(\|\xi^h\|_{F_h}, \|\eta^h\|_{F_h}, \|\theta^h\|_{F_h}) \le Ch^{3/2} \tag{2.5.14}$$

с некоторой положительной, не зависящей от h константой C, если выполнено одно из условий: либо на каждом из отрезков $[0, y_1], [y_1, y_2], \ldots, [y_m, 1]$ сетка является равномерной; либо сетка является квазиравномерной, т. е. при $h \to 0$ неравенство

$$|\Delta x_{k+1/2} - \Delta x_{k-1/2}| \le ch^2$$

с некоторой положительной константой C нарушается лишь ограниченное число раз. Перечень условий, при которых выполняется оценка (2.5.14), может быть продолжен. Однако мы ограничимся этими двумя, наиболее часто встречающимися на практике.

Заметим, что понижение гладкости любой из функций p, q и f на порядок приводит к оценке

$$\max(\|\xi^h\|_{F_h}, \|\eta^h\|_{F_h}, \|\theta^h\|_{F_h}) \le C_1h.$$

Рассмотренная нами разностная схема (2.5.11) применяется на практике весьма редко, так как явное интегрирование функций p, q и f может оказаться затруднительным. Поэтому вместо (2.5.11), как правило, используется ее упрощенный вариант:

$$(A^h\varphi^h)_k = \frac{1}{\Delta x_k}\left\{p_{k+1/2}\frac{\varphi^h_{k+1} - \varphi^h_k}{\Delta x_{k+1/2}} - p_{k-1/2}\frac{\varphi^h_k - \varphi^h_{k-1}}{\Delta x_{k-1/2}} - (q\Delta x)_k\varphi^h_k\right\},$$

$$(f^h)_k = \frac{1}{\Delta x_k}(f\Delta x)_k = f_k = \frac{f_{k+1/2}(x_k - x_{k-1/2}) + f_{k-1/2}(x_{k+1/2} - x_k)}{x_{k+1/2} - x_{k-1/2}},$$

$$k = 1, 2, \ldots, n-1.$$

Оказывается, что для этой простой схемы все сделанные выводы о величине погрешности аппроксимации полностью сохраняются, если, конечно, сохраняются соответствующие предположения о гладкости данных задачи.

Перейдем к исследованию сходимости разностного решения задачи (2.5.11), (2.5.12) к точному решению задачи (2.5.1) при сделанных предположениях о гладкости функций p, q и f. Для этого достаточно установить устойчивость схемы (2.5.11) и воспользоваться теоремой сходимости.

Приступая к доказательству устойчивости, оценим прежде всего при помощи неравенства Коши — Буняковского скалярное произведение (φ^h, f^h):

$$(\varphi^h, f^h) \le \|\varphi^h\|_{F_h} \|f^h\|_{F_h}, \tag{2.5.15}$$

где скалярное произведение понимается в следующем смысле:

$$(\chi, \psi) = \sum_{k=1}^{n-1} \Delta x_k \chi_k \psi_k, \quad \chi, \psi \in F_h.$$

Исследуем подробнее левую часть неравенства (2.5.15). Так как по предположению $q(x) \ge 0$ и $p(x) \ge p_0 > 0$, то

$$(\varphi^h, f^h) = (\varphi^h, A^h \varphi^h) =$$

$$= \sum_{k=1}^{n} \frac{(\varphi_k^h - \varphi_{k-1}^h)^2}{\int\limits_{x_{k-1}}^{x_k} \frac{dx}{p}} + \sum_{k=1}^{n-1} (\varphi_k^h)^2 \int\limits_{x_{k-1/2}}^{x_{k+1/2}} q \, dx \ge p_0 \sum_{k=1}^{n} \frac{(\varphi_k^h - \varphi_{k-1}^h)^2}{\Delta x_{k-1/2}} > 0. \tag{2.5.16}$$

Последнее неравенство вытекает из того факта, что вектор φ^h не нулевой, ибо φ^h есть решение неоднородной задачи (2.5.11) с невырожденной матрицей A^h.

Учитывая, что $\varphi_0^h = 0$, можно записать

$$\varphi_k^h = \sum_{j=1}^{k} (\varphi_j^h - \varphi_{j-1}^h) = \sum_{j=1}^{k} \frac{\varphi_j^h - \varphi_{j-1}^h}{\sqrt{\Delta x_{j-1/2}}} \sqrt{\Delta x_{j-1/2}},$$

откуда по неравенству Коши — Буняковского получим

$$(\varphi_k^h)^2 = \left(\sum_{j=1}^{k} \frac{\varphi_j^h - \varphi_{j-1}^h}{\sqrt{\Delta x_{j-1/2}}} \sqrt{\Delta x_{j-1/2}}\right)^2 \leq$$

$$\leq \left(\sum_{j=1}^{k} \frac{(\varphi_j^h - \varphi_{j-1}^h)^2}{\Delta x_{j-1/2}}\right) \left(\sum_{j=1}^{k} \Delta x_{j-1/2}\right) \leq \sum_{j=1}^{n} \frac{(\varphi_j^h - \varphi_{j-1}^h)^2}{\Delta x_{j-1/2}}.$$

Наконец,

$$\sum_{k=1}^{n-1} (\varphi_k^h)^2 \Delta x_k \leq \sum_{j=1}^{n} \frac{(\varphi_j^h - \varphi_{j-1}^h)^2}{\Delta x_{j-1/2}} \sum_{k=1}^{n-1} \Delta x_k < \sum_{j=1}^{n} \frac{(\varphi_j^h - \varphi_{j-1}^h)^2}{\Delta x_{j-1/2}}. \tag{2.5.17}$$

Из (2.5.15)—(2.5.17) получаем

$$\|\varphi^h\|_{F_h} \|f^h\|_{F_h} \geq (\varphi^h, f^h) \geq p_0 \sum_{j=1}^{n} \frac{(\varphi_j^h - \varphi_{j-1}^h)^2}{\Delta x_{j-1/2}} \geq p_0 \|\varphi^h\|_{F_h}^2,$$

так что

$$\|\varphi^h\|_{F_h} = \frac{1}{p_0} \|f^h\|_{F_h}.$$

Последнее неравенство означает по определению устойчивость разностного алгоритма. Следовательно, на основании теоремы сходимости (в норме (2.5.10)) получаем оценку

$$\|\varepsilon^h\|_{F_h} \leq K h^{3/2}, \quad \varepsilon^h = (\varphi)_h - \varphi^h,$$

где $K \geq 3C/p_0$ — некоторая положительная постоянная.

Покажем теперь, как привлекая некоторые сеточные аналоги теорем вложения можно улучшить эту оценку. Прежде всего отметим, что, как доказано выше (при $\varphi_0^h = \varphi_n^h = 0$),

$$(\varphi_k^h)^2 \leq \sum_{j=1}^{n} \frac{(\varphi_j^h - \varphi_{j-1}^h)^2}{\Delta x_{j-1/2}} = \sum_{j=1}^{n} \left(\frac{\varphi_j^h - \varphi_{j-1}^h}{\Delta x_{j-1/2}}\right)^2 \Delta x_{j-1/2}.$$

Следовательно, при $c_1 \leq \Delta x_{j-1/2}/\Delta x_j \leq c_2$, где $c_1,\ c_2$ — положительные постоянные, не зависящие от j, имеем

$$(\varphi_k^h)^2 \leq \sum_{j=1}^{n} \left(\frac{\varphi_j^h - \varphi_{j-1}^h}{\Delta x_{j-1/2}}\right)^2 \frac{\Delta x_{j-1/2}}{\Delta x_j} \Delta x_j \leq c_2 \sum_{j=1}^{n} \left(\frac{\varphi_j^h - \varphi_{j-1}^h}{\Delta x_{j-1/2}}\right)^2 \Delta x_j \equiv c_2 \|\varphi^h\|_{\mathring{W}_2^{1,h}}^2.$$

Откуда для сеточных функций получаем следующее соотношение (сеточный аналог теоремы вложения $\overset{\circ}{W}{}_2^1(0,1)$ в $C(0,1)$ в одномерном случае):

$$\|\varphi^h\|_{C_h} \equiv \max_{k=1,2,\ldots,n-1} |\varphi_k^h| \le c\|\varphi^h\|_{\overset{\circ}{W}{}_2^{1,h}}, \quad c = \text{const} < \infty.$$

Применим последнее неравенство для получения более точной оценки ошибки $\varepsilon^h = (\varphi)_h - \varphi^h$. Для этого запишем тождество

$$A^h\varepsilon^h = \xi^h + \eta^h + \theta^h$$

и умножим его скалярно на ε^h:

$$(A^h\varepsilon^h, \varepsilon^h) = (\xi^h + \eta^h + \theta^h, \varepsilon^h).$$

Тогда как (см. (2.5.16))

$$(A^h\varepsilon^h, \varepsilon^h) \ge p_0 \sum_{k=1}^{n} \frac{(\varphi_k^h - \varphi_{k-1}^h)^2}{\Delta x_{k-1/2}} \ge p_0 c_1 \|\varepsilon^h\|_{\overset{\circ}{W}{}_2^{1,h}},$$

$$|(\xi^h + \eta^h + \theta^h, \varepsilon^h)| = \left| \sum_{k=1}^{n-1} \Delta x_k (\xi_k^h + \eta_k^h + \theta_k^h)\varepsilon_k^h \right| \le$$

$$\le \|\varepsilon^h\|_{C_h} \sum_{k=1}^{n-1} \Delta x_k |\xi_k^h + \eta_k^h + \theta_k^h| \equiv \|\varepsilon^h\|_{C_h} \xi^h + \eta^h + \theta^h\|_{L_{1,h}},$$

то имеем

$$\|\varepsilon^h\|^2_{\overset{\circ}{W}{}_2^{1,h}} \le c\|\varepsilon^h\|_{C_h} \|\xi^h + \eta^h + \theta^h\|_{L_{1,h}}.$$

Привлекая теперь приведенную выше теорему вложения, получаем

$$\|\varepsilon^h\|^2_{\overset{\circ}{W}{}_2^{1,h}} \le c\|\xi^h + \eta^h + \theta^h\|_{L_{1,h}}.$$

Однако при сделанных ранее предположениях (о наличии необходимой гладкости решения и исходных данных и квазиравномерности сетки)

$$\|\xi^h + \eta^h + \theta^h\|_{L_{1,h}} \le 3Nmh^2 + cMh^2 \sum_{k=1}^{n-1} \Delta x_k,$$

где $c = \text{const} < \infty$. Следовательно, при достаточно малых h и конечном m приходим к искомой оценке

$$\|\varepsilon^h\|_{C_h} \le c\|\varepsilon^h\|_{\overset{\circ}{W}{}_2^{1,h}} \le O(h^2).$$

2.5.2. Вариационная форма интегрального тождества

В 2.5.1 был рассмотрен метод построения разностных уравнений на основе интегрального тождества. Покажем, что метод интегрального тождества можно рассматривать как один из вариационных методов. В самом деле, рассмотрим дифференциальное уравнение

$$-\frac{d}{dx}p(x)\frac{du}{dx} + q(x)u = f(x), \quad x \in (a,b), \tag{2.5.18}$$

с краевыми условиями

$$u(a) = u(b) = 0. \tag{2.5.19}$$

Предполагается, что

$$0 < p_0 \le p(x) \le p_1 < \infty, \quad p_0, p_1 = \text{const},$$

$$q(x) \ge 0,\ p(x), q(x) \in L_\infty(a,b),\ f(x) \in L_2(a,b).$$

Метод интегрального тождества применительно к решению задачи (2.5.18), (2.5.19) приводит к соотношению

$$\frac{u(x_k) - u(x_{k+1})}{\int\limits_{x_k}^{x_{k+1}} \frac{dx}{p(x)}} + \frac{u(x_k) - u(x_{k-1})}{\int\limits_{x_{k-1}}^{x_k} \frac{dx}{p(x)}} + \int\limits_{x_{k-1/2}}^{x_{k+1/2}} (qu - f)\, dx =$$

$$= -\frac{1}{\int\limits_{x_k}^{x_{k+1}} \frac{dx}{p(x)}} \int\limits_{x_k}^{x_{k+1}} \frac{dx}{p(x)} \int\limits_{x_{k+1/2}}^{x} (qu - f)\, d\xi +$$

$$+ \frac{1}{\int\limits_{x_{k-1}}^{x_k} \frac{dx}{p(x)}} \int\limits_{x_{k-1}}^{x_k} \frac{dx}{p(x)} \int\limits_{x_{k-1/2}}^{x} (qu - f)\, d\xi, \tag{2.5.20}$$

где $a = x_0 < x_{1/2} < x_1 < x_{3/2} < \ldots < x_{N-3/2} < x_{N-1} < x_{N-1/2} < x_N = b$ — некоторый набор точек. Используя теперь ту или иную аппроксимацию входящих в (2.5.20) интегралов, получаем соответствующую разностную схему.

Наша ближайшая цель — показать, что (2.5.20) можно записать в некоторой форме, которую будем называть «вариационной формой интегрального тождества» и которая будет близка к вариационным уравнения метода Галеркина, но по сравнению с последними она дает возможность использовать для построения приближенных решений разрывные базисные функции и позволяет доказать сходимость приближенных решений при достаточно общих предположениях о гладкости исходных данных.

Выполним ряд простых преобразований в (2.5.20). Пусть

$$\psi = qu - f, \quad \rho_k(x) = \int\limits_{x_k}^{x} \frac{dx'}{p(x')} \Bigg/ \int\limits_{x_k}^{x_{k+1}} \frac{dx'}{p(x')} \ .$$

Тогда

$$-\frac{1}{\int\limits_{x_k}^{x_{k+1}} \frac{dx}{p(x)}} \int\limits_{x_k}^{x_{k+1}} \frac{dx}{p(x)} \int\limits_{x_{k+1/2}}^{x} \psi(\xi)d\xi = -\frac{1}{\int\limits_{x_k}^{x_{k+1}} \frac{dx}{p(x)}} \int\limits_{x_k}^{x_{k+1}} d\left(\int\limits_{x_k}^{x} \frac{dx'}{p(x')}\right) \int\limits_{x_{k+1/2}}^{x} \psi(\xi)dx =$$

$$= -\int\limits_{x_{k+1/2}}^{x_{k+1}} \psi(\xi)d\xi + \frac{1}{\int\limits_{x_k}^{x_{k+1}} \frac{dx}{p(x)}} \int\limits_{x_k}^{x_{k+1}} \psi(\xi) \int\limits_{x_k}^{x} \frac{d\xi}{p(\xi)}\, dx =$$

$$= -\int\limits_{x_{k+1/2}}^{x_{k+1}} \psi(\xi)d\xi + \int\limits_{x_k}^{x_{k+1}} \rho_k(x)\psi(x)\, dx.$$

Аналогичным образом получаем, что

$$\frac{1}{\int\limits_{x_{k-1}}^{x_k} \frac{dx}{p(x)}} \int\limits_{x_{k-1}}^{x_k} \frac{dx}{p(x)} \int\limits_{x_{k-1/2}}^{x} \psi(\xi)d\xi = -\int\limits_{x_{k-1}}^{x_{k-1/2}} \psi(\xi)d\xi + \int\limits_{x_{k-1}}^{x_k} \widetilde{\rho}_k(x)\psi(x)\, dx,$$

где

$$\widetilde{\rho}_k(x) = \int\limits_{x}^{x_k} \frac{dx'}{p(x')} \Bigg/ \int\limits_{x_{k-1}}^{x_k} \frac{dx}{p(x)} \, .$$

С учетом этих преобразований тождество (2.5.20) примет вид

$$\frac{u(x_k) - u(x_{k+1})}{\int\limits_{x_k}^{x_{k+1}} \frac{dx}{p(x)}} + \frac{u(x_k) - u(x_{k-1})}{\int\limits_{x_{k-1}}^{x_k} \frac{dx}{p(x)}} +$$

$$+ \int\limits_{x_{k-1}}^{x_k} (1 - \widetilde{\rho}_k(x))\psi(x)\, dx + \int\limits_{x_k}^{x_{k+1}} (1 - \rho_k(x))\psi(x)\, dx = 0.$$

Пусть

$$Q_k(x) = \begin{cases} 1 - \int\limits_{x}^{x_k} \frac{d\xi}{p(\xi)} \Bigg/ \int\limits_{x_{k-1}}^{x_k} \frac{d\xi}{p(\xi)} \, , & x \in (x_{k-1}, x_k), \\ 1 - \int\limits_{x_k}^{x} \frac{d\xi}{p(\xi)} \Bigg/ \int\limits_{x_k}^{x_{k+1}} \frac{d\xi}{p(\xi)} \, , & x \in (x_k, x_{k+1}), \\ 0, & x \notin (x_{k-1}, x_{k+1}); \end{cases} \tag{2.5.21}$$

тогда тождество (2.5.20) запишется в виде

$$\frac{u(x_k) - u(x_{k+1})}{\int\limits_{x_k}^{x_{k+1}} \frac{dx}{p(x)}} + \frac{u(x_k) - u(x_{k-1})}{\int\limits_{x_{k-1}}^{x_k} \frac{dx}{p(x)}} + (qu, Q_k) = (f, Q_k), \tag{2.5.22}$$

$$k = 1, 2, \ldots, N-1,$$

где $(\varphi, \psi) = \int\limits_a^b \varphi\psi \, dx$, $\|\varphi\| = (\varphi, \varphi)^{1/2}$. Замечая, что

$$\left(p\frac{du}{dx}, \frac{dQ_k}{dx}\right) = \frac{u(x_k) - u(x_{k+1})}{\int\limits_{x_k}^{x_{k+1}} \frac{dx}{p(x)}} + \frac{u(x_k) - u(x_{k-1})}{\int\limits_{x_{k-1}}^{x_k} \frac{dx}{p(x)}},$$

перепишем соотношение (2.5.22) следующим образом:

$$\left(p\frac{du}{dx}, \frac{dQ_k}{dx}\right) + (qu, Q_k) = (f, Q_k), \quad k = 1, 2, \ldots, N-1. \tag{2.5.23}$$

Если теперь ввести в рассмотрение некоторый интерполянт $u_1(x)$ функции $u(x)$, для которого $u_1(x_i) = u(x_i),\ i = 0, 1, \ldots, N,$ и имеет смысл производная du_1/dx, то в силу равенства

$$\left(p\frac{du}{dx}, \frac{dQ_k}{dx}\right) = \left(p\frac{du_1}{dx}, \frac{dQ_k}{dx}\right)$$

выражение (2.5.23) эквивалентно следующему:

$$\left(p\frac{du_1}{dx}, \frac{dQ_k}{dx}\right) + (qu, Q_k) = (f, Q_k), \quad k = 1, 2, \ldots, N-1. \tag{2.5.24}$$

Итак, показано, что тождество (2.5.20) эквивалентно каждому из соотношений (2.5.22)—(2.5.24). Это позволяет рассматривать метод интегрального тождества как один из вариационных методов, а соотношения (2.5.22)—(2.5.24) использовать совместно с (2.5.20) для построения приближенных решений задачи при достаточно широком выборе базисных функций. Соотношение (2.5.23) есть не что иное, как известное вариационное равенство, которое в методе Галеркина применяется для приближенного решения на основе использования $Q_k(x)$ в качестве базисных функций. Из соотношения (2.5.22) видно, что в данном случае возможно использование разрывных базисных функций φ_i (с разрывами, не совпадающими с $x_i\ (i = 0, 1, \ldots, N)$). Применение тождества (2.5.24) позволяет достаточно просто получить ряд оценок в равномерной метрике.

Воспользуемся рассмотренными тождествами (2.5.22)—(2.5.24) для построения приближенных решений задачи (2.5.18), (2.5.19), выбирая базисные функции различными способами. Прежде всего заметим, что при замене переменных

$$y = \int_a^x \frac{d\xi}{p(\xi)}$$

задача (2.5.18), (2.5.19) принимает вид (при $\widetilde{p}(y) \equiv p(x(y))$, $\widetilde{q}(y) \equiv q(x(y))$, $\widetilde{f}(y) \equiv \widetilde{p}(y) f(x(y))$)

$$-\frac{d^2u}{dy^2} + \widetilde{p}(y)\widetilde{q}(y)u = \widetilde{f}(y), \quad 0 < y < T = \int\limits_a^b \frac{d\xi}{p(\xi)},$$

$$u(0) = u(T) = 0.$$

Следовательно, при $f(x) \in L_2(a,b)$ имеем $u(x) \in W_2^2(0,T)$. С учетом этого замечания приступим к построению приближенного решения. Будем искать его в виде

$$u^h(x) = \sum_{i=1}^{N-1} a_i Q_i(x),$$

где a_i определяются из однозначно разрешимой системы

$$\left(p\frac{du^h}{dx}, \frac{dQ_k}{dx}\right) + (qu^h, Q_k) = (f, Q_k), \quad k = 1, 2, \ldots, N-1,$$

то есть приближенное решение строится при помощи обычного метода Галеркина при базисных функциях $\{Q_i(x)\}$.

Оценим скорость сходимости $u^h(x)$ к точному решению при $h = \max\limits_i |x_i - x_{i-1}| \to 0$. Для этого рассмотрим сначала некоторые вопросы аппроксимации при помощи функций $Q_i(x)$, а также при помощи «функций-домиков»

$$\varphi_i(y) = \begin{cases} \dfrac{y - y_{i-1}}{y_i - y_{i-1}}, & y \in [y_{i-1}, y_i], \\ \dfrac{y_{i+1} - y}{y_{i+1} - y_i}, & y \in [y_i, y_{i+1}], \\ 0, & y \notin [y_{i-1}, y_{i+1}], \end{cases}$$

в которые переходят $Q_i(x)$ при замене переменной $y = \int\limits_\alpha^x d\xi/p(\xi)$.

Пусть на отрезке $[0,T]$ введена сетка $y_0 = 0 < y_1 < \ldots < y_{N-1} < y_N = T$, $H_i = y_i - y_{i-1}$, $H = \max\limits_i H_i$ и задана система «функций-домиков» $\{\varphi_i(y)\}_{i=1}^{N-1}$. Предположим, что для некоторой функции $u(y) \in W_2^2(0,T) \cap$

$\cap \overset{\circ}{W}{}_2^1(0,T)$ аппроксимирующая ее функция выбрана в виде

$$u_1(y) = \sum_{i=1}^{N-1} u(y_i)\varphi_i(y).$$

В 2.2.2 показано, что

$$\|u - u_I\|_{W_2^k(0,T)} \le CH^{2-k} \left\| \frac{d^2u}{dy^2} \right\|_{L_2(0,T)}, \quad k = 0, 1, \ldots.$$

Если, кроме того, $d^2u/dy^2 \in L_\infty(0,T)$, то имеем также

$$\|u - u_1\|_{C(0,T)} \le CH^2 \left\| \frac{d^2u}{dy^2} \right\|_{L_\infty(0,T)}.$$

Обратимся теперь к аппроксимации решения задачи (2.5.18), (2.5.19) при помощи функций $Q_i(x)$. Пусть

$$u_I(x) = \sum_{i=1}^{N-1} u(x_i)Q_i(x).$$

Тогда, пользуясь заменой $y = \int\limits_0^x \frac{d\xi}{p(\xi)}$ и полагая $y_i = \int\limits_a^{x_i} \frac{d\xi}{p(\xi)}$, $H_i = y_i - y_{i-1}$, $T = \int\limits_a^b \frac{d\xi}{p(\xi)}$, имеем

$$\int_a^b \frac{d\xi}{p(\xi)} \left| u(x) - \sum_{i=1}^{N-1} u(x_i)Q_i(x) \right|^2 \le \operatorname*{supvrai}_x p(x) \int_a^b \frac{dx}{p(x)} |u(x) -$$

$$- \sum_{i=1}^{N-1} u(x_i)Q_i(x)|^2 \le \int_0^T \left| u(y) - \sum_{i=1}^{N-1} u(y_i)\varphi_i(y) \right|^2 dy \le$$

$$\le cH_i^4 \int_0^T \left| \frac{d^2u}{dy^2} \right|^2 \le ch^4 \int_0^T \left| \frac{d^2u}{dy^2} \right|^2 dy.$$

Поскольку для рассматриваемой задачи

$$\int_0^T \left| \frac{d^2u}{dy^2} \right|^2 dy \le c \int_a^b \left| \frac{d}{dx} p(x) \frac{du}{dx} \right|^2 dy \le c \int_a^b |f(x)|^2,$$

$$\left\|\frac{d^2u}{dy^2}\right\|_{L_\infty(0,T)} \le c\left\|\frac{d}{dx}p(x)\frac{du}{dx}\right\|_{L_\infty(a,b)} \le c\|f\|_{L_\infty(a,b)},$$

то получаем следующую оценку аппроксимации в $L_2(a,b)$:

$$\|u-u_1\|_{L_2(a,b)} \le ch^2\|f\|_{L_2(a,b)}.$$

Аналогично приходим к оценке для $d(u-u_1)/dx$:

$$\int_a^b p\left|\frac{d(u-u_1)}{dx}\right|^2 dx = \int_a^b \left|\frac{d(u-u_1)}{dx/p(x)}\right|^2 \frac{dx}{p(x)} =$$

$$= \int_0^T \left|\frac{d(u-u_1)(y)}{dy}\right|^2 dy \le ch^2 \int_0^T \left|\frac{d^2u}{dy^2}\right|^2 dy \le ch^2\|f\|^2_{L_2(a,b)}.$$

Если в (2.5.18) предположить, что $f(x) \in L_\infty(a,b)$, то будем иметь $f(y) \in L_\infty(0,T)$ и $\dfrac{d^2u}{dy^2} \in L_\infty(0,T)$. И как следствие этого факта и полученных уже результатов получаем также оценку вида

$$\|u-u_1\|_{C(a,b)} \le c\max_i H_i^2 \left|\frac{d^2u}{dy^2}\right|_{L_\infty(0,T)} \le ch^2\|f\|_{L_\infty(a,b)}.$$

Теперь, зная порядок аппроксимации точного решения, можно оценить скорость сходимости приближенного решения $u^h(x) = \sum\limits_{i=1}^{N-1} a_iQ_i(x)$ к точному при $h \to 0$. Оценку в энергетической метрике или в метрике $W_2^1(a,b)$ (они здесь эквивалентны) получить просто. Для этого запишем тождество

$$\|u-u^h\|_L^2 \equiv \left(p\frac{d(u-u^h)}{dx}, \frac{d(u-u^h)}{dx}\right) + (q(u-u^h), u-u^h) =$$

$$= \left(p\frac{d(u-u^h)}{dx}, \frac{d(u-u_1)}{dx}\right) + (q(u-u^h), u-u_1),$$

где $u_1(x) = \sum\limits_{i=1}^{N-1} u(x_i)Q_i(x)$, и выполним простые оценки, применяя результаты об аппроксимации

$$\|u-u^h\|_L^2 \le \|u-u^h\|_L\|u-u_1\|_L;$$

$$\|u-u^h\|_L^2 \le \|u-u_1\|_L^2 \le ch^2 \int\limits_0^T \left|\frac{d^2u}{dy^2}\right|^2 dy \le ch^2\|f\|_{L_2(a,b)}^2.$$

Итак, доказано, что

$$\|u-u^h\|_L \le ch\|f\|_{L_2(a,b)},$$

$$\|u-u^h\|_{W_2^1(a,b)} \le ch\|f\|_{L_2(a,b)}$$

(так как $\|u-u^h\|_{W_2^1(a,b)} \le c\|u-u^h\|_L$).

Оценим скорость сходимости $u^h(x)$ к $u(x)$ в равномерной метрике. Воспользуемся для этого тем обстоятельством, что для точного решения справедливо соотношение (2.5.24), где

$$u_1 = \sum_{i=1}^{N-1} u(x_i)Q_i(x).$$

Тогда имеем

$$\left(p\frac{d(u_1-u^h)}{dx}, \frac{d(u_1-u^h)}{dx}\right) + (q(u-u^h), u_1-u^h) = 0,$$

$$\left(p\frac{d(u_1-u^h)}{dx}, \frac{d(u_1-u^h)}{dx}\right) + (q(u-u^h), u-u^h) =$$

$$= (q(u-u^h), u-u_1) \le (q(u-u^h), u-u^h)^{1/2}(q(u-u_1), u-u_1)^{1/2}.$$

Из последнего соотношения с учетом того, что $\|u-u_1\| \le ch^2$, следует оценка

$$\left(p\frac{d(u_1-u^h)}{dx}, \frac{d(u_1-u^h)}{dx}\right) + (q(u-u^h), u-u^h) \le ch^4.$$

А так как

$$\|u_1-u^h\|_{C(a,b)} \le C\left(p\frac{d(u_1-u^h)}{dx}, \frac{d(u_1-u^h)}{dx}\right)^{1/2},$$

то имеем также

$$\|u_1-u^h\|_{C(a,b)} \le ch^2.$$

Однако уже доказано, что если $f(x) \in L_\infty(a,b)$, то $\|u-u_1\|_{C(a,b)} \le ch^2\|f\|_{L_\infty(a,b)}$. Поэтому, применяя неравенство треугольника, получаем

оценку в равномерной метрике

$$\|u - u^h\|_{C(a,b)} \le ch^2.$$

Рассмотрим теперь еще два случая выбора базисных функций $\{\varphi_i^h(x)\}$. Пусть $h_i = x_{i+1/2} - x_{i-1/2}$, $h = \max\limits_i h_i$. Обозначим через $\varphi_i^h(x)$ характеристическую функцию интервала $(x_{i-1/2}, x_{i+1/2})$, через $\varphi_0^h(x)$ — интервала $(x_0, x_{1/2})$ и через $\varphi_N^h(x)$ — интервала $(x_{N-1/2}, x_N)$. Примем $\{\varphi_i^h\}_{i=0}^N$ в качестве базисных функций и будем искать приближенное решение в виде $u^h(x) = \sum\limits_{i=0}^{N} a_i\varphi_i^h(x)$, где $a_0 = a_N = 0$. Тогда $u^h(x) = \sum\limits_{i=1}^{N-1} a_i\varphi_i^h(x)$, где $\{a_i\}_{i=1}^{N-1}$ определяются из соотношения (см. (2.5.22))

$$\frac{u^h(x_k) - u^h(x_{k+1})}{\int\limits_{x_k}^{x_{k+1}} \frac{dx}{p(x)}} + \frac{u^h(x_k) - u^h(x_{k-1})}{\int\limits_{x_{k-1}}^{x_k} \frac{dx}{p(x)}} + (qu^k, Q_k) = (f, Q_k), \tag{2.5.25}$$

$$k = 1, 2, \ldots, N-1.$$

Для решения системы (2.5.25) справедлива априорная оценка вида (где $u_1^h = \sum\limits_{i=1}^{N-1} a_i Q_i(x)$)[11)]

$$\left(p\frac{du_1^h}{dx}, \frac{du_1^h}{dx}\right) + (qu^h, u^h) \le \frac{\text{const}\,\|f\|^2}{1 - O(h^2)} \tag{2.5.26}$$

(из (2.5.26) следует, что при достаточно малых h система (2.5.25) имеет единственное решение $\{a_i\}_{i=1}^{N-1}$), а также можно получить следующие оценки:

$$\left[\left(p\frac{d(u_1 - u_1^h)}{dx}, \frac{d(u_1 - u_1^h)}{dx}\right) + (q(u - u^h), u - u^h)\right]^{1/2} \le$$

$$\le \frac{O(h)}{1 - O(h)} \le ch, \tag{2.5.27}$$

$$\max_i |u(x_i) - u^h(x_i)| + (q(u - u^h), u - u^h)^{1/2} \le \frac{O(h)}{1 - Q(h)} \le ch$$

[11)] Доказательства приводимых в заключение данного раздела оценок осуществляются в целом аналогично проведенным выше и их можно найти в кн.: Г. И. Марчук Методы вычислительной математики. — М.: Наука, 1980.

при достаточно малых h. Заметим, что в приведенных выше рассуждениях нам достаточно предполагать, что $p(x),\ q(x),\ f(x) \in L_\infty(a,b)$. Эти ограничения в ряде случаев, вообще говоря, можно ослабить. Например, если потребовать от $f(x)$ лишь принадлежности к пространству $L_1(a,b)$.

Приближенное решение можно строить также при помощи координатных «функций-домиков» $\{\varphi_i^h(x)\}_{i=1}^{N-1}$, которые кусочно-линейны на $[a,b]$, причем $\varphi_i^h(x)$ равна нулю вне интервала (x_{i-1}, x_{i+1}) и $\varphi_i^h(x_i) = 1$. Решение ищется в виде $u^h(x) = \sum\limits_{i=1}^{N-1} a_i\varphi_i(x)$, где неизвестные a_i определяются из системы линейных алгебраических уравнений

$$\left(p\frac{du_1^h}{dx}, \frac{dQ_i}{dx}\right) + (qu^h, Q_i) = (f, Q_i), \quad i = 1, 2, \ldots, N-1, \tag{2.5.28}$$

где $u_1^h = \sum\limits_{i=1}^{N-1} a_i Q_i(x)$.

Нетрудно показать, что если $q(x) \in L_\infty(a,b)$, $f(x) \in L_2(a,b)$, $p(x)$ — функция класса $C^{(1')}$ с возможными разрывами первого рода в конечном числе узлов x_j сетки (не меняющемся при изменении самой сетки), то при достаточно малых h система (2.5.28) имеет единственное решение, при $h \to 0$ функция u^h сходится к $u(x)$ и справедливы оценки

$$\left[\left(p\frac{d(u_1 - u_1^h)}{dx}, \frac{d(u_1 - u_1^h)}{dx}\right) + (q(u - u^h), u - u^h)\right]^{1/2} \leq$$

$$\leq \frac{O(h^2)\|f\|}{1 - O(h^2)} \leq ch^2,$$

$$\max_i |u(x_i) - u^h(x_i)| + (q(u - u^h), u - u^h)^{1/2} \leq \frac{O(h^2)\|f\|}{1 - O(h^2)} \leq ch^2.$$

Также можно показать, что если $f(x) \in L_\infty(a,b)$, то

$$\|u - u^h\|_{C(a,b)} \leq \frac{O(h^2)}{1 - O(h^2)}\|f\|_{L_\infty(a,b)} \leq ch^2$$

при $h \to 0$.

Рассмотренный здесь подход к построению приближенных решений, основанный на вариационной форме интегрального тождества, можно применять к решению ряда других задач математической физики.

2.6. Построение схем для нестационарных задач проекционно-сеточным методом

При решении нестационарных задач проекционно-сеточный метод часто применяется для аппроксимации решения лишь по пространственным переменным, а приближение по t осуществляется обычным разностным методом (хотя не исключаются и другие способы приближения по временной переменной).

В настоящем разделе на примере параболического уравнения с одной пространственной переменной рассматривается одна из возможных форм применения проекционно-сеточных методов для решения нестационарных задач. Рассмотрим задачу

$$\frac{du}{dt} + Au = f, \tag{2.6.1}$$

$$u(x,0) = u_{(0)}, \tag{2.6.2}$$

где $f = f(x,t)$ при каждом t принадлежит $F = L_2(D)$, $x \in D = (a,b)$, $t \in (0,T)$, $u_{(0)} = u_{(0)}(x) \in L_2(D)$. Оператор A имеет вид

$$Au = -\frac{d}{dx}p(x)\frac{du}{dx} + q(x)u, \tag{2.6.3}$$

где $p(x),\ q(x)$ — ограниченные положительные функции на D с областью определения

$$D(A) = \left\{ v :\ v \in L_2,\ \frac{dv}{dx} \in L_2,\ Av \in L_2,\ v(a) = \frac{dv}{dx}(b) = 0 \right\}. \tag{2.6.4}$$

Через F_A будем обозначать энергетическое пространство, соответствующее A. Скалярное произведение и норма в F_A соответственно имеют вид

$$(u,v)_A = \int_a^b \left(p\frac{du}{dx}\frac{dv}{dx} + quv \right)\ dx,$$

$$\|u\|_A = (u,u)^{1/2} = \left(\int_a^b \left(p\left|\frac{du}{dx}\right|^2 + qu^2 \right)\ dx \right)^{1/2}.$$

Функции из F_A удовлетворяют условию $u(a) = 0$ (главное условие); в то же время среди них могут оказаться функции, не удовлетворяющие второму условию $\dfrac{du}{dx}(b) = 0$ (естественное условие).

Сформулируем обобщенную постановку задачи (2.6.1), (2.6.2). Назовем *обобщенным решением* задачи (2.6.1), (2.6.2) функцию $u(x,t)$, которая почти при каждом $t \in (0,T)$ принадлежит F_A, обладает производной $\partial u/\partial t \in L_2((0,T) \times D)$ и удовлетворяет почти всюду на $(0,T)$ соотношениям

$$(\partial u/\partial t, w)(t) + [u,w](t) = (f,w)(t), \tag{2.6.5}$$

$$(u(x,0), w) = (u_{(0)}, w) \tag{2.6.6}$$

при произвольных $w(x) \in F_A$.

При такой обобщенной постановке переменную t можно рассматривать как параметр, и для аппроксимации задачи по этой переменной уже возможно применение разностных методов.

Если, помимо введенных выше ограничений, предположить, что $u_{(0)} \in F_A$, то можно доказать существование единственного обобщенного решения $u \in W_2^1((0,T) \times D)$. Если же, кроме того, потребовать, чтобы $dp/dx \in L_\infty(a,b)$, то обобщенное решение будет обладать конечной нормой

$$|||u||| = \left[\int_0^T \left(\left\| \frac{\partial u}{\partial t} \right\|^2 + \left\| \frac{\partial^2 u}{\partial x^2} \right\|^2 + \left\| \frac{\partial u}{\partial x} \right\|^2 + \|u\|^2 \right) dt \right]^{1/2}. \tag{2.6.7}$$

Все эти ограничения в дальнейшем будем считать выполненными.

Для аппроксимации решения по x выберем кусочно-линейные функции $\{\varphi_i(x)\}$, построенные на сетке $a = x_0 < x_1 < \ldots < x_{N-1} < x_N = b$, $h_i = x_i - x_{i-1}$, $h = \max\limits_i h_i$, $h \le c \min\limits_i h_i$ по алгоритму, изложенному в 2.2.2.

Множество линейных комбинаций вида $v_h = \sum\limits_{i=1}^{N} a_i \varphi_i(x)$ образует подпространство $F^h \subset F_A$. Отметим, что $v \in F^h$ удовлетворяет главному краевому условию $v_h(a) = 0$.

Приближенное решение будем искать в виде

$$u_h(x,t) = \sum_{i=1}^{N} a_i(t) \varphi_i(x), \tag{2.6.8}$$

где коэффициенты являются уже функциями от $t \in (0,T)$ и определяются из системы обыкновенных дифференциальных уравнений

$$\left(\frac{\partial u_h}{\partial t}, \varphi_i\right)(t) + (u_h, \varphi_i)_A(t) = (f, \varphi_i)(t) \tag{2.6.9}$$

при начальных условиях

$$(u_h(x,0) - u_{(0)}, \varphi_i) = 0, \quad i = 1, 2, \ldots, N, \tag{2.6.10}$$

которые получены из (2.6.5), (2.6.6) по методу Галеркина. Задачу (2.6.9), (2.6.10) можно записать также в виде

$$\hat{B}\frac{da}{dt} + \hat{A}a = F(t), \tag{2.6.11}$$

$$\hat{B}a(0) = a_{(0)}, \tag{2.6.12}$$

где $a(t) = (a_1(t), \ldots, a_N(t))^T$, $F(t) = (F_1(t), \ldots, F_N(t))^T$, $F_i(t) = (f, \varphi_i)(t)$, $a_0 = (a_{(0),1}, \ldots, a_{(0),N})^T$, $a_{(0),i} = (u_{(0)}, \varphi_i)$, $\hat{B} = (B_{ij})$, $\hat{A} = (A_{ij})$, $B_{ij} = B_{ji} = (\varphi_i, \varphi_j)$, $A_{ij} = A_{ji} = (\varphi_i, \varphi_j)_A$. А поскольку $\hat{B}$ есть матрица Грама, соответствующая $\{\varphi_i\}$, то она невырожденная. Матрица $\hat{A}$ является положительно определенной. Поэтому согласно теории обыкновенных дифференциальных уравнений задача (2.6.11), (2.6.12) имеет единственное решение, а приближенное решение (2.6.8) существует и определено однозначно.

Получим одну из оценок погрешности $u - u_h$ при $u_{(0)} \equiv 0$. Запишем равенство

$$\left(\frac{\partial(u-u_h)}{\partial t}, u - u_h\right)(t') + \|u - u_h\|_A^2(t') =$$

$$= \left(\frac{\partial(u-u_h)}{\partial t}, u - v_h\right)(t') + (u - u_h, u - v_h)_A(t'),$$

где $t' \in [0,T]$, $v_h = \sum\limits_{i=1}^{N} b_i(t')\varphi_i(x)$, $b_i(t')$ — произвольная дифференцируемая функция, причем $b_i(0) = 0$. Проинтегрируем равенство по $t' \in (0,t]$, $t \in (0,T]$. Тогда, выполняя интегрирование по частям по t и применяя простые неравенства, получаем

$$\|u - u_h\|^2(t) + \int\limits_0^t \|u - u_h\|_A^2 dt' \leq$$

$$\leq c \left(\|u - v_h\|^2(t) + \int\limits_0^t \left\| \frac{\partial (u - v_h)}{\partial t'} \right\|^2 \, dt' + \int\limits_0^t \|u - v_h\|_A^2 \, dt' \right).$$

Следовательно,

$$\max_t \|u - u_h\|(t) + \left(\int\limits_0^T \|u - u_h\|_A^2 \, dt \right)^{1/2} \leq$$

$$\leq c \left(\max_t \|u - v_h\|^2(t) + \int\limits_0^T \left\| \frac{\partial (u - v_h)}{\partial t} \right\|^2 \, dt + \int\limits_0^T \|u - v_h\|_A^2 \, dt \right)^{1/2}.$$

Тогда, в силу произвольности v_h и аппроксимирующих свойств кусочно-линейных функций, при $u,\ \partial u/(\partial t \partial x),\ \partial^2 u/\partial x^2 \in L_2((0,T) \times D)$ справедлива оценка вида

$$\max_t \|u - u_h\|(t) + \left(\int\limits_0^T \|u - u_h\|_A^2 \, dt \right)^{1/2} \leq ch. \tag{2.6.13}$$

В заключение отметим, что задачу (2.6.11), (2.6.12) можно свести к задаче вида

$$\frac{d\psi}{dt} + \widetilde{A}\psi = \widetilde{f}, \quad t \in (0,T), \tag{2.6.14}$$

$$\psi(0) = g, \tag{2.6.15}$$

где исходные данные и решение ψ связаны с данными и решением a задачи (2.6.11), (2.6.12) с помощью следующих соотношений:

$$\widetilde{A} = \hat{B}^{-1/2} \hat{A} \hat{B}^{-1/2}, \quad \widetilde{f} = \hat{B}^{-1/2} F,$$

$$g = \hat{B}^{-1/2} a_{(0)}, \quad \psi = \hat{B}^{1/2} a.$$

Поэтому если для (2.6.14), (2.6.15) записать какую-либо разностную схему по t, то сразу же получим соответствующую схему для (2.6.11), (2.6.12). Так, например, если для численного решения (2.6.14), (2.6.15) применяется схема Кранка — Николсона

$$\frac{\psi^j - \psi^{j-1}}{\tau} + \widetilde{A} \frac{\psi^j + \psi^{j-1}}{2} = \widetilde{f}(t_{j-1/2}), \tag{2.6.16}$$

$$\psi^0 = g, \quad j = 1, 2, \ldots, J, \tag{2.6.17}$$

где $\tau = T/J$, $t_{j-1/2} = (t_j + t_{j-1/2})/2$, то для (2.6.11), (2.6.12) эта схема переписывается в следующей форме:

$$\hat{B}\frac{a^j - a^{j-1}}{\tau} + \hat{A}\frac{a^j + a^{j-1}}{2} = F(t_{j-1/2}), \tag{2.6.18}$$

$$\hat{B}a^0 = a_0, \quad j = 1, 2, \ldots, J. \tag{2.6.19}$$

Таким образом, проблема построения схем для (2.6.11), (2.6.12) сводится к аналогичной проблеме для (2.6.14), (2.6.15). Построение и исследование схем для численного решения задачи (2.6.14), (2.6.15) подробно изучаются в главе 5 настоящей книги.

Глава 3.

Интерполяция сеточных функций

Проблема интерполяции величин, заданных на дискретном множестве точек, на всю область определения функции непрерывного аргумента тесно связана с построением вариационно-разностных схем и непрерывного представления решений разностных задач. В самом деле, при построении разностных уравнений, как правило, осуществляется процесс дискретизации оператора и решения задачи с помощью подходящих методов проектирования. При этом решение разностной задачи обычно представляет собой приближенное решение исходной задачи на дискретном множестве точек. Предположим, что разностная задача решена и мы располагаем информацией о приближенном решении этой задачи. Дальнейшее связано с интерполяцией полученных данных на всю область определения решения исходной задачи. Естественно, что при такой интерполяции должны быть соблюдены некоторые условия, а именно: если решение разностного уравнения получено с определенной степенью точности, то порядок интерполяции данных должен согласовываться с порядком аппроксимации разностного уравнения и быть не ниже последнего. Если мы располагаем дополнительной информацией о погрешностях приближенного решения, то интерполирование приближенного решения можно осуществить не по точным данным, а с учетом возможных погрешностях в узлах. Тогда априорная информация о гладкости решения в некоторых случаях позволит даже уточнить приближенное решение задачи, полученное с помощью тех или иных

разностных методов. Конечно, проблема интерполяции данных имеет и самостоятельное значение.

Алгоритмы интерполяции функций по точным данным, определенным на дискретном множестве точек, как правило, основаны на использовании интерполяционных многочленов Лагранжа. При этом относительно интерполируемой функции $\varphi(x)$ вводится априорное предположение о том, что она обладает производными до некоторого порядка.

Другая, близкая к проблеме интерполяции, задача возникает в том случае, когда значения заданной функции $\varphi(x)$ известны в узловых точках x_k не точно, а с некоторой погрешностью, максимальная величина которой для каждой точки задается в качестве априорной информации. В этом случае задача состоит в построении такой кривой, которая бы в известном смысле наилучшим образом аппроксимировала функцию, заданную со случайными погрешностями в узловых точках. Такая задача обычно решается на основе метода наименьших квадратов.

В последние годы теория интерполяции обогатилась новыми методами, получившими название *сплайновых интерполяций*. Следует отметить, что *сплайном* обычно оказывается определенная в области D кусочно-полиномиальная функция, т. е. функция, для которой существует такое разбиение D на подобласти, что внутри каждого элемента разбиения функция представляет собой многочлен некоторой степени m. Кроме того, эта функция, как правило, непрерывна в области D вместе с производными до $(m-1)$-го порядка и имеет интегрируемую с квадратом производную порядка m. Наиболее употребительными в технике явились сплайны $-$ многочлены третьей степени.

3.1. Интерполяция функций одного переменного

3.1.1. Интерполяция функций одного переменного с помощью кубических сплайнов

Пусть на отрезке $[a, b]$ вещественной оси x задана сетка $a = x_0 < x_1 < \ldots < x_n = b$, в узлах которой заданы значения $\{f_k\}_{k=0}^N$ функции $f(x)$, определенной на $[a, b]$. Тогда задача кусочно-кубической интерполяции ставится следующим образом. На отрезке $[a, b]$ необходимо найти функцию $g(x)$, удовлетворяющую требованиям:

1) $g(x)$ принадлежит классу $C^{(2)}(a, b)$, т. е. непрерывна вместе со своими производными до второго порядка включительно;

2) на каждом из отрезков $[x_{k-1}, x_k]$ $g(x)$ является кубическим многочленом вида

$$g(x) \equiv g_k(x) = \sum_{l=0}^{3} a_l^{(k)} (x_k - x)^l, \quad k = 1, 2, \ldots, n; \tag{3.1.1}$$

3) в узлах сетки $\{x_k\}_{k=0}^n$ выполняются равенства

$$g(x_k) = f_k, \quad k = 0, 1, \ldots, n; \tag{3.1.2}$$

4) $g''(x)$ удовлетворяет граничным условиям

$$g''(a) = g''(b) = 0. \tag{3.1.3}$$

Достоинства выбранной нами интерполяции выяснятся в дальнейшем, когда мы установим естественное экстремальное свойство такой определенной функции $g(x)$.

Покажем, что поставленная задача на нахождение интерполирующей кусочно-кубической функции $g(x)$ имеет единственное решение. Для этого воспользуемся сформулированными выше условиями 1) — 4).

Так как вторая производная функции $g(x)$ непрерывна и линейна на каждом отрезке сетки $[x_{i-1}, x_i]$ $(i = 1, 2, \ldots, n)$, мы можем записать при $x_{i-1} \le x \le x_i$

$$g''(x) = m_{i-1} \frac{x_i - x}{h_i} + m_i \frac{x - x_{i-1}}{h_i}, \tag{3.1.4}$$

где $h_i = x_i - x_{i-1}$, $m_k = g''(x_k)$. Проинтегрируем дважды обе части равенства (3.1.4). Получим

$$g(x) = m_{i-1}\frac{(x_i - x)^3}{6h_i} + m_i\frac{(x - x_{i-1})^3}{6h_i} + A_i\frac{x_i - x}{h_i} + B_i\frac{x - x_{i-1}}{h_i}, \tag{3.1.5}$$

где A_i и B_i — некоторые константы интегрирования. Они вычисляются из условия $g(x_{i-1}) = f_{i-1}$, $g(x_i) = f_i$. Подставляя $x = x_i$ и $x = x_{i-1}$ в (3.1.5), получим

$$m_i\frac{h_i^2}{6} + B_i = f_i,$$

$$m_{i-1}\frac{h_i^2}{6} + A_i = f_{i-1}.$$

Окончательно имеем

$$g(x) = m_{i-1}\frac{(x_i - x)^3}{6h_i} + m_i\frac{(x - x_{i-1})^3}{6h_i} +$$

$$+ \left(f_{i-1} - \frac{m_{i-1}h_i^2}{6}\right)\frac{x_i - x}{h_i} + \left(f_i - \frac{m_i h_i^2}{6}\right)\frac{x - x_{i-1}}{h_i}, \tag{3.1.6}$$

$$g'(x) = -m_{i-1}\frac{(x_i - x)^2}{2h_i} + m_i\frac{(x - x_{i-1})^2}{2h_i} + \frac{f_i - f_{i-1}}{h_i} - \frac{m_i - m_{i-1}}{6}h_i. \tag{3.1.7}$$

Из (3.1.7) находим односторонние пределы производной в точках $x_1, x_2, \ldots, x_{n-1}$:

$$g'(x_i - 0) = \frac{h_i}{6}m_{i-1} + \frac{h_i}{3}m_i + \frac{f_i - f_{i-1}}{h_i},$$

$$g'(x_i + 0) = -\frac{h_{i+1}}{3}m_i - \frac{h_{i+1}}{6}m_{i+1} + \frac{f_{i+1} - f_i}{h_{i+1}}.$$

Согласно условию 1) функции $g''(x)$ и $g'(x)$ непрерывны на $[a, b]$. Из условия непрерывности $g'(x)$ в точках $x_1, x_2, \ldots, x_{n-1}$ получаем $n - 1$ уравнение

$$\frac{h_i}{6}m_{i-1} + \frac{h_i + h_{i+1}}{3}m_i + \frac{h_{i+1}}{6}m_{i+1} = \frac{f_{i+1} - f_i}{h_{i+1}} - \frac{f_i - f_{i-1}}{h_i}. \tag{3.1.8}$$

Дополняя эти уравнения из условий (3.1.3) равенствами $m_0 = m_n = 0$, получаем линейную алгебраическую систему для нахождения неизвестных $m_1, m_2, \ldots, m_{n-1}$:

$$Am = Hf. \tag{3.1.9}$$

Квадратная матрица A имеет вид

$$A = \begin{Vmatrix} \frac{h_1+h_2}{3} & \frac{h_2}{6} & 0 & \dots & 0 & 0 \\ \frac{h_2}{6} & \frac{h_2+h_3}{3} & \frac{h_3}{6} & \dots & 0 & 0 \\ 0 & \frac{h_3}{6} & \frac{h_3+h_4}{3} & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \frac{h_{n-1}}{6} & \frac{h_n+h_{n-1}}{3} \end{Vmatrix}; \qquad (3.1.10)$$

векторы m и f и прямоугольная матрица H таковы:

$$m = \begin{Vmatrix} m_1 \\ m_2 \\ \dots \\ m_{n-1} \end{Vmatrix}, f = \begin{Vmatrix} f_1 \\ f_2 \\ \dots \\ f_n \end{Vmatrix},$$

$$H = \begin{Vmatrix} \frac{1}{h_1} & \left(-\frac{1}{h_1}-\frac{1}{h_2}\right) & \frac{1}{h_2} & \dots & 0 & 0 \\ 0 & \frac{1}{h_2} & \left(-\frac{1}{h_2}-\frac{1}{h_3}\right) & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \left(\frac{1}{h_{n-1}}-\frac{1}{h_n}\right) & \frac{1}{h_n} \end{Vmatrix}. \qquad (3.1.11)$$

Матрица A симметрична, со строгим диагональным преобладанием. По теореме Гершгорина о локализации собственных значений она положительно определена и, разумеется, неособенная. Значит, коэффициенты $m_1, m_2, \dots, m_{n-1}$ определяются из системы (3.1.9) однозначно. Следовательно, сплайн-функция $g(x)$ также однозначно восстанавливается по формулам (3.1.6) и задача о нахождении кусочно-кубической функции $g(x)$ имеет единственное решение.

Кубические сплайн-функции обладают очень важным свойством, которое обусловливает высокую эффективность сплайн-интер-

поляции. А именно: рассмотрим на отрезке $[a, b]$ класс $W_2^2[a, b]$, состоящий из функций, имеющих суммируемые с квадратом вторые производные. Поставим задачу отыскания интерполяционной функции

$$u \in W_2^2[a, b], \quad u(x_k) = f_k, \quad k = 0, 1, \ldots, n, \tag{3.1.12}$$

которая минимизирует функционал

$$\Phi(u) = \int_a^b [u''(x)]^2 \, dx \tag{3.1.13}$$

на классе $W_2^2[a, b]$. Утверждается, что минимум такого функционала достигается на кусочно-кубической сплайн-функции $g(x)$, которую мы только что построили. В самом деле, рассмотрим величину

$$\Phi(u - g) = \int_a^b [u'' - g'']^2 \, dx. \tag{3.1.14}$$

Интегрируя по частям и используя свойства функций g и $u \in W_2^2$, получим

$$\Phi(u - g) = \Phi(u) - \Phi(g) - 2 \left[(u' - g')|_{x=a}^{x=b} - \int_a^b (u' - g') g''' \, dx \right] =$$

$$= \Phi(u) - \Phi(g) - 2 \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} (u' - g') g''' \, dx.$$

Но $g''' = c_k = \text{const}$ на отрезке $[x_{k-1}, x_k]$, поэтому

$$\Phi(u - g) = \Phi(u) - \Phi(g) + 2 \sum_{k=1}^{n} c_k \, (u - g)|_{x=x_{k-1}}^{x=x_k} = \Phi(u) - \Phi(g).$$

Отсюда и из (3.1.14) вытекает, что

$$\Phi(g) = \Phi(u) - \Phi(u - g) \leq \Phi(u) \tag{3.1.15}$$

для любой функции

$$u \in W_2^2, \quad u(x_k) = f_k, \quad k = 0, 1, \ldots, n.$$

Таким образом, на кусочно-кубической функции $g(x)$ реализуется минимум функционала (3.1.13). Нетрудно показать, что других точек минимума у функционала нет.

Основываясь на (3.1.12), (3.1.13), можно дать другое, эквивалентное определение *кусочно-кубической сплайн-функции*: это такая функция из класса $W_2^2[a,b]$, которая принимает в узлах сетки заданное значение и минимизирует функционал (3.1.13). Такое свойство сплайн-функции интересно тем, что функционал $\Phi(u)$ можно интерпретировать как аналог потенциальной энергии упругого стержня, закрепленного в точках плоскости (x_k, f_k), и на кубических сплайнах реализуется минимум этой энергии.

В рамках этого пункта мы ограничились рассмотрением кубических сплайнов, удовлетворяющих граничным условиям (3.1.3), которые представляют собой условия «свободного провисания» интерполяционной кривой в точках a и b. Однако на практике часто бывают известны наклоны интерполяционной кривой в граничных точках. Тогда становится естественным применение условий

$$g'(a) = f'_0, \quad g'(b) = f'_n. \tag{3.1.16}$$

Если же мы знаем кривизну кривой в точках a и b, то естественны условия

$$g''(a) = f''_0, \quad g''(b) = f''_n. \tag{3.1.17}$$

Если об интерполируемой функции известно априори, что она периодична с периодом $b-a$, то следует применить граничные условия

$$g'(a) = g'(b), \quad g''(a) = g''(b). \tag{3.1.18}$$

При этом, конечно, $f_0 = f_n$.

Каким же образом изменится система линейных алгебраических уравнений (3.1.10) при наличии такого ряда (граничных условий)? В простейшем случае (3.1.3) мы пополняли систему (3.1.8) равенствами $m_0 = m_n = 0$. Условия (3.1.16) приведут нас, с учетом (3.1.7), к равенствам

$$\frac{2}{3}m_0 + \frac{1}{3}m_1 = \frac{2}{h_1}\left(\frac{f_1 - f_0}{h_1} - f'_0\right),$$

$$\frac{1}{3}m_{n-1} + \frac{2}{3}m_n = \frac{2}{h}\left(f'_n - \frac{f_n - f_{n-1}}{h_n}\right);$$

условия (3.1.17) — к равенствам

$$m_0 = f''_0, \quad m_n = f''_n;$$

наконец, условия периодичности сплайна (3.1.18) — к равенствам

$$m_0 = m_n,$$

$$\frac{h_n}{6}m_{n-1} + \frac{h_n + h_1}{3}m_n = \frac{f_1 - f_n}{h_1} - \frac{f_n - f_{n-1}}{h_n}.$$

Этими равенствами и следует пополнить систему (3.1.8). Конечно, возможно комбинировать условия различных типов в точках a и b.

Следует отметить, что кубические сплайны с различными типами краевых условий все равно доставляют минимум функционалу (3.1.13), только уже не на всем классе функций $W_2^2[a,b]$, а на подмножестве этого класса, состоящем из функций, удовлетворяющих данному краевому условию.

3.1.2. Кусочно-кубическая интерполяция со сглаживанием

В этом пункте мы опять рассмотрим задачу о гладком восполнении функции, которая определена на сетке

$$a = x_0 < x_1 < \ldots < x_n = b.$$

Однако теперь значения функции $\widetilde{f}_k$ в узлах сетки возмущены некоторой погрешностью. В этом случае не имеет смысла строить интерполяционную функцию, которая в узлах в точности совпадает с заданными значениями. Более того, следует построить функцию, которая проходила бы вблизи заданных значений более «плавно», чем интерполяционная. Такие функции называют уже не интерполяционными, а *сглаживающими*.

Потребуем, чтобы искомая сглаживающая функция $g(x)$ минимизировала на классе $W_2^2[a,b]$ функционал

$$\Phi_1(u) = \int_a^b [u'']^2\,dx + \sum_{k=0}^{n} p_k[u(x_k) - \widetilde{f}_k]^2, \tag{3.1.19}$$

где p_k — некоторые положительные числа. В функционале $\Phi_1(u)$ скомбинированы интерполяционные условия прохождения кривой вблизи заданных значений и условие минимальности «изгибания» функции. Чем больше весовые коэффициенты p_k, тем больший вклад в функционал вносят интерполяционные условия, тем ближе к заданным значениям проходит сглаживающая функция.

Покажем, что решением вариационной задачи (3.1.19) является кубический сплайн, т. е. функция, удовлетворяющая требованиям 1), 2) и 4) из предыдущего пункта. Пусть $u_0 \in W_2^2[a,b]$ — решение задачи. Построим такой сплайн $g(x)$, что $g(x_k) = u_0(x_k)$ $(k = 0, 1, \ldots, n)$. Второе слагаемое в (3.1.19) одинаково для функций $g(x)$ и $u_0(x)$, поэтому

$$\int_a^b [u_0'']^2\,dx \le \int_a^b [g'']^2\,dx. \tag{3.1.20}$$

Но, как показано в предыдущем пункте, $g(x)$ — единственная функция, дающая при интерполировании $u_0(x)$ минимум выражения $\int_a^b [u'']^2\,dx$. Поэтому $u_0 \equiv g$.

Итак, минимум функционала $\Phi_1(u)$ достаточно искать в классе кубических сплайнов. Так как кубический сплайн однозначно определяется множеством его значений $\{\mu_k\}_{k=0}^n$, принимаемых в узлах $\{x_k\}_{k=0}^n$, то минимизация $\Phi_1(u)$ сводится к нахождению минимума функции от переменных $\mu_0, \mu_1, \ldots, \mu_n$.

Мы уже знаем, что $g''(x)$ — кусочно-линейная функция, причем

$$g''(x) = m_{k-1}\frac{x_k - x}{h_k} + m_k\frac{x - x_{k-1}}{h_k} \tag{3.1.21}$$

при $x \in [x_{k-1}, x_k]$, $m_k = g''(x_k)$, $k = 1, 2, \ldots, n-1$, $m_0 = m_n = 0$. Поэтому

$$\Phi_1(g) = \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} \left[m_{k-1} \frac{x_k - x}{h_k} + m_k \frac{x - x_{k-1}}{h_k} \right]^2 dx + \sum_{k=0}^{n} p_k (\mu_k - \widetilde{f}_k)^2. \quad (3.1.22)$$

Производя в формуле (3.1.22) интегрирование, получим, что

$$\sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} \left[m_{k-1} \frac{x_k - x}{h_k} + m_k \frac{x - x_{k-1}}{h_k} \right]^2 dx =$$

$$= \sum_{k=1}^{n} m_k \left[m_{k-1} \frac{h_k}{6} + \frac{h_k + h_{k+1}}{3} m_k + \frac{h_{k+1}}{6} m_{k+1} \right] = (Am, m). \quad (3.1.23)$$

Здесь A — известная матрица (3.1.10). Итак,

$$\Phi_1(g) = (Am, m) + \sum_{k=0}^{n} p_k (\mu_k - \widetilde{f}_k)^2. \quad (3.1.24)$$

Ввиду (3.1.9) m линейно выражается через $\mu = (\mu_0, \mu_1, \ldots, \mu_n)$, а потому $\Phi_1(g)$ есть положительно определенная форма от μ. Ее экстремумом может быть только минимум, необходимым условием которого является

$$\frac{\partial \Phi_1}{\partial \mu_s} = \frac{\partial}{\partial \mu_s} (Am, m) + 2p_s(\mu_s - \widetilde{f}_s) = 0, \quad s = 0, 1, \ldots, n.$$

Но матрица A не зависит от μ. Поэтому в силу (3.1.9)

$$\frac{\partial}{\partial \mu_s} (Am, m) = 2 \left(\frac{\partial (Am)}{\partial \mu_s}, m \right) = 2 \left(\frac{\partial (H\mu)}{\partial \mu_s}, m \right) =$$

$$= 2 \left(\frac{\partial \mu}{\partial \mu_s}, H^* m \right) = 2 (H^* m)_s.$$

Здесь H определяется формулой (3.1.11). Отсюда вытекает, что в векторной форме условие минимума имеет вид

$$H^* m + P\mu = P\widetilde{f}, \quad (3.1.25)$$

где $\widetilde{f} = (\widetilde{f}_0, \widetilde{f}_1, \ldots, \widetilde{f}_n)$, а P — диагональная матрица:

$$P = \begin{Vmatrix} p_0 & 0 & \ldots & 0 \\ 0 & p_1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & p_n \end{Vmatrix}. \tag{3.1.26}$$

Умножая (3.1.25) слева на HP^{-1}, получим, что

$$HP^{-1}H^*m + H\mu = H\widetilde{f},$$

или окончательно, учитывая (3.1.9),

$$(A + HP^{-1}H^*)m = H\widetilde{f}. \tag{3.1.27}$$

Матрица системы (3.1.27) пятидиагональна, симметрична и положительно определена. Систему (3.1.27) можно решить, например, методом исключения Гаусса. После того как вектор m определен, необходимо найти вектор сеточных значений сглаживающего сплайна по формуле, которая легко следует из (3.1.25):

$$\mu = \widetilde{f} - P^{-1}H^*m. \tag{3.1.28}$$

Затем по формуле (3.1.6) восстановить сплайн $g(x)$.

3.1.3. Гладкие восполнения

Расскажем теперь о другом способе гладкого восполнения сеточных функций, несколько отличном от методов теории сплайн-функций, однако тоже весьма эффективном с алгоритмической точки зрения.

Опишем очень кратко процесс построения интерполяционной функции произвольного класса гладкости C^p. Пусть $x_1 < x_2 < \ldots < x_{n-1} < x_n$ — некоторая фиксированная сетка, в узлах которой известны значения функции $f_1, f_2, \ldots, f_n$. Предположим, что n достаточно велико. Фиксируем целое число $p \ll n$. Сначала построим интерполяционную функцию класса гладкости C^p на отрезке сетки (x_1, x_2). Обозначим через $P_0(x)$ многочлен Лагранжа степени не выше p, совпадающий с заданными значениями в узлах сетки $x_1, x_2, \ldots, x_{p+1}$, а через $P_1(x)$

— многочлен Лагранжа также степени не выше p, принимающий заданные значения в узлах $x_2, x_3, \ldots, x_{p+2}$. Построим, далее, многочлен $Q_1(x)$ степени не выше $2p+1$, который удовлетворяет условиям

$$\left.\frac{d^k Q_1}{dx^k}\right|_{x=x_1} = \left.\frac{d^k P_0}{dx^k}\right|_{x=x_1}, \quad \left.\frac{d^k Q_1}{dx^k}\right|_{x=x_2} = \left.\frac{d^k P_1}{dx^k}\right|_{x=x_2}, \tag{3.1.29}$$

$$k = 0, 1, \ldots, p.$$

Очевидно, что такой многочлен $Q_1(x)$ существует и условиями (3.1.29) определен однозначно.

Пусть интерполяционная функция $g(x)$ на отрезке (x_1, x_2) равна $Q_1(x)$. В качестве иллюстрации ограничимся рассмотрением случая $p = 1$. Тогда на отрезке (x_1, x_2) многочлен третьей степени $Q_1(x)$ запишется в виде

$$Q_1(x) = a_0 + a_1(x - x_1) + a_2(x - x_1)^2 + a_3(x - x_1)^3;$$

его коэффициенты вычисляются по формулам

$$\begin{aligned} a_0 &= f_1, \\ a_1 &= \frac{f_2 - f_1}{x_2 - x_1}, \\ a_2 &= -\frac{1}{x_2 - x_1}\left(\frac{f_3 - f_2}{x_3 - x_2} - \frac{f_2 - f_1}{x_2 - x_1}\right), \\ a_3 &= \frac{1}{(x_2 - x_1)^2}\left(\frac{f_3 - f_2}{x_3 - x_2} - \frac{f_2 - f_1}{x_2 - x_1}\right). \end{aligned} \tag{3.1.30}$$

Интерполяционная функция $g(x)$ на отрезке (x_2, x_3) строится совершенно аналогично, если за начало отсчета принять точку x_2. Очевидно, что такой процесс построения $g(x)$ можно продолжить до интервала (x_{n-p-1}, x_{n-p}). Построенный интерполянт $g(x)$ есть кусочно-полиномиальная функция класса гладкости C^p; степень полиномов на интервалах сетки равна $2p+1$.

Если положить $p = 2$ (что соответствует гладкости кубических сплайнов), то мы будем иметь дело с многочленами пятой степени. Степень многочлена выше на два, но зато для построения $g(x)$ не нужно решать линейной алгебраической системы.

Построение интерполяционных функций многих переменных мало чем отличается от одномерного случая (к которому все и сводится).

Гладкие восполнения обладают хорошими аппроксимирующими свойствами, именно: при интерполировании функции многих переменных $f \in C^q$ восполнением $g(x)$ класса C^p, где $p \geq q$, для самой функции и ее производных справедливы равенства

$$\|D^k f - D^k g\|_C \leq C(p) h^{q-|k|} \sup_x \max_{|\alpha|=q} |D^\alpha f(x)|, \tag{3.1.31}$$

где мультииндекс $k = (k_1, k_2, \ldots, k_r)$ таков, что

$$|k| = \sum_{i=1}^{r} k_i < q, \quad D^k = \frac{\partial^{k_1+\ldots+k_r}}{\partial x^{k_1} \ldots \partial x^{k_r}},$$

$C(p)$ зависит только от p и не зависит от шага сетки h и функции f.

Гладкие восполнения обладают еще одним свойством. Если построить базисную функцию, т. е. функцию, равную единице в одном узле и нулю в остальных, то для любого p она оказывается финитной, причем «радиус» носителя не превышает p. Если же построить кусочно-кубическую сплайн-функцию, равную единице в одном узле и нулю в остальных, то она вообще не финитна. Это обстоятельство делает гладкие восполнения очень удобными в вариационных задачах, которые решаются методом конечных элементов.

3.1.4. Сходимость сплайн-функций

В рамках этого пункта мы проиллюстрируем на простейших примерах технику получения оценок сходимости кубических сплайн-функций и их производных. Ограничимся для простоты граничными условиями (3.1.16) и, чтобы не усложнять формул, только равномерными сетками. Эти ограничения по ходу доказательств не играют существенной роли.

Введем в рассмотрение одно понятие, которое здесь понадобится. Пусть на отрезке $[a, b]$ задана непрерывная функция $\varphi(x)$. Тогда величина

$$\omega(h, \varphi) = \sup_{\substack{x', x'' \in [a,b] \\ |x'-x''| \leq h}} |\varphi(x') - \varphi(x'')|, \tag{3.1.32}$$

представляющая собой максимальное изменение функции $\varphi(x)$ на отрезке длины h в рамках $[a,b]$, называется *модулем непрерывности* функции $\varphi(x)$.

Пусть теперь на $[a,b]$ задана дважды непрерывно дифференцируемая функция $f(x)$, которую мы будем интерполировать кубическим сплайном на сетке с шагом h. Как известно из 3.1.1, вторые производные m_i от сплайна в узловых точках удовлетворяют системе типа (3.1.9):

$$Am = Hf.$$

Трансформируем эту систему следующим образом: разделим обе части на h и вычтем из обеих частей вектор $\frac{1}{h^2}AHf$. Получим равенство

$$\frac{1}{h}Am - \frac{1}{h^2}AHf = \left(I - \frac{1}{h}A\right)\frac{Hf}{h}. \tag{3.1.33}$$

Пусть

$$d = \frac{1}{h}Hf \quad \text{и} \quad B = \frac{1}{h}A.$$

Тогда

$$Bm - Bd = (I - B)d. \tag{3.1.34}$$

Ясно, что j-я компонента d_j вектора d имеет вид

$$d_j = \frac{(f_{j+1} - f_j)/h - (f_j - f_{j-1})/h}{h}. \tag{3.1.35}$$

Следовательно, в некоторой промежуточной точке ξ_j из интервала (x_{j-1}, x_{j+1}) справедливо равенство

$$d_j = f''(\xi_j). \tag{3.1.36}$$

Из вида граничного условия (3.1.16) очевидно, что

$$d_0 = f''(\xi_0), \quad d_n = f''(\xi_n). \tag{3.1.37}$$

Если учесть теперь, что сумма коэффициентов в каждой строке матрицы B равна единице ($1/6 + 2/3 + 1/6$), то

$$\|(I - B)d\| \le \omega(h, f''). \tag{3.1.38}$$

Как известно, собственные числа матрицы расположены в объединении кругов комплексной плоскости с центрами, равными диагональным элементам матрицы, и радиусами, равными сумме модулей внедиагональных элементов каждой строки (или столбца). Применяя этот результат к матрице B^{-1}, получим, что

$$\|B^{-1}\| \leq 3. \tag{3.1.39}$$

Следовательно,

$$\|m - d\| \leq \|B^{-1}\|\|(I - B)d\| \leq 3\omega(h, f''). \tag{3.1.40}$$

Так как

$$|f''(x_j) - d_j| < \omega(h, f''), \tag{3.1.41}$$

то

$$|f'(x_j) - m_j| \leq 4\omega(h, f''). \tag{3.1.42}$$

Так как вторая производная $g''(x)$ от сплайна $g(x)$ кусочно-линейна, то

$$|f''(x) - g''(x)| \leq 5\omega(h, f''). \tag{3.1.43}$$

Ввиду равенства $g(x_j) = f(x_j)$ на каждом интервале (x_{j-1}, x_j) найдется точка η_j, для которой $f'(\eta_j) = g'(\eta_j)$. Следовательно,

$$|f'(x) - g'(x)| = \left| \int_{\eta_j}^{x} [f''(x) - g''(x)]\, dx \right| \leq 5h\omega(h, f''). \tag{3.1.44}$$

Повторяя интегрирование, находим

$$|f(x) - g(x)| \leq \frac{5}{2} h^2 \omega(h, f''). \tag{3.1.45}$$

Таким образом, мы получили оценки сходимости самого сплайна, его первой и второй производных. Отметим, что оценки сохраняются для граничных условий (3.1.17) и (3.1.18).

С возрастанием гладкости функции $f(x)$ оценки сходимости улучшаются. Однако при помощи кубических сплайнов нельзя добиться скорости функции выше $O(h^4)$, если, конечно, функция $f(x)$ не является кубическим многочленом.

3.2. Интерполяция функций двух и многих переменных

Проблема двумерной интерполяции с помощью кусочно-бикубических функций была предметом исследования многих авторов. Мы кратко рассмотрим здесь следующую модельную задачу.

Пусть $D = \{x_k, y_l : \ a \le x \le b, \ c \le y \le d\}$ — некоторый прямоугольник в плоскости (x, y). Построим в D сетку

$$D_h = \{x_k, y_l : \ a = x_0 < x_1 < \ldots < x_n = b;$$

$$c = y_0 < y_1 < \ldots < y_m = d\}.$$

При таких предположениях задача кусочно-бикубической интерполяции функции $f(x, y)$, заданной в точках D_h, заключается в построении функции $g(x, y)$, удовлетворяющей условиям:

1)

$$g(x, y) \in C^{(2)}(D); \tag{3.2.1}$$

2) в каждой ячейке сетки $g(x, y)$ является бикубическим многочленом вида

$$g(x, y) = g_{k,l}(x, y) = \sum_{i,j=0}^{3} a_{i,j}^{k,l} (x_k - x)^i (y_l - y)^j; \tag{3.2.2}$$

3) на сетке D_h $g(x, y)$ принимает заданные значения

$$g(x_k, y_l) = f_{kl}, \quad k = 0, 1, \ldots, n; \quad l = 0, 1, \ldots, m; \tag{3.2.3}$$

4) функция $g(x, y)$ удовлетворяет условию

$$\left.\frac{\partial^2 g}{\partial \nu^2}\right|_{\Gamma} = 0 \tag{3.2.4}$$

(здесь ν — внешняя нормаль к границе Γ области D).

Как построить такую функцию? Принципиально построение ничем не отличается от одномерного случая. Вспомним, что для вычисления одномерной сплайн-функции в любой точке по простым формулам (3.1.6) нужно знать значения самой функции и ее вторых производных в узловых точках. Для того чтобы найти эти вторые произ-

водные, нужно один раз решить линейную алгебраическую систему с трехдиагональной матрицей. Какие же предварительные вычисления нужно проделать, чтобы потом по явным формулам вычислять функцию в любой точке в двумерном случае?

Рассмотрим сначала одномерные задачи кубической сплайн-интерполяции на линиях сетки $y = y_j$ $(j = 0, 1, \ldots, m)$. Для этого решим $m+1$ линейных алгебраических систем типа (3.1.9). В результате мы найдем значения функции $g_{xx}(x, y)$ в узлах сетки D_h. Затем аналогично решим $n + 1$ задач одномерной сплайн-интерполяции на линиях $x = x_i$ $(i = 0, 1, \ldots, n)$ и найдем значение функции $g_{yy}(x, y)$ на D_h. Предположим теперь, что требуется посчитать значение $g(x)$ в некоторой точке (x, y). Пусть $x_{i-1} \le x \le x_i$, $y_{j-1} \le y \le y_j$.

Мы можем найти значения $g(x, y)$ в точках (x_i, y), (x_{i-1}, y) по формулам, аналогичным (3.1.6):

$$g(x_{i-1}, y) = N_{i-1,j-1}\frac{(y_j - y)^3}{6\tau_j} + N_{i-1,j}\frac{(y - y_{j-1})^3}{6\tau_j} +$$

$$+ \left(f_{i-1,j-1} - \frac{N_{i-1,j-1}\tau_j^2}{6}\right)\frac{y_j - y}{\tau_j} + \left(f_{i-1,j} - \frac{N_{i-1,j}\tau_j^2}{6}\right)\frac{y - y_{j-1}}{\tau_j}, \qquad (3.2.5)$$

$$g(x_i, y) = N_{i,j-1}\frac{(y_j - y)^3}{6\tau_j} + N_{i,j}\frac{(y - y_{j-1})^3}{6\tau_j} +$$

$$+ \left(f_{i,j-1} - \frac{N_{i,j-1}\tau_j^2}{6}\right)\frac{y_j - y}{\tau_j} + \left(f_{i,j} - \frac{N_{i,j}\tau_j^2}{6}\right)\frac{y - y_{j-1}}{\tau_j}. \qquad (3.2.6)$$

Здесь и далее

$$N_{ij} = g_{yy}(x_i, y_j), \quad M_{ij} = g_{xx}(x_i, y_j),$$

$$h_i = x_i - x_{i-1}, \quad \tau_j = y_j - y_{j-1}.$$

Если будут известны значения $g_{xx}(x_{i-1}, y)$ и $g(x_i, y)$, то по формулам типа (3.1.6) можно найти значение $g(x, y)$. Отметим, что функция $g_{xx}(x, y)$ является кусочно-кубической по y. Решим $m + 1$ одномерных задач на линиях $y = y_j$ для функции $g_{xx}(x, y)$, сеточные значения которой уже известны. В результате найдем на D_h функцию $g_{xxyy}(x, y)$. Обозначим $K_{ij} = g_{xxyy}(x_i, y_j)$. Тогда

$$g_{xx}(x_{i-1}, y) = K_{i-1,j-1}\frac{(y_j - y)^3}{6\tau_j} + K_{i-1,j}\frac{(y - y_{j-1})^3}{6\tau_j} +$$

$$+\left(M_{i-1,j-1}-\frac{K_{i-1,j-1}\tau_j^2}{6}\right)\frac{y_j-y}{\tau_j}+\left(M_{i-1,j}-\frac{K_{i-1,j}\tau_j^2}{6}\right)\frac{y-y_{j-1}}{\tau_j}, \qquad (3.2.7)$$

$$g_{xx}(x_i,y)=K_{i,j-1}\frac{(y_j-y)^3}{6\tau_j}+K_{i,j}\frac{(y-y_{j-1})^3}{6\tau_j}+$$

$$+\left(M_{i,j-1}-\frac{K_{i,j-1}\tau_j^2}{6}\right)\frac{y_j-y}{\tau_j}+\left(M_{i,j}-\frac{K_{i,j}\tau_j^2}{6}\right)\frac{y-y_{j-1}}{\tau_j}. \qquad (3.2.8)$$

Используя снова формулу (3.1.6), находим значение $g(x,y)$:

$$g(x,y)=g_{xx}(x_{i-1},y)\frac{(x_i-x)^2}{6h_i}+g_{xx}(x_i,y)\frac{(x-x_{i-1})^2}{6h_i}+$$

$$+\left(g(x_{i-1},y)-\frac{g_{xx}(x_{i-1},y)h_i^2}{6}\right)\frac{x_i-x}{h_i}+$$

$$+\left(g(x_i,y)-\frac{g_{xx}(x_i,y)h_i^2}{6}\right)\frac{x-x_{i-1}}{h_i}. \qquad (3.2.9)$$

Резюмируя все сказанное выше, можно оценить объем вычислений, необходимых для расчета бикубического сплайна $g(x,y)$. Прежде чем приступить к расчету функции $g(x,y)$ в интересующих нас точках, необходимо один раз решить $(n+1)+(m+1)+(m+1)=2m+n+3$ линейных алгебраических систем типа (3.1.9) и найти $N_{ij},\ M_{ij},\ K_{ij}$ $(i=0,1,\ldots,n;\ j=0,\ldots,m)$. Далее, для расчета функции $g(x,y)$ в одной точке области нужно пять раз выполнить вычисление по формулам типа (3.1.6), а именно: (3.2.5), (3.2.6), (3.2.7), (3.2.8), (3.2.9). Цепочку этих формул можно использовать для получения явного полиномиального выражения $g(x,y)$ в каждой ячейке сетки. Но для хранения массивов коэффициентов многочлена в оперативной памяти ЭВМ потребуется около $16mn$ ячеек, в то время как в нашем алгоритме требуется только $4mn$, хотя количество операции при расчете функции в точке возрастает тоже примерно в 4 раза.

В этом алгоритме можно поменять переменные местами, и тогда нужно будет решать $2n+m+3$ одномерных задач, но итоговый результат не изменится, как не изменится и общее количество операций.

Описанный алгоритм может быть легко обобщен на многомерные области типа параллелепипеда.

3.3. r-гладкое приближение функций многих переменных

В настоящем параграфе рассматривается один из возможных подходов к задаче восполнения функций многих переменных, заданных с ошибкой на хаотических сетках. Изложение относится главным образом к описанию алгоритмической стороны предлагаемого метода, теоретические же аспекты его еще не разработаны и освещаются эвристически. В основе метода лежит хорошо известное в анализе понятие разбиения единицы.

Говорят, что последовательность (φ_i) функций класса $C^r(V)$ ($V \subset \mathbf{R}^m$ — открытое множество) образует разбиение единицы на множестве $\Omega \subset V$, если

$$0 \le \varphi_i(x) \le 1 \quad \text{и} \quad \sum_i \varphi_i(x) = 1 \quad \text{для всех} \quad x \in \Omega.$$

Обычно требуется, чтобы функция φ_i обращалась в нуль вне заданного открытого множества $V_i \subset V$. Совокупность (V_i), очевидно, покрывает множество Ω. Разбиение единицы (φ_i) называют подчиненным этому покрытию.

Для прикладных целей естественны локально конечные покрытия стандартными множествами: m-мерными интервалами или шарами. Центры и диаметры этих множеств должны выбираться в соответствии со свойствами приближаемой функции $f : \Omega \to R$. В дальнейшем мы будем рассматривать покрытия m-мерными кубами, а подчиненные им разбиения единицы строить следующим образом.

Обозначим через $\xi^{(i)}$ центр куба V_i, а через $\delta^{(i)}$ — длину его ребра. Зададимся (стандартной) такой функцией $\psi : R^m \to R$ класса $C^r(R^m)$, что

$$\psi(x) > 0, \quad \text{если} \quad x \in \left(-\frac{1}{2}, \frac{1}{2}\right) \times \ldots \times \left(-\frac{1}{2}, \frac{1}{2}\right) = V_0,$$

$$\psi(x) = 0, \quad \text{если} \quad x \notin V_0.$$

Можно, например, положить (при $x = (x_1, \ldots, x_m) \in V_0$)

$$\psi(x) = \prod_{i=1}^{m} \cos^{r+1}(\pi x_i) \quad \text{или} \quad \psi(x) = \prod_{i=1}^{m} \left(\frac{1}{4} - x_i^2\right)^{r+1}$$

и т. п. Положим теперь

$$\psi_i(x) = \begin{cases} \psi\left(\dfrac{x-\xi^{(i)}}{\delta^{(i)}}\right), & x \in V_i \\ \\ 0, & x \notin V_i, \end{cases}$$

для каждого значения индекса i. Функции

$$\varphi_i(x) = \frac{\psi_i(x)}{\sum\limits_j \psi_j(x)}$$

являются функциями класса $C^r(V)$ $(V = \bigcup\limits_i V_i)$ и образуют разбиение единицы на Ω, подчиненное покрытию (V_i).

Предположим теперь, что для функции $f : \Omega \to R$ известна такая последовательность $\{f_i : V \to R\}$ функций класса $C^r(V)$ такая, что каждая из функций f_i «близка» к f на V_i. Пусть в равномерной норме для любого i

$$\|(f - f_i)v_i\| \le \varepsilon.$$

Тогда функция $\widetilde{f}$, определяемая равенством

$$\widetilde{f}(x) = \sum_i f_i(x)\varphi_i(x), \quad x \in \Omega, \tag{3.3.1}$$

будет «близка» к f на Ω. Действительно, в силу тождества

$$f(x) = \sum_i f(x)\varphi_i(x), \quad x \in \Omega,$$

получаем, что для любого $x \in \Omega$

$$|f(x) - \widetilde{f}(x)| = \left|\sum_i (f(x) - f_i(x))\varphi_i(x)\right| \le$$

$$\le \max_i \sup_{x \in V_i} |f(x) - f_i(x)| \sum_i \varphi_i(x) \le \varepsilon,$$

т. е.

$$\|f - \widetilde{f}\| \le \varepsilon.$$

Теперь используем разбиение единицы для восполнения функции f. Если она задана в узлах некоторой сетки $(x^{(j)} \in \Omega)$, то для каж-

дого i нужно построить по некоторой совокупности узлов $(x^{(j)})$, близких к центру $\xi^{(i)}$ куба V_i, ее локальное приближение $f_i(x)$, а затем эти приближения «склеить» по формуле (3.3.1). В дальнейшем мы будем считать, что узлы $x^{(i)}$ совпадают с центрами $\xi^{(i)}$ кубов V_i, а локальные приближения являются постоянными функциями, равными $f(x^{(i)})$.

Проведем оценку погрешности приближения для одного частного случая:

$$f(x) = x, \quad x \in R;$$

сетка — равномерная и бесконечная с узлами в точках $x_k = kh$ $(k = 0, \pm 1, \pm 2, \ldots)$; центры $\xi^{(k)}$ интервалов V_k совпадают с узлами сетки, а $\delta^{(k)} = (2N+1)h$, где N — некоторое натуральное число. Пусть разбиение единицы $\{\varphi_k\}$ принадлежит классу $C^r(R)$ $r \geq 3$, функция $\varphi_0(x)$ четна.

В этом случае остаток

$$\sigma(x) = x - \sum_k x_k \varphi_k(x)$$

представляет собой нечетную периодическую с периодом h функцию, равную нулю в целых и полуцелых узлах сетки. Поэтому достаточно рассмотреть случаи:

1) $0 \leq x \leq h/4$; 2) $h/4 \leq x \leq h/2$.

Рассмотрим случай 1). Так как $\varphi_k(x) = \varphi_0(x - x_k)$ и функция φ_0 четна, то

$$\widetilde{x} = \sum_k x_k \varphi_k(x) = \sum_k x_k \varphi_0(x - x_k) = \sum_k x_k [\varphi_0(x_k - x) - \varphi_0(x_k)].$$

Отсюда по формуле Тейлора имеем

$$\widetilde{x} = -x \sum_k x_k \varphi_0'(x_k) + \frac{x^2}{2} \sum_k x_k \varphi_0''(x_k) - \frac{x^3}{6} \sum_k x_k \varphi_0'''(x_k - \theta_k x).$$

Второе слагаемое в правой части этого равенства, очевидно, равно нулю, поэтому

$$\widetilde{x} = -x \sum_k x_k \varphi_0'(x_k) - \frac{x^3}{6} \sum_k x_k \varphi_0'''(x_k - \theta_k x). \tag{3.3.2}$$

Далее, применив к сумме

$$\overline{x} = -x \sum_k x_k \varphi_0'(x_k)$$

преобразование Абеля, получим

$$\overline{x} = xh \sum_{k=-N}^{N-1} \sum_{r=-N}^{k} \varphi_0'(rh) - xx_N \sum_{r=-N}^{N} \varphi_0'(rh) = xh \sum_{k=-N}^{N-1} \sum_{r=-N}^{k} \varphi_0'(rh).$$

Поскольку

$$h \sum_{r=-N}^{k} \varphi_0'(rh) = \int\limits_{-(N+1/2)h}^{(k+1/2)h} \varphi_0'(x)\, dx + \sigma_1(k) = \varphi_0\left(x_k + \frac{h}{2}\right) + \sigma_1(k),$$

где

$$|\sigma_1(k)| \le (\max_x |\varphi_0'''(x)|)\frac{h^3}{24}(k+N+1), \tag{3.3.3}$$

то

$$\overline{x} = x \sum_k \varphi_0\left(x_k + \frac{h}{2}\right) + x \sum_{k=-N}^{N-1} \sigma_1(k) = x + x \sum_{k=-N}^{N-1} \sigma_1(k). \tag{3.3.4}$$

Из формул (3.3.2)—(3.3.4) получаем оценку

$$|\sigma(x)| \le (\max_x |\varphi_0'''(x)|)\left[\frac{x^3 h}{6} \sum_{k=-N}^{N} |k| + \frac{xh^3}{24} \sum_{k=-N}^{N-1} (k+N+1)\right] =$$

$$= (\max_x |\varphi_0'''(x)|)h^4 \frac{3}{2} N\left(N + \frac{5}{9}\right) = C\frac{h^4 3\left(N + \frac{5}{9}\right)N}{2^7(2N+1)^3 h^3} < C\frac{h}{341N}. \tag{3.3.5}$$

Здесь C — максимальное значение модуля третьей производной функции $x(y) = \varphi_0(y(2N+1)h)$. Константа C зависит от N и убывает с ростом N.

В случае 2) приходим к тому же результату.

Значение полученной оценки состоит в том, что она дает один из возможных критериев выбора исходной функции ψ, по которой строится разбиение единицы, а именно: надо стремиться к тому, чтобы третья производная функции ψ была по возможности малой.

Аналогичные оценки можно получить для линейной функции и в случае нескольких переменных.

Поставим теперь основную задачу настоящего параграфа.

Пусть функция $f : \Omega \to R$ известна на нерегулярной сетке $(x^{(i)}\ (i = 1, 2, \ldots, n))$ с некоторой точностью ε (в равномерной норме). Требуется найти такую функцию $\widetilde{f}$ класса C^r, что выполняется условие аппроксимации

$$\|f - \widetilde{f}\| = \max_{i=1,2,\ldots,n} |f(x^{(i)}) - \widetilde{f}(x^{(i)})| \le \varepsilon, \tag{3.3.6}$$

а также условие сглаживания, которое мы сформулируем здесь эвристически, как требование «неосциллируемости» производных функции $\widetilde{f}$ некоторого порядка.

Для решения этой задачи мы предлагаем метод, основанный на многократном применении описанного выше способа восполнения с использованием некоторой последовательности разбиений единицы. При описании алгоритма будут указаны соотношения, которые, по нашему мнению, обеспечивают, по крайней мере в указанном выше эвристическом смысле, выполнение условия сглаживания.

Обозначим через L_0 оператор (сглаживающего приближения), действующий из пространства сеточных функций $\{f : x^{(i)} \to R\}$ в пространство дифференцируемых функций $C^r(V)$ по формуле

$$L_0(f)(x) = \sum_{i=1}^{n} f(x^{(i)})\varphi_i(x), \tag{3.3.7}$$

где (φ_i) — разбиение единицы на Ω описанного ранее вида, центры носителей функции φ_i являются узлами $x^{(i)}$ сетки.

Оценим норму отклонения f от $L_0(f)$. Имеем

$$\|f - L_0(f)\| = \max_{j=1,2,\ldots,n} |f(x^{(j)}) - \sum_{i=1}^{n} f(x^{(i)})\varphi_i(x^{(j)})| =$$

$$= \max_{j=1,2,\ldots,n} |f(x^{(j)})(1 - \varphi_j(x^{(j)})) - \sum_{i=1, i\ne j}^{n} f(x^{(i)})\varphi_i(x^{(j)})| \le$$

$$\|f\| \cdot \max \left| 1 - \varphi_j(x^{(j)}) - \sum_{i=1, i\ne j}^{n} \varphi_i(x^{(j)}) \right| \le$$

$$\le 2\|f\|[1 - \min_{j=1,2,\ldots,n} \varphi_j(x^{(j)})]. \tag{3.3.8}$$

Отсюда следует, что при однократном применении формулы (3.3.7) можно удовлетворить условию аппроксимации за счет выбора такого разбиения единицы, что величина

$$\lambda = \min_{j=1,2,\ldots,n} \varphi_j(x^{(j)}) \tag{3.3.9}$$

будет достаточно близка к единице. Но величина λ близка к единице, если число узлов, попадающих во множества V_i, мало либо функции φ_j «слишком куполообразны», что, как показывает оценка (3.3.5), ухудшает сглаживающие свойства оператора L_0.

Эти соображения указывают на то, что при однократном применении оператора L_0 невозможно, вообще говоря, получить решение поставленной задачи. Поэтому мы поступаем следующим образом.

Пусть $\{\varphi_i^{(k)}\|_{k=1,2,\ldots}$ — последовательность разбиений единицы, подчиненных (для определенности) одному и тому же покрытию множества Ω, а $(L_0^{(k)})$ — соответствующая последовательность операторов сглаживающего приближения, определяемых для каждого (φ_i^k) по формуле (3.3.7). Будем искать $\widetilde{f}$ в виде

$$\widetilde{f} = \sum_{k=1}^{k_0} L_0^{(k)}(f^{(k-1)}) \tag{3.3.10}$$

с некоторым k_0, где $\{f^{(k)}\}$ — последовательность сеточных функций, определяемая следующим образом:

$$f^{(k)}(x^{(j)}) = \begin{cases} f(x^{(j)}), & k = 0, \\ f^{(k-1)}(x^{(j)}) - L_0^{(k)}(f^{(k-1)})(x^{(j)}), & k > 0, \end{cases} \tag{3.3.11}$$

$j = 1, 2, \ldots, n$. Просуммировав эти равенства для $k = 0, 1, \ldots, k_0$, получим

$$f = \widetilde{f} + f^{(k_0)},$$

так что для выполнения условий аппроксимации необходимо, чтобы величина $\|f^{(k_0)}\|$ не превосходила величины ε. Это условие всегда может быть удовлетворено за счет специального выбора последовательности разбиений единицы $\{\varphi_i^{(k)}\}$. Действительно, если, к примеру, начиная с некоторого k_1, выполняется условие

$$\lambda^{(k)} = \min_{j=1,2,\ldots,n} \varphi_j^{(k)}(x^{(j)}) \geq \sigma > 1/2$$

для всех $k \geq k_1$, то в силу (3.3.8)

$$\|E - L_0^k\| \leq 2(1-\sigma) < 1,$$

а следовательно, $\|f^{(k)}\| \underset{k\to\infty}{\to} 0$.

Выбирая последовательность разбиений единицы $\{\varphi_j^{(k)}\}$ таким образом, чтобы величины $\lambda^{(k)}$ были по возможности малыми при условии монотонного убывания последовательности $\{\|f^{(k)}\|\}$ с некоторой заранее фиксированной скоростью $v > 1$ (т. е. $\|f^{(k)}\|/\|f^{(k+1)}\| \geq v$), можно добиться выполнения также и условия сглаживания (поскольку при этом «негладкий вклад» в сумму (3.3.10) будет вноситься уже малыми по норме значениями $f^{(k)}$).

Для практических целей последовательность разбиений единицы $\{\varphi_i^{(k)}\}$ естественно строить следующим образом. Если $\psi^{(k)}(x)$ — стандартная функция для разбиения $(\varphi_i^{(k)})$, то стандартную функцию $\psi^{(k+1)}(x)$ для разбиения $(\varphi_i^{(k+1)})$ выбираем в виде

$$\psi^{(k+1)}(x) = \psi^{(k)} g^{(k)}(x),$$

где множитель $g^{(k)}(x)$ можно взять либо в форме

$$g^{(k)}(x) = \exp\{-\mu_k(x_1^2 + x_2^2 + \ldots + x_m^2)\}, \quad \mu_k \geq 0,$$

либо в форме

$$g^{(k)}(x) = [1 + \mu_k(x_1^2 + x_2^2 + \ldots + x_m^2)]^{-1}, \quad \mu_k \geq 0,$$

и т. п.

Вычисления организуются так, что если $\|f^{(k-1)}\|/\|f^{(k)}\| \geq v$, то $\mu_k = 0$, а $\mu_k = \mu > 0$ в противном случае.

Таким образом, описанный алгоритм содержит три регулирующих параметра:

1) величину $\lambda^{(0)} = \lambda$, определяемую по (3.3.9);

2) скорость $v > 1$ уменьшения невязки $\|f^{(k)}\|$;

3) параметр μ, регулирующий скорость увеличения «дельтообразности» разбиений единицы.

Понятно, что оптимизацией по этим параметрам можно минимизировать функционалы, значения которых определяют то или иное понятие «осциллируемости» производных функции $\widetilde{f}$.

Замечание. Если взять последовательность разбиений единицы из одинаковых разбиений (φ_i), то в силу (3.3.10), (3.3.11) получим

$$\widetilde{f}(x) = \sum_{i=1}^{n} \left(\sum_{k=1}^{k_0} f^{(k-1)}(x^{(i)}) \varphi_i(x) \right).$$

Может возникнуть мысль — ограничиться одним набором разбиений единицы (φ_i) и искать решение $\widetilde{f}$ поставленной задачи в виде

$$\widetilde{f}(x) = \sum_{i=1}^{n} c_i \varphi_i(x), \tag{3.3.12}$$

определяя коэффициенты c_i из условия

$$\left\| f - \sum_{i=1}^{n} c_i \varphi_i(x) \right\| \leq \varepsilon \tag{3.3.13}$$

в предположении, что выполнено условие сглаживания.

В действительности эта постановка не отвечает существу дела, так как данные $f(x^{(i)})$ заданы с ошибкой.

Действительно, если

$$\det \|\varphi_i(x^{(j)})\| \neq 0, \tag{3.3.14}$$

то условию (3.3.13) можно удовлетворить с $\varepsilon = 0$, но тогда ошибки в «измеренных» $f(x^{(i)})$ распространятся и на полученное продолжение (3.3.12). Поэтому желательно стремиться к тому, чтобы оценка (3.3.13) по возможности «достигалась», а для этого естественно поставить условие

$$\max_{i=1,\ldots,n} |f(x^{(i)}) - c_i| \leq E, \quad E > 0, \tag{3.3.15}$$

которое ограничивало бы норму $\|f - c\|$. Но у нас нет никаких критериев выбора величины E, а задаваясь этой величиной произвольно (естественно задавать E «малой»), мы приходим к задаче, в общем неразрешимой (не говоря уже о том, что и условие (3.3.14) не всегда выполняется).

Из указанной противоречивости требований аппроксимации, сглаживания и условия (3.3.15) и возник описанный выше алгоритм.

3.4. Элементы общей теории сплайнов

Как мы показали, кубические сплайн-функции возникают при минимизации квадратичного функционала

$$\int_a^b [u''(x)]^2 \, dx,$$

который является аналогом энергии изгибания упругого стержня. В связи с этой задачей можно сказать, что развитие теории сплайн-функций шло в двух основных направлениях.

В первом направлении вместо оператора двукратного дифференцирования стали употреблять дифференциальный оператор L общей формы с переменными и даже разрывными коэффициентами и минимизировать функционал

$$\int_a^b [Lu(x)]^2 \, dx.$$

Это сразу сделало теорию сплайнов очень гибкой и приспособило ее для аппроксимации функций с минимизацией энергии, связанной с конкретной физической ситуацией.

Во втором направлении вместо простейшей ситуации, связанной с закреплением кривой в точках сетки, стали употреблять другие линейные функционалы, задание которых более естественно в конкретной задаче. Так появились сплайны четной степени, которые строятся по интегральным средним от функции, тригонометрические сплайны и другие популярные теперь конструкции. Конечно, все это сопровождалось развитием многомерной теории сплайнов. Такая аппроксимация функций сопровождалась, кроме того, сглаживанием.

Общее определение интерполирующего и сглаживающего сплайнов. Пусть X, Y — два гильбертовых пространства; $T: \ X \to Y$ — линейный ограниченный оператор, действующий из X в Y. Над про-

странством X задана система линейных ограниченных функционалов k_i $(i = 1, 2, \ldots, n)$, которая предполагается линейно независимой. *Интерполяционным сплайном* называется элемент $\sigma \in X$, удовлетворяющий двум условиям:

$$\begin{aligned} &1) \qquad k_i(\sigma) = (k_i, \sigma)_X = r_i, \quad i = 1, 2, \ldots, n; \\ &2) \qquad (T\sigma, T\sigma)_Y = \|T\sigma\|_Y^2 = \min. \end{aligned} \tag{3.4.1}$$

Здесь символами $(\cdot,\cdot)_X$, $(\cdot,\cdot)_Y$ обозначены скалярные произведения в пространствах X и Y соответственно; r_i $(i = 1, 2, \ldots, n)$ — наперед заданные числа. Значение функционала k_i на векторе σ заменено скалярным произведением согласно известной теореме Рисса о представлении линейного функционала в гильбертовом пространстве. Соответственно, *сглаживающим сплайном* назовем элемент $\sigma_\alpha \in X$, который реализует минимум функционала

$$\Phi_\alpha(u) = \alpha\|Tu\|_Y^2 + \sum_{i=1}^{n}[(k_i, u) - r_i]^2, \quad \alpha > 0. \tag{3.4.2}$$

Для задач (3.4.1) и (3.4.2) сформулированы общие теоремы существования и единственности сплайн-функции, которые находят сейчас применение при построении разнообразных сплайн-аппроксимаций. Имеется общий алгоритм построения интерполяционных и сглаживающих сплайнов. Показано, что в конечном итоге задачи (3.4.1) и (3.4.2) сводятся к решению некоторых алгебраических систем с симметричными положительно определенными матрицами.

Мы приведем в этом параграфе теорему сходимости в достаточно общей форме. Эту теорему можно применять как в классической ситуации сходимости сплайнов на прямоугольных сетках, так и в более сложных ситуациях при построении сплайнов на сгущающихся хаотических сетках или при анализе сходимости сплайнов, построенных по произвольной системе функционалов.

Прежде всего рассмотрим вопрос о сходимости интерполяционных сплайнов. Пусть X — гильбертово пространство; φ^* — элемент пространства X, который мы хотим аппроксимировать. В пространстве X рассмотрим базисную систему векторов $k_1, k_2, \ldots, k_n, \ldots$, вообще говоря, неортогональную. Если $T: \ X \to Y$ — линейный ограничен-

ный оператор, действующий в другое гильбертово пространство Y, то аппроксимирующий элемент будем искать как сплайн σ_n:

$$\begin{gathered}(\sigma_n, k_i)_X = (\varphi^*, k_i)_X = r_i, \quad i = 1, 2, \ldots, n, \\ \|T\sigma_n\|_Y^2 = \min.\end{gathered} \tag{3.4.3}$$

Если предположить теперь выполнение требований, гарантирующих существование и единственность сплайна σ_n для некоторого $n = n_0$, то сплайны σ_n будут существовать при любом $n \geq n_0$. Возникает вопрос о сходимости сплайнов σ_n к точному вектору φ^*. При выполнении требований теоремы существования и единственности и при условии, что ядро оператора T имеет конечную размерность, наблюдается сходимость сплайнов σ_n к φ^* и, что гораздо важнее на практике, последовательность $T\sigma_n$ сходится к $T\varphi^*$ в норме пространства Y.

Общая теорема сходимости может быть с успехом применена для получения оценок сходимости в терминах убывания шага сетки. Например, применение к сплайнам, состыкованным из многочленов степени $2n-1$, на отрезке $[a, b]$ дает следующие оценки аппроксимации функции $\varphi^* \in W_2^n$ и ее производных сплайном σ_N:

$$\|\sigma_N^{(i)} - \varphi^{*(i)}\|_{C[a,b]} = o(h^{n-i-1/2}), \quad i = 0, 1, \ldots, n-1, \tag{3.4.4}$$

$$\|\sigma_N^{(n)} - \varphi^{*(n)}\|_{L_2[a,b]} = o(1).$$

Символ $o(h^\alpha)$ означает, что погрешность стремится быстрее, чем h^α.

Общая теория сходимости может быть применена в ситуации, когда использование ранее известных методов не дает результата. Примером такой ситуации является анализ сходимости сплайнов, построенных по беспорядочно разбросанным в двумерной области интерполяционным узлам. Сплайн-функции в двумерной области строились из условия минимизации функционала

$$\Phi(u) = \int\limits_\Omega \left[\left(\frac{\partial^2 u}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 u}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 u}{\partial y^2} \right)^2 \right] d\Omega, \tag{3.4.5}$$

если функция u принимает в фиксированной совокупности узлов заданные значения. Если $\varphi^* \in W_2^2(\Omega)$, то, применяя общую теорию сходимости, можно получить оценку

$$\|\sigma_N - \varphi^*\|_{C[\Omega_\delta]} = o(h^{3/2}). \tag{3.4.6}$$

Здесь σ — сплайн; h — характерный параметр сгущения хаотической сетки, определяемый как параметр ε-сети, образуемой интерполяционными узлами в области Ω; Ω_δ — любая δ-внутренность области Ω, т. е. совокупность точек области Ω, отстоящих от границы на расстояние, большее δ.

Итак, мы кратко познакомились с теорией сходимости интерполяционных сплайнов. Однако задача об отыскании интерполяционного сплайна — это задача минимизации квадратичного функционала при линейных ограничениях, т. е. задача на условный минимум. В то же время задача об отыскании сглаживающего сплайна — это задача уже на безусловный минимум, и решать такую задачу иногда проще. Сглаживающий сплайн σ_α сходится к интерполяционному сплайну σ в норме пространства X при $\alpha \to 0$.

Следует отметить, что алгоритм построения интерполяционных и сглаживающих сплайнов, являющийся универсальным с теоретической точки зрения, на практике трудно применять уже в двумерных задачах. Поэтому приходится пользоваться приближенными методами построения, а именно: задача о построении интеполяционного сплайна σ заменяется задачей о нахождении сглаживающего σ_α, который определяется из условия минимума функционала

$$\Phi(u) = \alpha\|Tu\|_X^2 + \sum_{i=1}^{n}[(k_i, u) - r_i]^2, \quad \alpha > 0, \tag{3.4.7}$$

при малом α на всем пространстве X. Затем в пространстве X рассматривается конечномерное подпространство E_k, элементы которого аппроксимируют элементы X с устраивающей нас точностью, и отыскивается минимум $\Phi_\alpha(u)$ на $E_{k,t}$. Точка минимума называется *сплайном на подпространстве*.

Такой подход может быть с успехом применен в следующей задаче. Пусть Ω — прямоугольная область, в которой беспорядочно разбросаны интерполяционные узлы p_i $(i = 1, 2, \ldots, s)$. Необходимо по-

строить гладкую в области Ω интерполяционную функцию. Задача решается так: на область Ω наносим прямоугольную сетку Ω_h и связываем с этой сеткой конечномерное пространство гладких интерполянтов E_k.

Способы построения этих пространств мы уже рассматривали. Это могут быть *бикубические сплайны* или *интерполяции Рябенького* класса гладкости C^2. На этом пространстве будем решать задачу минимизации функционала

$$\Phi_\alpha(u) = \alpha \int\limits_\Omega [u_{xx}^2 + 2u_{xy}^2 + u_{yy}^2]\, d\Omega + \sum_{i=1}^{s} [u(P_i) - r_i]^2 \tag{3.4.8}$$

при малом $\alpha > 0$. Такая задача приводит к линейной алгебраической системе. Если в E_k удается выбрать базис, состоящий из финитных функций, то матрица системы оказывается *ленточной*. В качестве такого базиса могут быть взяты кубические B-сплайны или, что, на наш взгляд, удобнее, локальные базисные функции, рассмотренные в 3.1.3.

В заключение этого параграфа мы рассмотрим еще один практически важный вопрос — вопрос выбора параметра сглаживания. Существует несколько критериев выбора α. Мы остановимся на одном из них, именно на так называемом «критерии невязки». Этот способ выбора параметра естественен в широком круге задач.

Предположим, что решается задача сглаживания, т. е. задача нахождения элемента σ_α, принадлежащего гильбертову пространству X, который минимизирует функционал

$$\Phi_\alpha(u) = \alpha \|Tu\|_Y^2 + \sum_{i=1} [(k_i, u) - r_i]^2 \equiv \alpha \|Tu\|_Y^2 + \|Ku - r\|^2. \tag{3.4.9}$$

Здесь $K:\ X \to R^n$ — оператор, действующий по формуле

$$Ku = \{(u, k_1), \ldots, (u, k_n)\}, \tag{3.4.10}$$

а $r = (r_1, r_2, \ldots, r_n)$ — данный вектор. *Критерий невязки* состоит в том, чтобы выбрать α из условия

$$\|K\sigma_\alpha - r\| = \varepsilon, \tag{3.4.11}$$

где ε (фиксированное число) — «допустимый уровень невязки». Если ввести обозначение

$$\varphi(\alpha) = \|K\sigma_\alpha - r\|, \quad 0 < \alpha < +\infty, \tag{3.4.12}$$

то для определения α нужно решить нелинейное уравнение

$$\varphi(\alpha) = \varepsilon.$$

Сделаем замену переменных $\alpha = 1/\beta$ и рассмотрим уравнение

$$\varphi(\alpha) = \varphi(1/\beta) = \psi(\beta) = \varepsilon \tag{3.4.13}$$

относительно β. Отметим, что функция $\psi(\beta)$ строго монотонно убывает и уравнение $\psi(\beta) = \varepsilon$ имеет единственное решение при $\varepsilon < \varepsilon_0$, причем $\varepsilon_0 = \|Ke - r\|$, где e — решение задачи о наименьших квадратах

$$\|Ke - r\| = \min_{u \in \operatorname{Ker} T} \|Ku - r\|$$

на ядре оператора T. Для решения нелинейного уравнения (3.4.13) естетвенно применить метод Ньютона. Возведем обе части уравнения (3.4.13) в некоторую вещественную степень q:

$$\psi^q(\beta) = \varepsilon^q, \tag{3.4.14}$$

а при помощи q максимально увеличим скорость сходимости процесса Ньютона. Допустимыми значениями q, при которых итерации Ньютона сходятся при любом начальном приближении, являются $-1 \leq q < 0$, $q > 0$, причем максимальная скорость сходимости обеспечена при $q = -1$. Таким образом, необходимо решить уравнение

$$f(\beta) = 1/\varphi(1/\beta) = \varepsilon^{-1}. \tag{3.4.15}$$

Итерации Ньютона реализуются по формуле

$$\beta^{p+1} = \beta^p - \frac{f(\beta^p) - \varepsilon^{-1}}{f'(\beta^p)} = \beta^p - (\beta^p)^2 \frac{\varphi(1/\beta^p)}{\varphi'(1/\beta^p)} \frac{\varepsilon - \varphi(1/\beta^p)}{\varepsilon}. \tag{3.4.16}$$

Какой же объем вычислений нужно выполнить на каждом шаге метода Ньютона? Чтобы посчитать $\varphi(1/\beta^p) = \varphi(\alpha^p)$, необходимо

решить задачу сплайн-сглаживания при фиксированном параметре $\alpha = \alpha^p$, а это значит, что нужно решить некоторую алгебраическую систему. Кроме того, необходимо знать $\varphi'(\alpha^p)$. Для этого нужно решить ту же линейную алгебраическую систему, только с другой правой частью.

В заключение отметим, что развитая в общей форме теория сглаживания, конечно, полностью эквивалентна некоторому методу регуляризации некорректно поставленных задач. Например, мы можем считать, что K — вырожденная матрица, а интерполяционный сплайн u — нормальное решение совместной системы

$$\begin{gathered} Ku = f, \\ \|u\| = \min. \end{gathered} \tag{3.4.17}$$

Решение этой системы (интерполяционный сплайн) можно приблизить при $\alpha \to 0$ решением u_α регуляризованной системы (сглаживающим сплайном)

$$\begin{gathered} \alpha\|u_\alpha\|^2 + \|Ku_\alpha - f\|^2 = \min, \\ (\alpha I + K^*K)u_\alpha = K^*f. \end{gathered} \tag{3.4.18}$$

Глава 4.

Методы решения стационарных задач математической физики

Решение стационарных задач математической физики представляет собой более или менее самостоятельный раздел вычислительной математики, хотя решение многих стационарных задач с положительными операторами можно рассматривать как предельное при $t \to \infty$ решение нестационарной задачи. При решении стационарных задач методами асимптотического стационирования мы не обращаем внимания на промежуточные значения решения, поскольку они не представляют интереса, тогда как при решении нестационарных задач эти промежуточные значения имеют физический смысл. Вообще говоря, именно в этом состоит единство и различие этих классов задач. Проиллюстрируем сказанное на примере.

Пусть имеется задача

$$A\varphi = f,$$

где

$$A > 0, \quad \varphi \in \Phi \quad \text{и} \quad f \in F.$$

Вместо этой задачи рассмотрим нестационарную[1)]

$$\frac{\partial \psi}{\partial t} + A\psi = f,$$

[1)] В дальнейшем мы не будем специально указывать области определения операторов задач.

$$\psi = 0 \quad \text{при} \quad t = 0.$$

Представим φ, ψ и f в виде

$$\varphi = \sum_n \varphi_n u_n, \quad \psi = \sum_n \psi_n u_n, \quad f = \sum_n f_n u_n,$$

где

$$Au_n = \lambda_n u_n, \quad A^* u_n^* = \lambda_n u_n^*,$$

$$\varphi_n = (\varphi, u_n^*), \quad \psi_n = (\psi, u_n^*), \quad f_n = (f, u_n),$$

а $\{u_n\}$ и $\{u_n^*\}$ — биортогональный базис. Тогда известными приемами приходим к задачам для коэффициентов Фурье:

$$\lambda_n \varphi_n = f_n,$$

с одной стороны, и

$$\frac{d\psi_n}{dt} + \lambda_n \psi_n = f_n, \quad \psi_n(0) = 0,$$

с другой. Решая эти задачи, получим

$$\varphi = \sum_n \frac{f_n}{\lambda_n} u_n, \quad \psi = \sum_n \frac{f_n}{\lambda_n}(1 - e^{-\lambda_n t}) u_n.$$

В предположении вещественности спектра оператора A имеем $\lambda_n > 0$ $(n = 1, 2, \ldots)$. Отсюда следует, что

$$\lim_{t \to \infty} \psi = \varphi.$$

Если оператор стационарной задачи имеет спектр произвольной структуры, то в этом случае такой простой и прозрачной связи между решениями задач уже может не быть.

Разумеется, нестационарную задачу для новой функции ψ можно решать разностными методами по t, например

$$\frac{\psi^{j+1} - \psi^j}{\tau} + A\psi^j = f.$$

Тогда

$$\psi^{j+1} = \psi^j - \tau(A\psi^j - f).$$

Если нашей целью является решение стационарной задачи, то при определенном соотношении между τ и $\beta(A)$ имеем

$$\lim_{j\to\infty} \psi^j = \varphi.$$

Параметр τ может быть величиной как не зависящей, так и зависящей от j. Во всяком случае при решении стационарной задачи j удобно считать номером шага не временного, а итерационного.

Здесь имеет место еще одна особенность: в нестационарных задачах для обеспечения точности решения значения τ должны быть достаточно малы, в стационарных же — оптимальные итерационные параметры τ выбираются из условия минимальности числа итераций и могут принимать относительно большие значения.

4.1. Общие понятия теории итерационных методов

В дальнейшем на протяжении всей главы мы будем считать, что оператором A является квадратная матрица. Следовательно, исходная задача предполагается уже редуцированной к системе линейных алгебраических уравнений. При этом везде, за исключением 4.5, предполагается невырожденность матрицы A и всюду — вещественность участвующих векторов и матриц. Итак, пусь требуется решить систему

$$A\varphi = f, \tag{4.1.1}$$

где A — матрица, а φ и f — векторы.

Большинство итерационных методов, которые применяются для решения линейных систем, могут быть объединены общей формулой

$$B_j \frac{\varphi^{j+1} - \varphi^j}{\tau_j} = -(A\varphi^j - f), \tag{4.1.2}$$

где $\{B_j\}$ — последовательность невырожденных матриц, а $\{\tau_j\}$ — последовательность вещественных параметров. Если ввести обозначение $H_j = \tau_j B_j^{-1}$, то процесс (4.1.2) можно записать в эквивалентном виде:

$$\varphi^{j+1} = \varphi^j - H_j(A\varphi^j - f). \tag{4.1.3}$$

Векторы $\xi^j = A\varphi^j - f$ называются *векторами невязок* итерационного метода (4.1.2), а векторы $\psi = \varphi^j - \varphi^*$ ($\varphi^* = A^{-1}f$ — точное решение системы (4.1.1)) *называются векторами ошибок* этого метода. Вычитая из обеих частей соотношения (4.1.3) вектор φ^* и делая замену $f = A\varphi^*$, приходим к соотношению для последовательности векторов ошибок

$$\psi^{j+1} = T_j\psi^j, \tag{4.1.4}$$

где матрица

$$T_j = E - H_jA \tag{4.1.5}$$

называется *оператором j-го шага* итерационного метода (4.1.2). Умножая (4.1.4) на матрицу A, приходим к соотношению для последовательности векторов невязок

$$\xi^{j+1} = (E - AH_j)\xi^j. \tag{4.1.6}$$

Мы будем называть итерационный метод (4.1.2) *сходящимся*, если последовательность $\{\varphi^j\}$ сходится к точному решению φ^* системы (4.1.1) при любом начальном векторе, и *расходящимся* — в противном случае. Очевидно, что необходимым и достаточным условием сходимости итерационного метода (4.1.2) является сходимость последовательностей $\{\psi^j\}$ и $\{\xi^j\}$ к нулевому вектору для любых ψ^0 и ξ^0 ($A\psi^0 = \xi^0$).

Итерационный метод (4.1.2) называется *стационарным*, если матрица H_j не зависит от номера итерации (оператор T_j является постоянной матрицей), и *нестационарным* — в противном случае. Особо мы выделим класс циклических итерационных методов, которые могут быть отнесены как к стационарным, так и к нестационарным итерационным методам.

Циклическими мы будем называть итерационные методы, которые обладают свойством $H_j = H_{j+s}$ для любого $j \geq 0$ и некоторого фиксированного $s \geq 1$. Нетрудно видеть, что, объединяя каждые s последовательных итераций в одну, мы приходим к стационарному итерационному процессу вида

$$\varphi^{j+1} = \varphi^j - \widetilde{H}(A\varphi^j - f), \tag{4.1.7}$$

где $\widetilde{H}$ определяется из уравнения

$$E - \widetilde{H}A = \prod_{i=1}^{s-1}(E - H_i A). \tag{4.1.8}$$

С другой стороны, в первоначальной формулировке циклические итерационные методы относятся к нестационарным.

Одной из основных наших задач при исследовании итерационных методов будет проблема их оптимизации, т. е. выбор последовательности матриц $\{H_j\}$ из заданного класса с целью получения более эффективного вычислительного процесса. Целевой функцией, которую при этом мы должны минимизировать, является общее число арифметических и логических действий, необходимых для нахождения решения задачи с заданной точностью $\varepsilon > 0$. Полное решение сформулированной проблемы возможно лишь в исключительных случаях, поэтому на практике осуществляется минимизация некоторой функции

$$W(\{H_j\}, \varepsilon), \tag{4.1.9}$$

которая мажорирует сверху целевую функцию и в определенном смысле достаточно хорошо ее приближает. В предположении, что при любом H_j из заданного класса одна итерация метода (4.1.2) требует одно и то же число арифметических и логических действий W_0, оказывается целесообразным определить W соотношением

$$W = W_0 N(\{H_j\}, \varepsilon), \tag{4.1.10}$$

где $N(\{H_j\}, \varepsilon)$ — число итераций метода (4.1.2), достаточных для уменьшения некоторой нормы начальной ошибки ψ^0 в $1/\varepsilon$ раз. Иначе говоря, $N(\{H_j\}, \varepsilon)$ равно числу итераций метода (4.1.2), достаточных для решения системы (4.1.1) с точностью ε, если норма вектора ошибки ψ^0 равна единице. В дальнейшем на конкретных примерах мы покажем, как определяется функционал N и решается задача его минимизации.

4.2. Некоторые итерационные методы и их оптимизация

4.2.1. Простейший итерационный метод

Пусть матрица A системы

$$A\varphi = f \tag{4.2.1}$$

симметрична, положительно определена и известны границы ее спектра $\beta = \beta(A) \geq \alpha = \alpha(A) > 0$. Для решения этой системы применим итерационный метод

$$\varphi^{j+1} = \varphi^j - \tau(A\varphi^j - f) \tag{4.2.2}$$

с некоторым начальным вектором φ^0 и вещественным параметром τ. Для векторов ошибок $\psi^j = \varphi^j - \varphi^*$ этого метода выполняется соотношение

$$\psi^{j+1} = T_\tau \psi^j, \tag{4.2.3}$$

где

$$T_\tau = E - \tau A \tag{4.2.4}$$

— оператор шага.

Обозначим через $\{\lambda_n\}$ множество собственных чисел A, а через $\{u_n\}$ — полную ортонормированную систему соответствующих собственных векторов A. Тогда, разлагая векторы $\{\psi^j\}$ по ортонормированному базису $\{u_n\}$:

$$\psi^j = \sum_n \psi_n^j u_n, \quad \psi_n^j(\psi^j, u_n),$$

и используя (4.2.3), получим

$$\psi_n^{j+1} = (1 - \tau\lambda_n)\psi_n^j.$$

Отсюда следует, что для сходимости к нулю коэффициентов ψ_n^j при $j \to \infty$ и любом ψ^0 необходимо и достаточно выполнение неравенств

$$|1 - \tau\lambda_n| < 1,$$

а для сходимости к нулю евклидовой нормы вектора ψ^j:

$$\|\psi^j\|_2 = (\psi^j, \psi^j)^{1/2} = \left[\sum_n (\psi_n^j)^2\right]^{1/2},$$

необходимо и достаточно, чтобы выполнялось неравенство

$$q(\tau) = \beta(T_\tau) = \max_n |1 - \tau\lambda_n| < 1. \tag{4.2.5}$$

Исследуем величину $q(\tau)$. Рассматривая систему неравенств

$$-1 < 1 - \tau\lambda_n < 1, \quad \lambda_n \in [\alpha, \beta],$$

которая эквивалентна неравенству (4.2.5), приходим к выводу, что (4.2.5) будет выполняться только в случае

$$0 < \tau < \min_n \frac{2}{\lambda_n} = \frac{2}{\beta}, \tag{4.2.6}$$

т. е. при $\tau \in (0, 2/\beta)$. Таким образом, для определения области изменения значений τ, при которых метод (4.2.2) сходится, нам достаточно знать либо величину β, либо оценку сверху этой величины.

Перейдем к изучению проблемы оптимизации метода (4.2.2). Используя неравенство

$$\|\psi^j\|_2 \le \|T_\tau^j\|_2 \|\psi^0\|_2, \tag{4.2.7}$$

где в силу симметричности T_τ

$$\|T_j^i\|_2 = [\beta(T_\tau)]^j = [q(\tau)]^j, \tag{4.2.8}$$

нетрудно видеть, что для уменьшения $\|\psi^0\|_2$ в $1/\varepsilon$ $(\varepsilon < 1)$ раз достаточно провести N итераций, где N определяется из уравнения

$$\|T_\tau^N\|_2 = \varepsilon. \tag{4.2.9}$$

Отсюда, учитывая (4.2.8), получим

$$N = \left[\frac{\ln \varepsilon}{\ln[q(\tau)]}\right] + 1. \tag{4.2.10}$$

Так как число арифметических и логических действий W_0 на одну итерацию метода (4.2.2) не зависит от значения τ, то функционал W из (4.1.10) в данном случае определяется соотношением

$$W = W_0 \left(\left[\frac{\ln \varepsilon}{\ln[q(\tau)]} \right] + 1 \right). \tag{4.2.11}$$

(Здесь предполагается, что $\tau \in (0, 2/\beta)$.) Теперь легко видеть, что проблема оптимизации (минимизации W по τ) сводится к минимизации по $\tau \in (0, 2/\beta)$ функции $q(\tau)$. Решим эту задачу.

Простейший анализ показывает, что

$$q(\tau) = \max\{|1 - \tau\alpha|, |1 - \tau\beta|\}, \tag{4.2.12}$$

а оптимальное значение τ является решением уравнения

$$1 - \tau_{opt}\alpha = -(1 - \tau_{opt}\beta)$$

и вычисляется по формуле

$$\tau_{opt} = \frac{2}{\beta + \alpha}. \tag{4.2.13}$$

Подставляя это значение τ в (4.2.12), получим

$$q_{opt} = q(\tau_{opt}) = \frac{\beta - \alpha}{\beta + \alpha} = \frac{p - 1}{p + 1}, \tag{4.2.14}$$

где величина

$$p \equiv p(A) = \beta/\alpha \tag{4.2.15}$$

называется *числом обусловленности* симметричной и положительно определенной матрицы A.

Введем дополнительно величину

$$R(T_\tau) = -\ln q(\tau), \tag{4.2.16}$$

которую будем называть *асимптотической скоростью сходимости метода* (4.2.2). Очевидно, что $R^{-1}(T_\tau)$ равно числу итераций, достаточных, а при произвольном ψ^0 также необходимых для уменьшения $\|\psi^0\|_2$ в e раз, где $e = 2,7\ldots$ — основание натуральных логарифмов. Асимптотическая скорость сходимости оказывается удобным крите-

рием для сравнения различных методов, когда вопрос о числе арифметических и логических действий на итерацию не обсуждается. С учетом введенной величины (4.2.11) принимает вид

$$W = W_0 \left(\left[\frac{|\ln \varepsilon|}{R(T_\tau)} \right] + 1 \right). \tag{4.2.17}$$

В заключение остановимся на случае *плохо обусловленных матриц*, т. е. когда $p \gg 1$. Учитывая, что при $p \gg 1$

$$q_{opt} \approx 1 - \frac{2}{p}, \tag{4.2.18}$$

получаем

$$R(T_\tau) \approx 2/p, \tag{4.2.19}$$

и, следовательно,

$$W \approx W_0 |\ln \varepsilon| \frac{p}{2}. \tag{4.2.20}$$

Подобный анализ оказывается очень удобным с точки зрения качественного решения сравнения различных методов в случае плохо обусловленных матриц.

4.2.2. Сходимость и оптимизация стационарных итерационных методов

В предыдущем пункте было показано, что необходимым и достаточным условием сходимости итерационного метода (4.2.2) для любого начального приближения φ^0 является выполнение неравенства

$$\beta(T) = \max_n |\lambda_n(T)| < 1, \tag{4.2.21}$$

где $\lambda_n(T)$ — собственные числа оператора шага T. Сейчас мы докажем, что выполнение неравенства (4.2.21) является *необходимым и достаточным* условием сходимости стационарного метода

$$\varphi^{j+1} = \varphi^j - H(A\varphi^j - f)$$

с оператором шага

$$T = E - HA,$$

являющимися постоянной матрицей.

Пусть

$$J = \begin{Vmatrix} J_1 & 0 & \dots & 0 \\ 0 & J_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & J_k \end{Vmatrix}$$

— нормальная жорданова форма матрицы T (т. е. $T = SJS^{-1}$, где столбцами матрицы S являются собственные и корневые векторы матрицы T) с клетками Жордана

$$J_i = \begin{Vmatrix} \lambda_i & 1 & 0 & \dots & 0 & 0 \\ 0 & \lambda_i & 1 & \dots & 0 & 0 \\ 0 & 0 & \lambda_i & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \lambda_i & 1 \\ 0 & 0 & 0 & \dots & 0 & \lambda_i \end{Vmatrix}$$

порядка $k_i \geq 1$, которые соответствуют собственным числам $\lambda_i = \lambda_i(T)$ $(i = 1, 2, \dots, l)$. Тогда, так как

$$\psi^j = T^j \psi^0 = SJ^j S^{-1} \psi^0,$$

где $\{\psi^j\}$ — векторы ошибок, нетрудно видеть, что необходимым и достаточным условием сходимости стационарного итерационного метода является сходимость матриц T^j (и, следовательно, матриц J^j) к нулевой матрице при $j \to \infty$. Последнее выполняется только в том случае, если для любого i $(1 \leq i \leq l)$ матрицы J_i^j сходятся к нулевой матрице при $j \to \infty$.

Таким образом, для доказательства сформулированного утверждения о необходимом и достаточном условии сходимости стационарного итерационного метода нам достаточно показать, что матрицы J_i^j сходятся к нулевым матрицам при $j \to \infty$ только в случае $|\lambda_i| < 1$. Обозначим через $\alpha_{s,t}^{(j)}$ элементы матрицы J_j^i. Тогда путем непосред-

ственного перемножения матриц нетрудно показать, что при $j \geq k_i - 1$

$$\alpha_{s,t}^{(j)} = \begin{cases} 0, & \text{если} \quad s > t, \\ \lambda_i^j, & \text{если} \quad s = t, \\ C_j^{t-s} \lambda_i^{j-t+s}, & \text{если} \quad s < t, \end{cases}$$

где

$$C_j^{t-s} = \frac{j(j-1)\dots(j-t+s+1)}{(t-s)!}$$

есть биномиальные коэффициенты. Иначе говоря,

$$J_i^j = \begin{Vmatrix} \lambda_i^j & C_j^1 \lambda_i^{j-1} & \dots & C_j^{k_i-1} \lambda_i^{j-k_i+1} \\ 0 & \lambda_i^j & \dots & C_j^{k_i-2} \lambda_i^{j-k_i+2} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_i^j \end{Vmatrix}$$

для всех $j \geq k_i - 1$ и $1 \leq i \leq l$. С помощью элементарных выкладок легко показать, что $\alpha_{s,t}^{(j)}$ стремятся к нулю для всех $1 \leq s, t \leq k_i$ при $j \to \infty$ (соответственно, J_i^j сходятся к нулевой матрице) только в случае $|\lambda_i| < 1$. Утверждение доказано. Известно, что этот результат справедлив для любого стационарного итерационного метода.

В настоящем пункте мы рассмотрим общий подход к оптимизации стационарных итерационных методов с точки зрения их асимптотического поведения. Для этого нам потребуется известный факт из функционального анализа, который в случае матриц формулируется следующим образом.

Для любой квадратной матрицы T и любой матричной нормы $\|\cdot\|$ выполняется равенство

$$\lim_{k \to \infty} \|T^k\|^{1/k} = \beta(T). \tag{4.2.22}$$

Предположим теперь, что матрица H_τ стационарного итерационного метода

$$\varphi^{j+1} = \varphi^j - H_\tau (A\varphi - f) \tag{4.2.23}$$

зависит от параметров $\tau_1, \tau_2, \ldots, \tau_s$. Тогда если независимо от значений параметров каждая итерация метода требует одного и того же числа арифметических и логических действий W_0, то, согласно предыдущему (см. § 4.1), проблема оптимизации метода (4.2.23) заключается в минимизации по $\tau_1, \ldots, \tau_s$ функций

$$W = W_0 \cdot N(H_\tau, \varepsilon), \tag{4.2.24}$$

где $N(H_\tau, \varepsilon)$ — число итераций, достаточных для уменьшения некоторой нормы вектора ошибки ψ^0 в $1/\varepsilon$ раз.

Так как

$$\|\psi^j\| \le \|T_\tau^j\| \|\psi^0\|, \tag{4.2.25}$$

где $T_\tau = E - H_\tau A$, а $\|\cdot\|$ — некоторая норма, то уравнение для определения зависимости N от H_τ и ε имеет вид

$$\|T_\tau^N\| = \varepsilon. \tag{4.2.26}$$

Очевидно, что точное решение этого уравнения можно получать лишь в исключительных случаях, как это было, например, сделано в 4.2.1. Поэтому в конкретных ситуациях нужно ставить вопрос о его приближенном решении при тех или иных дополнительных предположениях.

Одним из возможных путей приближенного отыскания зависимости N от H_τ и ε является асимптотический подход, а именно: если предположить, что $\varepsilon \ll 1$ и, соответственно, $N \gg 1$, то из сформулированной выше теоремы будет следовать, что

$$\|T_\tau^N\| = [\|T_\tau^N\|^{1/N}]^N \approx [\beta(T_\tau)]^N,$$

$$N(H_\tau, \varepsilon) \approx \frac{|\ln \varepsilon|}{R(T_\tau)}, \tag{4.2.27}$$

где $R(T_\tau) = -\ln \beta(T_\tau)$ — асимптотическая скорость сходимости метода (4.2.23). Таки образом, функцию W можно определить соотношением

$$W = W_0 \frac{|\ln \varepsilon|}{R(T_\tau)}. \tag{4.2.28}$$

Заметим, что (4.2.17) является частным случаем (4.2.28).

Из полученной формулы (4.2.28) следует, что при сделанных предположениях проблема оптимизации метода (4.2.23) заключается в максимизации $R(T_\tau)$ (минимизации $\beta(T_\tau)$) по тем параметрам $\tau_1, \ldots, \tau_s$, для которых $\beta(T_\tau) < 1$.

Остановимся кратко на циклических итерационных методах с периодом $s \geq 1$. Как уже было показано в 4.1, циклический итерационный метод может быть приведен к обычному стационарному итерационному методу (4.1.7) с оператором шага

$$\widetilde{T}_\tau = \prod_{t=0}^{s-1}(E - H_i A). \tag{4.2.29}$$

Предполагая, что каждая итерация метода (4.1.7) требует sW_0 арифметических и логических действий, где W_0 — аналогичная величина для одного шага циклического метода, получим

$$W = W_0 \frac{|\ln \varepsilon|}{R(\widetilde{T}_\tau)}, \tag{4.2.30}$$

где

$$R(\widetilde{T}_\tau) = -\frac{1}{s} \ln \beta(\widetilde{T}_\tau).$$

Отсюда следует, что в случае циклических итерационных методов оптимизация этих методов при фиксированном s заключается в минимизации величины $\beta(\widetilde{T}_\tau)$.

4.2.3. Метод последовательной верхней релаксации

Во многих приложениях большую популярность приобрел итерационный метод, разработанный Янгом и Франкелом и называемый *методом последовательной верхней релаксации*. Проиллюстрируем основную идею этого метода на примере системы линейных алгебраических уравнений

$$A\varphi = f \tag{4.2.31}$$

с блочно-трехдиагональной матрицей

$$A = \begin{Vmatrix} E_1 & -S_1 & \dots & 0 & 0 \\ -R_2 & E_2 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & E_{k-1} & -S_{k-1} \\ 0 & 0 & \dots & -R_k & E_k \end{Vmatrix}, \tag{4.2.32}$$

где E_l — единичные матрицы порядка n_l, а R_l и S_l — некоторые матрицы порядка $n_l \times n_{l-1}$ и $n_l \times n_{l+1}$ соответственно. Отметим, что системы с матрицами такого вида часто возникают при редукции уравнений эллиптического типа к системам конечно-разностных или вариационно-разностных уравнений.

Представим матрицу A в виде

$$A = E - R - S, \tag{4.2.33}$$

где

$$R = \begin{Vmatrix} 0 & 0 & \dots & 0 & 0 \\ R_2 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & R_k & 0 \end{Vmatrix}, \quad S = \begin{Vmatrix} 0 & S_1 & \dots &) & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & S_{k-1} \\ 0 & 0 & \dots & 0 & 0 \end{Vmatrix}$$

и E — единичная матрица. Тогда если ввести матрицу

$$B = R + S, \tag{4.2.34}$$

систему (4.2.31) можно записать в виде

$$\varphi = B\varphi + f. \tag{4.2.35}$$

Далее мы везде будем предполагать, что итерационный метод

$$\varphi^{j+1} = B\varphi_j + f \tag{4.2.36}$$

сходится (т. е. $\beta(B) < 1$), все собственные числа матрицы B вещественные и она обладает полной системой собственных векторов.

Перепишем (4.2.31) в виде системы матричных уравнений

$$\Phi_l - R_l\Phi_{l-1} - S_l\Phi_{l+1} = F_l, \quad l = 1, 2, \ldots, k, \tag{4.2.37}$$

при условиях

$$R_1 = S_k = 0 \quad \text{и} \quad \Phi_0 = \Phi_{k+1} = 0,$$

где Φ_l и F_l — векторные компоненты векторов φ и f. Тогда метод последовательной верхней релаксации определяется формулами

$$\Phi_l^{j+1} = \Phi_l^j - \tau(\Phi_l^j - R_l\Phi_{l-1}^{j+1} - S_l\Phi_{l+1}^j - F_l), \quad l = 1, 2, \ldots, k, \tag{4.2.38}$$

где τ — некоторый вещественный параметр. Соотношения (4.2.38) можно также записать в виде

$$\Phi_l^{j+1} = \tau\Phi_l^{j+1/2} + (1-\tau)\Phi_l^j,$$

$$\Phi_l^{j+1/2} = R_l\Phi_{l-1}^{j+1} + S_l\Phi_{l+1}^j + F_l, \quad l = 1, 2, \ldots, k.$$

Из приведенных формул видно, что для каждого j процесс вычислений осуществляется последовательно с первой по k-ю группу компонент. При этом Φ_0^{j+1} и Φ_{k+1}^j полагаются равными нулю.

Перейдем от (4.2.38) к уравнениям для векторов ошибок $\psi_l^j = \Phi_l^j - \Phi_l^*$, где $\{\Phi_l^*\}$ — решение системы (4.2.37):

$$\psi_l^{j+1} = \psi_l^j - \tau(\psi_l^j - R_l\psi_{l-1}^{j+1} - S_l\psi_{l+1}^j), \quad l = 1, 2, \ldots, k. \tag{4.2.39}$$

Отсюда видно, что оператор шага метода (4.2.37) определяется равенством

$$T_\tau = (E - \tau R)^{-1}[(1-\tau)E + \tau S] \tag{4.2.40}$$

и зависит от одного параметра τ, а спектральная задача

$$T_\tau\psi = \lambda(T_\tau)\psi \tag{4.2.41}$$

легко преобразуется к виду

$$\lambda(T_\tau)\psi_l = \psi_l - \tau[\psi_l - \lambda(T_\tau)R_l\psi_{l-1} - S_l\psi_{l+1}], \quad l = 1, 2, \ldots, k. \tag{4.2.42}$$

Решение этой спектральной задачи будем искать в виде

$$\psi_l = [\lambda(T_\tau)]^{1/2} w_l, \tag{4.2.43}$$

где $\{w_l\}_{l=1}^k$ — векторные компоненты собственного вектора w спектральной задачи

$$Bw = \lambda(B)w, \tag{4.2.44}$$

а $w_0 = w_{k+1} = 0$. Подставляя (4.2.43) в (4.2.42) и учитывая (4.2.44), получим

$$\lambda^{l/2}(T_\tau)[\lambda(T_\tau) - \tau\lambda^{1/2}(T_\tau)\lambda(B) + \tau - 1]w_l = 0, \quad l = 1, 2, \ldots, k. \tag{4.2.45}$$

Так как нулевые значения $\lambda(T_\tau)$ с точки зрения сходимости метода последовательной верхней релаксации нас не интересуют, а среди $\{w_l\}$ есть хотя бы один ненулевой вектор, то из (4.2.45) получаем уравнение

$$\lambda(T_\tau) - \tau\lambda^{1/2}(T_\tau)\lambda(B) + \tau - 1 = 0,$$

которое связывает собственные числа матриц T_τ и B. Решив это уравнение, получим

$$\lambda^{1/2}(T_\tau) = \frac{\tau\lambda(B)}{2} \pm \sqrt{\frac{\tau^2\lambda^2(B)}{4} - \tau + 1}. \tag{4.2.46}$$

Прежде чем переходить к анализу этой формулы, заметим, что если $\lambda(B)$ — собственное число матрицы B, то величина $-\lambda(B)$ также является собственным числом матрицы B. Действительно, если $\lambda(B)w = Bw$, где w — собственный вектор матрицы B с векторными компонентами $\{w_p\}_{l=1}^k$, то вектор $\widetilde{w}$ с векторными компонентами $\widetilde{w}_p = (-1)^p w_p$ является собственным вектором B, соответствующим ее собственному числу $-\lambda(B)$, т. е.

$$-\lambda(B)\widetilde{w} = B\widetilde{w}. \tag{4.2.47}$$

Из доказанного и формулы (4.2.46) следует, что при анализе $\lambda^{1/2}(T_\tau)$ нам достаточно ограничиться случаем $\lambda(B) \geq 0$.

Исследуем величину $\lambda^{1/2}(T_\tau)$ как функцию параметра τ, считая, что $\lambda(B) < 1$ — некоторое фиксированное неотрицательное число матрицы B. Сначала рассмотрим случай $\tau \leq 0$. Как показывают несложные вычисления, при $\tau \leq 0$ неравенство

$$|\lambda^{1/2}(T_\tau)| = \left| \frac{\tau\lambda(B)}{2} \pm \sqrt{\frac{\tau^2\lambda^2(B)}{4} - \tau + 1} \right| < 1 \tag{4.2.48}$$

не имеет решения, и, следовательно, требование положительности параметра τ необходимо для сходимости метода последовательной верхней релаксации.

Ограничимся в дальнейшем случаем положительных значений τ. Тогда, учитывая, что сходимость метода определяется только максимальным по модулю собственным числом матрицы T_τ, нетрудно заметить, что теперь достаточно исследовать формулу

$$\lambda^{1/2}(T_\tau) = \frac{\tau\lambda(B)}{2} + \sqrt{\frac{\tau^2\lambda^2(B)}{4} - \tau + 1},$$

где при положительности подкоренного выражения значение квадратного корня берется положительным.

Далее, для значений $\tau \in [\tau_1, \tau_2]$, где

$$\tau_{1,2} = \frac{2}{\lambda^2(B)}[1 \pm \sqrt{1 - \lambda^2(B)}],$$

подкоренное выражение в (4.2.48) неположительно, и для этих значений имеем

$$|\lambda(T_\tau)| = |\lambda^{1/2}(T_\tau)|^2 = \left(\frac{\tau\lambda(B)}{2}\right)^2 - \left(\frac{\tau^2\lambda^2(B)}{4} - \tau + 1\right) = \tau - 1.$$

Если же $\tau \notin [\tau_1, \tau_2]$, то подкоренное выражение в (4.2.48) всегда положительно, а $\lambda^{1/2}(T_\tau)$ неотрицательно. Учитывая это, нетрудно показать, что при $\tau \in (0, \tau_1)$ неравенство

$$\lambda^{1/2}(T_\tau) \geq 1$$

не имеет решения, а при $\tau \geq \tau_2$ выполняется неравенство

$$\lambda^{1/2}(T_\tau) \geq \frac{\tau\lambda(B)}{2} \geq \frac{\tau_2\lambda(B)}{2} = \frac{1}{\lambda(B)}[1 + \sqrt{1 - \lambda^2(B)}] > 1.$$

Из доказанных фактов вытекает следующее утверждение: при сделанных выше предположениях условие

$$\tau \in (0, 2)$$

необходимо и достаточно для сходимости метода последовательной верхней релаксации (4.2.38).

Остановимся теперь на исследовании величины

$$\lambda^{1/2}(T_\tau) = \begin{cases} \dfrac{\tau\lambda(B)}{2} + \sqrt{\dfrac{\tau^2\lambda^2(B)}{4} - \tau + 1}, & \tau \le \tau_1, \\ \sqrt{\tau - 1}, & \tau \ge \tau_1, \end{cases}$$

как функции параметра τ. Так как для значений $\tau \in (0, \tau]$ получаем

$$\frac{\partial}{\partial\tau}[\lambda^{1/2}(T_\tau)] = \frac{\lambda(B)\lambda^{1/2}(T_\tau) - 1}{2\sqrt{\dfrac{\tau^2\lambda^2(B)}{4} - \tau + 1}} < 0$$

и $\lambda^{1/2}(T_\tau) = 1$ при $\tau = 0$, то графики $|\lambda^{1/2}(T_\tau)|$ для различных $\lambda(B)$ имеют следующий вид (см. рис. 4.1).

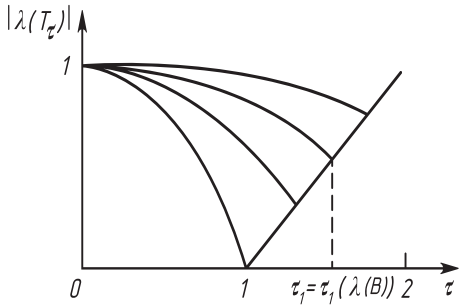


Рис. 4.1.

Поскольку

$$\frac{\partial}{\partial\tau}[\lambda^{1/2}(T_\tau)] = \begin{cases} \dfrac{\tau\lambda^{1/2}(T_\tau)}{2\sqrt{\dfrac{\tau^2\lambda^2(B)}{4} - \tau + 1}} > 0 & \text{при} \quad \tau < \tau_1, \\ 0 & \text{при} \quad \tau \ge \tau_1, \end{cases}$$

то для значений $\tau \in (0, 2)$

$$\beta(T_\tau) = \begin{cases} \dfrac{\tau\beta(B)}{2} + \sqrt{\dfrac{\tau^2\beta^2(B)}{4} - \tau + 1} & \text{при} \quad \tau < \tau_1, \\ \\ \tau - 1 & \text{при} \quad \tau \geq \tau_1. \end{cases}$$

Из последней формулы сразу следует, что минимальное значение $\beta(T_\tau)$ достигается при $\tau = \tau_1$. Таким образом, значение τ_{opt}, максимизирующее асимптотическую скорость сходимости $R(T_\tau)$ метода последовательной верхней релаксации, вычисляется по формуле

$$\tau_{opt} = \frac{2}{\beta^2(B)}[1 - \sqrt{1 - \beta^2(B)}] = \frac{2}{1 + \sqrt{1 - \beta^2(B)}}, \tag{4.2.49}$$

причем

$$R(T_{\tau_{opt}}) = -\ln(\tau_{opt} - 1) = -\ln\frac{1 - \sqrt{1 - \beta^2(B)}}{1 + \sqrt{1 - \beta^2(B)}}. \tag{4.2.50}$$

Проанализируем теперь формулы (4.2.49), (4.2.50) в случае плохо обусловленных матриц (предполагая дополнительно, что матрица A симметрична), т. е. когда $p(A) = \dfrac{\beta(A)}{\alpha(A)} \gg 1$. Так как одновременно с $\beta(B)$ собственным числом матрицы B вида (4.2.34) является величина $-\beta(B)$ (это доказано ранее), то из соотношения $B = E - A$ вытекает, что

$$\beta(B) = 1 - \alpha(A),$$

$$-\beta(B) = 1 - \beta(A).$$

Отсюда следует, что $\alpha(A) = 2 - \beta(A)$, а условие $p(A) \gg 1$ означает выполнение соотношений

$$\alpha(A) \approx \frac{2}{p},$$

$$\beta(B) \approx 1 - \frac{2}{p}.$$

Подставляя эти значения в (4.2.49), (4.2.50), асимптотически получим

$$\tau_{opt} \approx 2 - \frac{4}{\sqrt{p}}, \quad R(T_{\tau_{opt}}) \approx \frac{4}{\sqrt{p}},$$
$$\beta(T_{\tau_{opt}}) \approx 1 - \frac{4}{\sqrt{p}}, \quad W \approx W_0 \frac{|\ln \varepsilon|}{4}\sqrt{p}. \tag{4.2.51}$$

Таким образом, для плохо обусловленных матриц при оптимальных выборах параметров метод последовательной верхней релаксации асимптотически сходится в $2\sqrt{p}$ раз быстрее метода (4.2.2). Необходимо отметить, что для любого конкретного номера итерации (особенно на первых итерациях) сходимость метода последовательной верхней релаксации с оптимальным параметром медленнее, чем показывает третье из соотношений (4.2.51). Причиной этого является наличие в жордановой форме матрицы перехода $T_{\tau_{opt}}$ клетки порядка два.

4.2.4. Чебышевский итерационный метод

Рассмотрим систему линейных алгебраических уравнений

$$A\varphi = f \tag{4.2.52}$$

с симметричной и положительно определенной матрицей A. Для решения этой системы предложим итерационный метод

$$\varphi^{j+1} = \varphi^j - \tau_j(A\varphi - f), \tag{4.2.53}$$

который называют *методом Ричардсона первого порядка*. Проблема оптимизации этого метода заключается в выборе последовательности параметров $\{\tau_j\}$, обеспечивающих наибыстрейшую сходимость φ к точному решению системы (4.2.52).

Циклический итерационный метод (4.2.53) с длиной цикла $s \gg 1$ может быть описан (см. § 4.1) как стационарный итерационный метод вида

$$\varphi^{j+\frac{i}{s}} = \varphi^{j+\frac{i-1}{s}} - \tau_i(A\varphi^{j-\frac{i-1}{s}} - f), \quad i = 1, 2, \ldots, s, \tag{4.2.54}$$

с оператором шага

$$\widetilde{T}_\tau = \prod_{i=1}^{s}(E - \tau_i A). \tag{4.2.55}$$

Так как матрица A симметрична, то

$$\beta(\widetilde{T}_\tau) = \|\widetilde{T}_\tau\|_2 = \max_n \left|\prod_{i=1}^{s}(1 - \tau_j \lambda_n(A))\right|, \tag{4.2.56}$$

а согласно результатам 4.2.2, оптимизация метода (4.2.53) сводится к минимизации $\beta(\widetilde{T}_\tau)$ по параметрам $\tau_1, \ldots, \tau_s$. В случае $s = 1$ точное решение этой задачи было дано в 4.2.1. В случае же $s > 1$ точное решение задачи невозможно, так как простого знания границ спектра A для этого недостаточно, а определение всех собственных чисел A — более сложная задача, чем решение системы (4.2.52) простейшим итерационным методом. Поэтому с целью оптимизации метода осуществляют минимизацию функции

$$q_s(\tau) = \max_{\alpha \le \lambda \le \beta} \left|\prod_{i=1}^{s}(1 - \tau_i \lambda)\right|, \tag{4.2.57}$$

где $0 < \alpha = \alpha(A) \le \beta = \beta(A)$, которая мажорирует сверху величину $\beta(\widetilde{T}_\tau)$ и достаточно хорошо ее приближает на практике.

Наряду с задачей минимизации $q_s(\tau)$ рассмотрим задачу построения многочлена $\widetilde{P}_s(\lambda)$ степени s, являющегося решением задачи

$$\min_{P_s(\lambda) \in Q_s} \max_{\alpha \le \lambda \le \beta} |P_s(\lambda)| = \max_{\alpha \le \lambda \le \beta} \left|\widetilde{P}_s(\lambda)\right|, \tag{4.2.58}$$

где Q_s — множество всех $P_s(\lambda)$ степени s, удовлетворяющих условию $P_s(0) = 1$. Решение последней задачи было дано А. А. Марковым в виде

$$\widetilde{P}_s(\lambda) = \frac{T_s\left(\dfrac{\beta + \alpha - 2\lambda}{\beta - \alpha}\right)}{T_s\left(\dfrac{\beta + \alpha}{\beta - \alpha}\right)}, \tag{4.2.59}$$

где

$$T_s(x) = \begin{cases} \cos(s \arccos x), & x \in [-1, 1], \\ \\ \operatorname{ch}(s \operatorname{arcch} x), & x \notin [-1, 1], \end{cases} \tag{4.2.60}$$

— многочлен Чебышева степени s. Отсюда получаем

$$\widetilde{P}_s(\lambda) = \frac{\prod\limits_{i=1}^{s}(\lambda_i - \lambda)}{\prod\limits_{i=1}^{s}\lambda_i} = \left(1 - \frac{1}{\lambda_i}\lambda\right), \tag{4.2.61}$$

где λ_i — корни многочлена $T_s\left(\dfrac{\beta+\alpha-\lambda}{\beta-\alpha}\right)$ $(i = 1, 2, \ldots, s)$. При этом

$$\max_{\alpha\le\lambda\le\beta}\left|\widetilde{P}_s(\lambda)\right| = \frac{1}{T_s\left(\dfrac{\beta+\alpha}{\beta-\alpha}\right)}\max_{\alpha\le\lambda\le\beta}\left|T_s\left(\frac{\beta+\alpha-2\lambda}{\beta-\alpha}\right)\right| =$$

$$= \frac{1}{T_s\left(\dfrac{\beta+\alpha}{\beta-\alpha}\right)}\max_{-1\le t\le 1}|T_s(t)| = \frac{1}{T_s\left(\dfrac{\beta+\alpha}{\beta-\alpha}\right)}. \tag{4.2.62}$$

Вернемся теперь к проблеме оптимизации метода (4.2.53), т. е. к решению экстремальной задачи

$$q_s(\tau_{opt}) = \min_{\tau} q_s(\tau), \tag{4.2.63}$$

где τ обозначает последовательность параметров $\tau_1, \ldots, \tau_s$. Если обозначить через V_s множество многочленов $P_s(\lambda)$ степени s вида

$$P_s(\lambda) = \prod_{i=1}^{s}(1 - a_i\lambda), \tag{4.2.64}$$

то задача (4.2.63) может быть сформулирована следующим образом: найти многочлен $\hat{P}_s(\lambda)$, являющийся решением задачи

$$\min_{P_s(\lambda)\in V_s}\max_{\alpha\le\lambda\le\beta}|P_s(\lambda)| = \max_{\alpha\le\lambda\le\beta}\left|\hat{P}_s(\lambda)\right|. \tag{4.2.65}$$

Так как $V_s \subset Q_s$, то очевидно, что

$$\max_{\alpha\le\lambda\le\beta}\left|\widetilde{P}_s(\lambda)\right| \le \max_{\alpha\le\lambda\le\beta}\left|\hat{P}_s(\lambda)\right|. \tag{4.2.66}$$

Поскольку решением задачи (4.2.58) является многочлен $\widetilde{P}_s(\lambda)$ вида (4.2.64), т. е. принадлежащий V_s, то в (4.2.66) выполняется строгое равенство. Учитывая это, приходим к выводу, что в качестве $\hat{P}_s(\lambda)$ может быть выбран многочлен $\widetilde{P}_s(\lambda)$. Окончательно отметим, что вы-

бор

$$\hat{P}_s(\lambda) = \widetilde{P}_s(\lambda) \tag{4.2.67}$$

единственен, так как единственно решение задачи (4.2.58).

Если через $\kappa_s = (\sigma_1, \sigma_2, \ldots, \sigma_s)$, где $1 \leq \sigma_i \leq s$ и $\sigma_i \neq \sigma_k$ при $i \neq$ $\neq k$ $(i, k = 1, 2, \ldots, s)$, обозначить целочисленную перестановку порядка s, то из (4.2.62) следует, что для оптимального итерационного метода (4.2.54)

$$\tau_i = \frac{1}{\lambda_{\sigma_i}}, \quad i = 1, 2, \ldots, s, \tag{4.2.68}$$

где

$$\lambda_i = \frac{1}{2}[\beta + \alpha - (\beta - \alpha)x_i], \quad i = 1, 2, \ldots, s,$$

а x_i — корни многочлена $T_s(x)$. Итерационный метод (4.2.54) с выбранными по формуле (4.2.68) параметрами мы будем называть *чебышевским итерационным методом*.

Оценим теперь скорость сходимости чебышевского итерационного метода.

Покажем, что значение длины цикла s может быть выбрано из условия, чтобы s итераций метода (4.2.53) (одна итерация по методу (4.2.54)) обеспечивали уменьшение начальной ошибки в $1/\varepsilon$ раз. Рассмотрим уравнение

$$\varepsilon = \frac{1}{T_s\left(\dfrac{\beta + \alpha}{\beta - \alpha}\right)}.$$

Если ввести обозначения

$$t_0 = \frac{\beta + \alpha}{\beta - \alpha}, \quad \gamma = t_0 - \sqrt{t_0^2 - 1}$$

и использовать соотношения

$$T_s(t_0) = \frac{\left(t_0 + \sqrt{t_0^2 - 1}\right)^s + \left(t_0 + \sqrt{t_0^2 - 1}\right)^{-s}}{2},$$

$$t_0 - \sqrt{t_0^2 - 1} = \frac{1}{t_0 + \sqrt{t_0^2 - 1}},$$

то после несложных преобразований имеем

$$\varepsilon = \frac{2\gamma^s}{\gamma^{2s} + 1}.$$

Рассматривая последнее соотношение как квадратное уравнение относительно γ^s и учитывая, что $\gamma < 1$, получим

$$\gamma^s = \frac{1-\sqrt{1-\varepsilon^2}}{\varepsilon} = \frac{\varepsilon}{1+\sqrt{1-\varepsilon^2}}$$

и, следовательно,

$$s = \frac{\ln \dfrac{\varepsilon}{1+\sqrt{1-\varepsilon^2}}}{\ln \gamma}.$$

В случае $\varepsilon \ll 1$ и $p(A) \gg 1$ это выражение принимает вид

$$s \approx \frac{|\ln \varepsilon|\sqrt{p}}{2},$$

откуда следуют соотношения

$$R(\widetilde{T}_{\tau_{opt}}) \approx \frac{2}{\sqrt{p}},$$

$$W \approx W_0 |\ln \varepsilon| \frac{\sqrt{p}}{2}. \tag{4.2.69}$$

Из этих соотношений вытекает, что асимптотически по скорости сходимости чебышевский итерационный метод в $\sqrt{p}$ раз лучше простейшего одношагового метода, но в два раза уступает методу последовательной релаксации.

Как следует из изложенного, для реализации чебышевского итерационного метода нам необходимо знать границы α и β спектра матрицы A. Эта проблема весьма актуальна, поскольку для значений $x \notin [-1, 1]$ наблюдается быстрый рост многочленов Чебышева и, следовательно, ошибка в определении границ спектра может привести к медленной сходимости процесса.

Верхнюю границу спектра A, как правило, определяют с помощью теоремы Гершгорина. Наиболее же сложная проблема — определение нижней границы. В ряде случаев здесь возможны априорные оценки, но, как правило, приходится рассматривать дополнительный итерационный процесс, например метод Люстерника, описанный в первой главе, или метод минимальных итераций Ланцоша.

Второй важной проблемой чебышевского итерационного метода является проблема упорядочения параметров. Дело в том, что при произвольном порядке использования параметров τ_i в итерационном

процессе возникает неустойчивость при его численной реализации. Вопросами упорядочения параметров занимались многие исследователи, но лишь недавно эта проблема была решена в работах советских математиков. Опишем два простейших алгоритма упорядочения параметров.

Пусть $s = 2^r$, где r — некоторое положительное целое, а $n_{2^{r-1}} = (\sigma_1, \sigma_2, \ldots, \sigma_{s/2})$ — порядок номеров x_i, установленный для $s = 2^{r-1}$ в соответствии с описываемым алгоритмом. Потребуем теперь, чтобы в случае $s = 2^r$ порядок номеров x_i был следующим:

$$n_{2^r} = (\sigma_1, s + 1 - \sigma_1, \sigma_2, s + 1 - \sigma_2, \ldots, \sigma_{s/2}, s + 1 - \sigma_{s/2}).$$

Так, например, для $s = 16$ имеем

$$n_{16} = (1, 16, 8, 9, 4, 13, 5, 12, 2, 15, 7, 10, 3, 14, 6, 11).$$

Рассмотрим второй алгоритм. Пусть $s = 3^r$, где r — некоторое положительное целое, а $n_{3^{r-1}} = (\sigma_1, \sigma_2, \ldots, \sigma_{s/3})$ — порядок номеров x_i, установленный для $s = 3^{r-1}$ в соответствии со вторым алгоритмом. Тогда в случае $s = 3^r$ потребуем, чтобы порядок номеров x_i был следующим:

$$n_{3^r} = (\sigma_1, 2 \cdot 3^{r-1} + \sigma_1, 2 \cdot 3^{r-1} + 1 - \sigma_1, \ldots, 2 \cdot 3^{r-1} + 1 - \sigma_{s/3}).$$

Так, например, для $s = 9$ второй алгоритм дает

$$n_9 = (1, 7, 6, 3, 9, 4, 2, 8, 5).$$

При описанных алгоритмах операторы шага $T_i = E - \tau_i A$ с большой нормой достаточно равномерно распределяются среди операторов, уменьшающих норму ошибки.

Итерационный метод (4.2.53) был описан как оптимальный для заданного s. Новый класс итерационных процессов — устойчивые бесконечно продолжаемые оптимальные методы типа (4.2.53) с чебышевскими параметрами — позволяет продолжить после s заданных итераций метод (4.2.53) так, чтобы он был устойчивым и для некоторых $j = j_k$ $(j_k \to \infty, j_k > s)$ снова становился оптимальным. Изложим алгоритм построения параметров такого метода для одного

частного случая. Так как

$$\cos 3x = \cos x(2\cos 2x - 1),$$

то для многочленов $T_s(t)$ и $T_{3s}(t)$ справедливо соотношение

$$T_{3s}(t) = T_s(t)(2T_{2s}(t) - 1),$$

которое показывает, что множество корней многочлена $T_{3s}(t)$ состоит из множества (4.2.61) — корней многочлена $T_s(t)$ и множества корней многочлена $2T_{2s}(t) - 1$:

$$\widetilde{x}_i = \cos\frac{2i-1}{6s}\pi, \quad 2i \neq 1 (\mathrm{mod}\, 3), \quad 1 \leq i \leq 3s.$$

Следовательно, после s итераций, в которых в формулах (4.2.53) были использованы параметры (4.2.68), мы продолжим итерационный процесс далее, взяв в (4.2.68) за x_i (λ_i выражаются через x_i) соответствующим образом перемешанные параметры $\widetilde{x}_i$, то при $j = 3s$ снова получим оптимальный метод. Продолжая процесс образования оптимальных параметров аналогичным образом, мы получим бесконечную последовательность x_i, для которой метод (4.2.53) становится оптимальным при $j = 3^r s$.

Приведем формулы, определяющие порядок использования x_i, когда $s = 2$. Полагаем сначала $x_1 = -2^{-1/2}$, $x_2 = -x_1$. Пусть порядок последовательности x_i $(i = 1, \ldots, 2 \cdot 3^{r-1})$ построен. Отрезок последовательности x_i $(i = 2 \cdot 3^{r-1} + 1, \ldots, 2 \cdot 3^r)$ построим следующим образом: используя перестановку $n_{3^{r-1}}$, образуем величины

$$t_{j-1} = \sin\frac{2(\sigma_j + [\sigma_{j/2}]) - 1}{4 \cdot 3^r}\pi, \quad j = 1, 2, \ldots, 3^{r-1}.$$

Тогда

$$x_{2\cdot 3^{r-1}+4j+1} = -t_j, \quad x_{2\cdot 3^{r-1}+4j+2} = t_j,$$

$$x_{2\cdot 3^{r-1}+4j+3} = -\sqrt{1-t_j^2}, \quad x_{2\cdot 3^{r-1}+4j+4} = \sqrt{1-t_j^2},$$

$$j = 0, 1, \ldots, 3^{r-1} - 1.$$

Затем по перестановке n_{3^r} вычисляем x_i $(i = 2 \cdot 3^r + 1, \ldots, 2 \cdot 3^{r+1})$ и т. д.

Анализ чебышевского итерационного метода (4.2.53) показывает, что при его оптимизации нами нигде не используется свойство (4.2.56), а осуществляется минимизация функции $q_s(\tau)$, которая зависит только от границ спектра и при построении которой принимается во внимание только вещественность собственных чисел A. В силу сказанного, при оптимизации рассматриваемого метода достаточно ограничиться требованием положительности собственных чисел и полноты собственных векторов A и отказаться от требования симметричности.

Сделанные замечания позволяют построить теорию оптимизации для итерационных методов вида

$$\varphi^{j+1} = \varphi^j - \tau_j H(A\varphi^j - f) \tag{4.2.70}$$

в предположении, что все собственные числа матрицы HA вещественны и принадлежат отрезку $[\alpha, \beta]$, где $0 < \alpha = \alpha(HA) \le \beta = \beta(HA)$.

Покажем, что перечисленные требования относительно спектра матрицы HA будут выполнены, если матрицы H и A симметричны и положительно определены. С этой целью введем понятие положительного квадратного корня из симметричной и положительно определенной матрицы D. Пусть $\{v_n\}$ — полная ортонормированная система собственных векторов D, соответствующих ее собственным числам $\{d_n\}$. Тогда существует ортогональная матрица P, столбцами которой являются векторы $\{v_k\}$, такая, что

$$D = PD_0P^*,$$

где D_0 — диагональная матрица с собственными числами $\{d_n\}$ матрицы D по диагонали. Определим матрицу $D_0^{1/2}$ как диагональную матрицу с положительными числами $\{d_n^{1/2}\}$ по диагонали, удовлетворяющую равенству

$$D_0^{1/2}D_0^{1/2} = D_0.$$

Очевидно, что по матрице D_0 матрица $D_0^{1/2}$ определяется однозначно. Теперь матрица $D^{1/2}$ определяется соотношением

$$D^{1/2} = PD_0^{1/2}P^*.$$

Нетрудно видеть, что матрица $D^{1/2}$ симметрична, положительно определена (так как она симметрична и ее собственные числа $\{d_n^{1/2}\}$ положительны) и удовлетворяет равенствам

$$D^{1/2}D^{1/2} = [D^{1/2}]^2 = D.$$

Основываясь на изложенном, видим, что матрица HA подобна симметричной и положительно определенной матрице $S = A^{1/2}HA^{1/2}$:

$$A^{1/2}[HA][A^{1/2}]^{-1} = A^{1/2}HA^{1/2}.$$

Отсюда следует, что все собственные значения матрицы HA вещественны и положительны.

Покажем теперь, что при сделанных предположениях матрица HA обладает полной системой собственных векторов (это позволяет применять для решения системы не только циклические, но и бесконечно продолжаемые чебышевские итерационные методы). Обозначим через $\{v_n\}$ полную ортонормированную систему собственных векторов матрицы S, соответствующих ее положительным собственным числам $\{\mu_n\}$:

$$Sv_n = \mu_n v_n.$$

Умножая это соотношение на матрицу $[A^{1/2}]^{-1}$ и используя обозначение $w_n = [A^{-1/2}]^{-1}v_n$, получим

$$HAw_n = \mu_n w_n.$$

Таким образом, $\{\mu_n\}$ являются собственными числами матрицы HA, а $\{w_n\}$ — соответствующими им собственными векторами. Так как система $\{v_n\}$ образует базис в исходном пространстве векторов, а система $\{w_n\}$ получена из системы $\{v_n\}$ при помощи невырожденного преобразования, то $\{w_n\}$ также образует базис ,а следовательно, матрица HA обладает полной системой собственных векторов.

Рассмотрим еще один важный случай, когда собственные числа матрицы HA вещественны и положительны. Предположим, что матрица A имеет вид

$$A = \left\| \begin{array}{cc} E_1 & -S_1 \\ & \\ -R_2 & E_2 \end{array} \right\|, \qquad (4.2.71)$$

где E_1 и E_2 — единичные мартицы порядка n_1 и n_2 соответственно, а S_1 и R_2 являются матрицами порядка $n_1 \times n_2$ и $n_2 \times n_1$. Кроме того, предположим, что все собственные числа матрицы

$$B = \begin{Vmatrix} 0 & S_1 \\ R_2 & 0 \end{Vmatrix}$$

вещественны и меньше единицы по модулю (т. е. $|\lambda(B)| < 1$), а матрица H определена соотношением

$$H = \begin{Vmatrix} E_1 & 0 \\ -R_2 & E_2 \end{Vmatrix}^{-1} = \begin{Vmatrix} E_1 & 0 \\ R_2 & E_2 \end{Vmatrix}. \tag{4.2.72}$$

Такой выбор H соответствует методу последовательной верхней релаксации (4.2.38) с параметром $\tau = 1$ (см. 4.2.3). Очевидно, что любое собственное число матрицы

$$HA = \begin{Vmatrix} E_1 & -S_1 \\ 0 & E_2 - R_2 S_1 \end{Vmatrix}$$

либо равно единице, либо является собственным числом матрицы $E_2 - R_2 S_1$, все собственные числа которой вещественны и положительны. Для доказательства последнего факта достаточно заметить, что любое собственное число матрицы $R_2 S_1$ в то же время является собственным числом матрицы

$$B^2 = \begin{Vmatrix} S_1 R_2 & 0 \\ 0 & R_2 S_1 \end{Vmatrix}$$

(собственные числа B^2 положительны, так как они являются квадратами вещественных собственных чисел матрицы B) и справедливы неравенства

$$\lambda(R_2 S_1) = \lambda(B^2) = \lambda^2(B) < 1.$$

Таким образом, все собственные числа матрицы HA вещественны и положительны, что позволяет осуществлять оптимизацию ме-

тода (4.2.70) на основе разработанной выше теории выбора чебышевских параметров. Нужно отметить, что при дополнительном предположении симметричности матрицы A ($R_2 = S_1^*$) матрица HA обладает полной системой собственных векторов.

В заключение оценим скорость сходимости итерационного метода (4.2.70) с чебышевским выбором параметров, когда матрицы A и H определяются соотношениями (4.2.71) и (4.2.72), через величину

$$p = p(A) = \frac{\beta(A)}{\alpha(A)}$$

в предположении $p \gg 1$. Используя результаты из 4.2.3 и конкретный вид матрицы HA, нетрудно заметить, что

$$\alpha(HA) = 1 - \beta(B^2) = 1 - \beta^2(B) = 1 - (1 - \alpha(A))^2 =$$

$$= \alpha(A)(2 - \alpha(A)) \approx 2\alpha(A), \quad \beta(HA) = 1$$

и, следовательно,

$$P(HA) = \frac{\beta(HA)}{\alpha(HA)} \approx \frac{1}{2\alpha(A)} \approx \frac{1}{4}\frac{\beta(A)}{\alpha(A)} = \frac{1}{4}p.$$

Отсюда, согласно (4.2.69), для процесса (4.2.70) имеем

$$\begin{aligned} R(\widetilde{T}_{opt}) &\approx \frac{4}{\sqrt{p}}, \\ W &\approx W_0 |\ln \varepsilon| \frac{\sqrt{p}}{4}. \end{aligned} \tag{4.2.73}$$

Тем самым показано, что асимптотическая скорость сходимости рассматриваемого варианта чебышевского итерационного метода равна асимптотической скорости сходимости метода последовательной верхней релаксации с параметром τ_{opt}.

4.2.5. Сравнение скорости сходимости итерационных методов для систем разностных уравнений

В предыдущих пунктах мы установили оценки скорости сходимости ряда итерационных методов, используя число обусловленности матрицы A. Здесь мы рассмотрим применение этих итерацион-

ных методов для решения систем конечно-разностных уравнений, аппроксимирующих двумерные уравнения эллиптического типа.

Пусть система конечно-разностных уравнений записана в виде

$$\varphi_{k,l} = a_{k,l}\varphi_{k-1,l} + b_{k,l}\varphi_{k,l-1} + c_{k,l}\varphi_{k+1,l} + d_{k,l}\varphi_{k,l+1} + f_{k,l}, \tag{4.2.74}$$

где (k,l) пробегают некоторое множество индексов Q.

Предположим, что каждому узлу сеточной области, на которой осуществлялась аппроксимация задачи, соответствуют только одно значение $\varphi_{k,l}$ и только одно уравнение, т. е. число уравнений и число неизвестных равно числу узлов сетки. Тогда, чтобы перейти к матричной формулировке системы конечно-разностных уравнений, сначала осуществляют нумерацию узлов сетки, затем в соответствии с нумерацией узлов сетки нумеруют компоненты $\{\varphi_{k,l}\}$ и в такой же последовательности располагают конечно-разностные уравнения.

В качестве примера рассмотрим систему конечно-разностных уравненй (4.2.74), когда сеточной областью являются квадратная сетка, а областью определения решения — квадрат со стороной единица. В этом случае в (4.2.74) $k,l = 1,2,\ldots,m$, а порядок матрицы A соответствующей системы

$$A\varphi = f \tag{4.2.75}$$

равен $n = m^2$.

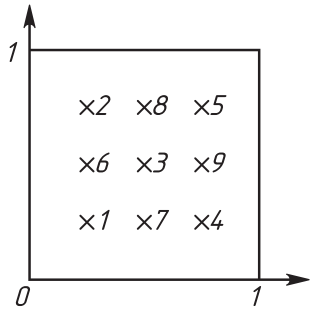


Рис. 4.2.

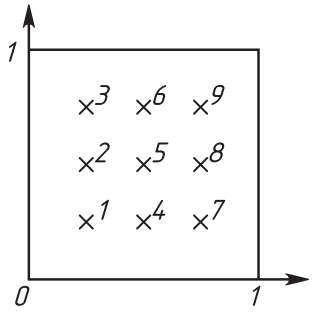


Рис. 4.3.

В качестве первого способа нумерации рассмотрим так называемый *«шахматный способ»*, когда все компоненты разбиваются на две группы: к первой относятся компоненты с четными значениями $k+l$, а ко второй — с нечетными значениями $k+l$. После этого нумеруются сначала компоненты первой группы, а затем — второй. В

конкретном случае $m = 3$ нумерация имеет вид, представленный на рис. 4.2.

Соответствующая матрица $A = A_1$ системы (4.2.75) определяется соотношением (которое справедливо уже для произвольного m)

$$A_1 = \begin{Vmatrix} E_1 & -S_1 \\ -R_2 & E_2 \end{Vmatrix}, \tag{4.2.76}$$

где порядок матрицы E_1 равен числу компонент в первой группе, а порядок E_2 — числу компонент во второй группе.

Второй способ нумерации для $m = 3$ приведен на рис. 4.3.

Объединяя в одну группу все компоненты при одинаковых значениях k, получим m групп компонент. Соответствующая матрица $A = A_2$ системы (4.2.75) имеет вид $(m = 3)$

$$A_2 = \begin{Vmatrix} A_{11} & A_{12} & 0 \\ A_{21} & A_{22} & A_{23} \\ 0 & A_{32} & A_{33} \end{Vmatrix}. \tag{4.2.77}$$

Умножая систему (4.2.75) с такой матрицей A на матрицу D^{-1}, где

$$D = \begin{Vmatrix} A_{11} & 0 & 0 \\ 0 & A_{22} & 0 \\ 0 & 0 & A_{33} \end{Vmatrix} \tag{4.2.78}$$

— блочно-диагональная матрица, приходим к системе

$$A_3\varphi \equiv \begin{Vmatrix} E_1 & -S_1 & 0 \\ -R_2 & E_2 & -S_2 \\ 0 & -R_3 & E_3 \end{Vmatrix}, \quad \varphi = F, \tag{4.2.79}$$

где

$$F = D^{-1}f,$$

$$S_l = -A_{ll}^{-1}A_{l,l+1},$$

$$R_l = -A_{ll}^{-1}A_{l,l-1}, \quad l = 1, 2, 3.$$

При произвольном $k = m$ матрица A_3 имеет вид (4.2.32).

Если операторы исходных эллиптических дифференциальных задач симметричны, положительно определены и аппроксимация осуществлялась либо с помощью интегральных соотношений баланса, либо вариационными методами, то построенные матрицы A обладают полной системой собственных векторов, а их собственные числа положительны. В исключительных случаях (например, прямоугольная область и постоянные коэффициенты) границы спектра могут быть вычислены точно, как это сделано в главе 1 для разностного оператора Лапласа в квадрате, либо оценены (иногда слишком грубо), либо вычислены с помощью какого-нибудь итерационного метода (методы Люстерника, Ланцоша и т. д.).

Таблица 4.1.

Методы	Асимптотическая скорость сходимости $R(T)$
Простейший итерационный метод	$\frac{2}{p} \approx \frac{\pi^2 h^2}{2}$
Чебышевский итерационный метод	$\frac{2}{\sqrt{p}} \approx \pi h$
Метод последовательной-верхней релаксации	$\frac{4}{\sqrt{p}} \approx 2\pi h$
Чебышевский итерационныйметод (4.2.70)—(4.2.72)	$\frac{4}{\sqrt{p}} \approx 2\pi h$

В качестве примера рассмотрим аппроксимацию задачи Дирихле для уравнения Пуассона в единичном квадрате с помощью обычных пятиточечных разностных отношений на квадратной сетке с ша-

гом $h \ll 1$. При этом мы ограничимся случаем, когда матрица A системы (4.2.75) имеет вид (4.2.76). Тогда, как показывают непосредственные расчеты,

$$\alpha(A) = 2\sin^2\frac{\pi h}{2} \approx \frac{\pi^2 h^2}{2},$$
$$\beta(A) = 2 - \alpha(A) \approx 2 - \frac{\pi^2 h^2}{2}, \tag{4.2.80}$$
$$p = p(A) = \frac{\beta(A)}{\alpha(A)} \approx \frac{4}{\pi^2 h^2}.$$

Объединим теперь данные по различным итерационным методам в одну большую таблицу, учитывая, что в силу $h \ll 1$ матрица A системы (4.2.75) плохо обусловлена.

4.3. Нестационарные итерационные методы

В этом параграфе будут рассмотрены нестационарные итерационные методы, осуществляющие последовательную минимизацию некоторого квадратичного функционала.

4.3.1. Теоремы сходимости

Пусть даны система линейных алгебраических уравнений

$$A\varphi = f \tag{4.3.1}$$

и квадратичный функционал

$$J(\varphi) = (D(\varphi - \varphi^*), \varphi - \varphi^*), \tag{4.3.2}$$

где $\varphi^* = A^{-1}f$ — точное решение системы (4.3.1), а D — симметричная положительно определенная матрица.

Так как $J(\varphi) > 0$ для любого $\varphi \neq \varphi^*$ и $J(\varphi^*) = 0$, то задача решения системы (4.3.1) эквивалентна задаче минимизации функционала (4.3.2), т. е. нахождения вектора φ^*, минимизирующего $J(\varphi)$. Если $D = A^*A$, то функционал

$$J(\varphi) = (A\varphi - f, A\varphi - f) = \|A\varphi - f\|_2^2 \tag{4.3.3}$$

называется *функционалом квадрата невязки*.

Если предположить, что A — симметричная положительно определенная матрица и $D = A$, то

$$J(\varphi) = J_1(\varphi) + (A\varphi^*, \varphi^*),$$

где

$$J_1(\varphi) = (A\varphi, \varphi) - 2(\varphi, f). \tag{4.3.4}$$

Таким образом, $J(\varphi)$ с точностью до константы $(A\varphi^*, \varphi^*) = (f, \varphi^*)$ совпадает с известным вариационным функционалом $J_1(\varphi)$. Заметим, что функционалы (4.3.3) и (4.3.4) не зависят от искомого вектора φ^*.

Сделанные замечания позволяют сформулировать новые принципы оптимизации итерационных методов, которые мы назовем вариационными принципами. Рассмотрим итерационный метод

$$\varphi^{j+1} = \varphi^j - H_\tau(A\varphi^j - f), \tag{4.3.5}$$

матрица H_τ которого зависит от параметров $\tau_1, \ldots, \tau_s$. Предположим, что для значений $\tau_1, \ldots, \tau_s$ из некоторого множества Q метод (4.3.5) сходится, причем последовательность значений $\{\varphi^j\}$ осуществляет последовательную минимизацию функционала (4.3.2). В предыдущем параграфе мы выбирали значения $\tau_1, \ldots, \tau_s$ из условия минимума спектрального радиуса $\beta(T_\tau)$ матрицы шага $T_\tau = E - H_\tau A$. В настоящем параграфе будут рассмотрены методы, для которых параметры $\tau_1, \ldots, \tau_s$ выбираются из условия максимальной минимизации на каждом шаге функционала (4.3.2).

Так как матрица D положительно определена, то нетрудно видеть, что соотношение

$$\|\psi\|_D = (D\psi, \psi)^{1/2} \tag{4.3.6}$$

определяет норму в пространстве векторов ошибок. Соответственно, соотношение

$$\|B\|_D = \sup_{\psi \neq 0} \frac{\|B\psi\|_D}{\|\psi\|_D} \tag{4.3.7}$$

определяет норму матрицы B. Введенную норму часто называют D-нормой. Ранее мы предположили, что при $(\tau_1, \ldots, \tau_s) \in Q$ функционал (4.3.2) с некоторой матрицей D последовательно минимизируется.

Это означает, что для $\psi^j = \varphi^j - \varphi^*$

$$\|T_\tau \psi^j\|_D < \|\psi^j\|_D. \tag{4.3.8}$$

Так как для стационарного итерационного метода матрица T_τ постоянна, то из последнего неравенства следует, что для любого вектора $\psi^j \neq 0$

$$\|T_\tau z^j\|_D < 1,$$

где $z^j = \psi^j / \|\psi^j\|_D$. Далее, поскольку множество $v = \{z : \ \|z\|_D = 1\}$ является ограниченным замкнутым множеством[2)], то величина

$$\sup_{\|z\|_D=1} \|T_\tau z\|_D = \|T_\tau\|_D$$

достигается на некотором векторе z_0. Отсюда и из (4.3.8) вытекает

$$\|T_\tau\|_D = \|T_\tau z_0\|_D < 1. \tag{4.3.9}$$

Итак, мы показали, что при сделанном предположении $(\tau_1, \ldots, \tau_s) \in Q : \ \|T_\tau\|_D < 1$.

Сформулируем теперь нестационарный метод, соответствующий методу (4.3.5):

$$\varphi^{j+1} = \varphi^j - H_j(A\varphi^j - f), \tag{4.3.10}$$

где $H_j = H(\tau_1^{(j)}, \ldots, \tau_s^{(j)})$ и параметры $\{\tau_i^{(j)}\}$ удовлетворяют уравнению

$$J(\varphi^j - H_j(A\varphi^j - f)) = \inf_{\tau_1,\ldots,\tau_s} J(\varphi^j - H_\tau(A\varphi^j - f)). \tag{4.3.11}$$

Если при любых $\tau_1, \ldots, \tau_s \in Q$, для которых метод (4.3.5) сходится, выполняется соотношение

$$\|T_\tau\|_D = \beta(T_\tau),$$

то асимптотическая скорость сходимости соответствующего нестационарного итерационного метода обычно не ниже, чем асимптотическая скорость стационарного метода с оптимальными параметрами.

[2)] Здесь и в дальнейшем используются обозначения вида $v = \{z : \ \|z\|_D = 1\}$, которые означают, что множество v состоит из элементов z, каждый из которых удовлетворяет условию $\|z\|_D = 1$.

Как было показано в 4.2.2,

$$R(T_\tau) = -\ln \beta(T_\tau). \tag{4.3.12}$$

С другой стороны,

$$\left[\frac{\left\|\prod_{j=1}^{k} T_j\psi^0\right\|_D}{\|\psi^0\|_D}\right]^{1/k} \le \|\widetilde{T}\|_D \le \|T_\tau\|_D,$$

где $\widetilde{T}$ — нелинейный оператор метода (4.3.10), (4.3.11). Отсюда получаем

$$R(\widetilde{T}) \ge R(T_\tau). \tag{4.3.13}$$

Так как это неравенство справедливо для любых $\tau_1, \ldots, \tau_s \in Q$, то утверждение доказано.

Из последнего утверждения следует, что если параметры оптимизации выбирать из условия минимума D-нормы оператора перехода T_τ, то вариационная оптимизация на любой итерации обеспечивает (по оценке) более быструю скорость сходимости.

4.3.2. Метод минимальных невязок

Выберем

$$D = A^*A \quad \text{и} \quad H = \tau E \tag{4.3.14}$$

и предположим, что $A = A^* > 0$.

В 4.2.1 было показано, что в этом случае при $\tau \in (0, 2/\beta(A))$ итерационный метод

$$\varphi^{j+1} = \varphi^j - \tau(A\varphi^j - f) \tag{4.3.15}$$

сходится. При этом простейший спектральный анализ показывает, что

$$\|E - \tau A\|_D = \|E - \tau A\|_2 = \beta(E - \tau A). \tag{4.3.16}$$

Рассмотрим соответствующий нестационарный процесс, называемый *методом минимальных невязок*:

$$\varphi^{j+1} = \varphi^j - \tau_j(A\varphi^j - f), \tag{4.3.17}$$

где

$$\tau_j = \frac{(A\xi^j, \xi^j)}{(A\xi^j, A\xi^j)} = \frac{(A\xi^j, \xi^j)}{\|A\xi^j\|_2^2} \tag{4.3.18}$$

($\xi^j = A\varphi^j - f$ — вектор невязки). Из (4.3.17) следует, что

$$\|(E - \tau_j A)\xi^j\|_D^2 = \|(E - \tau_j A)\xi^j\|_2^2 = \inf_\tau J(\varphi^j - \tau\xi^j) =$$

$$= \inf_\tau \|(E - \tau A)\xi^j\|_2^2 = \inf_\tau \{\|\xi^j\|_2^2 - 2\tau(A\xi^j, \xi^j) + \tau^2\|A\xi^j\|_2^2\}.$$

Очевидно, что τ_j находится из уравнения

$$\frac{\partial}{\partial\tau}\{\|\xi^j\|_2^2 - 2\tau(A\xi^j, \xi^j) + \tau^2\|A\xi^j\|_2^2\} = 0.$$

Так как метод (4.3.15) сходится для $\tau \in (0, 2/\beta)$, где $\beta = \beta(A)$, то, согласно утверждению из 4.2, метод минимальных невязок сходится, причем

$$\|\widetilde{T}\| \le \frac{\beta - \alpha}{\beta + \alpha}. \tag{4.3.19}$$

Из соотношения для последовательности норм невязок метода (4.3.17), (4.3.18)

$$\|\xi^{j+1}\|_2^2 = \|\xi^j\|_2^2 - \frac{(A\xi^j, \xi^j)^2}{\|A\xi^j\|_2^2}$$

легко видеть, что для монотонного убывания норм невязок в пространстве вещественных векторов достаточно положительной определенности матрицы A в этом пространстве (симметричность не обязательна). Это обстоятельство послужило основой для формулировки теоремы сходимости нестационарных итерационных методов и, в частности, доказательства сходимости метода минимальных невязок для систем уравнений с положительно определенными, но не симметричными матрицами.

Одной из важных в практическом отношении особенностей метода минимальных невязок является то, что на первых итерациях метод сходится значительно быстрее, чем асимптотически. Асимптотически ошибка метода наискорейшего спуска является линейной комбинацией только двух собственных векторов матрицы A, соответствующих ее собственным числам $\alpha(A)$ и $\beta(A)$. Аналогичная ситуация наблюдается в методе минимальных невязок. Таким образом, асимптотическое свойство итерационного метода выходить на свою наи-

худшую скорость сходимости (при фиксированных параметрах) сохраняется и для нестационарных методов.

Для ускорения сходимости метода минимальных невязок целесообразно время от времени использовать один шаг двухшагового метода минимальных невязок, формулы которого имеют вид

$$\varphi^{j+1} = \varphi^j - \tau_j(A\varphi^j - f) - \gamma_j A(A\varphi^j - f), \tag{4.3.20}$$

где τ_j и γ_j выбираются как решение системы двух уравнений:

$$\begin{aligned} &\frac{\partial}{\partial\tau_j}\|\xi^j - \tau_j A\xi^j - \gamma_j A^2\xi^j\|_2^2 = 0, \\ &\frac{\partial}{\partial\gamma_j}\|\xi^j - \tau_j A\xi^j - \gamma_j A^2\xi^j\|_2^2 = 0. \end{aligned} \tag{4.3.21}$$

4.3.3. Метод сопряженных градиентов

Определим в исходном пространстве векторов некоторую D-норму и некоторое подпространство G_s с базисом $\{g_i\}_{i=1}^s$. Тогда задача наилучшего приближения решения $\varphi^* = A^{-1}f$ системы $A\varphi = f$ на многообразии

$$U_s^0 = \varphi^0 + G_s = \{\varphi :\ \varphi = \varphi^0 + \psi,\ \psi \in G_s\}$$

формулируется следующим образом. Требуется найти такой вектор $\widehat{\psi} \in G_s$, что

$$\begin{aligned} \|\varphi^* - (\varphi^0 + \widehat{\psi})\|_D = \min_{\psi\in G_s}\|\varphi^* - (\varphi^0 + \psi)\|_D = \\ = \min_{\alpha_1,\alpha_2,\ldots,\alpha_s}\left\|\varphi^* - \varphi^0 - \sum_{i=1}^s \alpha_i g_i\right\|_D. \end{aligned} \tag{4.3.22}$$

Система уравнений для определения коэффициентов $\{\alpha_i^*\}$ разложения

$$\widehat{\psi} = \sum_{i=1}^s \alpha_i^* g_i$$

имеет вид

$$B\alpha = F,$$

где $B = (b_{ij})$ — матрица порядка s с элементами

$$b_{ij} = (g_i, g_j)_D, \quad i, j, = 1, 2, \ldots, s,$$

и $F = (F_1, \ldots, F_s)$ — вектор с компонентами

$$F_i = (\varphi^* - \varphi^0, g_i)_D, \quad i = 1, 2, \ldots, s.$$

Из предыдущего видно, что наиболее простым (с точки зрения реализации процесса) является случай, когда

$$(Dg_i, g_j) = \delta_{ij}\|g_i\|_D^2,$$

где δ_{ij} — символ Кронекера, т. е. когда $\{g_i\}$ будет D-ортогональным базисом пространства G_s. Если последнее условие выполнено, то

$$\alpha_i^* = \frac{(\varphi^* - \varphi^0, g_i)_D}{\|g_i\|_D^2} = \frac{(D(\varphi^* - \varphi^0), g_i)}{\|g_i\|_D^2}, \quad i = 1, 2, \ldots, s. \tag{4.3.23}$$

Достаточным условием осуществимости процесса (4.3.23) является требование, чтобы вектор $D\varphi^*$ был известен. Это требование выполняется, например, либо в случае $D = A$, если $A = A^* > 0$, либо в случае $D = A^*A$ для произвольной A.

Конкретизируем вариационную задачу (4.3.22) с целью изучения одного класса методов. Предположим, что подпространство G_s является линейной оболочкой системы линейно независимых векторов

$$\{A^i(\varphi^0 - \varphi^*)\}_{i=1}^s = \{A^{i-1}(A\varphi^0 - f)\}_{i=1}^s,$$

а матрица A симметрична и положительно определена. Как уже было указано выше, если в заданном подпространстве G_s мы сможем найти некоторый A-ортогональный базис $\{g_i\}_{i=1}^s$, то искомое приближение φ к вектору φ^* найдется по формулам

$$\widehat{\varphi} = \varphi^0 + \sum_{i=1}^{s} \alpha_i g_i,$$

$$\alpha_i = \frac{(\varphi^* - \varphi^0, g_i)_A}{(g_i, g_i)_A} = -\frac{(A\varphi^0 - f, g_i)}{(g_i, Ag_i)}, \quad i = 1, 2, \ldots, s. \tag{4.3.24}$$

Этот процесс можно записать еще следующим образом:

$$\begin{gathered} \varphi^k = \varphi^{k-1} - \alpha_k g_k, \\ \alpha_k = \frac{(\xi^{k-1}, g_k)}{(Ag_k, g_k)}, \quad k = 1, 2, \ldots, s, \end{gathered} \tag{4.3.25}$$

где $\xi^k = (A\varphi^k - f)$ — вектор невязки и $\widehat{\varphi} = \varphi^s$.

Наиболее известным способом построения базиса в пространствах типа G_s является процесс Шмидта. Однако для матриц A высокого порядка он требует большого числа арифметических действий и большой памяти ЭВМ при численной реализации. Для случая симметричных, но не положительно определенных матриц эффективным способом построения A^2-ортогонального базиса (т. е. когда $D = A^2$) в пространстве G_s является метод минимальных итераций Ланцоша. Самым экономичным из известных способов A-ортогонализации векторов $\{A^{i-1}(A\varphi^0 - f)\}_{i=1}^s$ для симметричных и положительно определенных матриц является *метод сопряженных градиентов*, формулы которого имеют вид

$$
\begin{aligned}
g_k &= \begin{cases} \xi^0, & \text{если} \quad k = 1, \\ \xi^{k-1} - b_k g_{k-1}, & \text{если} \quad k > 1, \end{cases} \\
b_k &= \frac{(A\xi^{k-1}, g_{k-1})}{(Ag_{k-1}, g_{k-1})}, \\
\varphi^k &= \varphi^{k-1} - \alpha_k g_k, \\
\alpha_k &= \frac{(\xi^{k-1}, g_k)}{(Ag_k, g_k)}, \quad k = 1, 2, \ldots, s,
\end{aligned}
\tag{4.3.26}
$$

где $\{\xi^k = A\varphi^k - f\}_{k=1}^s$ — векторы невязки.

Докажем, что построенные по этим формулам векторы $\{g_k\}_{k=1}^s$ образуют A-ортогональный базис пространства G_s, если векторы $\{A^{k-1}\xi^0\}_{k=1}^s$ линейно независимы. Сначала покажем, что все векторы $\{g_k\}_{k=1}^s$ будут ненулевыми. В самом деле, если в (4.3.26) провести последовательное исключение, то легко видеть, что

$$
g_k = A^{k-1}\xi^0 + \sum_{i=1}^{k-1} \beta_{ki} A^{i-1}\xi^0, \quad k = 1, 2, \ldots, s, \tag{4.3.27}
$$

с некоторыми коэффициентами $\{\beta_{ki}\}$. Поэтому соотношение $g_k = 0$ будет противоречить требованию линейной независимости векторов $\{A^{k-1}\xi^0\}_{k=1}^s$.

Теперь остается показать, что векторы $\{g_k\}_{k=1}^{s}$ A-ортогональны. Предположим, что для некоторого $k \geq 2$ выполняются соотношения (для $k = 1, 2$ их выполнение устанавливается непосредственной проверкой)

$$(Ag_k, g_j) = 0, \qquad j = 1, 2, \ldots, k-1,$$

$$(\xi^k, g_j) = 0, \qquad j = 1, 2, \ldots, k, \tag{4.3.28}$$

$$\alpha_j > 0, \qquad j = 1, 2, \ldots, k,$$

и докажем их справедливость на $(k+1)$-м шаге. Для этого нам понадобятся равенства

$$Ag_j = \frac{1}{\alpha_j}[g_j + b_j g_{j-1} - g_{j+1} - b_{j+1} g_j] = \varepsilon_j g_{j+1} + \beta_j g_j + \gamma_j g_{j-1},$$

$$j = 1, 2, \ldots, k-1 \tag{4.3.29}$$

(здесь полагается $g_0 = 0$). Чтобы вывести эти соотношения, достаточно воспользоваться равенствами, полученными из (4.3.26), и условиями

$$\xi^j = \xi^{j-1} - \alpha_j A g_j,$$

$$g_{j+1} = \xi^j - b_{j+1} g_j, \quad j = 1, 2, \ldots, k-1.$$

Так как в силу предположений

$$(Ag_{k+1}, g_j) = (A\xi^k - b_{k+1} A g_k, g_j) = (A\xi^k, g_j) - b_{k+1}(Ag_k, g_j) =$$

$$= (\xi^k, Ag_j) = (\xi^k, \varepsilon_j g_{j+1} + \beta_j g_j + \gamma_j g_{j-1}) = 0, \quad j = 1, 2, \ldots, k-1,$$

и по построению $(Ag_{k+1}, g_k) = 0$, то

$$(Ag_{k+1}, g_j) = 0 \tag{4.3.30}$$

для всех $j \leq k$. Далее, поскольку $(\xi^{k+1}, g_{k+1}) = 0$ по построению, то из предположений (4.3.28) и (4.3.30) следует, что

$$(\xi^{k+1}, g_j) = (\xi^k, g_j) - \alpha_{k+1}(Ag_{k+1}, g_j) = 0 \tag{4.3.31}$$

для любого $1 \le j \le k+1$. Объединяя (4.3.30) с (4.3.31) и учитывая неравенство

$$\alpha_{k+1} = \frac{(\xi^k, g_{k+1})}{(Ag_{k+1}, g_{k+1})} = \frac{(\xi^k, \xi^k)}{(Ag_{k+1}, g_{k+1})} > 0,$$

получаем, что на $(k+1)$-м шаге все соотношения (4.3.28) выполняются. Продолжая по индукции до s-го шага, приходим к выводу об A-ортогональности системы векторов $\{g_k\}$. Таким образом, метод сопряженных градиентов решает вариационную задачу (4.3.22).

В заключение обсудим случай вырождения, когда для некоторого $k \ge 1$ система векторов $\{A^{i-1}\xi^0\}_{i=1}^k$ линейно независима, а система $\{A^{i-1}\xi_0\}_{i=1}^{k+1}$ линейно зависима, т. е.

$$-\sum_{j=0}^{k} C_j A\xi^0 \equiv \sum_{j=0}^{k} C_j A^{j+1}(\varphi^* - \varphi^0) = 0 \tag{4.3.32}$$

с некоторыми коэффициентами $\{C_j\}_{j=0}^k$, среди которых есть отличные от нуля. Коэффициент C_0 отличен от нуля, так как в противном случае, умножая (4.3.32) на A^{-1}, получаем

$$\sum_{j=0}^{k} C_j A^j(\varphi^* - \varphi^0) = 0,$$

что противоречит линейной независимости системы векторов $\{A^j(\varphi^* - \varphi^0)\}_{j=1}^k$. Из (4.3.32) имеем

$$\varphi^* - \varphi^0 = -\frac{1}{C_0}A^{-1}\sum_{j=1}^{k} C_j A^{j+1}(\varphi^* - \varphi^0) = \sum_{j=1}^{k} \frac{C_j}{C_0} A^{j-1}\xi^0,$$

что означает $\varphi^* - \varphi^0 \in G_s$. Отсюда и из (4.3.28) следует, что приближение $\widehat{\psi}$ равно вектору $\varphi^* - \varphi^0$ ($\varphi^* = \varphi^0 + \widehat{\psi} = \varphi^k$). Иначе говоря, в данном случае метод сопряженных градиентов позволяет найти точное решение системы (4.3.1) уже на k-м шаге.

Метод сопряженных градиентов, как и другие способы ортогонализации, можно широко использовать для ускорения сходимости стационарных итерационных методов. Так, например, для итерационного процесса

$$\varphi^{k+1} = \varphi^k - B(A\varphi^k - f), \quad k = 1, 2, \ldots, \tag{4.3.33}$$

с симметричными и положительно определенными матрицами A и B формулы ускорения с помощью метода сопряженных градиентов имеют следующий вид:

$$g_k = \begin{cases} B\xi^0, & \text{если} \quad k = 1, \\ B\xi^{k-1} - b_k g_{k-1}, & \text{если} \quad k > 1, \end{cases}$$

$$b_k = \frac{(AB\xi^{k-1}, g_{k-1})}{(Ag_{k-1}, g_{k-1})}, \tag{4.3.34}$$

$$\varphi^k = \varphi^{k-1} - \alpha_k g_k,$$

$$\alpha_k = \frac{(\xi^{k-1}, g_k)}{(Ag_k, g_k)}, \quad k = 1, 2, \ldots, s.$$

Метод сопряженных градиентов по своей идее (теоретически) является прямым методом, поскольку при $s > n$, где n — порядок матрицы A, система векторов $\{A^i\xi^0\}_{i=0}^{s-1}$ всегда линейно зависима, и, следовательно, при некотором $k \le n$ процесс должен заканчиваться получением точного решения. С другой стороны, при реализации метода сопряженных градиентов на ЭВМ в случае матриц высокого порядка, как правило, уже через несколько десятков итераций из-за нелинейности возникает явление численной неустойчивости процесса ортогонализации, а реальный процесс перестает отражать свойства реализуемого метода. В силу отмеченного обстоятельства анализ метода сопряженных градиентов можно проводить как анализ нестационарного итерационного метода с длиной цикла s, т. е. через каждые s шагов по методу сопряженных градиентов начальное приближение выбирается заново. При такой постановке скорость сходимости метода может быть оценена через скорость сходимости циклического чебышевского итерационного метода. Действительно, для любого $s \ge 1$

$$\min_{\alpha_1,\alpha_2,\ldots,\alpha_s} \left\| \varphi^* - \varphi^0 - \sum_{i=1}^{s} \alpha_i A^{i-1}\xi^0 \right\|_A^2 \Big/ \|\varphi^* - \varphi^0\|_A^2 \le \|\widetilde{T}_s\|_A \le$$

$$\le \left\| E - \sum_{i=1}^{s} \beta_i A^i \right\|_2^2 = \left[\beta \left(E - \sum_{i=1}^{s} \beta_i A^i \right) \right]^2, \tag{4.3.35}$$

где $\widetilde{T}_s$ — оператор, преобразующий вектор ошибки за s шагов метода сопряженных градиентов, а $\{\beta_i\}_{i=1}^s$ — произвольные вещественные числа. В частности, если положить

$$E - \sum_{i=1}^{s} \beta_i A^i = \prod_{i=1}^{s} (E - \tau_i A) \tag{4.3.36}$$

и использовать оценку (4.3.12) настоящего параграфа, то приходим к выводу, что s-циклический метод сопряженных градиентов сходится примерно так же, как s-циклический чебышевский метод, и, следовательно, для него справедливы те же самые оценки скорости сходимости. Отметим две важные особенности метода сопряженных градиентов, проявляющиеся при решении конкретных вычислительных задач. Во-первых, реализация одного шага метода сопряженных градиентов требует большее (иногда значительно) число арифметических и логических действий по сравнению с одним шагом чебышевского итерационного метода. Во-вторых, при реализации (особенно на первых итерациях) метод сопряженных градиентов в соответствующей норме значительно быстрее минимизирует A-норму вектора ошибки, чем это показывает оценка. Для иллюстрации отмеченного факта выпишем соотношение, вытекающее из (4.3.22), для метода сопряженных градиентов (способа определения подпространства G_s):

$$(A\psi^k, \psi^k) \le \max_{m \le \lambda \le M} |\lambda P_k^2(\lambda)| \cdot (\psi^0, \psi^0), \tag{4.3.37}$$

где ψ^k — вектор ошибки на k-м шаге, ψ^0 — вектор начальной ошибки, а $P_k(\lambda)$ — многочлен степени k от λ, удовлетворяющий условию $P_k(0) = 1$. Теперь если положить $M = 1$ (этого всегда можно добиться нормировкой матрицы A) и $m = 0$ (без ограничения общности), то, полагая

$$P_k(\lambda) = (-1)^k \cos[(2k+1) \arccos \sqrt{\lambda}]/[(2k+1)\sqrt{\lambda}], \tag{4.3.38}$$

получим для метода сопряженных градиентов оценку

$$(A\psi^k, \psi^k) \le \frac{(\psi^0, \psi^0)}{(2k+1)^2}. \tag{4.3.39}$$

Заметим, что в силу $m = 0$ оценка (4.3.39) справедлива и в случае вырожденной матрицы A.

4.4. Метод расщепления

Среди итерационных методов решения стационарных задач математической физики широкое применение имеет метод переменных направлений. В настоящее время известно довольно большое число различных модификаций этого метода и схем его реализации. Метод переменных направлений основывается на специальных релаксационных процессах с возможностью редукции сложной задачи к последовательности простейших. Все методы такой редукции будем называть *методами расщепления*. Именно с этих позиций и рассмотрим ставший уже классическим *метод переменных направлений* (см. § 9.2).

Изложение будем вести на примере системы линейных алгебраических уравнений

$$A\varphi = f \tag{4.4.1}$$

в предположении, что

$$A = A_1 + A_2, \tag{4.4.2}$$

где A_1 и A_2 — положительно определенные матрицы. В 4.2 нами было показано, что использование вместо итерационного процесса

$$\varphi^{j+1} = \varphi^j - \tau(A\varphi^j - f) \tag{4.4.3}$$

итерационного процесса последовательной верхней релаксации

$$B(\varphi^{j+1} - \varphi^j) = -(A\varphi^j - f) \tag{4.4.4}$$

с матрицей B, зависящей от параметра τ, позволяет значительно повысить скорость сходимости метода практически без увеличения числа действий на итерацию. Недостатком метода оказалось жесткое ограничение на вид матрицы A и свойства ее спектра.

Рассмотрим другой класс методов, когда матрица $B = B_\tau$ определяется соотношением

$$B_\tau = \frac{1}{2\tau}(E + \tau A), \tag{4.4.5}$$

а матрица A предполагается симметричной и положительно определенной. В этом случае для векторов ошибок $\psi^j = \varphi^j - \varphi^*$ имеем

$$(E + \tau A)\psi^{j+1} = (E - \tau A)\psi^j,$$

или, что эквивалентно,

$$\psi^{j+1} = T_\tau \psi^j,$$

где

$$T_\tau = (E + \tau A)^{-1}(E - \tau A) = (E - \tau A)(E + \tau A)^{-1} \tag{4.4.6}$$

— оператор шага метода (4.4.4). Используя симметричность и положительную определенность матрицы, нетрудно видеть, что оператор T_τ определен для любого $\tau > 0$, симметричен и

$$\|T_\tau\|_2 = \beta(T_\tau) = \max_{|\lambda_n(A)|} \left| \frac{1 - \tau\lambda_n(A)}{1 + \tau\lambda_n(A)} \right|. \tag{4.4.7}$$

Здесь

$$0 < \alpha = \alpha(A) \le \lambda_n(A) \le \beta = \beta(A). \tag{4.4.8}$$

Введем и исследуем функцию

$$q(\tau) = \max_{\lambda_n(A)} |g(\tau, \lambda_n(A))|, \tag{4.4.9}$$

где

$$g(\tau, \lambda) = \frac{1 - \tau\lambda}{1 + \tau\lambda}. \tag{4.4.10}$$

Очевидно, что при $\tau \le 0$ функция $q(\tau) \ge 1$ и, следовательно, метод (4.4.4) расходится. Поэтому необходимым условием сходимости метода является требование $\tau > 0$. Предположим, далее, что $\tau > 0$. Тогда так как при $\tau, \lambda > 0$

$$g'_\lambda(\tau, \lambda) = -\frac{2\tau}{(1 + \tau\lambda)^2} < 0$$

и $\lambda_n(A) \in [\alpha, \beta]$, то

$$q(\tau) = \max_{\lambda_n(A)} |g(\tau, \lambda)| = \max\{|g(\tau, \alpha)|, |g(\tau, \beta)|\} =$$

$$= \max\left\{ \left| \frac{1 - \tau\alpha}{1 + \tau\alpha} \right|, \left| \frac{1 - \tau\beta}{1 + \tau\beta} \right| \right\} < 1. \tag{4.4.11}$$

Отсюда следует, что рассматриваемый метод (4.4.4) сходится для любого $\tau > 0$.

Метод (4.4.4) является стационарным итерационным методом. Поэтому его оптимизация (согласно 4.2.2) заключается в минимизации по τ величины $\beta(T_\tau)$, т. е. в решении экстремальной задачи

$$q(\tau_{opt}) = \min_{\tau>0} q(\tau). \tag{4.4.12}$$

Учитывая, что

$$[g(\tau,\lambda)]'_\tau = -\frac{2\lambda}{(1+\tau\lambda^2)} < 0 \tag{4.4.13}$$

для любых $\tau, \lambda > 0$, аналогично тому, как это было сделано в 4.2.1 для простейшего итерационного метода, можно показать, что значение $\tau = \tau_{opt}$ является решением уравнения

$$\frac{1-\tau\alpha}{1+\tau\alpha} = -\frac{1-\tau\beta}{1+\tau\beta} \tag{4.4.14}$$

и вычисляется по формуле

$$\tau_{opt} = \frac{1}{\sqrt{\alpha\beta}}. \tag{4.4.15}$$

Действительно, если $\tau = \tau_{opt}$, то

$$\frac{1-\tau_{opt}\alpha}{1+\tau_{opt}\alpha} = -\frac{1-\tau_{opt}\beta}{1+\tau_{opt}\beta} = \frac{1-\sqrt{\alpha/\beta}}{1+\sqrt{\alpha/\beta}} > 0$$

и, следовательно,

$$q(\tau_{opt}) = \frac{1-\sqrt{\alpha/\beta}}{1+\sqrt{\alpha/\beta}}. \tag{4.4.16}$$

Если же $\tau < \tau_{opt}$, то согласно (4.4.13)

$$q(\tau) = \frac{1-\tau\alpha}{1+\tau\alpha} > q(\tau_{opt}).$$

В случае $\tau > \tau_{opt}$

$$q(\tau) = -\frac{1-\tau\beta}{1+\tau\beta} > q(\tau_{opt}).$$

Оценим асимптотическую скорость сходимости метода (4.4.4), (4.4.5)

$$R(T_{\tau_{opt}}) = -\ln q(\tau_{opt}) \tag{4.4.17}$$

в случае плохо обусловленных матриц, т. е. в случае $p = p(A) \gg 1$. Согласно (4.4.16) имеем

$$\begin{gathered} q(\tau_{opt}) \approx 1 - \frac{2}{\sqrt{p}}, \\ R(T_{\tau_{opt}}) \approx \frac{2}{\sqrt{p}}. \end{gathered} \tag{4.4.18}$$

Из полученных соотношений вытекает, что предлагаемый метод по асимптотической скорости сходимости всего лишь в два раза уступает оптимальному методу последовательной верхней релаксации. Отмечая достоинства метода, нельзя не видеть его недостатка, заключающегося в том, что решение системы

$$Bz^{j+1} = -\xi^j$$

($z^{j+1} = \varphi^{j+1} - \varphi$) на каждом шаге может потребовать не меньшего числа арифметических действий, чем решение исходной системы. Поэтому возникает проблема замены оператора $B = B_\tau$ некоторым «близким» оператором так, чтобы сохранялась быстрота сходимости метода и каждая итерация требовала числа операций, сравнимого с числом операций на итерацию метода последовательной верхней релаксации. В большом числе случаев решение этой задачи достигается выбором

$$B = \frac{1}{2\tau}(E + \tau A_1)(E + \tau A_2), \tag{4.4.19}$$

который и будет рассматриваться в дальнейшем.

4.4.1. Коммутативный случай

Предположим, что матрицы A_1 и A_2 разложения (4.4.2) симметричны и обладают общей ортонормированной системой собственных

векторов $\{u_n\}$, т. е.

$$A_1 u_n = \lambda_n = \lambda_n(A_1) u_n,$$
$$A_2 u_n = \lambda_n = \lambda_n(A_2) u_n \tag{4.4.20}$$

для всех n и система $\{u_n\}$ образует ортонормированный базис в исходном пространстве векторов. Разложим произвольный вектор φ по системе $\{u_n\}$:

$$\varphi = \sum_n \varphi_n u_n, \quad \varphi_n = (\varphi, u_n).$$

Легко видеть, что

$$A_1 A_2 \varphi = A_1 \left(\sum_n \lambda_n(A_2) \varphi_n u_n \right) = \sum_n \lambda_n(A_1) \lambda_n(A_2) \varphi_n u_n =$$

$$= A_2 \sum_n \lambda_n(A_1) \varphi_n u_n = A_2 A_1 \varphi.$$

Так как вектор φ произволен, то отсюда следует коммутативность матриц A_1 и A_2 ($A_1 A_2 = A_2 A_1$).

Разложим теперь векторы ошибок $\psi^j = \varphi^j - \varphi^*$ метода расщепления

$$(E + \tau A_1)(E + \tau A_2)(\varphi^{j+1} - \varphi^j) = -2\tau(A\varphi^j - f) \tag{4.4.21}$$

по системе $\{u_n\}$:

$$\psi^j = \sum_n \psi_n^j u_n.$$

В результате, используя (4.4.21), получим

$$\psi_n^{j+1} = \frac{1 - \tau\lambda_{1,n}}{1 + \tau\lambda_{1,n}} \frac{1 - \tau\lambda_{2,n}}{1 + \tau\lambda_{2,n}} \psi_n^j, \tag{4.4.22}$$

где для простоты использовано обозначение $\lambda_{i,n} = \lambda_n(A_i)$ $(i = 1, 2)$.

Из (4.4.22) вытекает (это, вообще говоря, следует уже из коммутативности матриц A_1 и A_2), что оператор шага метода (4.4.21)

$$T_\tau = E - 2\tau(E + \tau A_2)^{-1}(E + \tau A_1)^{-1} A =$$

$$= (E + \tau A_2)^{-1}(E + \tau A_1)^{-1}(E - \tau A_1)(E - \tau A_2) \tag{4.4.23}$$

обладает полной ортонормированной системой собственных векторов $\{u_n\}$, а его собственные числа вещественны и вычисляются по

формуле

$$\lambda_n(T_\tau) = \frac{1-\tau\lambda_{1,n}}{1+\tau\lambda_{1,n}}\frac{1-\tau\lambda_{2,n}}{1+\tau\lambda_{2,n}}$$

через собственные числа матриц A_1 и A_2. Отсюда вытекает, что матрица T_τ симметрична и, следовательно,

$$\|T_\tau\|_2 = \beta(T_\tau) = \max_n \left|\frac{1-\tau\lambda_{1,n}}{1+\tau\lambda_{1,n}}\frac{1-\tau\lambda_{2,n}}{1+\tau\lambda_{2,n}}\right|. \tag{4.4.24}$$

В соответствии с изложенной в 4.2.2 теорией отимизации стационарных итерационных методов равенство (4.4.24) означает, что выбор параметра $\tau = \tau_{opt}$ из условия минимизации

$$\beta(T_{\tau_{opt}}) = \min_\tau \beta T_\tau \tag{4.4.25}$$

является оптимизацией не только с асимптотической точки зрения, но и для любого $\varepsilon > 0$, если требуется уменьшить в $1/\varepsilon$ раз именно евклидову норму вектора ошибки. Заметим также, что в силу положительности $\lambda_{1,n}$ и $\lambda_{2,n}$ неравенство $\beta(T_\tau) < 1$ выполняется только при $\tau > 0$, т. е. положительность τ является необходимым условием сходимости рассматриваемого метода расщепления.

Перейдем к решению задачи оптимизации (4.4.25). В общем случае это очень сложная проблема, которая не решена до настоящего времени. Поэтому вместо $\beta(T_\tau)$ мы будем минимизировать функцию

$$q^2(\tau) = \max_n \left|\frac{1-\tau\lambda_{1,n}}{1+\tau\lambda_{1,n}}\right| \max_n \left|\frac{1-\tau\lambda_{2,n}}{1+\tau\lambda_{2,n}}\right|, \tag{4.4.26}$$

которая мажорирует $\beta(T_\tau)$ сверху, достаточно хорошо ее приближает, а в ряде частных случаев даже совпадает с ней.

Предположим теперь, что

$$\widetilde{\alpha} = \alpha(A_1) = \alpha(A_2), \quad \widetilde{\beta} = \beta(A_1) = \beta(A_2). \tag{4.4.27}$$

Тогда так как максимум модуля функции

$$g(\tau,\lambda) = \frac{1-\tau\lambda}{1+\tau\lambda}$$

зависит только от границ интервала $[\widetilde{\alpha},\widetilde{\beta}]$, которому принадлежат величины $\lambda_{1,n}$ и $\lambda_{2,n}$ (это было доказано в начале параграфа), то при

сделанном предположении (4.4.27)

$$q^2(\tau) = [\max_{\widetilde{\alpha} \le \lambda \le \widetilde{\beta}} |g(\tau, \lambda)|]^2.$$

Следовательно, проблема оптимизации сводится к решению экстремальной задачи

$$q(\tau_{opt}) = \min_{\tau > 0} q(\tau), \tag{4.4.28}$$

которая была исследована ранее. Ее решение дается формулой (4.4.15):

$$\tau_{opt} = \frac{1}{\sqrt{\alpha\beta}}. \tag{4.4.29}$$

Отсюда имеем

$$\beta(T_{\tau_{opt}}) \le q^2(\tau_{opt}) = \left[\frac{1 - \sqrt{\widetilde{\alpha}/\widetilde{\beta}}}{1 + \sqrt{\widetilde{\alpha}/\widetilde{\beta}}}\right]^2. \tag{4.4.30}$$

Учитывая теперь, что

$$\begin{gathered} \alpha = \alpha(A) \ge \alpha(A_1) + \alpha(A_2) = 2\widetilde{\alpha}, \\ \beta = \beta(A) \le \beta(A_1) + \beta(A_2) = 2\widetilde{\beta} \end{gathered} \tag{4.4.31}$$

и, следовательно,

$$\widetilde{\alpha}/\widetilde{\beta} \le \alpha/\beta < 1,$$

а также используя монотонное убывание функции

$$u(t) = \frac{1 - \sqrt{t}}{1 + \sqrt{t}}$$

на отрезке $[0, 1]$ от единицы до нуля и неравенство (4.4.30), получаем

$$\beta(T_{\tau_{opt}}) \le u^2\left(\frac{\widetilde{\alpha}}{\widetilde{\beta}}\right) \le u^2\left(\frac{\alpha}{\beta}\right) = \left[\frac{p-1}{p+1}\right]^2,$$

где $p = \dfrac{\beta}{\alpha}$ — число обусловленности матрицы A.

Окончательно, считая, что $p \gg 1$, приходим к оценкам

$$\begin{gathered} \beta(T_{\tau_{opt}}) \leq 1 - \frac{4}{\sqrt{p}}, \\ R(T_{\tau_{opt}}) \geq \frac{4}{\sqrt{p}}. \end{gathered} \tag{4.4.32}$$

Сравнивая полученные оценки с данными, приведенными в таблице из 4.2.5, делаем вывод, что построенный метод с точки зрения скорости сходимости не хуже, чем лучший из рассмотренных выше стационарных итерационных методов — метод последовательной верхней релаксации. Заметим также, что предлагаемый метод расщепления с оптимальным параметром не менее чем в два раза быстрее по скорости сходимости метода (4.4.4), (4.4.5) и в то же время на практике оказывается значительно проще в реализации, поскольку одним из требований расщепления (4.4.2) является простота обращения матриц $E + \tau A_i$ $(i = 1, 2)$.

Одной из основных проблем метода расщепления является оптимизация многопараметрических схем вида

$$(E + \tau_j A_1)(E + \tau_j A_2)(\varphi^{j+1} - \varphi^j) = -2\tau_j(A\varphi^j - f). \tag{4.4.33}$$

Задача оптимизации ставится здесь следующим образом. Для заданного $s \geq 1$ требуется определить последовательность параметров $\{\tau_i\}_{i=1}^s$, минимизирующих некоторую норму или спектральный радиус оператора

$$T^{(s)} = \prod_{i=1}^{s} T_i, \tag{4.4.34}$$

где

$$T_i = (E + \tau_i A_2)^{-1}(E + \tau_i A_1)^{-1}(E - \tau_i A_1)(E - \tau_i A_2). \tag{4.4.35}$$

В коммутативном случае это эквивалентно задаче

$$\max_n \prod_{i=1}^{s} \left| \frac{1 - \tau_i \lambda_{1,n}}{1 + \tau_i \lambda_{1,n}} \frac{1 - \tau_i \lambda_{2,n}}{1 + \tau_i \lambda_{2,n}} \right| = \min_{\tau_1, \ldots, \tau_s}. \tag{4.4.36}$$

Аналогично однопараметрическому случаю эта задача заменяется минимизацией функции

$$q(\tau_1, \ldots, \tau_s) = \max_{\alpha \le \lambda \le \beta} \prod_{i=1}^{s} \left| \frac{1 - \tau_i \lambda}{1 + \tau_i \lambda} \right|, \tag{4.4.37}$$

где

$$\alpha = \min(\alpha(A_1), \alpha(A_2)),$$

$$\beta = \max(\beta(A_1), \beta(A_2)).$$

Очевидно, что

$$\beta(T^{(s)}) \le q^2(\tau_1, \ldots, \tau_s) < 1 \tag{4.4.38}$$

для любых $\tau_i > 0$ $(i = 1, 2, \ldots, s)$.

В задаче

$$q(\tau_1^0, \ldots, \tau_s^0) = \min_{\tau_1, \ldots, \tau_s} q(\tau_1, \ldots, \tau_s) \tag{4.4.39}$$

при специальном выборе длины цикла s и параметров $\check{\tau} = (\check{\tau}_1, \ldots, \check{\tau}_s)$, являющихся приближенным решением задачи (4.4.39), в случае плохо обусловленной матрицы A достигается оценка

$$\frac{C}{\ln p} \le R(T_\tau^{(s)}),$$

где $p = p(A)$, C — константа, не зависящая от s и p. Иначе говоря, по скорости сходимости этот метод является наилучшим из всех рассмотренных ранее.

Остановимся на еще одном подходе к решению проблемы оптимизации метода расщепления, ограничиваясь для простоты однопараметрической схемой (4.4.21). Из предыдущего следует, что при любом $\tau > 0$ метод расщепления на всех итерациях обеспечивает подавление любого коэффициента ψ_n^i разложения вектора ошибки

$$\psi^j = \sum_n \psi_n^j u_n,$$

причем при оптимальном выборе τ обеспечивается наилучшее равномерное подавление коэффициентов по всем значениям n. Предположим теперь, что для собственных значений $\lambda_n \in [\widetilde{m}, \Delta]$, где

$$\widetilde{m} = \min(\alpha(A_1), \alpha(A_2)) < \Delta,$$

$$\Delta \ll \widetilde{M} = \max(\beta(A_1), \beta(A_2)),$$

соответствующие коэффициенты $\{\psi_n^0\}$ в разложении вектора начальной ошибки ψ^0 значительно преобладают над остальными коэффициентами. В этом случае представляется целесообразным несколько итераций по методу (4.4.21) провести со значением параметра τ, выбираемым по формуле

$$\tau = \frac{1}{\sqrt{\widetilde{m}\Delta}}.$$

Такой выбор обеспечит наилучшее равномерное подавление выделенных преобладающих коэффициентов разложения за несколько первых итераций, после чего мы можем продолжить процесс со значением τ, выбираемым по формуле (4.4.29).

В практической ситуации преобладающими коэффициентами оказываются коэффициенты разложения по собственным векторам матрицы, соответствующим минимальным собственным значениям, причем (с некоторой идеализацией) часто коэффициенты можно даже считать монотонно убывающими с ростом λ_n. Физически этот факт соответствует преобладающему влиянию крупномасштабных процессов (малые значения λ_n) по сравнению с мелкомасштабными (большие значения λ_n). Очевидно, что конкретный выбор величины Δ должен осуществляться в зависимости от конкретных априорных сведений о задаче.

4.4.2. Некоммутативный случай

Перейдем к рассмотрению метода расщепления

$$(E + \tau A_1)(E + \tau A_2)(\varphi^{j+1} - \varphi^j) = -2\tau(A\varphi^j - f) \tag{4.4.40}$$

с некоммутирующими положительно определенными матрицами A_1 и A_2. С этой целью введем симметричную и положительно определенную матрицу

$$D_\tau = D_{1,\tau}^* D_{1,\tau} \tag{4.4.41}$$

и векторы

$$z^j = D^{1,\tau}\psi^j. \tag{4.4.42}$$

Здесь

$$D_{1,\tau} = (E + \tau A_1)^{-1} A \tag{4.4.43}$$

и $\psi^j = \varphi^j - \varphi^*$ — векторы ошибок. Очевидно, что

$$z^j = (E + \tau A_1)^{-1}\xi^j,$$

где $\xi^j = A\varphi^j - f$ — векторы невязок.

С помощью стандартных преобразований легко показать, что

$$\psi^{j+1} = T_\tau \psi^j, \tag{4.4.44}$$

где

$$T_\tau = E - 2\tau(E + \tau A_2)^{-1}(E + \tau A_1)^{-1}A \tag{4.4.45}$$

— оператор шага метода (4.4.40), и

$$z^{j+1} = \widetilde{T}_\tau z^j, \tag{4.4.46}$$

где

$$\widetilde{T}_\tau = T_{1,\tau}T_{2,\tau}, \tag{4.4.47}$$

$$T_{i,\tau} = (E - \tau A_i)(E + \tau A_i)^{-1}, \quad i = 1, 2. \tag{4.4.48}$$

Заметим, что при выводе равенств (4.4.46)—(4.4.48) мы использовали соотношения

$$\widetilde{T}_\tau = (E + \tau A_1)^{-1}AT_\tau A^{-1}(E + \tau A_1) =$$

$$= (E + \tau A_1)^{-1}[(E + \tau A_1)(E + \tau A_2) - 2\tau A](E + \tau A_2)^{-1} =$$

$$= (E + \tau A_1)^{-1}(E - \tau A_1)(E - \tau A_2)(E + \tau A_2)^{-1} =$$

$$= (E - \tau A_1)(E + \tau A_1)^{-1}(E - \tau A_2)(E + \tau A_2)^{-1}.$$

Таким образом, для любого вектора ψ ($z = D_{1,\tau}\psi,\ x = T_{2,\tau}z$) справедливы формулы

$$\|T_\tau\psi\|_{D_\tau} = \|\widetilde{T}_\tau z\|_2 = \frac{\|T_{1,\tau}x\|_2}{\|x\|_2}\frac{\|T_{2,\tau}z\|_2}{\|z\|_2}\|\psi\|_{D_\tau}, \tag{4.4.49}$$

$$\|T_\tau\|_{D_\tau} = \|\widetilde{T}_\tau\|_2 \le \|T_{1,\tau}\|_2\|T_{2,\tau}\|_2. \tag{4.4.50}$$

Исследуем $\|T_{i,\tau}\|_2$ $(i = 1, 2)$, предполагая, что для любого u выполняются неравенства

$$(A_i u, u) \geq m(u, u),$$
$$(A_i u, A_i u) \leq M(A_i u, u), \quad i = 1, 2. \tag{4.4.51}$$

По определению

$$\|T_{i,\tau}\|_2^2 = \sup_{v \neq 0} \frac{\|T_{i,\tau} v\|_2^2}{\|v\|_2^2} = \sup_{u \neq 0} \frac{\|(E - \tau_i A_i)u\|_2^2}{\|(E + \tau_i A_i)u\|_2^2} =$$

$$= \sup_{u \neq 0} \frac{(u, u) - 2\tau(A_i u, u) + \tau^2 (A_i u, A_i u)}{(u, u) + 2\tau(A_i u, u) + \tau^2 (A_i u, A_i u)} =$$

$$= 1 - 4\tau \inf_{u \neq 0} \frac{(A_i u, u)}{(u, u) + 2\tau(A_i u, u) + \tau^2 (A_i u, A_i u)}, \tag{4.4.52}$$

где $u = (E + \tau A_i)^{-1} v$. Из приведенных соотношений видно, что отношение

$$\frac{\|T_{i,\tau} v\|_2}{\|v\|_2}$$

для любого ненулевого вектора v и любого i либо больше единицы при $\tau < 0$, либо меньше единицы при $\tau > 0$. Отсюда, согласно равенствам (4.4.49), следует, что необходимым условием сходимости метода расщепления (4.4.40) является положительность τ. В дальнейшем везде будем предполагать, что $\tau > 0$.

Оценим $\|T_{i,\tau}\|_2$, используя неравенства (4.4.51) и соотношения (4.4.52):

$$\|T_{i,\tau}\|^2 \leq 1 - 4\tau \inf_{u \neq 0} \frac{(A_i u, u)}{(u, u) + 2\tau(A_i u, u) + \tau^2 M(A_i u, A_i u)} =$$

$$= 1 - 4\tau \inf_{\|w\|_2 = 1} \frac{(A_i w, w)}{1 + (2\tau + M\tau^2)(A_i w, w)} \leq 1 - 4\tau \inf_{t \geq m} \frac{t}{1 + (2\tau + M\tau^2)t} =$$

$$= 1 - 4\tau \frac{m}{1 + (2\tau + M\tau^2)m} = \frac{1 - 2m\tau + mM\tau^2}{1 + 2m\tau + mM\tau^2} < 1.$$

Здесь нами были использованы обозначения

$$w = \frac{u}{\|u\|_2}, \quad t = (A_i w, w)$$

и тот факт, что функция

$$u(t) = \frac{t}{1+at}$$

монотонно возрастает при $t \geq 0$ для любого положительного a. Итак, мы доказали, что при сделанных предположениях (4.4.51) справедлива оценка

$$\|\widetilde{T}_\tau\|_{D_\tau} \leq \|T_{1,\tau}\|_2 \|T_{2,\tau}\|_2 \leq \frac{1-2m\tau+mM\tau^2}{1+2m\tau+mM\tau^2} < 1$$

для любого $\tau > 0$.

С целью приближенной оптимизации метода (4.4.40) минимизируем по τ функцию

$$q(\tau) = \frac{1-2m\tau+mM\tau^2}{1+2m\tau+mM\tau^2} = 1 - \frac{4m\tau}{1+2m\tau+mM\tau^2}. \tag{4.4.53}$$

Так как $q(0) = q(+\infty) = 1$, а при $\tau \in (0, +\infty)$ функция $q(\tau)$ бесконечно дифференцируема и

$$0 \leq q(\tau) < 1,$$

то значение $\tau = \tau_{opt}$, минимизирующее $q(\tau)$, является решением уравнения

$$\frac{dq(\tau)}{d\tau} = -4m\frac{1-mM\tau^2}{(1+2m\tau+mM\tau^2)^2} = 0.$$

Отсюда получаем

$$\tau_{opt} = \frac{1}{\sqrt{mM}}, \tag{4.4.54}$$

причем

$$q(\tau_{opt}) = \frac{1-\sqrt{m/M}}{1+\sqrt{m/M}}, \tag{4.4.55}$$

$$R(T_{\tau_{opt}}) \geq -\ln\left[\frac{1-\sqrt{m/M}}{1+\sqrt{m/M}}\right]. \tag{4.4.56}$$

Чтобы проанализировать полученную оценку, предположим, что матрицы A_1 и A_2 симметричны,

$$\alpha(A_1) = \alpha(A_2) = \frac{1}{2}\alpha(A),$$

$$\beta(A_1) = \beta(A_2) = \frac{1}{2}\beta(A).$$

Тогда нетрудно видеть, что

$$m = \frac{1}{2}\alpha(A), \quad M = \frac{1}{2}\beta(A)$$

и, следовательно,

$$q_{\tau_{opt}} = \frac{\sqrt{p}-1}{\sqrt{p}+1}.$$

Таким образом, в случае плохо обусловленных систем ($p = p(A) \gg 1$) мы имеем оценку

$$R(T_{\tau_{opt}}) \geq \frac{2}{\sqrt{p}}. \tag{4.4.57}$$

В то же время непосредственный спектральный анализ (аналогично тому, как это было сделано в 4.4.1) показывает, что при $\tau = \tau_{opt}$

$$\|T_{i,\tau_{opt}}\|_2 = \max_{m \leq \lambda \leq M} \left| \frac{1-\tau_{opt}\lambda}{1+\tau_{opt}\lambda} \right| = \frac{1-\sqrt{m/M}}{1+\sqrt{m/M}} \tag{4.4.58}$$

и, следовательно,

$$\beta(T_{\tau_{opt}}) \leq \left[\frac{1-\sqrt{m/M}}{1+\sqrt{m/M}} \right]^2 \leq \left[\frac{\sqrt{p}-1}{\sqrt{p}+1} \right]^2. \tag{4.4.59}$$

В случае плохо обусловленных систем имеем соответственно

$$\begin{gathered} \beta(T_{\tau_{opt}}) \leq 1 - \frac{4}{\sqrt{p}}, \\ R(T_{\tau_{opt}}) \geq \frac{4}{\sqrt{p}}, \end{gathered} \tag{4.4.60}$$

т. е., с точки зрения скорости сходимости, метод расщепления (4.4.40) при $\tau = \tau_{opt}$ не медленнее, чем любой из рассмотренных ранее стационарных итерационных методов. Необходимо отметить, что несовпадение оценок (4.4.57) и (4.4.60) вызвано тем обстоятельством, что первый способ построения оценок скорости сходимости является более общим, а использование дополнительных свойств матриц A_1 и A_2 позволяет эти оценки значительно улучшать.

В заключение остановимся кратко еще на одном важном подклассе схем метода расщепления (4.4.40). Пусть матрица A системы

$A\varphi = f$ симметрична и положительно определена,

$$A = A_1 + A_2,$$
$$A_1 = A_2^*. \tag{4.4.61}$$

Так как нами предполагается вещественность матриц A_1 и A_2, то в пространстве вещественных векторов при выполнении условий (4.4.61) для любого ψ выполняются равенства

$$(A_1\psi, \psi) = (A_2\psi, \psi) = \frac{1}{2}(A\psi, \psi).$$

Отсюда следует, что вместе с матрицей A всегда будут положительно определены матрицы A_1 и A_2 и, согласно доказанному выше, метод расщепления (4.4.40) с такими матрицами будет сходиться при любом $\tau > 0$.

Наиболее часто построение матриц A_1 и A_2 по заданной матрице A осуществляется следующим образом. Сначала матрица A представляется в блочном виде

$$A = \begin{Vmatrix} A_{11} & A_{12} & \dots & A_{1k} \\ A_{21} & A_{22} & \dots & A_{2k} \\ \dots & \dots & \dots & \dots \\ A_{k1} & A_{k2} & \dots & A_{kk} \end{Vmatrix}$$

с квадратными диагональными подматрицами A_{ii}, затем в качестве A_1 выбирается блочно-треугольная матрица вида

$$A_1 = \begin{Vmatrix} \frac{1}{2}A_{11} & 0 & \dots & 0 \\ A_{21} & \frac{1}{2}A_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ A_{k1} & A_{k2} & \dots & \frac{1}{2}A_{kk} \end{Vmatrix}$$

и A_2 полагается равной A_1^*. Простейшим частным случаем такого способа построения A_1 и A_2 является выбор $A_{ii} = a_{ii}$, т. е. равными диагональным элементам A. Тогда матрица A_1 является нижней треугольной матрицей, а A_2 — верхней треугольной.

4.4.3. Вариационная и чебышевская оптимизация методов расщепления

В предыдущем пункте было показано, что для итерационного метода расщепления

$$(E + \tau A_1)(E + \tau A_2)(\varphi^{j+1} - \varphi^j) = -\gamma_j(A\varphi^j - f) \tag{4.4.62}$$

в случае $\gamma_j = 2\tau$ выполняются неравенства

$$q_j = \frac{\|(E - \tau_j H_\tau A)\psi^j\|_{D_\tau}}{\|\psi^j\|_{D_\tau}} \le \|T_\tau\|_{D_\tau} < 1 \tag{4.4.63}$$

при любом $\tau > 0$, где

$$H_\tau = (E + \tau A_2)^{-1}(E + \tau A_1)^{-1},$$

$$T_\tau = E - 2\tau H_\tau A$$

и ψ^j — вектор ошибки. Согласно результатам 4.3.1 это означает, что нестационарный итерационный метод (4.4.62) с параметрами γ_j, являющимися решением уравнения

$$\frac{d}{d\gamma_j}\|(E - \gamma_j H_\tau A)\psi^j\|_{D_\tau} = 0, \tag{4.4.64}$$

т. е. выбираемыми из условия наибольшей минимизации на каждом шаге функционала $J(\varphi) = (D_\tau(\varphi^* - \varphi), \varphi^* - \varphi) = \|\varphi^* - \varphi\|^2_{D_\tau}$, будет сходиться. При этом если обозначить через $T_{\tau,\gamma}$ оператор шага предлагаемого итерационного метода, то при любом $\tau > 0$ выполняется оценка

$$\|T_{\tau,\gamma}\|_{D_\tau} < \|T_\tau\|_{D_\tau} < 1. \tag{4.4.65}$$

Решая уравнение (4.4.64), находим выражение для γ_j:

$$\gamma_j = \frac{(H_\tau A\psi^j, \psi^j)_{D_\tau}}{\|H_\tau A\psi^j\|^2_{D_\tau}} = \frac{((E+\tau A_1)^{-1}AH_\tau\xi^j, (E+\tau A_1)^{-1}\xi^j)}{\|(E+\tau A_1)^{-1}AH_\tau\xi^j\|_2^2}, \tag{4.4.66}$$

где $\xi^j = A\varphi^j - f$ — вектор невязки.

Остановимся теперь на дополнительной оптимизации метода расщепления (4.4.40) в предположении, что $H_\tau = H_\tau^*$. Как следует из предыдущих пунктов настоящего параграфа, такая ситуация возникает либо в коммутативном случае, если дополнительно $A_i = A_i^*$ $(i = 1, 2)$, либо в случае $A_1 = A_2^*$:

$$H_\tau = (E+\tau A_2)^{-1}(E+\tau A_1)^{-1} = (E+\tau A_1^*)^{-1}(E+\tau A_1)^{-1} =$$

$$= [(E+\tau A_1)^{-1}]^*(E+\tau A_1)^{-1} = H_\tau^*.$$

Основываясь на сделанном предположении и выводах 4.2.4, приходим к заключению, что параметры $\{\gamma_j\}$ циклического итерационного метода

$$\varphi^{j+1} = \varphi^j - \gamma_j H_\tau(A\varphi^j - f) \tag{4.4.67}$$

с длиной цикла s можно выбрать, исходя из общей теории чебышевских циклических итерационных методов, развитой в 4.2.4. При этом величины $\alpha = \alpha(A)$ и $\beta = \beta(A)$ необходимо соответственно заменить на

$$m = \alpha(H_\tau A),$$

$$M = \beta(H_\tau A).$$

Для построения оценок величин m и M или для их вычисления могут быть использованы либо оценки нормы оператора $T_\tau = E - 2\tau H_\tau A$, либо дополнительные итерационные методы типа метода Люстерника.

Оценим, например, величины m и M в коммутативном случае, предполагая, что $\tau = \tau_{opt}$ (см. формулу (4.4.29)) и A — плохо обуслов-

ленная матрица. Тогда из первого неравенства (4.4.32) имеем

$$m = \alpha(H_\tau A) \geq \frac{1}{2\tau}\frac{4}{\sqrt{p(A)}},$$
$$M = \beta(H_\tau A) \leq \frac{1}{2\tau}\left(2 - \frac{4}{\sqrt{p(A)}}\right). \tag{4.4.68}$$

Используя эти неравенства, приходим к соотношениям

$$p(H_\tau A) = \frac{\beta(H_\tau A)}{\alpha(H_\tau A)} \leq \frac{1}{2}(\sqrt{p(A)} - 2) \approx \frac{\sqrt{p(A)}}{2}. \tag{4.4.69}$$

Обращаясь теперь к оценкам (4.2.69) и учитывая, что $p(H_\tau A) \gg 1$ при $p(A) \gg 1$, получаем для метода (4.4.67) в случае $s \gg 1$, что

$$R(T_{\tau,\gamma}) \approx \frac{2}{\sqrt{p(H_\tau A)}} \approx \frac{2\sqrt{2}}{\sqrt[4]{p(A)}}. \tag{4.4.70}$$

Таким образом, использование чебышевского итерационного метода позволяет повысить эффективность метода расщепления в $\sqrt{p(A)/2}$ раз.

Другой подход к ускорению сходимости метода расщепления в случае $A_1 = A_2^*$ основан на использовании метода сопряженных градиентов. В силу того, что матрица H_τ симметрична и положительно определена, можно непосредственно переписать формулы (4.3.34):

$$g_k = \begin{cases} H_\tau \xi^0, & k = 1, \\ H_\tau \xi^{k-1} - b_k g_{k-1}, & k > 1, \end{cases}$$
$$b_k = \frac{(AH_\tau \xi^{k-1}, g_{k-1})}{(Ag_{k-1}, g_{k-1})}, \tag{4.4.71}$$
$$\varphi^k = \varphi^{k-1} - \alpha_k g_k,$$
$$\alpha_k = \frac{(\xi^{k-1}, g_k)}{(Ag_k, g_k)}, \quad k = 1, 2, \ldots, s.$$

Отметим, что асимптотическая скорость сходимости метода (4.4.71) будет примерно такая же, как метода (4.4.67) при любом s-циклическом выборе параметров $\{\gamma_j\}$.

Пример. Проиллюстрируем применение изученных выше итерационных алгоритмов к решению конечно-разностных уравнений для двумерного уравнения Пуассона (см. 1.5.3). Рассмотрим систему (1.5.28) из 1.5.3 и применим для ее решения итерационный алгоритм вида

$$\varphi^{j+1} = \varphi^j - \tau_j(\Lambda\varphi^j - F), \tag{4.4.72}$$

который можно записать в следующем виде:

$$\varphi^{j+1} = \varphi^j - \tau_j\xi^j, \quad \xi^j = \Lambda\varphi^j - F.$$

Итерационный процесс (4.4.72) необходимо продолжать до тех пор, пока не будет выполнено

$$\|\varphi^j - \varphi\| \le \varepsilon,$$

где ε — априорная константа. Эта оценка имеет место, если

$$\|\xi^j\| \le \alpha(\Lambda)\varepsilon. \tag{4.4.73}$$

Скорость сходимости итерационного процесса, вообще говоря, может быть повышена, если вместо итерационного процесса (4.4.72) рассматривается метод последовательных приближений вида

$$\varphi^{j+1} = \varphi^j - \tau_j B^{-1}(\Lambda\varphi^j - F), \tag{4.4.74}$$

где

$$B = (E + \sigma\Lambda_1)(E + \sigma\Lambda_2),$$

$$\sigma = \frac{2}{\sqrt{\alpha\beta}}, \quad \alpha = \alpha(\Lambda); \quad \beta = \beta(\Lambda).$$

Найдем границы спектра оператора $B^{-1}\Lambda$. Поскольку матрицы Λ_1 и Λ_2 имеют общий базис, то собственные числа λ_1 матрицы Λ_1 и λ_2 матрицы Λ_2 будут связаны с собственными числами задачи

$$B^{-1}\Lambda u = \lambda(B^{-1}\Lambda)u$$

следующим соотношением:

$$\lambda(B^{-1}\Lambda) = \frac{\lambda_1 + \lambda_2}{\left(1 + \dfrac{2\lambda_1}{\sqrt{\alpha\beta}}\right)\left(1 + \dfrac{2\lambda_2}{\sqrt{\alpha\beta}}\right)}, \tag{4.4.75}$$

$$\alpha/2 \le \lambda_1 \le \beta/2, \quad \alpha/2 \le \lambda_2 \le \beta/2.$$

Выражение (4.4.75) приведем к виду

$$\lambda(B^{-1}\Lambda) = \frac{\sqrt{\alpha\beta}}{2} f(x, y), \tag{4.4.76}$$

где

$$x = \frac{2\lambda_1}{\sqrt{\alpha\beta}}, \quad y = \frac{2\lambda_2}{\sqrt{\alpha\beta}}, \quad f(x, y) = \frac{x + y}{(1 + x)(1 + y)}. \tag{4.4.77}$$

Таким образом, для того чтобы определить границы спектра матрицы $B^{-1}\Lambda$, нам достаточно определить наименьшее и наибольшее значения функции $f(x, y)$ в квадрате

$$\sqrt{\alpha/\beta} \le x, y \le \sqrt{\beta/\alpha}.$$

Анализируя производные этой функции

$$\frac{\partial f}{\partial x} = \frac{1 - y}{1 + y}\frac{1}{(1 + x)^2}, \quad \frac{\partial f}{\partial y} = \frac{1 - x}{1 + x}\frac{1}{(1 + y)^2}, \tag{4.4.78}$$

нетрудно показать, что максимальное значение $f(x, y)$ принимает в двух угловых точках:

$$\max_{\sqrt{\alpha/\beta} \le x,y, \le \sqrt{\beta/\alpha}} f(x, y) = f\left(\sqrt{\frac{\alpha}{\beta}}, \sqrt{\frac{\beta}{\alpha}}\right) = f\left(\sqrt{\frac{\beta}{\alpha}}, \sqrt{\frac{\alpha}{\beta}}\right), \tag{4.4.79}$$

а минимальное — в двух других:

$$\min_{\sqrt{\alpha/\beta} \le x,y, \le \sqrt{\beta/\alpha}} f(x, y) = f\left(\sqrt{\frac{\alpha}{\beta}}, \sqrt{\frac{\alpha}{\beta}}\right) = f\left(\sqrt{\frac{\beta}{\alpha}}, \sqrt{\frac{\beta}{\alpha}}\right). \tag{4.4.80}$$

Отсюда и из (4.4.76) следует, что

$$\alpha(B^{-1}\Lambda) = \frac{\alpha}{(1 + \sqrt{\alpha/\beta})^2}, \quad \beta(B^{-1}\Lambda) = \frac{1}{2}\sqrt{\frac{\alpha}{\beta}}\frac{(\alpha + \beta)}{(1 + \sqrt{\alpha/\beta})^2}. \tag{4.4.81}$$

Это значит, что асимптотически при $\beta(\Lambda) \gg \alpha(\Lambda)$ имеем

$$p(B^{-1}\Lambda) = \frac{\beta(B^{-1}\Lambda)}{\alpha(B^{-1}\Lambda)} = \frac{1}{2}\frac{\alpha+\beta}{\sqrt{\alpha\beta}} \cong \frac{1}{2}\sqrt{\frac{\beta}{\alpha}} = \frac{1}{2}[p(\Lambda)]^{1/2}, \tag{4.4.82}$$

$$R = \frac{2}{\sqrt{p(B^{-1}\Lambda)}} = 2^{3/2}p^{-1/4}(\Lambda). \tag{4.4.83}$$

Приведем теперь схему реализации итерационного процесса (4.4.74):

$$\begin{gathered} \xi^j = \Lambda\varphi^j - F, \\ (E+\sigma\Lambda_1)\xi^{j+1/2} = \xi^j, \\ (E+\sigma\Lambda_2)\xi^{j+1} = \xi^{j+1/2}, \\ \varphi^{j+1} = \varphi^j - \tau_j\xi^{j+1}, \end{gathered} \tag{4.4.84}$$

где τ_j вычисляется на основе избранного метода оптимизации.

В покомпонентном представлении второе и третье уравнения системы (4.4.84) расписываются следующим образом: сначала решается задача при $l = 1, 2, \ldots, N-1$ (при $\sigma_1 = \sigma/h^2$):

$$\begin{gathered} (1+2\sigma_1)\xi_{1,l}^{j+1/2} - \sigma_1\xi_{2,l}^{j+1/2} = \xi_{1,l}^j, \\ -\sigma_1\xi_{k-1,l}^{j+1/2} + (1+2\sigma_1)\xi_{k,l}^{j+1/2} - \sigma_1\xi_{k+1,l}^{j+1/2} = \xi_{k,l}^j, \\ -\sigma_1\xi_{N-2,l}^{j+1/2} + (1+2\sigma_1)\xi_{N-1,l}^{j+1/2} = \xi_{N-1,l}^j; \end{gathered} \tag{4.4.85}$$

затем решается задача при $k = 1, 2, \ldots, N-1$:

$$\begin{gathered} (1+2\sigma_1)\xi_{k,1}^{j+1} - \sigma_1\xi_{k,2}^{j+1} = \xi_{k,1}^{j+1/2}, \\ -\sigma_1\xi_{k,l-1}^{j+1} + (1+2\sigma_1)\xi_{k,l}^{j+1} - \sigma_1\xi_{k,l+1}^{j+1} = \xi_{k,l}^{j+1/2}, \\ -\sigma_1\xi_{k,N-2}^{j+1} + (1+2\sigma_1)\xi_{k,N-1}^{j+1} = \xi_{k,N-1}^{j+1/2}. \end{gathered} \tag{4.4.86}$$

Уравнения (4.4.85) и (4.4.86) решаются с помощью метода факторизации.

Существенное отличие схемы реализации разностного аналога задачи Неймана от задачи Дирихле состоит в том, что в задаче Ней-

мана $\alpha(\Lambda) = 0$. Поэтому правая часть F и приближенное решение на каждом шаге φ должны быть ортогональны к вектору с одинаковыми компонентами. Это значит, что перед осуществлением нового шага итераций необходимо из каждой компоненты вектора ξ вычесть постоянную составляющую

$$\frac{1}{(N+1)^2}\sum_{k,l}\xi^j_{k,l}.$$

При такой процедуре ортогонализации, когда векторы с одинаковыми компонентами исключаются из элементов исходного гильбертова пространства, переводя его в подпространство Φ, и пробные векторы выбираются из этого подпространства, в качестве нижней границы спектра можно взять наименьшее ненулевое собственное число. Вспомним теперь выражения (1.5.46), (1.5.47) из 1.5.3 для собственных чисел разностного аналога оператора Лапласа в задаче Неймана и для границ спектра. Тогда можем сделать вывод о том, что параметр σ следует выбрать по формуле

$$\sigma = \frac{2}{\sqrt{\alpha^*\beta}}, \tag{4.4.87}$$

а вместо оценки для окончания итерационного процесса (4.4.73) необходимо принять следующую:

$$\|\xi^j\| < \alpha^*(\Lambda)\varepsilon. \tag{4.4.88}$$

В остальном алгоритм численного решения задачи Неймана для уравнения Пуассона не отличается от рассматриваемого алгоритма решения задачи Дирихле.

4.5. Итерационные методы для систем с вырожденными матрицами

Рассмотрим систему линейных алгебраических уравнений

$$A\varphi = f \tag{4.5.1}$$

с симметричной и положительно полуопределенной матрицей A. Через $\{u_n\}$ обозначим систему ортонормированных собственных векторов A, соответствующих ее собственным числам $\{\lambda_n\}$, а через $\ker A$ обозначим нуль-пространство матрицы A, т. е. множество таких векторов ψ, что $A\psi = 0$. Мы будем предполагать, что размерность $\ker A$ равна m, а нулевыми собственными числами являются $\{\lambda_n\}_{n=1}^m$ (соответственно $u_n \in \ker A$ для $n = 1, 2, \ldots, m$).

Разложим векторы f и φ системы (4.5.1) по базису $\{u_n\}$:

$$f = \sum_n f_n u_n,$$

$$\varphi = \sum_n \varphi_n u_n.$$

Подставляя эти разложения в (4.5.1) получим

$$\lambda_n \varphi_n = f_n$$

и, в частности,

$$\lambda_n \varphi_n = f_n \quad \text{для} \quad n = 1, 2, \ldots, m.$$

Так как последние равенства возможны только в случае $f_n = 0$ $(n = 1, 2, \ldots, m)$, то условием совместности системы является требование ортогональности f к $\ker A$ $(f \perp \ker A)$, или, что эквиалентно, $(f, u_n) = 0,\ n = 1, 2, \ldots, m$.

В случае, когда система (4.5.1) несовместна, под ее решением часто понимают решение φ^* соответствующей совместной системы

$$A\varphi = \widetilde{f}, \tag{4.5.2}$$

где

$$\widetilde{f} = \sum_{n>m} f_n u_n,$$

которое при этом называют *обобщенным решением* системы (4.5.1). Обобщенные решения, как нетрудно показать, минимизируют евклидову норму вектора невязки. Если матрица A системы (4.5.1) произвольна, то вектор φ^* называется обобщенным решением этой системы, если он является решением задачи

$$\|A\varphi^* - f\|_2 = \min_{\varphi} \|A\varphi - f\|_2.$$

4.5.1. Случай совместной системы

Предположим, что система (4.5.1) совместна и для ее решения применяется стационарный итерационный метод

$$B(\varphi^{j+1} - \varphi^j) = -(A\varphi^j - f) \tag{4.5.3}$$

с симметричной и положительно определенной матрицей B. Исследуем условия сходимости этого метода. Если ввести векторы $\psi^j = \varphi^j - \varphi^*$, где φ^* — некоторое произвольное фиксированное решение системы (4.5.1), то итерационный процесс для векторов $\{\psi^j\}$ может быть записан в виде

$$\psi^{j+1} = T\psi^j, \tag{4.5.4}$$

где через

$$T = E - B^{-1}A \tag{4.5.5}$$

обозначен оператор шага итерационного метода (4.5.3). Введем теперь векторы $z^j = B^{1/2}\psi^j$. Умножая (4.5.4) на матрицу $B^{1/2}$, получим

$$z^{j+1} = \widetilde{T}z^j, \tag{4.5.6}$$

где

$$\widetilde{T} = E - B^{-1/2}AB^{-1/2} \tag{4.5.7}$$

— симметричная матрица, причем матрица $S = B^{-1/2}AB^{-1/2}$ симметрична и положительно полуопределена (здесь $B^{-1/2} = [B^{1/2}]^{-1}$).

Обозначим через $\{v_n\}$ полную ортонормированную систему собственных векторов матрицы $\widetilde{T}$, соответствующих ее собственным числам $\{\mu_n\}$, и разложим векторы z по этой системе:

$$z^j = \sum_n z_n^j v_n. \tag{4.5.8}$$

Подставляя (4.5.8) в (4.5.6), приходим к соотношениям

$$z_n^{j+1} = \mu_n z_n^j = \mu_n^{j+1} z_n^0.$$

Отсюда следует, что условия

$$|\mu_n| \le 1,$$
$$\mu_n \ne 1 \tag{4.5.9}$$

необходимы для сходимости метода (4.5.3) при любых начальных приближениях φ^0. Действительно, если $|\mu_n| > 1$ или $\mu_n = -1$, то последовательность z_n^j соответственно либо расходится, либо не сходится ни к какой конкретной величине для любого $z_n^0 \ne 0$.

Докажем, что условия (4.5.9) достаточны для сходимости последовательности $\{\varphi^j\}$ к некоторому решению $\widetilde{\varphi}^*$ системы (4.5.1) при любом начальном приближении φ^0. Очевидно, что неравенство $|\mu_n| < 1$ всегда обеспечивает сходимость z_n^j к нулю при $j \to \infty$, а в случае $\mu_n = 1$

$$z_n^{j+1} = z_n^j.$$

Предположим теперь, что выполнены условия (4.5.9) и кратность $\mu_n = 1$ равна s (можно доказать $s = m$), причем $\mu_1 = \mu_2 = \ldots = \mu_s = 1$. Тогда нетрудно видеть, что

$$\lim_{j\to\infty} z = z^\infty = \sum_{n=1}^{s} z_n^0 v_n.$$

Из равенства $\widetilde{T} z^\infty = z^\infty$ (так как $\widetilde{T} v_n = v_n$ для $n = 1, 2, \ldots, s$) вытекает, что

$$B^{-1/2} A B^{-1/2} z^\infty = 0,$$

и, следовательно,

$$B^{-1/2} z^\infty \in \ker A. \tag{4.5.10}$$

С другой стороны, имеем

$$B^{-1/2} z^\infty = \lim_{j\to\infty} B^{-1/2} z^j = \lim_{j\to\infty} \psi^j = \psi^\infty. \tag{4.5.11}$$

Окончательно, объединяя (4.5.10) и (4.5.11), приходим к выводу, что последовательность $\{\varphi^j\}$ итерационного метода (4.5.3) сходится к некоторому вектору

$$\varphi^\infty = \varphi^* + \psi^\infty, \tag{4.5.12}$$

который в силу $\psi^\infty \in \ker A$ является решением системы (4.5.1), т. е. $\varphi^\infty = \widetilde{\varphi}^*$, и зависит от выбора начального приближения φ^0.

4.5.2. Случай несовместной системы

Рассмотрим два подхода к решению системы (4.5.1) в случае ее несовместности. Предположим сначала, что матрицы B и A итерационного метода (4.5.3) обладают общей полной системой ортонормированных собственных векторов $\{u_n\}$, т. е.

$$\begin{aligned} Bu_n &= \nu_n u_n, \\ Au_n &= \lambda_n u_n, \end{aligned} \tag{4.5.13}$$

и для этого итерационного метода выполнены условия (4.5.9). Последнее, как было показано выше, означает, что метод (4.5.3) сходится для соответствующих совместных систем. Раскладывая векторы φ^j и f по базису $\{u_n\}$:

$$\varphi^j = \sum_n \varphi_n^j u_n,$$

$$f = \sum_n f_n u_n,$$

получим

$$\varphi_n^{j+1} = \left(1 - \frac{\lambda_n}{\nu_n}\right)\varphi_n^j + \frac{1}{\nu_n} f_n = \left(1 - \frac{\lambda_n}{\nu_n}\right)^{j+1} \varphi_n^0 + \frac{1}{\nu_n} \sum_{k=0}^{j} \left(1 - \frac{\lambda_n}{\nu_n}\right)^k f_n,$$

откуда, используя условия (4.5.9), в свою очередь имеем

$$\lim_{j\to\infty} \varphi_n^j = \frac{1}{\lambda_n} f_n$$

для $\lambda \neq 0$ $(n > m)$ и

$$\varphi_n^j = \frac{j}{\nu_n} f_n + \varphi_n^0$$

для λ_n $(n \leq m)$. Таким образом, последовательность φ^j расходится, если $f_n \neq 0$ хотя бы для одного $n \leq m$. Образуем наряду с последовательностью $\{\varphi^j\}$ последовательность

$$z^j = \varphi^j - j(\varphi^{j+1} - \varphi^j) \tag{4.5.14}$$

и разложим векторы z по базису $\{u_n\}$:

$$z^j = \sum_n z_n^j u_n.$$

В результате получим

$$z_n^j = \begin{cases} \varphi_n^j - j\left[\dfrac{\lambda_n}{\nu_n}\left(1-\dfrac{\lambda_n}{\nu_n}\right)^j \varphi_n^0 + \dfrac{1}{\nu_n}\left(1-d\dfrac{\lambda_n}{\nu_n}\right)^j f_n\right], & n > m, \\ \varphi_n^0, & n \le m, \end{cases}$$

и, следовательно,

$$\lim_{j\to\infty} z_n^j = \begin{cases} \dfrac{1}{\lambda_n} f_n, & n > m, \\ \varphi_n^0, & n \le m. \end{cases} \tag{4.5.15}$$

Таким образом, вектор

$$z^\infty = \lim_{j\to\infty} z^j \tag{4.5.16}$$

существует и является решением совместной системы (4.5.2).

Рассмотрим другой подход к решению несовместных систем на примере системы (4.5.1). Предположим, что размерность нуль-пространства матрицы A равна единице $(m = 1)$, матрица B итерационного процесса (4.5.3) симметрична и положительно определена и выполнены условия (4.5.9). Очевидно, что нуль-пространства матриц A и $B^{-1}A$ совпадают, все собственные числа матрицы $B^{-1}A$ вещественны и она обладает полной системой собственных векторов (последнее доказывается в 4.2.4).

Обозначим через $\{v_n\}$ систему собственных векторов матрицы $B^{-1}A$, а через $\{v_n^*\}$ — систему собственных векторов матрицы $(B^{-1}A)^* = AB^{-1}$. Известно, что в случае полноты эти системы образуют биортонормированный базис в исходном пространстве векторов, т. е.

$$(v_i, v_j^*) = \begin{cases} 0, & i \ne j, \\ 1, & i = j. \end{cases}$$

Будем считать, что $v_1 \in \ker(B^{-1}A) = \ker A$. Разложим векторы x^j итерационного метода

$$x^{j+1} = x^j - B^{-1}Ax^j, \qquad x^0 = B^{-1}f \tag{4.5.17}$$

по системе $\{v_n\}$. Получим

$$x^j = \sum_n x_n^j v_n, \quad x_n^j = (x^j, v_n^*).$$

Тогда, согласно (4.5.9),

$$x^\infty = x_1^0 v_1.$$

Предположение, что $x_1^0 = (B^{-1}f, v_1^*) = 0$, будет означать, что система

$$B^{-1}A\varphi = B^{-1}f$$

совместна и, следовательно, совместна исходная система, что противоречит предположению о ее несовместности. Таким образом, $x_1^0 \neq 0$ и вектор

$$v^\infty = x_1^0 v_1 \in \ker A.$$

Ортогонализируя вектор f к вектору v^∞:

$$\widetilde{f} = f - \frac{(f, v^\infty)}{(v^\infty, v^\infty)} v^\infty, \tag{4.5.18}$$

приходим к системе

$$A\varphi = \widetilde{f}, \tag{4.5.19}$$

которая совместна и эквивалентна исходной системе в смысле множества обобщенных решений.

4.5.3. Метод фиктивных областей

Пусть дана система линейных алгебраических уравнений

$$A_0\varphi_0 = f_0 \tag{4.5.20}$$

с симметричной положительно определенной матрицей A_0. Наряду с этой системой рассмотрим систему

$$A\varphi = f \tag{4.5.21}$$

с матрицей A вида

$$A = \begin{Vmatrix} A_0 & 0 \\ 0 & A_1 \end{Vmatrix}, \tag{4.5.22}$$

где A_1 — симметричная положительно полуопределенная матрица,

$$\varphi = \begin{Vmatrix} \varphi_0 \\ \varphi_1 \end{Vmatrix} \quad \text{и} \quad f = \begin{Vmatrix} f_0 \\ 0 \end{Vmatrix}.$$

Очевидно, что A — симметричная положительно полуопределенная матрица и для любого решения

$$\varphi = \begin{Vmatrix} \varphi_0 \\ \varphi_1 \end{Vmatrix}$$

системы (4.5.21) вектор φ_0 является решением системы (4.5.20).

Для решения системы (4.5.21) применим стационарный итерационный метод

$$\varphi^{j+1} = \varphi^j - H(A\varphi^j - f) \tag{4.5.23}$$

с симметричной и положительно определенной матрицей H и оператором шага

$$T = E - HA, \tag{4.5.24}$$

собственные числа которого удовлетворяют условию (4.5.9). Как уже было показано в 4.5.1, скорость сходимости этого метода в случае совместной системы (4.5.21) (а эта система будет всегда совместна) определяется максимальным по модулю, не равным единице собственным числом оператора T. Можно показать, что при любой симметричной и положительно определенной матрице H оптимальным с точки зрения минимизации этой величины является выбор

$$A_1 = 0.$$

Проиллюстрируем предлагаемый метод на конкретном примере решения системы пятиточечных уравнений, аппроксимирующих задачу Дирихле для уравнения Пуассона в L-образной области D (рис. 4.4). На этом рисунке крестиками отмечены узлы, в которых заданы пятиточечные уравнения с учетом граничного условия, а кружочками — узлы, в которых заданы уравнения с нулевыми коэффициентами (это соответствует $A_1 = 0$).

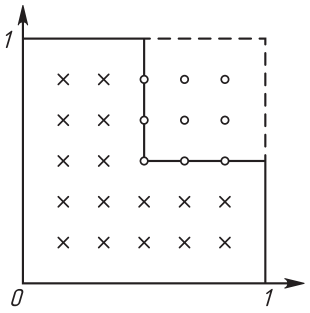


Рис. 4.4.

Исследуем случай, когда в качестве матрицы $B = \tau H^{-1}$ выбирается пятиточечный разностный аналог оператора Лапласа во всех узлах прямоугольной сетки с шагом h, покрывающих квадрат. При этом для максимального собственного числа матрицы T (не равного 1) устанавливается оценка ($h \ll 1$)

$$|\lambda_{max}| \leq 1 - Ch, \tag{4.5.25}$$

где C — константа, не зависящая от h.

Докажем оценку (4.5.25). Пусть $h = 1/(n+1), k = (n+1)/2$. Введем нумерацию узлов сетки так, чтобы сначала нумеровались узлы, лежащие внутри L-образной области, затем остальные. Выписывая последовательно пятиточечные уравнения

$$\frac{-u_{i-1j} + 2u_{ij} - u_{i+1j}}{h^2} + \frac{-u_{ij-1} + 2u_{ij} - u_{ij+1}}{h^2} = f_{ij} \tag{4.5.26}$$

согласно введенной нумерации и дополняя их уравнениями с нулевыми коэффициентами, приходим к матричной задаче

$$Au = f$$

вида (4.5.21), (4.5.22) с матрицей A порядка n^2:

$$A = \begin{Vmatrix} A_0 & 0 \\ \\ 0 & 0 \end{Vmatrix}.$$

Матрица разностного аналога оператора Лапласа для введенной нумерации узлов сетки имеет следующий вид:

$$B = \begin{Vmatrix} A_0 & -C \\ \\ -C & B_1 \end{Vmatrix},$$

где B_1 — симметричная положительно определенная матрица.

Минимальное отличное от нуля собственное чило матрицы HA находится из уравнения

$$\tau A\varphi = \lambda B\varphi, \tag{4.5.27}$$

где $\varphi = (\varphi_0, \varphi_1)^T$ — соответствующий собственный вектор. Из этого уравнения легко получить, что

$$\varphi_1 = B_1^{-1} C^T \varphi_0. \tag{4.5.28}$$

Умножая (4.5.27) скалярно на φ и учитывая (4.5.28), имеем

$$\lambda = \tau \frac{(A\varphi, \varphi)}{(B\varphi, \varphi)} = \tau \frac{(A_0\varphi_0, \varphi_0)}{(A_0\varphi_0, \varphi_0) - (B^{-1}C^T\varphi_0, C^T\varphi_0)} \geq \tau, \tag{4.5.29}$$

так как $(B\varphi, \varphi) > 0$ и $(B^{-1}C^T\varphi_0, C^T\varphi_0) \geq 0$.

Предположим теперь, что $\lambda = \rho$ есть максимальное собственное число задачи (4.5.27), а φ — соответствующий ему собственный век-

тор. Введем квадратные матрицы порядка n:

$$A_n = \frac{1}{h^2}\begin{Vmatrix} 2 & -1 & 0 & \dots & 0 & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & 0 & \dots & 0 & -1 & 2 \end{Vmatrix}, \quad \widehat{A}_n = \begin{Vmatrix} A_{k-1} & 0 \\ 0 & 0 \end{Vmatrix},$$

а также векторы длины n, образованные из компонент вектора φ:

$$\varphi^j = \begin{Vmatrix} \varphi_{i1} \\ \varphi_{i2} \\ \dots \\ \varphi_{in} \end{Vmatrix}, \quad i = 1, 2, \dots, n, \quad \psi^j = \begin{Vmatrix} \varphi_{1j} \\ \varphi_{2j} \\ \dots \\ \varphi_{nj} \end{Vmatrix}, \quad j = 1, 2, \dots, n.$$

Умножая (4.5.27) скалярно на вектор φ, получим

$$\rho = \frac{(A\varphi, \varphi)}{(H^{-1}\varphi, \varphi)} = \tau \frac{(A\varphi, \varphi)}{(B\varphi, \varphi)} =$$

$$= \tau \frac{\sum\limits_{i=1}^{k-1}[(A_n\varphi^i, \varphi^i) + (A_n\psi^i, \psi^i)] + \sum\limits_{i=1}^{n}[(\widehat{A}_n\varphi^i, \varphi^i) + (\widehat{A}_n\psi^i, \psi^i)]}{\sum\limits_{i=1}^{n}[(A_n\varphi^i, \varphi^i) + (A_n\psi^i, \psi^i)]}.$$

Заметим, что если какое-либо из скалярных произведений, стоящих в знаменателе, обращается в нуль, то и соответствующий вектор φ^i или ψ^i также нулевой. Так как неравенство

$$\left(\sum_{i=1}^{m} a_i \Big/ \sum_{i=1}^{m} b_i\right) \le \max_{1 \le i \le m} \frac{a_i}{b_i}$$

справедливо для любых $a_i \geq 0$, $b_i > 0$, то на его основе приходим к неравенству

$$\rho \leq \tau \max \left\{ 1, \frac{(\widehat{A}_n g, g)}{(A_n g, g)} \right\}, \tag{4.5.30}$$

где ненулевой вектор g совпадает с одним из векторов φ^i, ψ^i $(i = 1, 2, \ldots, n)$.

Величина $(\widehat{A}_n g, g)/(A_n g, g)$ не превышает спектрального радиуса матрицы $G = A_n^{-1}\widehat{A}_n$. Матрица G легко вычисляется:

$$G = \left\| \begin{array}{cc} X & 0 \\ & \\ Y & 0 \end{array} \right\|,$$

где $X = (A_{k-1} - CA_k^{-1}C^T)^{-1}A_{k-1}$, $Y = A_k^{-1}C^T X$, а C — прямоугольная $[(k-1) \times k]$-матрица с единственным отличным от нуля элементом в нижнем левом углу, равным $1/h^2$.

Отсюда следует, что спектральный радиус матрицы G совпадает с максимальным собственным числом задачи

$$A_{k-1}v = \mu(A_{k-1} - CA_k^{-1}C^T)v.$$

Вычисляя явное произведение $CA_k^{-1}C^T$, легко свести эту задачу к следующей:

$$\begin{gathered} \frac{1}{k+1}\varphi_{k-1} = \left(1 - \frac{1}{\mu}\right)\varphi_1, \\ \frac{2}{k+1}\varphi_{k-1} = \left(1 - \frac{1}{\mu}\right)\varphi_2, \\ \ldots \\ \frac{k-1}{k+1}\varphi_{k-1} = \left(1 - \frac{1}{\mu}\right)\varphi_{k-1}. \end{gathered}$$

Отсюда сразу получим

$$\mu = \frac{k+1}{2} = h^{-1}\frac{1+2h}{4}.$$

Учитывая (4.5.30), приходим к выводу, что

$$\rho \leq \frac{\tau}{h}\frac{1+2h}{4}. \tag{4.5.31}$$

Выбирая $\tau = 8/(1+6h)$, получаем оценку (4.5.25) с постоянной $C \approx 8$.

4.6. Итерационные методы при неточных входных данных

Рассмотрим операторное уравнение

$$A\varphi = f, \tag{4.6.1}$$

где A — положительно определенная матрица и f — заданный вектор.

До сих пор при рассмотрении методов решения уравнений вида (4.6.1) мы предполагали, что матрица A и вектор f заданы точно и, таким образом, требуется найти решение уравнения (4.6.1) при точных входных данных. Однако при решении практических задач очень часто приходится иметь дело не с точными входными данными, а с приближенными, т. е. вместо уравнения (4.6.1) — с уравнением

$$A^h\varphi^h = f^h, \tag{4.6.2}$$

где индекс h показывает, что входные данные зависят либо от погрешности аппроксимации, либо от различных статистических погрешностей и случайных ошибок. Будем предполагать, что ошибки в аппроксимации оператора и вектора нам известны. Иначе говоря, априори заданы оценки вида

$$\|(A - A^h)\varphi\| \le \xi(h), \quad \|f - f^h\| \le \eta(h). \tag{4.6.3}$$

Попытаемся решить задачу (4.6.1), имея в распоряжении уравнение (4.6.2) и априорную информацию (4.6.3). Поскольку наши результаты тривиально распространяются на большинство рассмотренных итерационных процессов, ограничимся описанием алгоритма приближенного решения задачи (4.6.1) на основе простейшей схемы. Рассмотрим итерационный процесс

$$[\varphi^h]^{j+1} = [\varphi^h]^j - \tau(A^h[\varphi^h]^j - f^h), \quad [\varphi^h]^0 = 0, \tag{4.6.4}$$

где параметр τ удовлетворяет условию

$$q = \|E - \tau A^h\| < 1. \tag{4.6.5}$$

Возникает вопрос о том, как долго следует продолжать итерационный процесс, если заранее известно, что входные данные заданы с погрешностями в виде (4.6.3). Естественно предположить, что при заданных погрешностях последовательные приближения следует продолжать до тех пор, пока ошибка итерационного процесса не станет приблизительно равной ошибке, возникающей от аппроксимации. Если ограничиться таким номером итерации, при котором эти ошибки различной природы являются одинаковыми, то ошибка в приближенном решении окажется неулучшаемой. Более того, если матрица A плохо обусловлена и, следовательно, обратная матрица $[A^h]^{-1}$ может отличаться от A^{-1} очень значительно, то попытка продлить итерационный процесс (4.6.4) может привести не к улучшению, а, наоборот, к существенному ухудшению результата. Именно поэтому возникает задача: по заданным погрешностям во входных данных найти оптимальное число итераций, при котором происходит согласование всякого рода погрешностей.

Проведем следующий анализ. Рассмотрим формально уравнения (4.6.1) и (4.6.2) относительно неизвестных. Получим

$$\varphi = A^{-1}f, \quad \varphi^h = [A^h]^{-1}f^h. \tag{4.6.6}$$

С помощью этих соотношений запишем тождество

$$\varphi^h - \varphi = [A^h]^{-1}[f^h - f + (A - A^h)\varphi]. \tag{4.6.7}$$

Отсюда следует:

$$\|\varphi^h - \varphi\| \le \|[A^h]^{-1}\|[\|f - f^h\| + \|(A - A^h)\varphi\|],$$

или, с учетом априорных сведений (4.6.3),

$$\|\varphi^h - \varphi\| \le \|[A^h]^{-1}\|\|[\xi(h) + \eta(h)]\|. \tag{4.6.8}$$

Рассмотрим итерационный процесс (4.6.4). Нетрудно получить равенство

$$\varphi^h - [\varphi^h] = (E - \tau A^h)[A^h]^{-1} f^h,$$

откуда

$$\|\varphi^h - [\varphi^h]^j\| \le q^j \|[A^h]^{-1}\| \|f^h\|. \tag{4.6.9}$$

Но в силу неравенства треугольника имеем

$$\|\varphi^h - [\varphi^h]^j\| \le \|[\varphi^h]^j - \varphi^h\| + \|\varphi^h - \varphi\|,$$

откуда с учетом (4.6.8) и (4.6.9) получаем, что

$$\|\varphi^h - [\varphi^h]^j\| \le q^j \|[A^h]^{-1}\| \|f^h\| + \|[A^h]^{-1}\| [\xi(h) + \eta(h)]. \tag{4.6.10}$$

Первое слагаемое в правой части неравенства (4.6.10) дает оценку погрешности итерационного процесса, а второе слагаемое оценивает погрешность за счет ошибок во входных данных. Потребуем, чтобы эти ошибки были одинаковы:

$$q^j \|[A^h]^{-1}\| \|f^h\| = \|[A^h]^{-1}\| [\xi(h) + \eta(h)]. \tag{4.6.11}$$

Мы получили уравнение для номера $j = j_0$ итерации, на которой процесс следует закончить:

$$j_0 = \left[\frac{1}{\ln q} \ln \frac{\xi(h) + \eta(h)}{\|f^h\|} \right]. \tag{4.6.12}$$

Следует отметить, что в формуле (4.6.12) норма обратного оператора отсутствует. Это существенно упрощает вычисление оптимального числа итераций.

Мы видим, что формула (4.6.12), кроме априорной информации о $\xi(h), \eta(h)$ и $\|f^h\|$, еще содержит $q = \|E - \tau A^h\|$. В случае, когда норма является евклидовой, $\|\cdot\|_2$, эта величина может быть найдена с помощью максимального собственного числа оператора T^*T, где $T = E - \tau A^h$:

$$q = \sqrt{\beta(T^*T)}.$$

Для вычисления верхней границы спектра оператора T следует воспользоваться методом, изложенным в 1.1.

4.7. Прямые методы решения конечно-разностных уравнений

В последнее время появилось значительное количество работ по применению прямых методов для решения систем конечно-разностных уравнений. В связи с этим в первую очередь следует отметить метод Фурье, который применялся для решения разностных уравнений и раньше, но, как правило, в очень редких случаях. Это объясняется тем, что по количеству арифметических операций, необходимых для решения задачи, дискретный метод Фурье уступает другим методам: большая часть работы приходилась на расчет системы собственных функций, а затем на нахождение коэффициентов ряда Фурье и его суммы.

Весьма эффективным для решения систем конечно-разностных уравнений является метод циклической редукции. В сущности, метод циклической редукции является оригинальной модификацией метода исключения Гаусса и частным случаем метода факторизации.

4.7.1. Быстрое преобразование Фурье

Идея быстрого преобразования Фурье высказывалась неоднократно, но только недавно был изложен алгоритм, приведший к значительному уменьшению количества необходимых операций, что стимулировало большой интерес к методу.

Пусть имеется функция дискретного аргумента $f(k)$, где параметр $k = 0, 1, \ldots, N-1$. Представим эту функцию в виде конечного ряда (т. е. суммы) Фурье:

$$f(k) = \sum_{n=0}^{N-1} A(n) W^{kn},$$
$$A(n) = \frac{1}{N} \sum_{k=0}^{N-1} f(k) W^{-kn}. \tag{4.7.1}$$

Здесь введено следующее обозначение для главного корня N-й степени из единицы:

$$W = e^{i2\pi/N}.$$

Назовем операцией выполнение подряд двух действий в комплексной арифметике, а именно сложения и умножения. Тогда из (4.7.1) следует, что при заданных $A(n)$ и W^{kn} потребуется N^2 операций для нахождения $f(k)$.

Идея состоит в том, что если N не является простым, то можно значительно уменьшить число операций, представив (4.7.1) в виде кратной суммы.

В самом деле, рассмотрим случай $N = N_1 \cdot N_2$, где N_1 и N_2 — натуральные числа. Представим k и n в виде

$$\begin{aligned} &k = k_1 N_2 + k_2; \quad k_1 = 0, 1, \ldots, N_1 - 1; \quad k_2 = 0, 1, \ldots, N_2 - 1; \\ &n = n_1 + n_2 N_1; \quad n_1 = 0, 1, \ldots, N_1 - 1; \quad n_2 = 0, 1, \ldots, N_2 - 1. \end{aligned} \tag{4.7.2}$$

Так как

$$W^{k_1 n_2 N_1 N_2} = (W^N)^{k_1 n_2} = 1,$$

то

$$W^{k n_2 N_1} = W^{k_2 n_2 N_1}$$

и

$$f(k) = f(k_1, k_2) = \sum_{n_1=0}^{N_1-1} \left[\sum_{n_2=0}^{N_2-1} A(n_1, n_2) W^{k_2 n_2 N_1} \right] W^{k n_1}. \tag{4.7.3}$$

Следовательно, нахождение суммы ряда (4.7.1) сводится к нахождению двойной суммы (4.7.3) или, что то же, к последовательному нахождению сумм рядов

$$A_1(n_1, k_2) = \sum_{n_2=0}^{N_2-1} A(n_1, n_2) W^{k_2 n_2 N_1}, \tag{4.7.4}$$

$$f(k_1, k_2) = \sum_{n_1=0}^{N_1-1} A_1(n_1, k_2) W^{k n_1}. \tag{4.7.5}$$

Но из (4.7.2) и (4.7.4) следует, что для нахождения A_1 требуется $N N_2$ операций. Зная A_1, с помощью (4.7.5) находим $f(k)$, применяя $N N_1$ операций. Следовательно, всего потребуется $N(N_1 + N_2)$ операций. Чем больше N, тем значительнее уменьшается число операций.

Легко видеть, что если N_1 — простое, а N_2 — составное число, то это преобразование можно применить к сумме (4.7.4), в которой n_1 является параметром, и еще уменьшить число операций, предста-

вив N_2 в виде произведения. И вообще, если $N = N_1 \times N_2 \times \ldots \times N_m$, то вместо N^2 операций мы придем к $N(N_1 + \ldots + N_m)$ операциям, причем наибольшее уменьшение получается при $N_i = 2, 3$ или 4. Если, например, $N = 256 = 2^8$, то число операций уменьшится в $256/(8 \times 2) = 16$ раз, а для $N = 243 = 3^5$ — в $243/(5 \times 3) = 16,2$ раза.

С точки зрения программирования, наиболее удобен случай $N_i = 2$ $(i = 1, 2, \ldots, m)$, хотя имеются экономичные варианты и при других N_i $(N_i = 4,\ 8)$. Рассмотрим этот случай: $N = 2^m$. Для получения соответствующих формул можно положить $N_1 = 2,\ N_2 = 2^{m-1}$ и получить суммы типа (4.7.4) и (4.7.5), а затем продолжить этот процесс. Имеем

$$k = \overline{k_{m-1}k_{m-2}\ldots k_1k_0} \equiv k_{m-1}2^{m-1} + k_{m-2}2^{m-2} + \ldots + k_2 2 + k_0,$$

$$n = \overline{n_{m-1}n_{m-2}\ldots n_1n_0} \equiv n_{m-1}2^{m-1} + n_{m-2}2^{m-2} + \ldots + n_1 2 + n_0,$$

где k_i и n_i равны 0 или 1. Тогда

$$f(k_{m-1}, \ldots k_0) = \sum_{n_0=0}^{1} \left\{ \sum_{n_1=0}^{1} \left[\ldots \sum_{n_{m-1}=0}^{1} (A(n_{m-1}, \ldots, n_0) W^{k n_{m-1} 2^{m-1}}) \ldots W^{k n_1 2} \right] W^{k n_0} \right\}. \tag{4.7.6}$$

Так как

$$W^{k n_{m-1} 2^{m-1}} = W^{k_0 n_{m-1} 2^{m-1}},$$

$$W^{k n_{m-2} 2^{m-2}} = W^{\overline{k_1 k_0} n_{m-2} 2^{m-2}}$$

и т. д., то нахождение кратной суммы (4.7.6) сводится к последовательному вычислению m сумм:

$$A_1(k_0, n_{m-2}, \ldots, n_0) = \sum_{n_{m-1}=0}^{1} A(n_{m-1}, \ldots n_0) W^{k_0 n_{m-1} 2^{m-1}},$$

$$A_2(k_1, k_0, n_{m-3}, \ldots, n_0) = \sum_{n_{m-2}=0}^{1} A_1(k_0, n_{m-2}, \ldots n_0) W^{\overline{k_1 k_0} n_{m-2} 2^{m-2}},$$

$$\ldots \tag{4.7.7}$$

$$A_m(k_{m-1}, \ldots, k_0) = \sum_{n_0=0}^{1} A_{m-1}(k_{m-2}, \ldots k_0, n_0) W^{\overline{k_{m-1} \ldots k_0} n_0},$$

$$f(k) = A_m(k_{m-1}, \ldots k_0).$$

Следует отметить, что быстрое преобразование Фурье весьма эффективно используется в корреляционном анализе обработки статистических данных для случайных величин $f(k)$ $(k = 0, 1, \ldots, N-1)$.

Рассмотрим теперь задачу Дирихле для уравнения

$$\begin{aligned} &-\Delta\varphi + \mu\varphi = f \quad \text{в} \quad D, \\ &\varphi = 0 \quad \text{на} \quad \partial D. \end{aligned} \tag{4.7.8}$$

Здесь μ — заданная константа, а f — заданная в $D = \{0 \le x \le 1,\ 0 \le y \le 1\}$ функция, обладающая необходимой гладкостью.

Поставим в соответствие задаче (4.7.8) ее разностный аналог

$$\begin{aligned} &\frac{4\varphi_{k,l} - \varphi_{k-1,l} - \varphi_{k+1,l} - \varphi_{k,l-1} - \varphi_{k,l+1}}{h^2} + \mu\varphi_{k,l} = f_{k,l} \quad \text{в} \quad D_h, \\ &\varphi_{k,l} = 0 \quad \text{на} \quad D_h, \\ &0 \le k \le \frac{1}{h} = N, \quad 0 \le l \le \frac{1}{h} = N. \end{aligned} \tag{4.7.9}$$

Если $\mu \ge 0$, то решения задач (4.7.8) и (4.7.9) существуют и единственны. В случае $\mu < 0$ требования существования решений задач (4.7.8) и (4.7.9) накладывают дополнительные ограничения на μ и f.

Предположим, что решения задач (4.7.8) и (4.7.9) существуют и единственны.

Введем обозначения

$$\varphi_l = \begin{Vmatrix} \varphi_{1,l} \\ \ldots \\ \varphi_{N-1,l} \end{Vmatrix}, \quad f_l = \begin{Vmatrix} f_{1,l} \\ \ldots \\ f_{N-1,l} \end{Vmatrix}, \quad l = 1, 2, \ldots, N-1,$$

$$A = \begin{Vmatrix} 2 & -1 & 0 & \dots & 0 & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & 0 & \dots & 0 & -1 & 2 \end{Vmatrix},$$

где A — матрица порядка $N-1$. Обозначим через E единственную матрицу того же порядка.

Перепишем задачу (4.7.9) в следующем виде:

$$\begin{aligned} &B\varphi_1 - \varphi_2 = h^2 f_1, \\ &-\varphi_{l-1} + B\varphi_l - \varphi_{l+1} = h^2 f_l, \quad l = 2, \dots, N-2, \\ &-\varphi_{N-2} + B\varphi_{N-1} = h^2 f_{N-1}, \end{aligned} \tag{4.7.10}$$

где $B = A + (2 + \mu h^2)E$.

Отметим, что матрицы B и A имеют общий базис собственных векторов и решение полной проблемы собственных значений

$$Au^{(m)} = \lambda_m(A)u^{(m)}$$

имеет вид

$$\lambda_m(A) = 2\left(1 - \cos\frac{m\pi}{N}\right), \quad u_k^{(m)} = \sqrt{\frac{2}{N}}\sin\frac{m\pi k}{N},$$

где $u_k^{(m)}$ — компонента с номером k собственного вектора $u^{(m)}$, $k = 1, 2, \dots, N-1$; $m = 1, 2, \dots, N-1$. Множитель $\sqrt{2/N}$ введен из условия нормировки

$$\|u^{(m)}\|^2 = \sum_{k=1}^{N-1}(u_k^{(m)})^2 = 1.$$

Так как векторы $u^{(m)}$ образуют ортонормированный базис в $(N-1)$-мерном пространстве, то векторы φ_l и f_l $(l = 1, 2, \dots, N-1)$ можно

представить в виде

$$\varphi_l = \sum_{m=1}^{N-1} \Phi_{m,l} u^{(m)}, \quad f_l = \sum_{m=1}^{N-1} F_{m,l} u^{(m)}. \tag{4.7.11}$$

Подставляя эти выражения в систему (4.7.10) и умножая обе части на вектор $u^{(m)}$, получим для каждого фиксированного m систему уравнений с трехдиагональной матрицей:

$$\begin{aligned} &\lambda_m \Phi_{m,1} - \Phi_{m,2} = F_{m,1}, \\ &-\Phi_{m,l-1} + \lambda_m \Phi_{m,l} - \Phi_{m,l+1} = F_{m,l}, \quad l = 2, \ldots, N-2, \\ &-\Phi_{m,N-2} + \lambda_m \Phi_{m,N-1} = F_{m,N-1}. \end{aligned} \tag{4.7.12}$$

Здесь $\lambda_m = \lambda_m(B) = \lambda_m(A) + 2 + \mu h^2$.

Таким образом, для того чтобы решить систему (4.7.10), достаточно вычислить $N-1$ раз коэффициенты Фурье векторов f_l, решить $N-1$ систему с трехдиагональными матрицами вида (4.7.12), определяющими коэффициенты Фурье векторов φ_l $(l = 1, 2, \ldots, N-1)$, и вычислить φ_l по (4.7.11). Разложение в ряд Фурье можно осуществлять, используя быстрое преобразование Фурье. Для этого формулы, определяющие коэффициенты Фурье $F_{m,l}$ вектора f_l, можно записать в следующем виде:

$$F_{m,l} = \sqrt{\frac{2}{N}} \sum_{n=1}^{N-1} f_{n,l} \sin \frac{m\pi n}{N} = \sqrt{\frac{2}{N}} \sum_{n=1}^{2N-1} f_{n,l} \sin \frac{2m\pi n}{2N},$$

где $f_{0,l} = f_{N,l} = \ldots = f_{2N-1,l} = 0$. Обозначим через $\widetilde{w}$ значение главного корня степени $M = 2N$ из единицы, тогда

$$F_{m,l} = \sqrt{\frac{2}{N}} \operatorname{Im} \left(\sum_{n=0}^{M-1} f_{n,l} \widetilde{w}^{nm} \right), \quad m = 1, 2, \ldots, N-1,$$

и для вычисления сумм можно непосредственно применять описанный алгоритм. Аналогичным образом осуществляется и вычисление векторов φ_l.

Описанный алгоритм прямого решения уравнения Гельмгольца применим не только к условиям Дирихле, но и к граничному условию Неймана и условию периодичности функции $\varphi(x, y)$ на границах

квадрата для уравнения

$$a(y)\frac{\partial^2\varphi}{\partial x^2}+\frac{\partial}{\partial y}\left(b(y)\frac{\partial\varphi}{\partial y}\right)-\mu(y)\varphi=f(x,y).$$

Обратим внимание на то, что при применении ЭВМ, в которых округление результата арифметических действий производится путем простого отбрасывания лишних цифр, при достаточно больших N может происходить значительное уменьшение точности конечного результата[3)].

4.7.2. Метод циклической редукции

Снова рассмотрим систему (4.7.10) и предположим, что $N=2^{k+1}$. Напомним, что этот случай является наиболее подходящим для применения описанного алгоритма для вычисления конечных рядов Фурье. Тем не менее при $N=2^{k+1}$ существует прямой метод решения системы (4.7.10), сравнимый по количеству арифметических операций с алгоритмом, основанным на использовании разложений в ряд Фурье. Этим методом является метод циклической редукции, и для его реализации не требуется знание собственных векторов и значений матрицы B. Последнее является несомненным преимуществом метода циклической редукции.

Идея метода состоит в том, что при четном N из системы уравнений (4.7.10):

$$-\varphi_{l-1}+B\varphi_l-\varphi_{l+1}=h^2f_l,\quad l=1,2,\ldots,N-1,$$

можно получить систему уравнений аналогичного вида, содержащую φ_l только с четным l.

Выпишем последовательно три матричных уравнения из (4.7.10):

$$-\varphi_{l-2}+B\varphi_{l-1}-\varphi_l=h^2f_{l-1},$$

$$-\varphi_{l-1}+B\varphi_l-\varphi_{l+1}=h^2f_l,$$

$$-\varphi_l+B\varphi_{l+1}-\varphi_{l+2}=h^2f_{l+1}$$

[3)] См., например, Т. Канеко, Б. Лю [13], К. Сегет [13].

при четном l. Умножив обе части второго уравнения на матрицу B и сложив затем эти три уравнения, получим

$$-\varphi_{l-2} + B^{(1)}\varphi_l - \varphi_{l+2} = h^2 f_l^{(1)}, \quad l = 2, 4, \dots N-2, \tag{4.7.13}$$

где $B^{(1)} = B^2 - 2E,\ f^{(1)} = f_{l-1} + Bf_l + f_{l+1},\ l = 2, 4, \dots, N-2$.

Для простоты записи будем полагать, что $\varphi_0 = \varphi_N = f_0 = f_N = 0$.

Для каждого нечетного l, при известных φ_l с четными l, получаем систему уравнений

$$B\varphi_l = h^2 f_l + \varphi_{l-1} + \varphi_{l+1}, \quad l = 1, 3, \dots, N-1. \tag{4.7.14}$$

Описанный способ понижения порядка системы матричных уравнений называется *редукцией*. Отметим, что систему уравнений (4.7.13) можно решить с помощью разложения в ряд Фурье, так как матрица $B^{(1)}$ имеет общий базис собственных векторов с матрицами B и A.

Вернемся к системе (4.7.13), состоящей из $N/2-1$ матричного уравнения. Так как $N = 2^{k+1}$, то к этой системе снова можно применить редукцию и получить систему такого же вида из $N/4-1$ матричного уравнения:

$$-\varphi_{4(l-1)} + B^{(2)}\varphi_{4l} - \varphi_{4(l+1)} = h^2 f_{4l}^{(2)}, \quad l = 1, 2, \dots, N/4-1, \tag{4.7.15}$$

где

$$B^{(2)} = (B^{(1)})^2 - 2E,$$

$$f_{4l}^{(2)} = f_{4l-2}^{(1)} + f_{4l+2}^{(1)} + B^{(1)} f_{4l}^{(1)}.$$

Для каждого четного l, не кратного четырем, при известных φ_i с l, кратными четырем, получим систему уравнений

$$B^{(1)}\varphi_l = h^2 f^{(1)} + \varphi_{l-2} + \varphi_{l+2}, \quad l = 2, 6, 10, \dots, N-2. \tag{4.7.16}$$

Применив циклически редукцию k раз, придем к уравнению

$$B^{(k)}\varphi_{2^k} = h^2 f_{2^k}^{(k)}; \tag{4.7.17}$$

остальные неизвестные определяются путем последовательного решения систем

$$B^{(r)}\varphi_l = h^2 f_l^{(r)} + \varphi_{l-2^r} + \varphi_{l+2^r}, \tag{4.7.18}$$

$$l = (2i+1)2^r, \quad i = 0, 1, \ldots, 2^{k-r} - 1, \quad r = k-1, k-2, \ldots, 1, 0.$$

Матрицы $B^{(r)}$ удовлетворяют соотношениям

$$B^{(0)} = B, \quad B^{(r+1)} = (B^{(r)})^2 - 2E, \quad r = 0, 1, \ldots, k-1, \tag{4.7.19}$$

а векторы $f_l^{(r)}$ определяются по формулам

$$f^(0)_l = f_l, \quad l = 1, \ldots, N-1;$$

$$f_l^{(r+1)} = f_{l-2^r}^{(r)} + f_{l+2^r}^{(r)} + B^{(r)} f_l^{(r)}, \tag{4.7.20}$$

$$l = j \cdot 2^{r+1}, \quad j = 1, 2, \ldots, 2^{k-r} - 1, \quad r = 0, 1, \ldots, k-1.$$

Матрицу $B^{(r)}$ можно представить в виде произведения 2^r трехдиагональных матриц. Действительно, рассмотрим последовательность многочленов

$$P_1(b) = b,$$

$$P_{2^{r+1}}(b) = (P_{2^r}(b))^2 - 2, \quad r = 0, 1, \ldots.$$

Если $b = 2\cos\varphi$, то

$$P_{2^r}(b) = 2\cos 2^r\varphi.$$

Следовательно, величины

$$b_i = 2\cos\frac{(2i-1)\pi}{2^{r+1}}, \quad i = 1, 2, \ldots, 2^r,$$

являются корнями многочлена $P_{2^r}(b)$ и нужное нам разложение имеет вид

$$B^{(r)} = \prod_{i=1}^{2^r} \left(B - 2\cos\frac{(2i-1)\pi}{2^{r+1}} E \right). \tag{4.7.21}$$

Таким образом, в явном вычислении матрицы $B^{(r)}$ нет необходимости и для ее обращения на векторе достаточно выполнить 2^r прогонок для трехдиагональных матриц (см. 4.7.3).

Следует отметить, что при вычислениях по формулам (4.7.20) компоненты векторов $f_l^{(r)}$ чрезвычайно быстро увеличиваются, что

является следствием быстрого роста собственных значений матриц $B^{(r)}$, и это приводит к неустойчивости счета. Рассмотрим устойчивый (с вычислительной точки зрения) алгоритм реализации метода циклической редукции.

В этом алгоритме векторы $f_l^{(r)}$ из (4.7.20) представляются в виде

$$\begin{aligned} &h^2 f_l^{(r)} = q_l^{(r)} - B^{(r)} p_l^{(r)}, \\ &l = i \cdot 2^r, \quad i = 1, 2, \ldots, 2^{k+1-r} - 1. \end{aligned} \tag{4.7.22}$$

Подставив (4.7.22) в (4.7.20), получим формулы для последовательного вычисления векторов $p_l^{(r)}$ и $q_l^{(r)}$:

$$\begin{aligned} &p_l^{(0)} = 0, \quad q_l^{(0)} = h^2 f_l^{(0)}, \quad l = 1, 2, \ldots, 2^{k+1} - 1; \\ &p_l^{(r+1)} = p_l^{(r)} + (B^{(r)})^{-1}(p_{l-2^r}^{(r)} + p_{l+2^r}^{(r)} - q_l^{(r)}), \\ &q_l^{(r+1)} = q_{l-2^r}^{(r)} + q_{l+2^r}^{(r)} - p_l^{(r+1)}, \\ &l = i \cdot 2^{r+1}, \quad i = 1, 2, \ldots, 2^{k-r} - 1. \end{aligned} \tag{4.7.23}$$

Учитывая (4.7.22), можно переписать системы матричных уравнений (4.7.17), (4.7.18) в виде

$$B^{(r)}(\varphi_l + p_l^{(r)}) = q_l^{(r)} + \varphi_{l-2^r} + \varphi_{l+2^r}, \tag{4.7.24}$$

$$l = (2i+1)2^r, \quad i = 0, 1, \ldots, 2^{k-r} - 1, \quad r = k, k-1, \ldots, 1, 0.$$

Метод циклической редукции применим не только для решения уравнения Гельмгольца с условием Дирихле, но и к граничному условию Неймана и условию периодичности функции $\varphi(x, y)$ на границах квадрата для уравнений более сложного вида[4)].

4.7.3. Факторизация разностных уравнений

За последнее время в математической литературе появилось большое число работ, посвященных прямым методам решения конечно-

[4)] См., например, А. А. Самарский, Е. С. Николаев [8].

разностных уравнений. Для широкого класса одномерных уравнений эффективным оказался метод факторизации.

По-видимому, в настоящее время рассматривать вопрос о приоритете в разработке этого метода не стоит, поскольку здесь мы имеем дело с типичным возрождением старых хороших идей с расстановкой акцентов в них на новых местах, выбор которых обусловлен современным состоянием и возможностями вычислительной математики и техники.

Сущность метода может быть продемонстрирована на простейшей задаче диффузии:

$$p\frac{d^2\varphi}{dx^2} - q\varphi = -f(x),$$

$$\frac{d\varphi}{dx} = 0 \quad \text{при} \quad x = 0,$$

$$\varphi = 0 \quad \text{при} \quad x = 1,$$

где p и q — некоторые положительные константы, а f — заданная непрерывная функция аргумента x.

Непрерывное решение сформулированной задачи требуется найти в области $0 \le x \le 1$.

Рассмотрим оператор

$$L = p\frac{d^2}{dx^2} - q$$

и представим его в виде произведения двух операторов:

$$L = p\left(\frac{d}{dx} + \beta\right)\left(\frac{d}{dx} - \alpha\right),$$

где α и β — функции, которые следует найти из условия

$$p\left(\frac{d}{dx} + \beta\right)\left(\frac{d}{dx} - \alpha\right) \equiv \frac{d^2}{dx^2} - q.$$

Для выполнения этого тождества достаточно потребовать, чтобы

$$\alpha = \beta, \quad \frac{d\beta}{dx} + \beta^2 = \frac{q}{p}.$$

Введем в рассмотрение новую функцию z по формуле

$$z = -\frac{d\varphi}{dx} + \beta\varphi.$$

Тогда уравнение диффузии сведется к эффективной системе уравнений первого порядка

$$\frac{d\beta}{dx} + \beta^2 = \frac{q}{p},$$

$$\frac{dz}{dx} + \beta z = \frac{f}{p},$$

$$\frac{d\varphi}{dx} - \beta\varphi = -z.$$

Эту систему уравнений будем назвать факторизованной.

Присоединим к ней граничные условия

$$\beta = 0 \quad \text{при} \quad x = 0,$$

$$z = 0 \quad \text{при} \quad x = 0,$$

$$\varphi = 0 \quad \text{при} \quad x = 1.$$

Непосредственно убеждаемся, что при таком выборе «начальных» данных задачи Коши для нашей системы уравнений граничные условия в задаче диффузии: $d\varphi/dx = 0$ при $x = 0$ и $\varphi = 0$ при $x = 1$ — будут выполнены. Таким образом, вместо краевой задачи для уравнения диффузии, мы пришли к задаче Коши для трех обыкновенных дифференциальных уравнений первого порядка, которые решаются последовательно одно за другим. При этом для устойчивости счета вычисление β и z следует вести в сторону возрастающих значений x, а φ — в сторону убывающих. Именно в такой редукции краевой задачи для уравнения второго порядка к задачам Коши с уравнениями первого порядка и состоит *метод аналитической факторизации*. Указанный метод весьма просто обобщается на случай более сложных задач диффузии с кусочно-разрывными функциями p, q и f и различными краевыми условиями.

Переходим к рассмотрению разностных уравнений. В этом случае алгоритм факторизации, или, как его обычно называют, прогонки, определяется следующим образом. Запишем исходное разност-

ное уравнение в виде

$$-a_i\varphi_{i-1} + p_i\varphi_i - c_i\varphi_{i+1} = f_i, \quad i = 1, 2, \ldots, n. \tag{4.7.25}$$

Будем полагать, что граничные условия для искомой функции φ уже использованы, так что[5)]

$$a_1 = 0, \quad c_n = 0, \quad p_i \geq a_i + c_i. \tag{4.7.26}$$

Исходное уравнение (4.7.25) запишем в векторно-матричной форме:

$$A\Phi = F, \tag{4.7.27}$$

где $\Phi = (\varphi_1, \varphi_2, \ldots, \varphi_n)^T$, $F = (f_1, f_2, \ldots, f_n)^T$. Представим уравнение (4.7.27) в виде

$$KS_1S_2\Phi = F, \tag{4.7.28}$$

где

$$S_1 = \begin{Vmatrix} 1 & 0 & \ldots & 0 & 0 & 0 \\ -\alpha_2 & 1 & \ldots & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & -\alpha_{n-1} & 1 & 0 \\ 0 & 0 & \ldots & 0 & -\alpha_n & 1 \end{Vmatrix},$$

$$S_2 = \begin{Vmatrix} 1 & -\xi_1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & -\xi_2 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & 0 & 0 & \ldots & 1 & -\xi_{n-1} \\ 0 & 0 & \ldots & 0 & 0 & 1 \end{Vmatrix},$$

[5)] Граничные условия для функции φ, записанные в дискретной форме, можно рассматривать и как самостоятельные уравнения.

а K — диагональная матрица, играющая роль коэффициента.

Обозначив $S_2\Phi = Z, K^{-1} = \Gamma$, получим систему уравнений

$$S_1 Z = \Gamma F, \quad S_2\Phi = Z. \tag{4.7.29}$$

Запишем уравнение для Z и Φ покомпонентно:

$$-\alpha_i Z_{i-1} + Z_i = \gamma_i f_i, \tag{4.7.30}$$

$$\varphi_i - \xi_i \varphi_{i+1} = Z_i. \tag{4.7.31}$$

Для отыскания коэффициентов $\alpha_i,\ \xi_i, \gamma_i$ подставим в уравнение (4.7.30) выражения для Z_{i-1} и Z_i из (4.7.31):

$$-\alpha_i \varphi_{i-1} + (1 + \alpha_i \xi_{i-1})\varphi_i - \xi_i \varphi_{i+1} = \gamma_i f_i.$$

Сопоставляя результат с исходным уравнением (4.7.25), получим

$$\begin{gathered} -\alpha_i = \gamma_i a_i, \quad \xi_i = \gamma_i c_i, \\ \gamma_i = (p_i - a_i \gamma_{i-1} c_{i-1})^{-1}. \end{gathered} \tag{4.7.32}$$

Уравнения (4.7.30), (4.7.31) перепишем окончательно в виде

$$\begin{gathered} Z_i = \gamma_i (a_i Z_{i-1} + f_i), \\ \varphi_i = \gamma_i c_i \varphi_{i+1} + Z_i. \end{gathered} \tag{4.7.33}$$

Известно, что для счетной устойчивости уравнений (4.7.9), (4.7.10) достаточно выполнения условий $a_k \geq 0,\ c_k \geq 0,\ p_k > 0,\ a_k + c_k \leq p_k$, причем в последнем условии строгое неравенство должно иметь место хотя бы для одного значения k.

Заметим, что обычно формулы для решения разностного уравнения (4.7.25) записывают в следующем виде:

$$\begin{gathered} \beta_{i+1} = (p_i - a_i \beta_i)^{-1} c_i, \\ z_{i+1} = (p_i - a_i \beta_i)^{-1} (a_i z_i + f_i), \\ \varphi_i = \beta_{i+1} \varphi_{i+1} + z_{i+1}, \quad i = 1, 2, \ldots, n, \end{gathered} \tag{4.7.34}$$

при условии

$$\beta_1 = z_1 = 0, \quad \varphi_n = z_{n+1}.$$

В дальнейшем метод факторизации, иногда именуемый «прогонкой», был обобщен на случай системы обыкновенных линейных дифференциальных уравнений первого порядка с произвольными линейными ограничениями, содержащими как частный случай многоточечные и краевые условия.

Алгоритм (4.7.33), (4.7.34) остается в силе, если функции φ_i и f_i, входящие в (4.7.25), — векторы, а коэффициенты $a_i,\ c_i,\ p_i$ — матрицы. (Не следует только в формулах (4.7.33), (4.7.34) менять местами множители.)

С учетом условий (4.7.26) рекуррентные формулы (4.7.33), (4.7.34) позволяют получить решение.

Метод факторизации (4.7.33), называемый методом векторной прогонки, применительно к двумерному разностному уравнению эллиптического типа оказывается эффективным лишь в том случае, если по одной из переменных число узловых точек невелико.

Рассмотрим простой и достаточно эффективный метод решения двумерных и трехмерных разностных уравнений эллиптического типа — *метод неполной факторизации Булеева*. Идея метода заключается в следующем.

Пусть имеется, например, двумерное разностное уравнение

$$-a_{ik}\varphi_{i-1,k} - c_{ik}\varphi_{i+1,k} - b_{ik}\varphi_{i,k-1} - d_{ik}\varphi_{i,k+1} + p_{ik}\varphi_{ik} = f_{ik} \tag{4.7.35}$$

$$i = 1, 2, \ldots, m; \quad k = 1, 2, \ldots, n,$$

$$a_{1k} = c_{mk} = b_{i1} = d_{in} = 0, \quad p_{ik} \geq a_{ik} + c_{ik} + b_{ik} + d_{ik}. \tag{4.7.36}$$

Запишем его в векторно-матричной форме:

$$A\Phi = F,$$

где $\Phi = (\varphi_1, \varphi_2, \ldots, \varphi_N)^T,\ F = (f_1, f_2, \ldots, f_N)^T,\ N = mn$. Прибавим к левой и правой частям уравнения вектор $B\Phi$. Получим

$$(A + B)\Phi = F + B\Phi. \tag{4.7.37}$$

Матрицу B выберем так, чтобы матрица $A+B$ могла быть представлена в виде произведения двух простых матриц S_1 и S_2 с единичными элементами на главной диагонали и диагональной:

$$A+B=KS_1S_2.$$

Уравнение (4.7.37) по аналогии с уравнением (4.7.27) заменяется системой

$$\begin{aligned} S_1Z&=\Gamma(F+B\Phi),\\ S_2\Phi&=Z, \end{aligned} \tag{4.7.38}$$

которая будет решаться методом последовательных приближений.

Заменим исходное двумерное уравнение (4.7.35) системой уравнений вида

$$\begin{aligned} Z_{ik}&=\alpha_{ik}Z_{i-1,k}+\beta_{ik}Z_{i,k-1}+\gamma_{ik}[f_{ik}+D_{ik}(\varphi)-q_{ik}\varphi_{ik}],\\ \varphi_{ik}&=\xi_{ik}\varphi_{i+1,k}+\eta_{ik}\varphi_{i,k+1}+Z_{ik}. \end{aligned} \tag{4.7.39}$$

(Индекс итерационного шага опускается; $D_{ik}(\varphi)=(D\Phi)_{ik}$ с матрицей $D=KB+\operatorname{diag}q_{ik}$, $\operatorname{diag}q_{ik}$ — диагональная матрица с элементами q_{ik}.)

Для коэффициентов $\alpha_{ik},\ \beta_{ik},\ \xi_{ik},\ \eta_{ik},\ \gamma_{ik}$ получим формулы, аналогичные (4.7.32), (4.7.33):

$$\alpha_{ik}=\gamma_{ik}a_{ik},\quad \beta_{ik}=\gamma_{ik}b_{ik},\quad \xi_{ik}=\gamma_{ik}c_{ik},\quad \eta_{ik}=\gamma_{ik}d_{ik},$$

$$\gamma_{ik}=(p_{ik}-q_{ik}-a_{ik}c_{i-1,k}\gamma_{i-1,k}-b_{ik}d_{i,k-1}\gamma_{i,k-1})^{-1},$$

а для итерируемого оператора $D_{ik}(\varphi)$ — выражение

$$D_{ik}(\varphi)=a_{ik}\eta_{i-1,k}\varphi_{i-1,k+1}+b_{ik}\xi_{i,k-1}\varphi_{i+1,k-1}.$$

Коэффициент q_{ik} принимается равным $\theta_{ik}(a_{ik}\eta_{i-1,k}+b_{ik}\xi_{i,k-1}),\ 0\le\theta<1$.

Для обеспечения сходимости итерационного процесса (4.7.39) при $\theta>0$ к этим уравнениям добавляется простая итерация

$$\widetilde{\varphi}_{ik}^{(l)}=\frac{1}{\gamma_{ik}p_{ik}}(\alpha_{ik}\varphi_{i-1,k}^{(l)}+\beta_{ik}\varphi_{i,k-1}^{(l)}+\xi_{ik}\varphi_{i+1,k}^{(l)}+\eta_{ik}\varphi_{i,k+1}^{(l)}\gamma_{ik}f_{ik}), \tag{4.7.40}$$

чтобы для получения $(l+1)$-го приближения функции φ_{ik} в уравнение для $Z_{ik}^{(l+1)}$ подставлялось l-е приближение функции $\widetilde{\varphi}_{ik}$. Если задача плохо обусловлена, то квадратную скобку в (4.7.39) для точек сетки, отвечающих правой и верхней границе области, следует заменить на

$$f_{ik} + D_{ik}(\varphi) - q_{ik}\varphi_{ik} + \kappa_{ik}(a_{ik} + b_{ik})\varphi_{ik}, \tag{4.7.41}$$

где $0 \le \kappa \le 1$.

Методом индукции нетрудно показать, что коэффициенты α_{ik}, β_{ik}, ξ_{ik}, δ_{ik}, η_{ik} удовлетворяют условиям

$$\xi_{ik} + \delta_{ik} \le 1,$$

$$\alpha_{ik} + \beta_{ik} \le \frac{a_{ik} + b_{ik}}{c_{ik} + d_{ik} + (1 - \theta_{ik})(\alpha_{ik}\eta_{i-1,k} + \beta_{ik}\xi_{i,k-1}) + \kappa_{ik}(a_{ik} + b_{ik})}.$$

Выбором параметров θ_{ik} и κ_{ik} всегда можно добиться выполнения условия $\alpha_{in} + \beta_{in} \le 1$, т. е. пространственной счетной устойчивости схемы (4.7.39).

Задача Неймана решается по схеме (4.7.39), (4.7.41) без закрепления искомой функции в какой-либо точке с $\kappa_{in} > 0$ в выражении (4.7.41). При этом после каждой итерации по схеме (4.7.39), (4.7.41) из получаемого приближения вычитается его среднее значение по всей рассматриваемой области.

Рассмотрим другую схему неполной факторизации, которая определяется следующими соотношениями:

$$\begin{gathered} Z_{ik} = \alpha_{ik}Z_{i-1,k} + \gamma_{ik}[f_{ik} + D_{ik}(\varphi) - q_{ik}\varphi_{ik}], \\ \varphi_{ik} - \beta_{ik}\varphi_{i,k-1} - \delta_{ik}\varphi_{i,k+1} = \xi_{ik}\varphi_{i+1,k} + Z_{ik}. \end{gathered} \tag{4.7.42}$$

При этом для коэффициентов α_{ik}, ξ_{ik}, β_{ik}, δ_{ik}, γ_{ik} и для $D_{ik}(\varphi)$ получаются выражения

$$\alpha_{ik} = \gamma_{ik}a_{ik}, \quad \beta_{ik} = \gamma_{ik}b_{ik}, \quad \xi_{ik} = \gamma_{ik}c_{ik}, \quad \delta_{ik} = \gamma_{ik}d_{ik},$$

$$\gamma_{ik} = (p_{ik} - q_{ik} - a_{ik}c_{i-1,k}\gamma_{i-1,k})^{-1},$$

$$D_{ik}(\varphi) = a_{ik}(\beta_{i-1,k}\varphi_{i-1,k-1} + \delta_{i-1,k}\varphi_{i-1,k-1}).$$

Коэффициент q_{ik} принимается равным $\theta a_{ik}(\beta_{i-1,k} + \delta_{i-1,k})$, где $0 \le \theta \le 1$.

В точках, отвечающих правой границе области, квадратную скобку в (4.7.42) полезно заменить на

$$f_{ik} + D_{ik}(\varphi) - q_{ik}\varphi_{ik} + \kappa_{ik}a_{ik}\varphi_{ik}, \quad 0 \le \kappa_{ik} \le 1.$$

Коэффициенты системы (4.7.42) удовлетворяют условиям

$$\beta_{ik} + \delta_{ik} + \xi_{ik} \le 1,$$

$$\alpha_{ik} \le \frac{a_{ik}}{b_{ik} + d_{ik} + c_{ik} + (1-\theta)(\beta_{i-1,n} + \delta_{i-1,k})a_{ik} + \kappa_{ik}a_{ik}}.$$

Естественно, возникает вопрос: как уменьшить относительный вес итерируемого выражения в уравнениях вида (4.7.42), т. е. как получить достаточно простой факторизуемый разностный оператор $A+B$, «близкий» к заданному оператору A?

Итак, будем искать систему уравнений, эквивалентную исходному двумерному уравнению (4.7.35), в следующем виде:

$$\begin{gathered} Z_{ik} = \alpha_{ik}Z_{i-1,k} + \mu_{ik}Z_{i-1,k-1} + \nu_{ik}Z_{i-1,k+1} + \gamma_{ik}[f_{ik} + D_{ik}(\varphi) - s_{ik}\varphi_{ik}], \\ \varphi_{ik} - \beta_{ik}\varphi_{i,k-1} - \delta_{ik}\varphi_{i,k+1} = \xi_{ik}\varphi_{i+1,k} + Z_{ik}, \end{gathered} \tag{4.7.43}$$

где $\alpha_{ik},\ \beta_{ik},\ \delta_{ik},\ \xi_{ik},\ \mu_{ik},\ \nu_{ik},\ \gamma_{ik}$ — пока неизвестные коэффициенты.

Сопоставляя уравнение, эквивалентное системе (4.7.43), с исходным уравнением (4.7.35), получим выражение для $D_{ik}(\varphi)$:

$$\gamma_{ik}D_{ik}(\varphi) = (\alpha_{ik}\beta_{i-1,k} - \mu_{ik})\varphi_{i-1,k-1} + (\alpha_{ik}\delta_{i-1,k} - \nu_{ik})\varphi_{i-1,k+1} +$$

$$+ \mu_{ik}\beta_{i-1,k-1}\varphi_{i-1,k-2} + \nu_{ik}\delta_{i-1,k+1}\varphi_{i-1,k+2},$$

а также пять соотношений, связывающих эти коэффициенты с коэффициентами уравнения (4.7.35).

Поскольку неизвестных коэффициентов семь, а соотношений пять, то два коэффициента могут быть взяты произвольно. Положим

$$\mu_{ik} = \alpha_{ik}\beta_{i-1,k}, \quad \nu_{ik} = \alpha_{ik}\delta_{i-1,k}. \tag{4.7.44}$$

Тогда главные члены итерируемого выражения $D_{ik}(\varphi)$ обратятся в нуль, а соотношения для $\gamma_{ik}D_{ik}(\varphi)$ и коэффициентов $\alpha_{ik},\ \beta_{ik},\ \delta_{ik},\ \xi_{ik},\ \gamma_{ik}$ примут вид

$$\alpha_{ik} = \gamma_{ik}e_{ik},$$

$$\xi_{ik} = \gamma_{ik}c_{ik},$$

$$\beta_{ik} = \gamma_{ik}b_{ik} + \alpha_{ik}\beta_{i-1,k}\xi_{i-1,k-1}, \tag{4.7.45}$$

$$\delta_{ik} = \gamma_{ik}d_{ik} + \alpha_{ik}\delta_{i-1,k}\xi_{i-1,k+1},$$

$$\gamma_{ik} = (p_{ik} - s_{ik} - e_{ik}\xi_{i-1,k})^{-1},$$

$$\gamma_{ik}D_{ik}(\varphi) = \alpha_{ik}(\beta_{i-1,k}\beta_{i-1,k-1}\varphi_{i-1,k-2} + \delta_{i-1,k}\delta_{i-1,k+1}\varphi_{i-1,k+2}), \tag{4.7.46}$$

где

$$e_{ik} = a_{ik}(1 - \beta_{i-1,k}\varphi_{i-1,k-1} - \delta_{i-1,k}\beta_{i-1,k+1})^{-1}. \tag{4.7.47}$$

Пусть значение коэффициента $\gamma_{ik}s_{ik}$ пропорционально сумме коэффициентов в операторе $\gamma_{ik}D_{ik}(\varphi)$:

$$\gamma_{ik}s_{ik} = \theta\alpha_{ik}(\beta_{i-1,k}\beta_{i-1,k-1} + \delta_{i-1,k}\delta_{i-1,k+1}), \quad 0 \le \theta \le 1. \tag{4.7.48}$$

С учетом (4.7.48) и (4.7.47) коэффициент γ_{ik} будет вычисляться по формуле

$$\gamma_{ik} = [p_{ik} - e_{ik}(\xi_{i-1,k} + \theta\beta_{i-1,k}\beta_{i-1,k-1} + \theta\delta_{i-1,k}\delta_{i-1,k+1})]^{-1}. \tag{4.7.49}$$

Коэффициенты уравнения (4.7.43) вычисляются в следующей последовательности:

$$e_{ik},\ \gamma_{ik},\ \alpha_{ik},\ \xi_{ik},\ \beta_{ik},\ \delta_{ik}.$$

Можно считать, что по формуле (4.7.49) осуществляется явная прогонка для вектора

$$\gamma_i = (\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{in}).$$

Таблица 4.2.

θ	0	$0,3$	$0,5$	$0,6$	$0,7$	$0,8$	$0,85$	$0,9$	$0,95$	$1,0$
q	$0,86$	$0,82$	$0,76$	$0,72$	$0,65$	$0,47$	$0,43$	$0,61$	$1,0$	

Перепишем окончательно уравнение (4.7.43) в следующем виде:

$$Z_{ik} = \alpha_{ik}(Z_{i-1,k} + \beta_{i-1,k}Z_{i-1,k-1} + \delta_{i-1,k}Z_{i-1,k+1}) +$$

$$+ \gamma_{ik}f_{ik} + \alpha_{ik}[\beta_{i-1,k}\beta_{i-1,k-1}(\varphi_{i-1,k-2} - \theta\varphi_{ik}) + \tag{4.7.50}$$

$$+ \delta_{i-1,k}\delta_{i-1,k+1}(\varphi_{i-1,k+2} - \theta\varphi_{ik}),$$

$$\varphi_{ik} - \beta_{ik}\varphi_{i,k-1} - \delta_{ik}\varphi_{i,k+1} = \xi_{ik}\varphi_{i+1,k} + Z_{ik}.$$

Уравнения (4.7.50) решаются методом последовательных приближений. Сначала осуществляется явная прогонка для вектора $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{in})$, а затем неявная прогонка в обратном направлении для вектора $\Phi_i = (\varphi_{i1}, \varphi_{i2}, \ldots, \varphi_{in})$.

Решение уравнения Пуассона в квадрате при произвольных граничных условиях показывает, что итерационный процесс (4.7.50) сходится при значениях параметра θ из интервала $0 \le \theta \le 1$, причем наилучшая сходимость имеет место при $\theta = 0,8 \; - \; 0,9$.

Для примера в таблице 4.2 приведена зависимость нормы оператора шага итерационного процесса (4.7.50), обозначенной через q, от параметра θ для задачи Дирихле:

$$\frac{\partial^2\varphi}{\partial x^2} + \frac{\partial^2\varphi}{\partial y^2} = -2(2 - x^2 - y^2),$$

$$\varphi = 0 \quad \text{при} \quad x = \pm 1, \quad \varphi = 0 \quad \text{при} \quad y = \pm 1$$

при шаге сетки $\Delta x = \Delta y = 1/12$ (529 сеточных узлов).

Согласно результатам, представленным в таблице 4.2, одна итерация в этой задаче по схеме (4.7.50) при $\theta = 0,7;\ 0,8$ или $0,9$ равносильна соответственно $44,\ 77$ или 72 простым итерациям.

Таким образом, при оптимальных значениях параметра θ схема (4.7.50) в приведенной выше задаче Дирихле оказывается конкурентоспособной с лучшими методами переменных направлений при постоянном значении итерационного параметра τ.

Как и в схеме (4.7.42), на правой границе области к итерируемому выражению $\gamma_{ik}D_{ik}(\varphi) - \gamma_{ik}\sin\varphi_{ik}$ (4.7.43) полезно добавить слагаемое $\kappa_{ik}\alpha_{ik}\varphi_{ik}$. Тогда методом индукции можно показать, что коэффициенты системы (4.7.43) удовлетворяют условиям

$$\beta_{ik} + \delta_{ik} + \xi_{ik} \le 1,$$

$$\alpha_{ik} + \mu_{ik} + \nu_{ik} \le \frac{a_{ik}(1 + \beta_{i-1,k} + \delta_{i-1,k})}{(1 - \sigma_{ik})(b_{ik} + d_{ik} + c_{ik}) + [(1 - \theta)\lambda_{ik} + \rho_{ik} + \kappa_{ik}]a_{ik}},$$

где

$$\lambda_{ik} = \beta_{i-1,k}\beta_{i-1,k-1} + \delta_{i-1,k}\delta_{i-1,k+1},$$

$$\sigma_{ik} = \beta_{i-1,k}\delta_{i-1,k-1} + \delta_{i-1,k}\beta_{i-1,k+1},$$

$$\rho_{ik} = \beta_{i-1,k}\xi_{i-1,k-1} + \delta_{i-1,k}\xi_{i-1,k+1}.$$

В применении к плохо обусловленным задачам схемы (4.7.43) или (4.7.50) по скорости несколько уступают методам переменных направлений.

Задача Неймана эффективно решается по схеме (4.7.50) с $\theta = 0,8 - 0,9$ без закрепления искомой функции в какой-либо точке.

Естественно, что первая итерация по схеме (4.7.50) будет близка к точному решению задачи, если положить θ равным единице или близким к единице. Поэтому при решении трудоемкой задачи с использованием схемы (4.7.50) первую итерацию или несколько итераций имеет смысл сделать при θ, равном единице, или близком к единице, а затем перейти к оптимальному значению θ.

Следует отметить, что произвольность формы рассматриваемой двумерной области и произвольность граничных условий для искомой функции φ при использовании схем неполной факторизации не вносят никаких затруднений. Необходимо лишь, чтобы граничное условие, записанное в разностной форме, содержало искомую функцию в пределах стандартного пятиточечного шаблона, т. е. имело бы вид, аналогичный (4.7.35), причем два отличных от нуля периферийных коэффициента этого уравнения, если они положительны, в сумме не превышали бы коэффициент p_{ik}.

Как было отмечено выше, рассмотренные методы неполной факторизации оказались эффективными для решения задач Дирихле. Попытки получения хороших схем для решения плохо обусловленных задач привели к формулировке усовершенствованных схем

неполной факторизации. В операторной формулировке одна из них состоит в следующем:

$$z_{ik} = \alpha_{ik} z_{i-1,k} + \gamma_{ik}[f_{ik} + D_{ik}(\varphi) + E_{ik}(\varphi)],$$
$$\varphi_{ik} - \beta_{ik}\varphi_{i,k-1} - \delta_{ik}\varphi_{i,k+1} = \zeta_{ik}\varphi_{i+1,k} + z_{ik}, \tag{4.7.51}$$

где $\alpha_{ik},\ \beta_{ik},\ \zeta_{ik},\ \delta_{ik},\ \gamma_{ik}$ — неопределенные коэффициенты, а оператор E содержит только функции φ с пятиточечного шаблона. Положим

$$\gamma_{ik}E_{ik}(\varphi) = \alpha_{ik}\beta_{i-1,k}(\kappa\varphi_{i-1,k} + \eta\varphi_{i,k-1} + \sigma_i\omega\varphi_{i+1,k} - \theta_i\varphi_{ik}) +$$
$$+ \alpha_{ik}\delta_{i-1,k}(\kappa\varphi_{i-1,k} + \eta\varphi_{i,k+1} - \sigma_i\omega\varphi_{i+1,k} - \theta_i\varphi_{ik}), \tag{4.7.52}$$

где $\kappa,\ \eta,\ \omega,\ \theta_i$ — пока неопределенные параметры[6)],

$$\sigma_i = \begin{cases} 1 & \text{при} \quad i = 1, 2, \ldots, m-1, \\ 0 & \text{при} \quad i = m. \end{cases}$$

Система (4.7.51) эквивалентна одному уравнению:

$$(1 + \alpha_{ik}\zeta_{i-1,k})\varphi_{ik} - \alpha_{ik}\varphi_{i-1,k} - \zeta_{ik}\varphi_{i+1,k} - \beta_{ik}\varphi_{i,k-1} -$$
$$- \delta_{ik}\varphi_{i,k+1} + \alpha_{ik}\beta_{i-1,k}\varphi_{i-1,k-1} + \alpha_{ik}\delta_{i-1,k}\varphi_{i-1,k+1} -$$
$$- \gamma_{ik}D_{ik}(\varphi) + \gamma_{ik}E_{ik}(\varphi) = \gamma_{ik}f_{ik}. \tag{4.7.53}$$

Из сопоставления (4.7.52), (4.7.35) следует, что

$$\gamma_{ik}D_{ik}(\varphi) = \alpha_{ik}\beta_{i-1,k}\varphi_{i-1,k-1} + \alpha_{ik}\delta_{i-1,k}\varphi_{i-1,k+1}. \tag{4.7.54}$$

Слагаемые в выражении (4.7.54), содержащие φ_{ik}, служат для «усиления» главной диагонали факторизующейся матрицы $(A + D - E)$, так что в итоге линейная комбинация $\kappa\varphi_{i-1,k} + \eta\varphi_{i,k-1} + \sigma_i\omega\varphi_{i+1,k}$ ком-

[6)] Чтобы не нарушалась общая формулировка для $E_{ik}(\varphi)$ в правом столбце из-за введения множителя $\sigma_m = 0$, можно граничные условия справа, имеющие вид

$$-a_{m+1,k}\varphi_{m,k} + \varphi_{m+1,k} = f_{m+1,k},$$

использовать как самостоятельные уравнения. При этом на вертикали $i = m + 1$ функции $D_{ik}(\varphi)$, $E_{ik}(\varphi)$ и z_{ik} не вычисляются, а уравнения (4.7.51) начинают решать на m-м столбце после исключения из них функции $\varphi_{m+1,k}$ с помощью условий (4.7.55).

пенсирует сумму $\varphi_{i-1,k-1}+\theta_i\varphi_{ik}$, а линейная комбинация $\kappa\varphi_{i-1,k}+\eta\varphi_{i,k+1}+\sigma_i\omega\varphi_{i+1,k}$ компенсирует сумму $\varphi_{i-1,k+1}+\theta_i\varphi_{ik}$.

Потребуем, чтобы итерационные параметры $\kappa,\ \eta,\ \omega,\ \theta_i$ удовлетворяли условию

$$1+\theta_i=\kappa+\eta+\sigma_i\omega+\varepsilon, \tag{4.7.55}$$

где ε — малая положительная величина или нуль.

Сопоставление уравнений (4.7.52) и (4.7.35) с учетом (4.7.54) приводит к следующим соотношениям, связывающим коэффициенты системы (4.7.51) с коэффициентами (4.7.35):

$$\begin{gathered}1+\alpha_{ik}\xi_{i-1,k}-\theta_i\alpha_{ik}(\beta_{i-1,k}+\delta_{i-1,k})=\gamma_{ik}p_{ik},\\ \alpha_{ik}-\kappa\alpha_{ik}(\beta_{i-1,k}+\delta_{i-1,k})=\gamma_{ik}a_{ik},\\ \beta_{ik}-\eta\alpha_{ik}\beta_{i-1,k}=\gamma_{ik}b_{ik},\\ \delta_{ik}-\eta\alpha_{ik}\delta_{i-1,k}=\gamma_{ik}d_{ik},\\ \xi_{ik}-\sigma_i\omega\alpha_{ik}(\beta_{i-1,k}+\delta_{i-1,k})=\gamma_{ik}c_{ik}\end{gathered} \tag{4.7.56}$$

или

$$\gamma_{ik}=[p_{ik}+(1-\kappa\beta_{i-1,k}-\kappa\delta_{i-1,k})^{-1}a_{ik}(\theta_i\beta_{i-1,k}+\theta_i\delta_{i-1,k}-\xi_{i-1,k})]^{-1}. \tag{4.7.57}$$

Окончательно итерационную схему (4.7.51) с учетом (4.7.53) и (4.7.54) можно записать в виде

$$\begin{gathered}z_{ik}^{l}=\alpha_{ik}z_{i-1,k}^{(l)}+\gamma_{ik}f_{ik}+\\ +\alpha_{ik}[\beta_{i-1,k}(\varphi_{i-1,k-1}^{(l)}-\eta\varphi_{i,k-1}^{(l-1)})+\delta_{i-1,k}(\varphi_{i-1,k+1}^{(l-1)}-\eta\varphi_{i,k+1}^{(l-1)})+\\ +(\beta_{i-1,k}+\delta_{i-1,k})(\theta_i\varphi_{ik}^{(l-1)}-\kappa\varphi_{i-1,k}^{(l-1)}-\sigma_i\omega\varphi_{i+1,k}^{(l-1)})],\end{gathered} \tag{4.7.58}$$

$$\varphi_{ik}^{(l)}-\beta_{ik}\varphi_{i,k-1}^{(l)}-\delta_{ik}\varphi_{i,k+1}^{l}=\xi_{ik}\varphi_{i+1,k}^{(l)}+z_{ik}^{(l)}. \tag{4.7.59}$$

Уравнение (4.7.59) решается методом одномерной прогонки:

$$\begin{gathered}\rho_{ik}=(1-\beta_{ik}\delta_{i,k-1}\rho_{i,k-1})^{-1},\\ w_{ik}=\rho_{ik}(\beta_{ik}w_{i,k-1}+\xi_{ik}\varphi_{i+1,k}+z_{ik}),\end{gathered} \tag{4.7.60}$$

$$\varphi_{ik} = \rho_{ik}\delta_{ik}\varphi_{i,k+1} + w_{ik}. \tag{4.7.61}$$

Естественно, что в итерационной схеме (4.7.51) нужно стремиться к тому, чтобы в операторе $D_{ik}(\varphi) - E_{ik}(\varphi)$ сумма коэффициентов была близкой к нулю, а сумма модулей коэффициентов была как можно меньшей по сравнению с коэффициентом a_{ik} исходного уравнения (4.7.35).

Будем называть оператор $(A + B)_{ik}$ близким к оператору A_{ik} по модулю невязки, если сумма модулей коэффиициентов оператора B_{ik} мала по сравнению со значениями коэффициента a_{ik} оператора A_{ik}.

В рассматриваемой здесь схеме близость оператора $(A + B)_{ik}$ к оператору A_{ik} по модулю невязки обеспечивается соотношением (4.7.55).

Оптимальное значение θ должно быть соизмеримым с коэффициентом при $\varphi_{i-1,k-1}$ и $\varphi_{i-1,k+1}$, т. е. должно быть величиной порядка единицы. Если стремиться к тому, чтобы коэффициенты $\beta_{ik},\ \delta_{ik}$ и ξ_{ik} были приближенно пропорциональны коэффициентам $b_{ik},\ d_{ik}$ и c_{ik}, то параметры η и ω следует связать приближенным соотношением $2\omega \approx \eta$.

Как показал опыт решения задач диффузии и конвективного переноса в области прямоугольной формы, оптимальными значениями параметров $\kappa,\ \eta,\ \omega$ и ε являются

$$\kappa = 0,5 \div 1, \quad \eta = 1, \quad \omega = 0 \div 0,4, \quad \varepsilon = 0,$$

причем $\kappa + \omega = 1,0 \div 1,1$. Для уравнения Пуассона оптимальным можно считать следующий набор $\kappa = \eta = 1,\ \omega = 0,1,\ \varepsilon = 0$.

Исследуем теперь пространственную устойчивость схемы (4.7.51). Ради общности будем полагать, что в схеме (4.7.51)

$$\gamma_{ik}D_{ik}(\varphi) - \gamma_{ik}(\varphi) =$$

$$= d_{ik}[(\beta_{i-1,k-1} - \eta\varphi_{i,k-1}) + \delta_{i-1,k}(\varphi_{i-1,k-1} - \eta\varphi_{i,k+1})] +$$

$$+ \alpha_{ik}(\beta_{i-1,k} + \delta_{i-1,k})(\theta_i\varphi_{ik} - \kappa\varphi_{i-1,k} - \sigma_i\omega\varphi_{i+1,k}) + s_{ik}\alpha_{ik}\varphi_{ik}.$$

Методом индукции нетрудно показать, что коэффициенты $\alpha_{ik},\ \beta_{ik},\ \delta_{ik},\ \xi_{ik}$ удовлетворяют условиям

$$\beta_{ik} + \delta_{ik} + \xi_{ik} \le 1, \tag{4.7.62}$$

$$\alpha_{ik} = \frac{a_{ik}}{(1-\kappa_{i-1,k})(l_{ik}+d_{ik}+p_{ik}) + [(\eta+\sigma_i\omega+\varepsilon)\nu_{ik}+s_{ik}]a_{ik}}. \qquad (4.7.63)$$

Выбором параметров $\kappa,\ \eta,\ \omega,\ \varepsilon,\ s$ всегда может быть обеспечена пространственная счетная устойчивость схемы. Условия (4.7.62), (4.7.63) легко позволяют строить итерационную схему с заранее поставленными требованиями на коэффициенты системы (4.7.51).

4.8. Асимптотический анализ алгоритмов решения задач[7)]

В свое время с появлением быстродействующих ЭВМ последовательного действия эффективный прямой алгоритм решения систем уравнений с трехдиагональной матрицей, называемый прогонкой, оказал революционизирующее влияние на развитие методов решения стационарных и нестационарных задач математической физики — было построено качественно новое семейство методов, называемых методами расщеплений.

Далее, применение операторов, эквивалентных по спектру или близких, но более простых по сравнению с оператором решаемой задачи, позволило существенно ускорить сходимость итерационных методов. Во многих из этих методов роль «базового» алгоритма, ускоряющего сходимость итераций, уже играют алгоритмы быстрого решения краевых задач для уравнения Пуассона, заданного в прямоугольнике.

Идею конструирования эффективных алгоритмов для решения задач можно развить дальше в таком направлении, чтобы реализация этих алгоритмов разбивалась на каком-то достаточно продолжительном этапе на взаимно независимые вычисления. Это обстоятельство может служить естественной основой «распараллеливания» алгоритмов на ЭВМ со специальной организацией вычислений.

Пусть для аппроксимации исходной задачи мы воспользовались разностным или проекционно-сеточным методом, а теперь полученную систему алгебраических уравнений требуется решить подходящим прямым или итерационным алгоритмом. Арсенал таких алгоритмов, как мы видели выше, в настоящее время достаточно богат.

[7)] См. книгу В. И. Лебедева, Н. С. Бахвалова, В. И. Агошкова, О. В. Бабурина, А. В. Князева, В. П. Шутяева [3].

Однако здесь возникают и требуют своего решения следующие вопросы:

а) сравнительный анализ эффективности алгоритмов и возможности их распараллеливания;

б) анализ параллельных итерационных алгоритмов при нескольких характерных соотношениях между числом неизвестных и числом микропроцессоров (что позволит выявить тенденции в изменении эффективности каждого параллельного итерационного метода с ростом числа микропроцессоров);

в) нахождение почти предельного распараллеливания итерационных алгоритмов.

Решение этих вопросов при рассмотрении ряда известных алгоритмов в применении к эллиптическим задачам осуществляется при реализации их на идеализированной синхронной вычислительной системе, которую будем обозначать АП.

Пусть она состоит из p идентичных арифметических микропроцессоров с памятью, доступной каждому из них. Каждый процессор за единичный интервал времени, называемый *тактом*, может исполнять одну из бинарных арифметических операций либо простаивать. Пусть не учитывается время, требуемое для выполнения вспомогательных операций, передачи данных между процессорами и памятью. Будем предполагать, что в процессе вычислений не возникают конфликты относительно памяти и нет задержки выполнения команд устройством управления. Все исходные данные перед началом вычислений считаются находящимися в памяти. Вопросы загруженности процессоров, устранения конфликтов в памяти, улучшения характера обменов между процессорами и устойчивости алгоритмов в рамках подобной модели ЭВМ — АП также не рассматриваются. Анализ алгоритмов, осуществляемый при сделанных предположениях, будем называть *асимптотическим анализом* (учитывая при этом, что p может быть достаточно большим по величине).

4.8.1. Оценки некоторых алгоритмов линейной алгебры

Обозначим через T_p число тактов, необходимых для реализации алгоритма с помощью p процессоров. Эту величину называют высо-

той алгоритма. Величина $R_p = T_1/T_p$ называется ускорением, $E_p = R_p/p$ — эффективностью, а $C_p = pT_p$ — ценой алгоритма.

Приведем некоторые известные параллельные алгоритмы с оценками для T_p. При умножении полной $N \times N$-матрицы A на вектор $\vec{b}$ размерности N

$$T_p = O\left(\frac{N^2}{p} + \ln p\right). \tag{4.8.1}$$

Если матрица A содержит в каждой строке фиксированное, не зависящее от N число ненулевых элементов, то

$$T_p = O\left(\frac{N}{p}\right). \tag{4.8.2}$$

При умножении BA двух $N \times N$-матриц A и B известны алгоритмы, при которых

$$T_p = O\left(\frac{N}{p}\right) \tag{4.8.3}$$

при условии, что матрицы A, B в каждом столбце и в каждой строке имеют лишь фиксированное число ненулевых элементов.

Для нахождения решения системы N линейных уравнений

$$Au = f \tag{4.8.4}$$

с

$$A = D - L - U, \tag{4.8.5}$$

где D — диагональная, L — нижняя, а U — верхняя треугольная матрицы с ненулевыми диагональными элементами, известны следующие оценки для T_p:

если $U = 0$, то

$$T_p = \begin{cases} O(N^{3/2}) & \text{при} \quad p = N^{1/2}, \\ O(N) & \text{при} \quad p = N, \\ O(\ln^2 N) & \text{при} \quad p = N^3; \end{cases} \tag{4.8.6}$$

если A — трехдиагональная, то, используя метод Яненко для распараллеливания прогонки, имеем

$$T_p = O\left(\frac{N}{p} + \ln p\right); \tag{4.8.7}$$

если A — двухдиагональная, $U = 0$, то используя алгоритм рекурсивного сдваивания Когге — Стоуна, получаем

$$T_p = O\left(\frac{N}{p} + \ln p\right). \tag{4.8.8}$$

Ниже при исследовании итерационных алгоритмов решения системы из N линейных алгебраических уравнений

$$Au = f \tag{4.8.9}$$

предполагается, что

$$1 \le p \le N^3, \tag{4.8.10}$$

ибо уже при $p \asymp N^4/\ln N$[8)] существуют прямые методы решения системы (4.8.9) с использованием элементов обратной матрицы за $T_p \asymp \lg^2 \mathrm{N}$ тактов, т. е. применение в этом случае итерационных методов, вообще говоря, теряет смысл.

4.8.2. Анализ вычислительных алгоритмов решения модельной задачи

Анализ вычислительных алгоритмов в рамках сформулированной идеализированной модели ЭВМ — АП рассмотрим на примере модельной задачи в стандартной области. Так, пусть $D \subset \mathbf{R}^2$ есть квадрат $D = \{0 < x < 1;\ 0 < y < 1\}$, а также рассматривается задача

$$\begin{gathered} \Lambda u \equiv -\frac{\partial}{\partial x}\left(a_1 \frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial y}\left(a_2 \frac{\partial u}{\partial y}\right) + qu = f \quad \text{в} \quad D, \\ u|_{\partial D} = 0, \end{gathered} \tag{4.8.11}$$

где $a_1,\ a_2,\ q$ — достаточно гладкие функции, $0 \le q_0 \le q \le q_1$, $q_i = \mathrm{const}$, a_i — положительные функции в D, а f — кусочно-непрерывная функ-

[8)] Здесь и далее $a \asymp b$ обозначает соотношение $a = O(b)$.

ция. Задачу (4.8.11) аппроксимируем методом конечных разностей с использованием сетки $x_i = ih$, $y_j = jh$, $h = 1/(n+1)$ и привлечением известных пятиточечных аппроксимаций дифференциального оператора. В результате получим систему линейных алгебраических уравнений (4.8.9), где A — $N \times N$-матрица, $A = A^* > 0$, $u,\ f$ — N-мерные векторы, $N = n^2$.

Будем предполагать N таким, что норма разности между точным решением задачи и решением системы (4.8.9), проинтерполированным полилинейным образом с узлов сетки на D, есть величина ε порядка $N^{-\gamma}$ ($\varepsilon \asymp N^{-\gamma}$), где $0 < \gamma \leq 2$ — фиксированное число. Систему (4.8.9) будем решать итерационным методом с той же точностью ε в норме, представляющей разностный аналог нормы $\|\cdot\|_{W_2^1}$. Поскольку затрагиваемое при этом число итераций является функцией от $\lg \varepsilon^{-1} \sim \gamma \lg N$, то в оценках его по порядку можно условно считать $\gamma = 1$.

В применении к решению (4.8.9) был рассмотрен ряд одношаговых методов, которые в единой форме записи имеют вид

$$u^{k+1} = u^k - \alpha_{k+1} H_k (Au^k - f), \tag{4.8.12}$$

и трехчленных методов вида

$$u^{k+1} = u^k - \alpha_{k-1} B(Au^k - f) - \beta_{k+1}(u^k - u^{k-1}). \tag{4.8.13}$$

Рассматривались также: метод с аппроксимацией обратного оператора; варианты многосеточного метода; метод итерирования по центральным крестам, предложенный Н. С. Бахваловым; метод Монте-Карло; метод дискретного быстрого преобразования Фурье (ДБПФ, $a_i \equiv 1$). Через Δ_h^{-1} в дальнейшем обозначаем оператор, обратный к разностному оператору Лапласа для задачи Дирихле. Через m, M обозначим границы спектра оператора A или BA.

Приведем теперь результаты исследования распараллеленных итерационных алгоритмов решения (4.8.9) с привлечением описанной выше ЭВМ — АП из p процессоров (см. таблице 4.3).

В качестве примера проведем здесь анализ распараллеливания некоторых из перечисленных алгоритмов для $1 \leq p \leq N^2$. Для этого

матрицу A разностной задачи запишем в виде

$$A = D - L - U,$$

где D — диагональная, L — нижняя, а U — верхняя строго треугольные матрицы. Через k_0 будем обозначать число необходимых итераций в каждом методе для достижения точности $\varepsilon = O(N^{-1})$.

1. *Метод простой итерации* имеет вид

$$u^{k+1} = D^{-1}((L+U)u^k + f), \quad k = 1, 2, \ldots, \quad u^0 = 0. \tag{4.8.14}$$

Для этого метода $k_0 = O(N \ln N)$, $T_1 = O(Nk_0) = O(N^2 \ln N)$. При наличии p процессоров на одну итерацию здесь потребуется (см. (4.8.2)) $O(N/p)$ тактов для вычисления $D^{-1}((L+U)u^k - f)$. Тогда для достижения точности $\varepsilon = O(1/N)$ нам необходимо $T_p = O\left(\dfrac{N^2}{p} \ln N\right)$ тактов. Таким образом,

$$R_p = T_1/T_p \sim p, \quad E_p = R_p/p \sim 1.$$

Таблица 4.3.

Методы	p = 1	$p \asymp N^{1/2}$		$p \asymp N$		$p \asymp N^3$	
	T_1	T_p	E_p	T_p	E_p	T_p	E_p
1. Метод с аппроксимацией обратного оператора:							
а) $C_0 = \dfrac{2}{M+m} I$,	$N^3 \ln N$	$N^{5/2} \ln N$	1	$N^2 \ln N$	1	$\ln^2 N$	$\ln^{-1} N$
б) $C_0 = \alpha \Delta_h^{-1}$	$N^3 \ln \ln N$	$N^{5/2} \ln \ln N$	1	$N^2 \ln \ln N$	1	$\ln N \ln \ln N$	$\ln^{-1} N$
2. Метод простой итерации	$N^2 \ln N$	$N^{3/2} \ln N$	1	$N \ln N$	1	$\ln^2 N$	$N^{-1} \ln^{-1} N$
3. Метод Гаусса — Зейделя	$N^2 \ln N$	$N^{3/2} \ln N$	1	$N \ln N$	1	$\ln^2 N$	$N^{-1} \ln^{-1} N$
4. SOR:							
а) естественная нумерация,	$N^{3/2} \ln N$	$N \ln^2 N$	$\ln^{-1} N$	$N^{1/2} \ln^2 N$	$\ln^{-1} N$	$\ln^2 N$	$N^{-3/2} \ln^{-1} N$
б) σ-упорядочивание	$N^{3/2} \ln N$	$N \ln N$	1	$N^{1/2} \ln N$	1	$\ln^2 N$	$N^{-3/2} \ln^{-1} N$
5. SLOR:							
а) естественная нумерация	$N^{3/2} \ln N$	$N \ln N$	1	$N^{1/2} \ln^2 N$	$\ln^{-1} N$	$\ln^2 N$	$N^{-3/2} \ln^{-1} N$
б) специальная нумерация	$N^{3/2} \ln N$	$N \ln^2 N$	$\ln^{-1} N$	$N^{1/2} \ln^2 N$	$\ln^{-1} N$	$\ln^2 N$	$N^{-3/2} \ln^{-1} N$

Методы	p = 1	$p \asymp N^{1/2}$		$p \asymp N$		$p \asymp N^3$	
	T_1	T_p	E_p	T_p	E_p	T_p	E_p
6. Метод сопряженных градиентов:							
а) скалярный случай,	$N^{3/2}\ln N$	$N\ln N$	1	$N^{1/2}\ln^2 N$	$\ln^{-1} N$	$\ln^{1/2} N$	N^{-2}
б) с применением Δ_h^{-1}	$N\ln^2 N$	$N^{1/2}\ln^2 N$	1	$\ln^2 N$	1	$\ln^2 N$	N^{-2}
7. Чебышевский двухслойный итерационный метод:							
а) скалярный случай,	$N^{3/2}\ln N$	$N\ln N$	1	$N^{1/2}\ln N$	1	$\ln^2 N$	$N^{-3/2}\ln^{-1} N$
б) с применением Δ_h^{-1}	$N\ln^2 N$	$N^{1/2}\ln^2 N$	1	$\ln^2 N$	1	$\ln N\ln\ln N$	$N^{-2}\dfrac{\ln N}{\ln\ln N}$
8. Чебышевский трехчленный итерационный метод:							
а) скалярный случай,	$N^{3/2}\ln N$	$N\ln N$	1	$N^{1/2}\ln N$	1	$\ln^2 N$	$N^{-3/2}\ln^{-1} N$
б) с применения Δ_h^{-1}	$N\ln^2 N$	$N^{1/2}\ln N$	1	$ln^2 N$	1	$\ln N\ln\ln N$	$N^{-2}\dfrac{\ln N}{\ln\ln N}$

Методы	p = 1	$p \asymp N^{1/2}$		$p \asymp N$		$p \asymp N^3$	
	T_1	T_p	E_p	T_p	E_p	T_p	E_p
9. Блочный чебышевский метод:							
а) σ-упорядочивание,	$N^{3/2}\ln N$	$N\ln N$	1	$N^{1/2}\ln N$	1	$\ln^2 N$	$N^{-3/2}\ln^{-1} N$
б) специальная нумерация узлов	$N^{3/2}\ln N$	$N\ln^2 N$	$\ln^{-1} N$	$N^{1/2}\ln^2 N$	$\ln^{-1} N$	$\ln^2 N$	$N^{-3/2}\ln^{-1} N$
10. (SSOR):							
а) точечный,	$N^{5/4}\ln N$	$N^{3/4}\ln N$	1	$N^{3/4}\ln N$	$N^{-1/2}$	$\ln^2 N$	$N^{-7/4}\ln^{-1} N$
б) блочный	$N^{5/4}\ln N$	$N^{3/4}\ln^2 N$	$\ln^{-1} N$	$N^{3/4}\ln^2 N$	$N^{-1/2}\ln^{-1} N$	$\ln^2 N$	$N^{-7/4}\ln^{-1} N$
11. МПН:							
а) коммутативный случай,	$N\ln^2 N$	$N^{1/2}\ln^2 N$	1	$\ln^3 N$	$\ln^{-1} N$	$\ln N\ln\ln N$	$N^{-2}\dfrac{\ln N}{\ln\ln N}$
б) некоммутативный случай	$N^{3/2}\ln N$	$N\ln^2 N$	$\ln^{-1} N$	$N^{1/2}\ln^2 N$	$\ln^{-1} N$	$\ln^2 N$	$N^{-3/2}\ln^{-1} N$
12. Многосеточный метод:							
а) вариант 12а,	$N\ln N$	$N^{1/2}\ln N$	1	$2^{c'\sqrt{\ln N}}$	$\ln N 2^{-c'\sqrt{\ln N}}$	—	—
б) варианты 12в, г	N	$N^{1/2}$	1	$\ln^2 N$	$\ln^{-2} N$	—	—
13. Метод крестов	N	$N^{1/2}$	1	$\ln^3 N$	$\ln^{-3} N$	—	—
14. Метод Монте-Карло	$N^2\ln N$	$N^{3/2}\ln N$	1	$N\ln N$	1	$\ln N$ $(p \geq N^2)$	N^{-1}
15. ДБПФ	$N\ln N$	$N^{1/2}\ln N$	1	$\ln N$	1	—	—

2. *Метод Гаусса — Зейделя* записывается в виде

$$u^{k+1} = (D - L)^{-1}(Uu^k + f), \quad k = 1, 2, \ldots. \tag{4.8.15}$$

Для этого метода $k_0 = O(N \ln N)$, $T_1 = O(Nk_0) = O(N^2 \ln N)$. Используя вид матриц $D,\ L$ и U, этот метод можно записать в покомпонентной форме:

$$u^{k+1} = a_{ij}u^{k+1}_{i-1,j} + b_{ij}u^{k+1}_{i,j-1} + c_{ij}u^k_{i+1,j} + d_{i,j+1}u^k_{i,j+1} + f_{ij}, \tag{4.8.16}$$

$$i, j = 1, 2, \ldots, n, \quad k = 1, 2, \ldots, k_0,$$

где $a_{ij},\ b_{ij},\ c_{ij},\ d_{ij}$ — известные коэффициенты, которые определяются из вида $D,\ U,\ L$.

В данном случае целесообразно проводить распараллеливание не только по индексам i, j, но и по k.

Рассмотрим алгоритм реализации (4.8.16) с помощью p процессоров, каждый из которых обрабатывает одну компоненту a^{k+1}_{ij} за $t = 8$ тактов и которые работают в следующей последовательности:

а) сначала вычисляется первое приближение для u_{11};

б) затем вычисляются второе приближение для u_{11} и первое приближение для $u_{21},\ u_{12}$;

в) далее вычисляются первое приближение для следующей диагонали прямоугольной сетки ($u_{13},\ u_{22}, u_{31}$), второе приближение для $u_{21},\ u_{12}$ и третье приближение для u_{11} и т. д.

Вычисления разобьем на несколько шагов.

Шаг 1. Обрабатываем один узел за $8 < 1/p >$ тактов, где $< 1/p >$ — наименьшее целое, превосходящее $1/p$.

Шаг 2. Обрабатываем три узла за $8 < 3/p >$ тактов.

Шаг 3. Обрабатываем 6 узлов $(3 + 2 + 1)$ за $8 < 6/p >$ тактов.

...

Шаг $(n-1)$. Обрабатываем $(n-1)+\ldots+2+1)$ узлов за $\left\langle \dfrac{n(n-1)}{2p} \right\rangle$ тактов.

...

Шаг $(2n-1)$. Обрабатываем $(1+\ldots+n-1+n+\ldots+2+1)$ узлов за $\left\langle \dfrac{2n \cdot n}{2p} \right\rangle$ тактов.

Общее число тактов t', необходимое для $(2n-1)$ шагов, равно

$$t' = 8\sum_{k=1}^{n} \left\langle \frac{k(k+1)}{2p} \right\rangle + 8\sum_{k=1}^{n-1} \left\langle \frac{\dfrac{n(n+1)}{2} + \sum\limits_{k'=k}^{n-1} k'}{p} \right\rangle \cong O\left(\frac{n^3}{p}\right).$$

После этого нужно еще проделать k_0 шагов, на каждом из которых будем обрабатывать $\dfrac{(n-1)n}{2} + \dfrac{n(n+1)}{2}$ узлов за $O(N/p)$ тактов, где $N = n^2$.

В итоге приходим к выражению для T_p:

$$T_p = k_0 O\left(\frac{N}{p}\right) + t' \cong O\left(\frac{N^2}{p}\ln N + \frac{N^{3/2}}{p}\right) = O\left(\frac{N^2}{p}\ln N\right).$$

Здесь $R_p \sim p$, $E_p \sim 1$.

3. *Метод чебышевского ускорения* рассмотрим в виде

$$u^{k+1} = u^k - \alpha_{k+1}(Au^k - f), \quad k = 1, 2, \ldots. \tag{4.8.17}$$

Для этого метода $k_0 = O(N^{1/2}\ln N)$, $T_1 = O(N^{3/2}\ln N)$. При наличии p процессоров на одну итерацию метода потребуется (см. (4.8.2)) $O(N/p)$ тактов для вычисления $u^k - \alpha_{k+1}(Au^k - f)$. Тогда для k_0 итераций имеем

$$T_p = O\left(\frac{N^{3/2}}{p}\ln N\right).$$

Здесь $R_p \sim p$, $E_p \sim 1$.

4. *Метод последовательной верхней релаксации с оптимальным выбором параметра* (SOR). При σ-упорядочивании узлов сетки (в шахматном порядке) этот метод записывается в виде

$$\begin{aligned} u_1^{k+1/2} &= D_1^{-1}(C_1 u_2^k + f_1), \\ u_1^{k+1} &= u_1^k + \omega(u_1^{k+1/2} - u_1^k), \\ u_2^{k+1/2} &= D_2^{-1}(C_2 u_1^{k+1} + f_2), \\ u_2^{k+1} &= u_2^k + \omega(u_2^{k+1/2} - u_2^k), \end{aligned} \tag{4.8.18}$$

где D_i — диагональные матрицы.

Для этого метода $k_0 = O\left(N^{1/2}\ln N\right)$, $T_1 = O(N^{3/2}\ln N)$. При наличии p процессоров вычисления проводим по формулам (4.8.18) сначала для всех узлов, когда $i + j$ четно, затрачивая $O(N/p)$ тактов, а потом для всех узлов, когда $i + j$ нечетно, затрачивая тоже $O(N/p)$ тактов. Для k_0 итераций метода имеем

$$T_p = O\left(\frac{N^{3/2}}{p}\ln N\right).$$

Здесь $R_p \sim p$, $E_p \sim 1$.

5. *Метод последовательной верхней релаксации по линиям* (SLOR).

Используя вид матриц D, L, U, этот метод можно записать в покомпонентной форме:

$$a_{ij}u^{k+1}_{i-1,j} + b_{ij}u^{k+1}_{ij} + c_{ij}u^{k+1}_{i+1,j} = \alpha_{ij}u^{k}_{i,j-1} + \beta_{ij}u^{k}_{i,j+1} + \gamma_{ij}u^{k}_{ij} +$$

$$+ \zeta^{k}_{i-1,j}u^{k}_{i-1,j} + \eta_{ij}u^{k}_{i+1,j} + \theta_{ij}f^{k}_{ij}, \quad i,j = 1,2,\ldots,n, \quad k = 1,2,\ldots,k_0. \tag{4.8.19}$$

Здесь $k_0 = O(N^{1/2}\ln N)$, $T_1 = O(N^{3/2}\ln N)$. При наличии p процессоров вычисления можно провести следующим образом.

Для вычисления U^{k+1}_{ij} при фиксированном j с помощью $\sqrt{p}$ процессоров требуется $O\left(\sqrt{\dfrac{N}{p}} + \ln p\right)$ тактов, поскольку мы решаем трехдиагональную систему порядка $n = \sqrt{N}$, используя метод Яненко (см. (4.8.7)). Эту процедуру нужно повторить $\sqrt{N}$ раз ($j = 1,2,\ldots,n-1$), а с помощью $\sqrt{p}$ процессоров мы это проделаем за $\sqrt{N/p}$ шагов. В итоге получаем выражение для T_p:

$$T_p = k_0 O\left(\sqrt{\frac{N}{p}}\left(\sqrt{\frac{N}{p}} + \ln p\right)\right) = O\left(N^{1/2}\left(\frac{N}{p} + \sqrt{\frac{N}{p}}\ln p\right)\right)\ln N,$$

или

$$T_p = O\left(\frac{N^{3/2}}{p}\ln N + \frac{N}{\sqrt{p}}\ln p \ln N\right).$$

Здесь $R_p \sim p$, $E_p \sim 1$.

6. *Метод симметричной верхней релаксации* (SSOR) с чебышевскими параметрами (метод Шелдона) записывается в виде

$$u^{k+1/3} = u^k + \omega(D^{-1}(Lu^{k+1/3} + Uu^k + f) - u^k),$$

$$u^{k+2/3} = u^{k+1/3} + \omega(D^{-1}(Uu^{k+2/3} + Uu^{k+1/3} + f) - u^{k+1/3}), \tag{4.8.20}$$

$$u^{k+1} = u^k + \alpha_{k+1}(u^{k+2/3} - u^k).$$

Для этого метода $k_0 = O(N^{1/4}\ln N)$, $T_1 = O(N^{5/4}\ln N)$.

При наличии p процессоров, $1 \le p \le \sqrt{N}$, на одну итерацию метода потребуется:

а) $O\left(\sqrt{N}\dfrac{\sqrt{N}}{p}\right)$ тактов на определение $u^{k+1/3}$, вычисляя за $O(\sqrt{N}/p)$ тактов все неизвестные $u^{k+1/3}_{ij}$ с одинаковой суммой индексов $s = i + j$, где $s = 2,\ldots,2n$;

б) $O\left(\sqrt{N}\dfrac{\sqrt{N}}{p}\right)$ тактов на аналогичное вычисление $u^{k+2/3}$, здесь $s = 2n,\ldots,2$;

в) $O(N/p)$ тактов для вычисления u^{k+1} по u^k и $u^{k+2/3}$. В итоге получаем

$$T_p = k_0 O\left(\frac{N}{p}\right) = O\left(\frac{N^{5/4}}{p}\ln N\right).$$

Здесь

$$R_p \sim 1, \quad E_p \sim 1.$$

Если $p = \sqrt{n}$, то $T_p = O(N^{3/4}\ln N)$. Если же $p > \sqrt{N}$, то порядок числа тактов можно уменьшить за счет применения модифицированного алгоритма Когге — Стоуна. Мы ограничиваемся лишь рассмотренной реализацией метода.

7. *Метод переменных направлений* (МПН) имеет вид

$$\begin{aligned}(E + \tau_1^k L_x)u^{k+1/2} &= (E - \tau_1^k L_y)u^k + \tau_1^k f,\\ (E + \tau_2^k L_y)u^{k+1} &= (E - \tau_2^k L_x)u^{k+1/2} + \tau_2^k f,\end{aligned} \tag{4.8.21}$$

где L_x, L_y — трехдиагональные матрицы.

Будем рассматривать только коммутативный случай, когда $L_x L_y = L_y L_x$. В этом случае

$$k_0 = O(\ln^2 N), \quad T_1 = O(N\ln^2 N).$$

При наличии p процессоров на одну итерацию метода потребуется:

а) $O\left(\sqrt{\frac{N}{p}}\left(\sqrt{\frac{N}{p}} + \ln p\right)\right)$ тактов для отыскания $u^{k+1/2}$ (при этом нам потребуется $\sqrt{N/p}$ раз одновременно решать $\sqrt{p}$ трехдиагональных систем порядка $\sqrt{N}$, для чего мы можем использовать метод Яненко (см. (4.8.7)));

б) $O\left(\sqrt{\frac{N}{p}}\left(\sqrt{\frac{N}{p}} + \ln p\right)\right)$ тактов для отыскания аналогичным образом u^{k+1}.

В итоге для k_0 итераций метода имеем

$$T_p = k_0 O\left(\sqrt{\frac{N}{p}}\left(\sqrt{\frac{N}{p}} + \ln p\right)\right) = O\left(\sqrt{\frac{N}{p}}\left(\sqrt{\frac{N}{p}} + \ln p\right)\ln^2 N\right).$$

Здесь $R_p \sim p$, $E_p \sim 1$.

На основе приведенных результатов можно высказать ряд предположений. Так, например, при $p \asymp N^{1/2}$ можно предположить, что: 1) явные «точечные» методы 4б, 7а, 8а, 9а естественно предельно и просто распараллеливаются, все они имеют одинаковые спектральные радиусы оператора перехода. Очевидно, что методу 4б стоит предпочесть метод 9а, а последний имеет лучшие характеристики по сравнению с методами 7а, 8а (при

условии возможности σ-упорядочивания). Эти методы просто конвейеризуются; 2) как и при $p = 1$, методы 6б, 7б, 8б имеют одинаковые по порядку число тактов T_p. Однако при $p \asymp N^{1/2}$ (а также при $p \asymp N$) метод переменных направлений 1а уступает по числу тактов методам 6б, 7б, 8б; 3) методы 4а, 4б, 5а, 5б, 7а, 8а, 9а, 9б, 11б при $p = 1$ имеют одинаковое (по порядку) число тактов T_1. Однако параллельные алгоритмы этих методов при $p \asymp N^{1/2}$ (а также при $p \asymp N$) имею разные T_p; 4) в методах 4б, 7а, 8а, 9а T_p в $\ln N$ раз меньше, чем в методах 4а, 5а, 5б, 9б, 11б. Это связано с тем, что во всех рассматриваемых методах присутствует алгоритм решения вспомогательных систем с треугольными матрицами, в блочных методах — алгоритм решения систем уравнений с трехдиагональными матрицами.

Если, например, рассматривается $p \asymp N$, то: 1) методы 6б, 7б, 8б, 11а, 12 уже имеют почти предельное распараллеливание ($T_p \asymp \lg^\gamma N$, где γ — константа); 2) идеальную характеристику для T_p имеет ДБПФ; 3) как правило, блочные методы (с числом неизвестных в блоке $N^{1/2}$) уступают соответствующим точечным методам за счет «лишнего» $\lg N$, возникающего при распараллеливании решения вспомогательных систем; 4) наличие в реальных системах ЭВМ спецпроцессора для осуществления ДБПФ для отыскания решения дискретной задачи для оператора Лапласа в прямоугольнике часто дает возможность существенно ускорить процесс решения задачи.

На основе приведенных результатов можно высказать и ряд других предположений о возможности распараллеливания того или иного итерационного алгоритма. Несмотря на асимптотичность проведенного анализа, его выводы могут подсказать исследователю правильный выбор эффективного алгоритма и в реальной ситуации.

Глава 5.

Методы решения нестационарных задач

В качестве основного объекта исследования рассматривается эволюционная задача математической физики

$$\frac{\partial \varphi}{\partial t} + A\varphi = f \text{ в } D \times D_t, \quad \varphi = g \text{ в } D \text{ при } t = 0, \tag{5.0.1}$$

где $A \geqslant 0$, а решение φ , как и функции f и g, обладает необходимой гладкостью. Будем предполагать, что на границе области ∂D решения задачи удовлетворяют некоторым граничным условиям[1)].

5.1. Разностные схемы второго порядка аппроксимации с операторами, зависящими от времени

Рассмотрим эволюционное уравнение

$$\frac{\partial \varphi}{\partial t} + A\varphi = 0 \text{ в } D \times D_t, \quad \varphi = g \text{ в } D \text{ при } t = 0, \tag{5.1.1}$$

Разностное уравнение, соответствующее уравнению (5.0.1), запишем в виде

$$\frac{\varphi^{j+1} - \varphi^j}{\tau} + A\frac{\varphi^{j+1} + \varphi^j}{2} = 0, \varphi^0 = g. \tag{5.1.2}$$

Нетрудно проверить, что при достаточной гладкости решения задача (5.1.1) аппроксимируется приближенной задачей (5.1.2) со вторым поряд-

[1)] Всюду в этой главе, если не оговорено специально, оператор A считается уже приведенным к конечно-разностному виду.

ком по τ. Разностная схема (5.1.2) обычно называется *схемой Кранка — Николсона*. Любопытно отметить, что схема (5.1.2) является результатом попеременного применения схем первого порядка точности, явной и неявной, записанных для интервалов $t_j \leqslant t \leqslant t_{j+\frac{1}{2}}$ и $t_{j+\frac{1}{2}} \leqslant t \leqslant t_{j+1}$ соответственно (если A — линейный оператор, не зависящий от t):

$$\begin{aligned} \frac{\varphi^{j+1/2}-\varphi^j}{\tau/2} + A\varphi^j = 0, \\ \frac{\varphi^{j+1}-\varphi^{j+1/2}}{\tau/2} + A\varphi^{j+1} = 0. \end{aligned} \tag{5.1.3}$$

Исключая из системы разностных уравнений неизвестный $\varphi^{j+1/2}$, приходим к схеме Кранка — Николсона.

Предположим, что оператор

$$\frac{\varphi^{j+1}-\varphi^j}{\tau} + \Lambda^j \frac{\varphi^{j+1}+\varphi^j}{2} = 0, \quad \varphi^0 = g, \tag{5.1.4}$$

$$(\Lambda^j \varphi,\ \varphi) \geqslant 0 \tag{5.1.5}$$

для любых функций из подпространства Ф. Уравнение (5.1.4) разрешим относительно φ^{j+1}. Получим

$$\varphi^{j+1} = \left(E + \frac{\tau}{2}\Lambda^j\right)^{-1} \left(E - \frac{\tau}{2}\Lambda^j\right) \varphi^j, \tag{5.1.6}$$

или

$$\varphi^{j+1} = T^j \varphi^j, \tag{5.1.7}$$

где T^j — оператор шага:

$$T^j = \left(E + \frac{\tau}{2}\Lambda^j\right)^{-1} \left(E - \frac{\tau}{2}\Lambda^j\right). \tag{5.1.8}$$

Для доказательства счетной устойчивости норму оператора шага T^j можно не оценивать. Умножив (5.1.4) скалярно на $(\varphi^{j+1}+\varphi^j)/2$, получим

$$\frac{(\varphi^{j+1},\ \varphi^{j+1}) - (\varphi^j,\ \varphi^j)}{2\tau} + \left(\Lambda^j \frac{\varphi^{j+1}+\varphi^j}{2},\ \frac{\varphi^{j+1}+\varphi^j}{2}\right) = 0. \tag{5.1.9}$$

Поскольку по предположению оператор Λ^j положительно полуопределенный (см. (5.1.5)), то

$$\|\varphi^{j+1}\| \leqslant \|\varphi^j\|, \tag{5.1.10}$$

т. е. устойчивость схемы обеспечена.

Тем не менее для анализа разностных схем важное значение имеет оценка нормы оператора шага. Чтобы получить эту оценку, воспользуемся

соотношением (5.1.10) и определением нормы оператора (см. 5.0.1). Тогда приходим к неравенству

$$\|T^j\| \leqslant 1. \tag{5.1.11}$$

Отметим, что неравенство (5.1.11) можно непосредственно получить с помощью леммы Келлога (см. 5.0.1), воспользовавшись (5.1.8) и тем, что $\Lambda^j \geqslant 0$ и $\tau > 0$. Если оператор Λ^j кососимметрический, т. е. имеет место равенство

$$(\Lambda^j \varphi,\ \varphi) = 0,$$

то вместо (5.1.10) имеем равенство

$$\|\varphi^{j+1}\| = \|\varphi^j\|. \tag{5.1.12}$$

Аналогично предыдущему можно показать, что в этом случае

$$\|T^j\| = 1. \tag{5.1.13}$$

Рассмотрим теперь вопрос об аппроксимации схемы Кранка — Никольсона при зависимости оператора A от времени. Определим оператор H равенством

$$H\varphi \equiv \frac{\partial \varphi}{\partial t} + A\varphi \tag{5.1.14}$$

и оператор H_τ — равенством

$$(H_\tau \varphi)^j = \frac{(\varphi)^{j+1} - (\varphi)^j}{\tau} + \Lambda^j \frac{(\varphi)^{j+1} + (\varphi)^j}{2}, \tag{5.1.15}$$

где $(\varphi)^j$ — проекция функции φ на сетку D_τ. Далее, введем в рассмотрение норму

$$\|(H_\tau \varphi)\|_{C_\tau} = \max_{t_j \in D_\tau} \|(H_\tau \varphi)^j\|, \tag{5.1.16}$$

где $\|\cdot\|$ — некоторая норма в пространстве, которому принадлежит $(H_\tau \varphi^j)$. Для оценки нормы (5.1.16) разложим решение исходного уравнения (5.0.1) в ряд Тейлора. Тогда будем иметь

$$(\varphi)^{j+1} = (\varphi)^j + \tau(\varphi_t)^j + \frac{\tau^2}{2}(\varphi_{tt})^j + \ldots. \tag{5.1.17}$$

Принимая во внимание очевидные соотношения

$$\varphi_t = -A\varphi, \quad \varphi_{tt} = A^2\varphi - A_t\varphi, \tag{5.1.18}$$

где $A_t = \partial A/\partial t$, преобразуем ряд Тейлора (5.1.17) к виду

$$(\varphi)^{j+1} = (\varphi)^j - \tau A^j(\varphi)^j + \frac{\tau^2}{2}[(A^j)^2(\varphi)^j - A_t^j(\varphi)^j] - \dots . \tag{5.1.19}$$

Подставим (5.1.19) в (5.1.16). Тогда с учетом (5.1.15) получим

$$\|(f - H_\tau\varphi)\|_{C_\tau} = \max_{t_j}\left\|\Lambda^j(\varphi)^j - A^j(\varphi)^j + \frac{\tau}{2}\{(A^j)^2 - A_t^j - \Lambda^j A^j\}(\varphi)^j + O(\tau^2)\right\|, \tag{5.1.20}$$

где f^j — правая часть уравнения (5.0.1), в данном случае равная нулю.

Если в качестве аппроксимирующего оператора Λ^j выбрать

$$\Lambda^j = A^j = A(t_j), \tag{5.1.21}$$

то из (5.1.20) следует, что

$$\|(f - H_\tau\varphi)\|_{C_\tau} = \frac{\tau}{2}\max_{t_j \in D_\tau}\|A_t^j(\varphi)^j\| + O(\tau^2),$$

и мы имеем первый порядок аппроксимации. Заметим, что в частном случае, когда A не зависит от t, аппроксимация в форме (5.1.21) обеспечивает второй порядок по τ. Предположим теперь, что аппроксимирующий оператор выбран в виде

$$\Lambda^j = A^j + \frac{\tau}{2}A_t^j. \tag{5.1.22}$$

В этом случае будем иметь

$$\|(f - H_\tau\varphi)\|_{C_\tau} = O(\tau^2).$$

Отметим, что аппроксимация схемой Кранка — Николсона также будет второго порядка по τ, если оператор Λ^j выбрать в виде

$$\Lambda^j = A^{j+1/2} \tag{5.1.23}$$

или

$$\Lambda^j = \frac{1}{2}(A^{j+1} + A^j). \tag{5.1.24}$$

В различных приложениях, особенно при численном решении квазилинейных уравнений, применяется одна из трех описанных форм (5.1.22), (5.1.23) или (5.1.24) аппроксимации оператора A, обеспечивающих второй порядок точности.

5.2. Неоднородные уравнения эволюционного типа

В предыдущем параграфе были рассмотрены однородные уравнения. Рассмотрим теперь неоднородные уравнения

$$\begin{aligned} \frac{\partial \varphi}{\partial t} + A\varphi = f \text{ и } D \times D_t, \\ \varphi = g \text{ при } t = 0. \end{aligned} \tag{5.2.1}$$

Разностная аппроксимация задачи (5.2.1) на основе схемы Кранка — Николсона в предположениях, сформулированных в 4.1, имеет вид

$$\begin{aligned} \frac{\varphi^{j+1} - \varphi^j}{\tau} + \Lambda^j \frac{\varphi^{j+1} + \varphi^j}{2} = f', \\ \varphi^0 = g, \end{aligned} \tag{5.2.2}$$

где

$$f^j = f(t_{j+1/2}).$$

Нетрудно убедиться, что разностная задача (5.2.2) аппроксимирует (5.2.1) со вторым порядком по τ. Решение задачи (5.2.2) запишем в виде

$$\varphi^{j+1} = T^j \varphi^j + \tau \left(E + \frac{\tau}{2}\Lambda^j\right)^{-1} f^j. \tag{5.2.3}$$

В 5.1 в случае однородного уравнения было показано, что при $\Lambda^j \geqslant 0$ имеет место оценка

$$\|T^j\| \leqslant 1. \tag{5.2.4}$$

Из уравнения (5.2.3) следует, что

$$\|\varphi^{j+1}\| \leqslant \|T^j\| \|\varphi^j\| + \tau \left\| \left(E + \frac{\tau}{2}\Lambda^j\right)^{-1} \right\| \|f^j\|. \tag{5.2.5}$$

Для установления устойчивости воспользуемся оценкой (1.1.25) из главы 1. Поскольку $\tau > 0$ и

$$(\Lambda^j \varphi^j,\ \varphi^j) \geqslant 0, \tag{5.2.6}$$

то

$$\left\| \left(E + \frac{\tau}{2}\Lambda^j\right)^{-1} \right\| \leqslant 1. \tag{5.2.7}$$

Следовательно, с учетом (5.2.4) и (5.2.7) неравенство (5.2.5) преобразуется к виду

$$\|\varphi^{j+1}\| \leqslant \|\varphi^j\| + \tau \|f^j\|. \tag{5.2.8}$$

Полагая

$$\|\varphi^0\| = \|g\|, \quad \|f\| = \max_j \|f^j\|,$$

с помощью рекуррентного соотношения (5.2.8) получаем

$$\|\varphi^j\| \leqslant \|g\| + j\tau\|f\|, \quad j\tau \leqslant \text{const}. \tag{5.2.9}$$

Таким образом, соотношение (5.2.9) показывает устойчивость разностной схемы. Кроме того, это соотношение является априорной оценкой нормы решения.

5.3. Методы расщепления нестационарных задач

Во многих случаях, когда требуется решить сложную задачу математической физики, оказывается возможным свести ее к последовательному решению задач более простых, эффективно решаемых с помощью ЭВМ. Редукция сложных задач к более простым обычно возможна в тех случаях, когда исходный положительно полуопределенный оператор задачи представим в виде суммы положительно полуопределенных простейших операторов. Такие методы будем называть *методами расщепления*.

Первоначально методы расщепления формулировались и теоретически обосновывались для простейших задач с коммутирующими положительно определенными операторами. Как теперь стало ясным, для таких задач методы расщепления, введенные в рассмотрение различными авторами, по существу оказались либо эквивалентными и отличающимися только схемами реализации, либо близкими.

В дальнейшем круг нетривиальных задач, решаемых с помощью методов расщепления, существенно расширился, и к настоящему времени методы расщепления стали мощным аппаратом решения весьма сложных задач математической физики. Поскольку теория методов расщепления особенно полно разработана в случае, когда исходный оператор задачи представим в виде суммы двух более простых, то именно с рассмотрения этого случая мы и начнем изложение вопроса. Наиболее универсальным для приложения является, по нашему мнению, метод покомпонентного расщепления. Это обстоятельство, надеемся, будет учтено читателем при изучении материала данной главы.

Итак, рассмотрим эволюционное уравнение

$$\frac{\partial \varphi}{\partial t} + A\varphi = f \text{ в } D \times D_t, \tag{5.3.1}$$

$$\varphi = g \text{ в } D \text{ при } t = 0,$$

где оператор $A \geqslant 0$ не зависит от времени и представим в виде

$$A = A_1 + A_2 \tag{5.3.2}$$

при условии, что

$$A_1 \geqslant 0, \quad A_2 \geqslant 0. \tag{5.3.3}$$

Предположим, что решение задачи (5.3.1) обладает необходимой гладкостью. Там, где это необходимо для доказательства, будем предполагать, что задача (5.3.1) уже редуцирована к разностному виду и, следовательно, операторами A, A_1 и A_2 являются матрицы.

5.3.1. Метод стабилизации

Рассмотрим разностную схему решения задачи (5.3.1)—(5.3.3) в предположении, что $f = 0$:

$$\left(E + \frac{\tau}{2}A_1\right)\left(E + \frac{\tau}{2}A_2\right)\frac{\varphi^{j+1} - \varphi^j}{\tau} + A\varphi^j = 0, \quad \varphi^0 = g. \tag{5.3.4}$$

Нетрудно показать, что при достаточной гладкости решения задача (5.3.4) аппроксимирует исходную задачу (5.3.1)—(5.3.3) с точностью до величин второго порядка малости по τ. В самом деле, уравнение (5.3.4) с помощью алгебраических преобразований приводится к виду

$$\left(E + \frac{\tau^2}{4}A_1A_2\right)\frac{\varphi^{j+1} - \varphi^j}{\tau} + A\frac{\varphi^{j+1} + \varphi^j}{2} = 0, \quad \varphi^0 = g. \tag{5.3.5}$$

Нетрудно заметить, что разностное уравнение (5.3.5) при достаточной гладкости решения совпадает по порядку аппроксимации со схемой Кранка — Николсона

$$\frac{\varphi^{j+1} - \varphi^j}{\tau} + A\frac{\varphi^{j+1} + \varphi^j}{2} = 0, \quad \varphi^0 = g, \tag{5.3.6}$$

т. е. имеет второй порядок аппроксимации по τ.

Переходим теперь к анализу устойчивости разностного уравнения (5.3.4). С этой целью преобразуем уравнение (5.3.4) к виду

$$\left(E + \frac{\tau}{2}A_1\right)\left(E + \frac{\tau}{2}A_2\right)\varphi^{j+1} = \left(E - \frac{\tau}{2}A_1\right)\left(E - \frac{\tau}{2}A_2\right)\varphi^j, \quad \varphi^0 = g. \tag{5.3.7}$$

Разрешим разностное уравнение (5.3.7) относительно φ^{j+1}. Получим

$$\varphi^{j+1} = \left(E + \frac{\tau}{2}A_2\right)^{-1}\left(E + \frac{\tau}{2}A_1\right)^{-1}\left(E - \frac{\tau}{2}A_1\right)\left(E - \frac{\tau}{2}A_2\right)\varphi^j. \tag{5.3.8}$$

От неизвестной φ^j перейдем к ψ^j по следующей формуле

$$\psi^j = \left(E + \frac{\tau}{2}A_2\right)\varphi^j. \tag{5.3.9}$$

Тогда для новой неизвестной ψ^j приходим к соотношению

$$\psi^{j+1} = T\psi^j, \tag{5.3.10}$$

где T — оператор шага:

$$T = \left(E + \frac{\tau}{2}A_1\right)^{-1}\left(E - \frac{\tau}{2}A_1\right)\left(E - \frac{\tau}{2}A_2\right)\left(E + \frac{\tau}{2}A_2\right)^{-1}. \tag{5.3.11}$$

С помощью (5.3.10) получим оценку

$$\|\psi^{j+1}\| \leqslant \|T\|\|\psi^j\|. \tag{5.3.12}$$

Оценим норму оператора T:

$$\|T\| \leqslant \|T_1\|\|T_2\|,$$

где

$$T_\alpha = \left(E - \frac{\tau}{2}A_\alpha\right)\left(E + \frac{\tau}{2}A_\alpha\right)^{-1}, \quad \alpha = 1,\ 2. \tag{5.3.13}$$

Здесь использовано свойство

$$\left(E + \frac{\tau}{2}A_\alpha\right)^{-1}\left(E - \frac{\tau}{2}A_\alpha\right) = \left(E - \frac{\tau}{2}A_\alpha\right)\left(E + \frac{\tau}{2}A_\alpha\right)^{-1},$$

которое следует из очевидного тождества

$$\left(E + \frac{\tau}{2}A_\alpha\right)\left(E + \frac{\tau}{2}A_\alpha\right)^{-1} = E. \tag{5.3.14}$$

В самом деле, умножив (5.3.14) слева на

$$\left(E + \frac{\tau}{2}A_\alpha\right)^{-1}\left(E - \frac{\tau}{2}A_\alpha\right)$$

и использовав проверяемую непосредственным умножением коммутативность операторов

$$\left(E - \frac{\tau}{2}A_\alpha\right) \text{ и } \left(E + \frac{\tau}{2}A_\alpha\right),$$

приходим к доказываемому свойству.

Таким образом, задача установления устойчивости свелась к оценке норм операторов T_α. Применяя лемму Келлога для оценки норм операторов T_1 и T_2 в соотношении (5.3.13), приходим к выводу, что

$$\|T\| \leqslant 1 \tag{5.3.15}$$

и, следовательно,

$$\|\psi^{j+1}\| \leqslant \|\psi^j\| \tag{5.3.16}$$

Однако нашей конечной целью является установление устойчивости исходной разностной задачи (5.3.4). Поэтому воспользовавшись соотношением (5.3.9), перепишем (5.3.16) в виде

$$\left\|\left(E+\frac{\tau}{2}A_2\right)\varphi^{j+1}\right\| \leqslant \left\|\left(E+\frac{\tau}{2}A_2\right)\varphi^{j}\right\| \tag{5.3.17}$$

и введем обозначение

$$\left\|\left(E+\frac{\tau}{2}A_2\right)\varphi\right\| = (C_2\varphi,\ \varphi)^{1/2} = \|\varphi\|_{C_2}, \tag{5.3.18}$$

где

$$C_\alpha = \left(E+\frac{\tau}{2}A_\alpha^*\right)\left(E+\frac{\tau}{2}A_\alpha\right), \quad \alpha = 1,\ 2.$$

Нетрудно видеть, что $C_\alpha > 0$ и $\|\cdot\|_{C_2}$ действительно является нормой.

Таким образом, в данной норме имеет место условие абсолютной устойчивости

$$\|\varphi^{j+1}\|_{C_2} \leqslant \|\varphi^j\|_{C_2}. \tag{5.3.19}$$

Итак, мы приходим к выводу, что если $A_1 \geqslant 0$, $A_2 \geqslant 0$ и элементы этих матриц не зависят от времени, то при достаточной гладкости решения задачи (5.3.1) разностная схема (5.3.4) абсолютно устойчива и аппроксимирует исходную задачу со вторым порядком по τ.

В заключение заметим, что разностная схема метода стабилизации допускает удобную реализацию на ЭВМ. В самом деле, разностное уравнение (5.3.4) можно записать в виде

$$\begin{gathered} F^j = A\varphi^j, \quad \left(E+\frac{\tau}{2}A_1\right)\xi^{j+1/2} = -F^j, \\ \left(E+\frac{\tau}{2}A_2\right)\xi^{j+1} = \xi^{j+1/2}, \quad \varphi^{j+1} = \varphi^j + \tau\xi^{j+1}. \end{gathered} \tag{5.3.20}$$

Здесь $\xi^{j+1/2}$ и ξ^{j+1} — некоторые вспомогательные величины, обеспечивающие редукцию задачи (5.3.4) к последовательности простейших задач (5.3.20). Заметим, что первое и последнее уравнения из (5.3.20) являются явными соотношениями. Это значит, что обращать операторы нужно лишь

во втором и третьем уравнениях, в которых присутствуют только простейшие операторы A_1 и A_2.

Рассмотрим теперь неоднородную задачу

$$\frac{\partial \varphi}{\partial t} + A\varphi = f, \quad \varphi = g \text{ при } t = 0, \tag{5.3.21}$$

где $A = A_1 + A_2,\ A_1 \geqslant 0,\ A_2 \geqslant 0$. В этом случае схема метода стабилизации запишется следующим образом:

$$\left(E + \frac{\tau}{2}A_1\right)\left(E + \frac{\tau}{2}A_2\right)\frac{\varphi^{j+1} - \varphi^j}{\tau} + A\varphi^j = f^j, \quad \varphi^0 = g, \tag{5.3.22}$$

где

$$f^j = f(t_{j+1/2}). \tag{5.3.23}$$

При условии (5.3.23) разностная задача (5.3.22) аппроксимирует исходную задачу (5.3.21) со вторым порядком по τ.

Исследуем устойчивость разностной схемы. С этой целью преобразуем уравнение (5.3.22) к виду

$$\psi^{j+1} = T\psi^j + \tau\left(E + \frac{\tau}{2}A_1\right)^{-1} f^j, \tag{5.3.24}$$

где

$$\psi^j = \left(E + \frac{\tau}{2}A_2\right)\varphi^j. \tag{5.3.25}$$

Из уравнения (5.3.24) следует, что

$$\|\psi^{j+1}\| \leqslant \|T\|\|\psi^j\| + \tau\left\|\left(E + \frac{\tau}{2}A_1\right)^{-1}\right\|\|f^j\|. \tag{5.3.26}$$

Поскольку для случая однородного уравнения уже было установлено, что

$$\|T\| \leqslant 1,$$

то имеем

$$\|\psi^{j+1}\| \leqslant \|\psi^j\| + \tau\left\|\left(E + \frac{\tau}{2}A_1\right)^{-1}\right\|\|f^j\|. \tag{5.3.27}$$

Легко видеть, что

$$\|f^j\| = \left\|\left(E + \frac{\tau}{2}A_2\right)^{-1}\left(E + \frac{\tau}{2}A_2\right)f^j\right\| \leqslant \left\|\left(E + \frac{\tau}{2}A_2\right)^{-1}\right\|\left\|\left(E + \frac{\tau}{2}A_2\right)f^j\right\|. \tag{5.3.28}$$

С учетом соотношений (5.3.18), (5.3.25) и (5.3.28) получим из (5.3.27) неравенство

$$\|\varphi^{j+1}\|_{C_2} \leqslant \|\varphi^j\|_{C_2} + \tau \left\| \left(E + \frac{\tau}{2} A_1\right)^{-1} \right\| \left\| \left(E + \frac{\tau}{2} A_2\right)^{-1} \right\| \|f^j\|_{C_2}. \tag{5.3.29}$$

Воспользуемся оценкой

$$\left\| \left(E + \frac{\tau}{2} A_\alpha\right)^{-1} \right\| \leqslant 1, \tag{5.3.30}$$

установленной при $A_\alpha \geqslant 0$ в главе 1 (см. (1.1.25)). В результате получим, что

$$\|\varphi^{j+1}\|_{C_2} \leqslant \|\varphi^j\|_{C_2} + \tau \|f^j\|_{C_2}. \tag{5.3.31}$$

Отсюда с помощью рекуррентных соотношений приходим к оценке

$$\|\varphi^j\|_{C_2} \leqslant \|g\|_{C_2} + j\tau \|f\|_{C_2}, \tag{5.3.32}$$

где

$$\|f\|_{C_2} = \max_j \|f^j\|_{C_2}. \tag{5.3.33}$$

Таким образом, если $A_1 \geqslant 0,\ A_2 \geqslant 0$ и элементы матриц A_1, A_2 не зависят от времени, то при достаточной гладкости решения φ и функции f задачи (5.3.1) разностная схема (5.3.22) абсолютно устойчива и аппроксимирует исходную задачу со вторым порядком точности по τ.

В заключение еще раз подчеркнем, что приведенное выше доказательство справедливо только в том случае, когда исходный оператор A не зависит от времени.

5.3.2. Метод предиктор-корректор

Рассмотрим метод расщепления, называемый *методом предиктор-корректор*. Смысл этого метода состоит в следующем. Весь интервал $0 \leqslant t \leqslant T$ разбивается на частичные промежутки, и в пределах каждого из элементарных промежутков $t_j \leqslant t \leqslant t_{j+1}$ задача (5.3.1) решается в два приема. Сначала по схеме первого порядка точности и с довольно значительным «запасом» устойчивости находится приближенное решение задачи в момент времени $t_{j+1/2} = t_j + \tau/2$, этот этап обычно называется предиктором. После этого на всем интервале $(t_j,\ t_{j+1})$ расписывается исходное уравнение со вторым порядком аппроксимации, которое служит корректором. Существенно именно то, что при конструкции корректора используется «грубое» решение при $t_{j+1/2}$, найденное с помощью предиктора.

Схема предиктор-корректор может быть записана, например, в форме

$$\begin{aligned} \frac{\varphi^{j+1/4}-\varphi^j}{\tau/2}+A_1\varphi^{j+1/4}&=0,\\ \frac{\varphi^{j+1/2}-\varphi^{j+1/4}}{\tau/2}+A_2\varphi^{j+1/2}&=0,\\ \frac{\varphi^{j+1}-\varphi^j}{\tau}+A\varphi^{j+1/2}&=0,\\ \varphi^0&=g. \end{aligned} \tag{5.3.34}$$

Изучим схему предиктор-корректор более внимательно. Прежде всего из первых двух уравнений (5.3.34) исключим вспомогательную функцию $\varphi^{j+1/4}$. Тогда систему (5.3.34) можно привести к следующей:

$$\begin{aligned} \left(E+\frac{\tau}{2}A_1\right)\left(E+\frac{\tau}{2}A_2\right)\varphi^{j+1/2}&=\varphi^j,\\ \frac{\varphi^{j+1}-\varphi^j}{\tau}+A\varphi^{j+1/2}&=0. \end{aligned} \tag{5.3.35}$$

Исключив из этих уравнений $\varphi^{j+1/2}$, получим

$$\frac{\varphi^{j+1}-\varphi^j}{\tau}+A\left(E+\frac{\tau}{2}A_2\right)^{-1}\left(E+\frac{\tau}{2}A_1\right)^{-1}\varphi^j=0,\quad \varphi^0=g. \tag{5.3.36}$$

Исследуем вопрос об аппроксимации. С этой целью перепишем уравнение (5.3.36) в виде

$$\left(E+\frac{\tau}{2}A_1\right)\left(E+\frac{\tau}{2}A_2\right)\frac{\varphi^{j+1}-\varphi^j}{\tau}+\Lambda\varphi^j=0,$$

где

$$\Lambda=\left(E+\frac{\tau}{2}A_1\right)\left(E+\frac{\tau}{2}A_2\right)A\left(E+\frac{\tau}{2}A_2\right)^{-1}\left(E+\frac{\tau}{2}A_1\right)^{-1}.$$

Раскладывая правую часть последнего соотношения в ряд по степеням τ в предположении, что

$$\frac{\tau}{2}\|A_\alpha\|\leqslant 1,$$

получим, что

$$\Lambda=A+O(\tau^2).$$

С помощью оценки, использованной для метода стабилизации, приходим к выводу, что метод предиктор-корректор имеет второй порядок аппроксимации по τ.

Исследуем устойчивость этого метода. С этой целью уравнение (5.3.36) запишем в виде

$$\left(E+\frac{\tau}{2}A_1\right)\left(E+\frac{\tau}{2}A_2\right)\frac{\Phi^{j+1}-\Phi^j}{\tau}+A\Phi^j=0, \tag{5.3.37}$$

где

$$\Phi^j = \left(E + \frac{\tau}{2}A_2\right)^{-1}\left(E + \frac{\tau}{2}A_1\right)^{-1}\varphi^j. \tag{5.3.38}$$

Разностное уравнение (5.3.37) устойчиво, так как

$$\|\Phi^{j+1}\|_{C_2} \leqslant \|\Phi^j\|_{C_2}. \tag{5.3.39}$$

Отсюда с учетом соотношений (5.3.38) и (5.3.18) получим

$$\left\|\left(E + \frac{\tau}{2}A_1\right)^{-1}\varphi^{j+1}\right\| \leqslant \left\|\left(E + \frac{\tau}{2}A_1\right)^{-1}\varphi^j\right\|, \tag{5.3.40}$$

или

$$\|\varphi^{j+1}\|_{C_1^{-1}} \leqslant \|\varphi^j\|_{C_1^{-1}}, \tag{5.3.41}$$

где

$$C_1^{-1} = \left(E + \frac{\tau}{2}A_1^*\right)^{-1}\left(E + \frac{\tau}{2}A_1\right)^{-1}.$$

Таким образом, устойчивость в метрике (5.3.41) доказана. Следовательно, если матрицы $A_1 \geqslant 0$, $A_2 \geqslant 0$ и элементы этих матриц не зависят от времени, то при достаточной гладкости решения φ задачи (5.3.1) разностная схема (5.3.34) является абсолютно устойчивой и аппроксимирует исходную задачу со втором порядком точности по τ.

В случае неоднородной задачи метод предиктор-корректор сформулируем следующим образом:

$$\begin{aligned} \frac{\varphi^{j+1/4} - \varphi^j}{\tau/2} + A_1\varphi^{j+1/4} &= f^j, \\ \frac{\varphi^{j+1/2} - \varphi^{j+1/4}}{\tau/2} + A_2\varphi^{j+1/2} &= 0, \\ \frac{\varphi^{j+1} - \varphi^j}{\tau} + A\varphi^{j+1/2} &= f^j, \end{aligned} \tag{5.3.42}$$

где

$$f^j = f(t_{j+1/2}).$$

При выборе f^j в указанной форме можно показать, что (5.3.42) аппроксимирует исходную задачу со вторым порядком аппроксимации по τ. Устойчивость схемы (5.3.42) устанавливается следующим образом. Исключая $\varphi^{j+1/2}$ и $\varphi^{j+1/4}$, получим

$$\frac{\varphi^{j+1} - \varphi^j}{\tau} + A\left(E + \frac{\tau}{2}A_2\right)^{-1}\left(E + \frac{\tau}{2}A_1\right)^{-1}\left(\varphi^j + \frac{\tau}{2}f^j\right) = f^j. \tag{5.3.43}$$

Введем обозначение

$$\psi^j = \left(E + \frac{\tau}{2}A_1\right)^{-1}\left(\varphi^j + \frac{\tau}{2}f^j\right).$$

Тогда соотношение (5.3.43) приводится к виду

$$\frac{\psi^{j+1} - \psi^j}{\tau} + \left(E + \frac{\tau}{2}A_1\right)^{-1} A\left(E + \frac{\tau}{2}A_2\right)^{-1}\psi^j = \left(E + \frac{\tau}{2}A_1\right)^{-1}\frac{f^j + f^{j+1}}{2}.$$

Отсюда

$$\psi^{j+1} = \left[E - \left(E + \frac{\tau}{2}A_1\right)^{-1}\tau A\left(E + \frac{\tau}{2}A_2\right)^{-1}\right]\psi^j + \tau\left(E + \frac{\tau}{2}A_1\right)^{-1}\left(\frac{f^j + f^{j+1}}{2}\right). \tag{5.3.44}$$

Поскольку имеет место соотношение

$$\begin{aligned}
&E - \left(E + \frac{\tau}{2}A_1\right)^{-1}\tau A\left(E + \frac{\tau}{2}A_2\right)^{-1} = \\
&\quad = \left(E + \frac{\tau}{2}A_1\right)^{-1}\left[\left(E + \frac{\tau}{2}A_1\right)\left(E + \frac{\tau}{2}A_2\right) - \tau A\right]\left(E + \frac{\tau}{2}A_2\right)^{-1} = \\
&\quad\quad = \left(E + \frac{\tau}{2}A_1\right)^{-1}\left[\left(E - \frac{\tau}{2}A_1\right)\left(E - \frac{\tau}{2}A_2\right)\right]\left(E + \frac{\tau}{2}A_2\right)^{-1} = \\
&\quad\quad\quad = \left[\left(E + \frac{\tau}{2}A_1\right)^{-1}\left(E - \frac{\tau}{2}A_1\right)\right]\left[\left(E - \frac{\tau}{2}A_2\right)\left(E - \frac{\tau}{2}A_2\right)^{-1}\right],
\end{aligned}$$

то, согласно лемме Келлога и (1.1.25) из главы 1, получаем оценку

$$\left\|\varphi^j + \frac{\tau}{2}f^j\right\|_{C_1^{-1}} \leqslant \left\|g + \frac{\tau}{2}f^0\right\|_{C_1^{-1}} + \tau j\|f\|_{C_1^{-1}}, \tag{5.3.45}$$

где

$$\|f\|_{C_1^{-1}} = \max_j \|f^j\|_{C_1^{-1}},$$

т. е. при $0 \leqslant t_j \leqslant T$ снова имеем устойчивость разностной схемы.

Таким образом, если $A_1 \geqslant 0$, $A_2 \geqslant 0$ и элементы матриц A_1, A_2 не зависят от времени, то при достаточной гладкости решения и правой части f задачи (5.3.1) разностная схема (5.3.42) абсолютно устойчива и позволяет получить решение второго порядка точности по τ.

В заключение обратим внимание на то, что хотя разностная схема (5.3.34) абсолютно устойчива, но входящиая в нее как часть разностная схема для корректора, рассмотренная отдельно, может быть абсолютно неустойчивой. Покажем это. Для упрощения выкладок рассмотрим случай, когда A — разностный аналог двумерного оператора Лапласа, D — единичный квадрат, на границе которого решение равно нулю.

В этом случае корректор имеет вид

$$\frac{\varphi^{j+1}_{k,\,l}-\varphi^{j-1}_{k,\,l}}{2\tau}-\frac{\varphi^{j}_{k+1,\,l}-2\varphi^{j}_{k,\,l}+\varphi^{j}_{k-1,\,l}}{h^2}-\frac{\varphi^{j}_{k,\,l+1}-2\varphi^{j}_{k,\,l}+\varphi^{j}_{k,\,l-1}}{h^2}=0$$

(так как теперь предиктор не участвует в вычислениях, то для удобства рассматриваются только целые значения j).

Для этой разностной задачи укажем решение φ, для которого не выполняется неравенство (5.3.16) из определения устойчивости. Такое решение будем искать в виде

$$\varphi^{j}_{k,\,l}=\lambda^j\sin m\pi kh\cdot\sin p\pi lh,$$

где j в левой части есть индекс, а в правой части — показатель степени. Подставляя это выражение в разностное уравнение, приходим к следующему характеристическому уравнению:

$$\lambda^2+8a_{mp}\lambda-1=0,$$

где

$$a_{mp}=\frac{\tau}{h^2}\left(\sin^2\frac{m\pi h}{2}+\sin^2\frac{p\pi h}{2}\right).$$

Взяв в качестве λ величину

$$\lambda=-4a_{mp}-\sqrt{1+16a^2_{mp}},$$

получаем, что при $\tau\to 0$ (так что $\tau/h^2=$const)

$$\|\varphi^j\|/\|\varphi^0\|=|\lambda|^j\to\infty,$$

т. е. схема абсолютно неустойчива.

Таким образом, несмотря на то, что разностная схема, используемая в качестве корректора, является абсолютно неустойчивой, но запас устойчивости, которым обладает предиктор, достаточен для абсолютной устойчивости в целом.

5.3.3. Метод покомпонентного расщепления

Метод стабилизации и метод предиктор-корректор эквивалентны по точности и абсолютно устойчивы при условии, что $A_\alpha\geqslant 0$. Следует, однако, иметь в виду одно ограничение на операторы A_α, которое мы сделали: они не зависят от времени. Это ограничение позволило нам довести до конца анализ устойчивости при конструктивном предположении только положи-

тельной полуопределенности операторов A_α. К сожалению, в случае зависимости этих операторов от времени анализ устойчивости в предполагаемой форме осуществить, вообще говоря, не удается. Исключением является *метод покомпонентного расщепления*, к формулировке которого мы приступаем.

Пусть в (5.3.1)—(5.3.3) $f = 0,\ A_1(t) \geqslant 0$ и $A_2(t) \geqslant 0$. Рассмотрим аппроксимации этих матриц на интервале $t_j \leqslant t \leqslant t_{j+1}$ в форме

$$\Lambda_\alpha^j = A_\alpha(t_{j+1/2})$$

в предположении, что их элементы имеют достаточную гладкость. Построим систему разностных уравнений, состоящую из последовательности простейших схем Кранка — Николсона[2)]:

$$\begin{aligned} \frac{\varphi^{j+1/2} - \varphi^j}{\tau} + \Lambda_1^j \frac{\varphi^{j+1/2} + \varphi^j}{2} = 0, \\ \frac{\varphi^{j+1} - \varphi^{j+1/2}}{\tau} + \Lambda_2^j \frac{\varphi^{j+1} + \varphi^{j+1/2}}{2} = 0. \end{aligned} \tag{5.3.46}$$

Система разностных уравнений (5.3.46) при исключении вспомогательных функций $\varphi^{j+1/2}$ может быть приведена к одному уравнению

$$\varphi^{j+1} = T^j \varphi^j, \tag{5.3.47}$$

где

$$T^j = \left(E + \frac{\tau}{2}\Lambda_2^j\right)^{-1} \left(E - \frac{\tau}{2}\Lambda_2^j\right) \left(E + \frac{\tau}{2}\Lambda_1^j\right)^{-1} \left(E - \frac{\tau}{2}\Lambda_1^j\right). \tag{5.3.48}$$

Изучим сначала проблемы аппроксимации. Для этой цели разложим оператор T^j по степеням τ, предполагая, что

$$\frac{\tau}{2}\|\Lambda_\alpha^j\| < 1.$$

В результате получим, что

$$T^j = E - \tau\Lambda^j + \frac{\tau^2}{2}((\Lambda_1^j)^2 + 2\Lambda_2^j\Lambda_1^j + (\Lambda_2^j)^2) - \ldots. \tag{5.3.49}$$

Если операторы Λ_α^j коммутируют, т. е. $\Lambda_1^j\Lambda_2^j = \Lambda_2^j\Lambda_1^j$, то разложение (5.3.49) можно записать в виде

$$T^j = E - \tau\Lambda^j + \frac{\tau^2}{2}(\Lambda^j)^2 - \ldots. \tag{5.3.50}$$

[2)] Теоретическое обоснование и модификации схемы даны автором книги в докладе на симпозиуме по численным методам решения уравнений с частными производными, состоявшемся в США в 1970 г. (SYNSPADE-1970).

Таким образом, если $A_1(t) \geqslant 0$, $A_2(t) \geqslant 0$, то при достаточной гладкости элементов этих матриц и решения φ задачи (5.3.1)—(5.3.3) разностная схема (5.3.46) абсолютно устойчива (это сразу следует из справедливого, согласно лемме Келлога, неравенства $\|T^j\| < 1$) и аппроксимирует исходное уравнение (5.3.1) со вторым порядком по τ в случае, если Λ_1^j и Λ_2^j коммутируют, и с первым порядком, если не коммутируют.

Теперь операторы $A_1(t)$ и $A_2(t)$ будем аппроксимировать не на интервале $t_j \leqslant t \leqslant t_{j+1}$, как в (5.3.46), а на интервале $t_{j-1} \leqslant t \leqslant t_{j+1}$. Положим

$$\Lambda_\alpha^j = A_\alpha(t_j).$$

Рассмотрим следующие две системы разностных уравнений:

$$\begin{aligned} \frac{\varphi^{j-1/2} - \varphi^{j-1}}{\tau} + \Lambda_1^j \frac{\varphi^{j-1/2} + \varphi^{j-1}}{2} = 0, \\ \frac{\varphi^{j} - \varphi^{j-1/2}}{\tau} + \Lambda_2^j \frac{\varphi^{j} + \varphi^{j-1/2}}{2} = 0 \end{aligned} \tag{5.3.51}$$

и

$$\begin{aligned} \frac{\varphi^{j+1/2} - \varphi^{j}}{\tau} + \Lambda_2^j \frac{\varphi^{j+1/2} + \varphi^{j}}{2} = 0, \\ \frac{\varphi^{j+1} - \varphi^{j+1/2}}{\tau} + \Lambda_1^j \frac{\varphi^{j+1} + \varphi^{j+1/2}}{2} = 0. \end{aligned} \tag{5.3.52}$$

Цикл вычислений состоит именно в поочередном применении разностных схем (5.3.51), (5.3.52). Аналогично предыдущему можно показать, что на полном цикле вычислений с помощью (5.3.51) и (5.3.52) имеем

$$\varphi^{j+1} = T^j \varphi^{j-1}, \tag{5.3.53}$$

где

$$\begin{aligned} T^j = \left(E + \frac{\tau}{2}\Lambda_1^j\right)^{-1} \left(E - \frac{\tau}{2}\Lambda_1^j\right) \left(E + \frac{\tau}{2}\Lambda_2^j\right)^{-1} \left(E - \frac{\tau}{2}\Lambda_2^j\right) \times \\ \times \left(E + \frac{\tau}{2}\Lambda_2^j\right)^{-1} \left(E - \frac{\tau}{2}\Lambda_2^j\right) \left(E + \frac{\tau}{2}\Lambda_1^j\right)^{-1} \left(E - \frac{\tau}{2}\Lambda_1^j\right) = \\ = E - 2\tau\Lambda^j + \frac{(2\tau)^2}{2}(\Lambda^j)^2 - \ldots. \end{aligned}$$

Если оператор шага T^j сравним с оператором шага схемы Кранка — Николсона

$$\frac{\varphi^{j+1} - \varphi^{j-1}}{2\tau} + \Lambda^j \frac{\varphi^{j+1} + \varphi^{j-1}}{2} = 0,$$

то можно установить, что с точностью до величины τ^2 операторы шага T^j для двуциклической схемы расщепления и схемы Кранка — Николсона, примененной к удвоенному интервалу по времени, совпадают независимо от того, являются операторы A_α коммутирующими или нет. Таким образом,

этот прием снимает весьма сильное требование коммутативности операторов.

Переходим к обсуждению вопроса о счетной устойчивости метода. Из соотношения (5.3.47) следует, что

$$\|\varphi^{j+1}\| \leqslant \|T^j\|\|\varphi^j\|.$$

Поскольку, как было показано выше,

$$\|T^j\| \leqslant 1$$

при $A_\alpha \geqslant 0$, то

$$\|\varphi^{j+1}\| \leqslant \|\varphi^j\|. \tag{5.3.54}$$

Отсюда непосредственно вытекает, что

$$\|\varphi^j\| \leqslant \|g\|. \tag{5.3.55}$$

Если рассматривается двуциклический метод, то на каждом шаге цикла имеет место оценка вида (5.3.54). Это означает, что и двуциклический метод абсолютно устойчив.

Таким образом, если $f = 0$, $A_1(t) \geqslant 0$ и $A_2(t) \geqslant 0$, то при достаточной гладкости решения φ задачи (5.3.1)—(5.3.3) и элементов матриц $A_1(t)$ и $A_2(t)$ система разностных уравнений (5.3.51), (5.3.52) абсолютно устойчива, а схема (5.3.53) аппроксимирует исходное уравнение (5.3.1) со вторым порядком по τ.

Будем искать решение неоднородной задачи с помощью двуциклического полного расщепления. С этой целью рассмотрим систему разностных уравнений вида (5.3.51), (5.3.52), записанных в более удобной форме:

$$\begin{aligned}
\left(E + \frac{\tau}{2}\Lambda_1^j\right)\varphi^{j-1/2} &= \left(E - \frac{\tau}{2}\Lambda_1^j\right)\varphi^{j-1},\\
\left(E + \frac{\tau}{2}\Lambda_2^j\right)(\varphi^j - \tau f^j) &= \left(E - \frac{\tau}{2}\Lambda_2^j\right)\varphi^{j-1/2},\\
\left(E + \frac{\tau}{2}\Lambda_2^j\right)\varphi^{j+1/2} &= \left(E - \frac{\tau}{2}\Lambda_2^j\right)(\varphi^j + \tau f^j),\\
\left(E + \frac{\tau}{2}\Lambda_1^j\right)\varphi^{j+1} &= \left(E - \frac{\tau}{2}\Lambda_1^j\right)\varphi^{j+1/2},
\end{aligned} \tag{5.3.56}$$

где $f^j = f(t_j)$. Разрешая эти уравнения относительно φ^{j+1}, получим

$$\varphi^{j+1} = T^j\varphi^{j-1} + 2\tau T_1^j T_2^j f^j, \tag{5.3.57}$$

где

$$T^j = T_1^j T_2^j T_2^j T_1^j, \tag{5.3.58}$$

$$T_\alpha^j = \left(E + \frac{\tau}{2}\Lambda_\alpha^j\right)^{-1}\left(E - \frac{\tau}{2}\Lambda_\alpha^j\right). \tag{5.3.59}$$

С помощью разложения по степеням малого параметра τ придем к соотношению

$$\varphi^{j+1} = \left[E - 2\tau\Lambda^j + \frac{(2\tau)^2}{2}(\Lambda^j)^2\right]\varphi^{j-1} + 2\tau(E - \tau\Lambda^j)f^j + O(\tau^3), \tag{5.3.60}$$

которое в свою очередь преобразуем к виду

$$\frac{\varphi^{j+1} - \varphi^{j-1}}{2\tau} + \Lambda^j(E - \tau\Lambda^j)\varphi^{j-1} = (E - \tau\Lambda^j)f^j + O(\tau^2). \tag{5.3.61}$$

Исключим φ^{j-1}, используя разложение решения в ряд Тейлора в окрестности точки t_{j-1}. С точностью до τ^2 будем иметь

$$\varphi^j = \varphi^{j-1} + \left(\frac{\partial\varphi}{\partial t}\right)^{j-1}\tau + O(\tau^2). \tag{5.3.62}$$

Производную $\partial\varphi/\partial t$ исключим с помощью соотношения

$$\left(\frac{\partial\varphi}{\partial t}\right)^{j-1} = -\Lambda^j\varphi^{j-1} + f^j + O(\tau). \tag{5.3.63}$$

Подставим (5.3.63) в (5.3.62). Тогда

$$\varphi^j = (E - \tau\Lambda^j)\varphi^{j-1} + \tau f^j + O(\tau^2).$$

Отсюда

$$(E - \tau\Lambda^j)\varphi^{j-1} = \varphi^j - \tau f^j + O(\tau^2). \tag{5.3.64}$$

Подставим соотношение (5.3.64) в (5.3.61). В результате будем иметь

$$\frac{\varphi^{j+1} - \varphi^{j-1}}{2\tau} + \Lambda^j\varphi^j = f^j + O(\tau^2). \tag{5.3.65}$$

Очевидно, что уравнение (5.3.65) аппроксимирует исходное уравнение (5.3.1) на интервале $t_{j-1} \leqslant t \leqslant t_{j+1}$ со вторым порядком по τ. Таким образом, нами найдена разностная аппроксимация неоднородного эволюционного уравнения второго порядка с помощью двуциклического метода.

Устойчивость метода доказывается в энергетической норме элементарно. В самом деле, оценим (5.3.57) по норме

$$\|\varphi^{j+1}\| \leqslant \|T^j\|\|\varphi^{j-1}\| + 2\tau\|T_1^j\|\|T_2^j\|\|f^j\|. \tag{5.3.66}$$

Выше было установлено, что $\|T_\alpha^j\| \leqslant 1$, следовательно,

$$\|T^j\| \leqslant \|T_1^j\|\|T_2^j\|\|T_2^j\|\|T_1^j\| \leqslant 1.$$

Поэтому

$$\|\varphi^{j+1}\| \leqslant \|\varphi^{j-1}\| + 2\tau\|f^j\|. \tag{5.3.67}$$

С помощью рекуррентного соотношения (5.3.63) получим, что

$$\|\varphi^j\| \leqslant \|g\| + \tau j\|f\|, \tag{5.3.68}$$

где

$$\|f\| = \max_j \|f^j\|.$$

Из соотношения (5.3.68) следует счетная устойчивость схемы на любом конечном временном интервале.

Систему уравнений (5.3.56) можно записать также в следующей эквивалентной форме:

$$\begin{aligned}
\left(E + \frac{\tau}{2}\Lambda_1^j\right)\varphi^{j-2/3} &= \left(E - \frac{\tau}{2}\Lambda_1^j\right)\varphi^{j-1},\\
\left(E + \frac{\tau}{2}\Lambda_2^j\right)\varphi^{j-1/3} &= \left(E - \frac{\tau}{2}\Lambda_2^j\right)\varphi^{j-2/3},\\
\varphi^{j+1/3} &= \varphi^{j-1/3} + 2\tau f^j,\\
\left(E + \frac{\tau}{2}\Lambda_2^j\right)\varphi^{j+2/3} &= \left(E - \frac{\tau}{2}\Lambda_2^j\right)\varphi^{j+1/3},\\
\left(E + \frac{\tau}{2}\Lambda_1^j\right)\varphi^{j+1} &= \left(E - \frac{\tau}{2}\Lambda_1^j\right)\varphi^{j+2/3}.
\end{aligned} \tag{5.3.69}$$

Исключая неизвестные величины с дробными индексами, приходим к разрешенному уравнению вида

$$\varphi^{j+1} = T_1^j T_2^j T_2^j T_1^j \varphi^{j-1} + 2\tau T_1^j T_2^j f^j, \tag{5.3.70}$$

которое совпадает с (5.3.57). В некоторых случаях запись уравнений в форме (5.3.69) предпочтительнее, чем в форме (5.3.56).

Итак, если $A_1(t) \geqslant 0$, $A_2(t) \geqslant 0$, то при достаточной гладкости решения φ, функции $f(t)$ и элементов матриц $A_1(t)$, $A_2(t)$ система разностных уравнений (5.3.56) абсолютно устойчива на интервале $0 \leqslant t \leqslant T$ и аппроксимирует исходное уравнение со вторым порядком по τ.

Рассмотрим двуциклический метод расщепления в применении к двумерной задаче теплопроводности:

$$\begin{gathered}\frac{1}{c^2}\frac{\partial\varphi}{\partial t}-\Delta\varphi=f \text{ в } D\times D_t,\\ \varphi=g \text{ или } \frac{\partial\varphi}{\partial n}=g \text{ на } \partial D\times D_t,\end{gathered} \tag{5.3.71}$$

$$\varphi=s(x,\ y) \text{ в } D \text{ при } t=0, \tag{5.3.72}$$

где D — квадрат $\{0\leqslant x\leqslant 1,\ 0\leqslant y\leqslant 1\}$.

Прежде всего редуцируем эту задачу к конечно-разностной по переменным $(x,\ y)$. В результате преобразований, аналогичных проведенным при рассмотрении двумерной задачи Дирихле, приходим к следующей задаче:

$$\begin{gathered}\frac{1}{c^2}\frac{d\varphi}{dt}-(\Lambda_1+\Lambda_2)\varphi=F,\\ \varphi=s \text{ при } t=0,\end{gathered} \tag{5.3.73}$$

где матрицы $\Lambda_1,\ \Lambda_2$ и вектор-функции φ и F определяются так же, как и в задаче (1.5.24) из 1.5.3. Разница состоит только в том, что компоненты векторов φ и F теперь будут зависеть от переменной t. Заметим, что в случае граничных условий Дирихле для уравнения теплопроводности имеем

$$\Lambda_1>0,\ \Lambda_2>0, \tag{5.3.74}$$

а в случае граничных условий Неймана имеем

$$\Lambda_1\geqslant 0,\ \Lambda_2\geqslant 0. \tag{5.3.75}$$

Разбивая отрезок $0\leqslant t\leqslant T$ точками t_j на интервалы и полагая $\tau=c^2\Delta t$, рассмотрим аппроксимацию задачи (5.3.73) на основе метода покомпонентного расщепления. Тогда на отрезке $t_{j-1}\leqslant t\leqslant t_{j+1}$ будем иметь

$$\begin{aligned}\left(E+\frac{\tau}{2}\Lambda_1\right)\varphi^{j-1/2}&=\left(E-\frac{\tau}{2}\Lambda_1\right)\varphi^{j-1},\\ \left(E+\frac{\tau}{2}\Lambda_2\right)\varphi^{j}&=\left(E-\frac{\tau}{2}\Lambda_2\right)\varphi^{j-1/2}+\tau f^j,\\ \varphi^{j+1/3}&=\varphi^{j-1/3}+2\tau f^j,\\ \left(E+\frac{\tau}{2}\Lambda_2\right)\varphi^{j+1/2}&=\left(E-\frac{\tau}{2}\Lambda_2\right)\varphi^{j}+\tau f^j,\\ \left(E+\frac{\tau}{2}\Lambda_1\right)\varphi^{j+1}&=\left(E-\frac{\tau}{2}\Lambda_1\right)\varphi^{j+1/2}.\end{aligned} \tag{5.3.76}$$

Заметим, что правые части в (5.3.76) в покомпонентном виде вычисляются по явным схемам. Таким образом, мы имеем дело с системой уравнений в покомпонентном виде, представляемых в форме трехточечных схем,

аналогичных (4.4.84) из 4.4.3, где роль σ теперь играет параметр $\tau/2$, а роль ξ_k^j — компоненты правых частей уравнений (5.3.76). Аналогично решается задача теплопроводности при граничных условиях Неймана.

5.3.4. Некоторые общие замечания. Попеременно-треугольный метод

Прежде всего сопоставим рассмотренные в настоящем параграфе методы расщепления в предположении, что A, A_1, A_2 не зависят от времени, $A_1 \geqslant 0$, $A_2 \geqslant 0$ и $A_1 A_2 = A_2 A_1$. Это именно тот простейший случай, который рассмотрен в литературе весьма полно. Нетрудно видеть, что все схемы расщепления для эволюционной задачи (5.3.1) при $f = 0$, формально разрешенные на каждом шаге относительно искомого решения, эквивалентны друг другу и, таким образом, являются лишь различными схемами реализации. Эти схемы, в частности, могут быть приведены к виду

$$\varphi^{j+1} = T\varphi^j,$$

где

$$T = \left(E + \frac{\tau}{2}A_1\right)^{-1}\left(E - \frac{\tau}{2}A_1\right)\left(E + \frac{\tau}{2}A_2\right)^{-1}\left(E - \frac{\tau}{2}A_2\right).$$

Схемы расщепления для неоднородной эволюционной задачи оказываются эквивалентными только по порядку аппроксимации. Это значит, что для неоднородных задач разные схемы расщепления, обладая вторым порядком аппроксимации по τ, будут приводить в пределах погрешности порядка τ^2 к различным результатам, отличающимся на величину $O(\tau^2)$.

Второе замечание касается использования схем расщепления в случае, когда $A_1 \geqslant 0$, $A_2 \geqslant 0$ не зависят от времени и $A_1 A_2 \neq A_2 A_1$. Как было показано, в этом случае для приближенного решения эволюционных задач могут быть использованы все три рассмотренные схемы расщепления: метод стабилизации, метод предиктор-корректор и метод покомпонентного расщепления. Все эти методы хотя и эквивалентны по порядку точности, однако существенно различаются, поскольку даже в случае однородной эволюционной задачи они оказываются не тождественными друг другу, т. е. имеют несовпадающие операторы шага T. Сейчас еще трудно дать рекомендации о сферах наиболее эффективного применения той или иной схемы, поскольку этот вопрос изучен недостаточно. Однако уже сам факт, что для решения одной и той же задачи можно использовать три различных (независимых) метода, позволяет с большей уверенностью различными путями подходить к решению сложных задач.

Третье замечание относится к самому общему случаю, когда $A_1 \geqslant 0$, $A_2 \geqslant 0$, $A_1 A_2 \neq A_2 A_1$ и операторы A_1 и A_2 зависят от времени. В этом случае лучше использовать метод покомпонентного расщепления, который в двуциклической форме приводит к решению задачи со вторым порядком аппроксимации по τ.

В заключение данного пункта рассмотрим *попеременно-треугольный метод* (см. Ильин В. П. [15], Самарский А. А. [15]).

Метод переменных направлений, предложенный Писманом, Дугласом, Рэчфордом, обычно связывают с разбиением оператора A на одномерные операторы A_α, при этом на каждом дробном шаге для решения уравнения используется метод прогонки. Однако распространение данного метода на задачи с тремя пространственными переменными встречает трудности. С другой стороны, можно отказаться от требования одномерности оператора A на такие, которые позволили бы экономично реализовать решение задачи на каждом шаге и сохранили бы основные преимущества метода переменных направлений. Таким расщеплением является разбиение матричного оператора $A = A^*$ на такие две матрицы A_1 и A_2, что $A_1 = A_2^*$, $A_1 + A_2 = A$. Если после этого формально записать схемы переменных направлений, то получим схемы попеременно-треугольного метода.

Сформулируем этот метод. Рассмотрим эволюционную задачу

$$\begin{aligned} \frac{d\varphi}{dt} + A\varphi = f \text{ в } D_t, \\ \varphi = g \text{ при } t = 0, \end{aligned} \tag{5.3.77}$$

где $D_t = \{t : 0 < t < T\}$, а $A = A(t)$ есть квадратная матрица, представляющая конечномерную аппроксимацию оператора соответствующей исходной задачи по всем переменным, исключая t. Оператор A из (5.3.77) действует в гильбертовом пространстве F со скалярным произведением $(u,\ v)$ и нормой $\|u\| = (u,\ u)^{1/2}$. Предполагается, что A представим в виде суммы двух треугольных положительно определенных матриц:

$$A = A_1 + A_2, \quad A_1 = (a_{ik}^{(1)}), \quad A_2 = (a_{ik}^{(2)}), \tag{5.3.78}$$

$$a_{ik}^{(1)} = 0, \quad k > i, \quad a_{ii}^{(1)} + a_{ii}^{(2)} = a_{ii}, \tag{5.3.79}$$

$$(A_\alpha u,\ u) \geqslant c\|u\|^2, \quad c = \text{const} > 0, \quad \alpha = 1,\ 2.$$

Если A — симметричная матрица, то $A_1^* = A_2$, $A_2^* = A_1$, и можно принять $a_{ii}^{(1)} = a_{ii}^{(2)} = a_{ii}/2$. Запишем схему попеременно-треугольного метода:

$$\frac{\varphi^{j+1/2} - \varphi^j}{\tau} + \frac{1}{2}(A_1^{j+1/2}\varphi^{j+1/2} + A_2^j \varphi^j) = \frac{1}{2} f^j,$$

$$\frac{\varphi^{j+1}-\varphi^j}{\tau}+\frac{1}{2}(A_1^{j+1/2}\varphi^{j+1/2}+A_2^{j+1}\varphi^{j+1})=\frac{1}{2}f^j, \qquad (5.3.80)$$
$$j=0,\ 1,\ \ldots; \quad \varphi^0=g,$$

где

$$A_1^{j+1/2}=A_1(t_{j+1/2}), \quad A_2^j=A_2(t_j), \quad f^j=f(t_{j+1/2}), \quad \tau=t_{j+1}-t_j.$$

Легко заметить, что для решения первого уравнения из (5.3.80) надо обратить треугольную матрицу $\left(E+\frac{\tau}{2}A_1^{j+1/2}\right)$, а для решения второго уравнения — обратить треугольную матрицу $\left(E+\frac{\tau}{2}A_2^{j+1/2}\right)$. (Отметим, что если A_α — самосопряженные, то схема (5.3.80) есть обобщение метода переменных направлений.)

Если выполнены условия (5.3.79) и $A_2(t)\varphi(t),\ d\varphi/dt\in C^{(1,1)}[0,T]$, то схема (5.3.80) имеет второй порядок точности по τ:

$$\|\varphi^j-\varphi(t_j)\|\leqslant c\tau^2, \quad j=1,\ 2,\ \ldots,$$

где c — положительная постоянная, независящая от τ.

Рассмотрим одну из конкретных реализаций попеременно-треугольного метода для двумерного уравнения теплопроводности

$$\frac{\partial\varphi}{\partial t}-\mathrm{div}(D(x,\ y)\mathrm{grad}\varphi)=0. \qquad (5.3.81)$$

Запишем следующую разностную пятиточечную схему:

$$\frac{\varphi_{kl}^{j+1}-\varphi_{kl}^j}{\Delta t_j}+\Lambda\varphi_{kl}^{j+1}=0. \qquad (5.3.82)$$

Здесь и далее $h_x=h_y=h$, $\Delta t_j=t_{j+1}-t_j$,

$$\Lambda\varphi_{kl}=\frac{1}{h^2}(p_{kl}\varphi_{kl}-a_{kl}\varphi_{k-1,l}-b_{kl}\varphi_{k,l-1}-c_{kl}\varphi_{k+1,l}-d_{kl}\varphi_{k,l+1}), \qquad (5.3.83)$$

а коэффициенты a, b, c, d, p могут быть переменными. Предполагается также, что граничные условия, при которых рассматривается уравнение (5.3.81), учитываются значениями коэффициентов разностного уравнения.

Схема попеременно-треугольного метода имеет вид

$$\frac{\varphi_{kl}^{j+1/2}-\varphi_{kl}^j}{\Delta t_j}+\frac{1}{2h^2}\left(\frac{p_{kl}}{2}\varphi_{kl}^{j+1/2}-a_{kl}\varphi_{k-1,\,l}^{j+1/2}-b_{kl}\varphi_{k,\,l-1}^{j+1/2}+\frac{p_{kl}}{2}\varphi_{kl}^j-c_{kl}\varphi_{k+1,\,l}^j-d_{kl}\varphi_{k,\,l+1}^j\right)=0,$$
$$\frac{\varphi_{kl}^{j+1}-\varphi_{kl}^{j+1/2}}{\Delta t_j}+\frac{1}{2h^2}\left(\frac{p_{kl}}{2}\varphi_{kl}^{j+1/2}-a_{kl}\varphi_{k-1,\,l}^{j+1/2}-b_{kl}\varphi_{k,\,l-1}^{j+1/2}+\frac{p_{kl}}{2}\varphi_{kl}^{j+1}-c_{kl}\varphi_{k+1,\,l}^{j+1}-d_{kl}\varphi_{k,\,l+1}^{j+1}\right)=0. \qquad (5.3.84)$$

Условие пространственной устойчивости по отношению к накоплению ошибок на каждом временном шаге имеет вид неравенств

$$\begin{aligned} 1+\frac{\Delta t_j}{4h^2}p_{kl} &\geqslant \frac{\Delta t_j}{2h^2}(|a_{kl}|+|b_{kl}|), \\ 1+\frac{\Delta t_j}{4h^2}p_{kl} &\geqslant \frac{\Delta t_j}{2h^2}(|c_{kl}|+|d_{kl}|), \end{aligned} \tag{5.3.85}$$

которые должны выполняться для всех, кроме, может быть, некоторых точек сетки. Если ввести операторы Λ_1, Λ_2 в виде

$$\begin{aligned} \Lambda_1\varphi_{kl} &= \frac{1}{h^2}\left(\frac{p_{kl}}{2}\varphi_{kl} - a_{kl}\varphi_{k-1,\,l} - b_{kl}\varphi_{k,\,l-1}\right), \\ \Lambda_2\varphi_{kl} &= \frac{1}{h^2}\left(\frac{p_{kl}}{2}\varphi_{kl} - c_{kl}\varphi_{k+1,\,l} - d_{kl}\varphi_{k,\,l+1}\right), \end{aligned} \tag{5.3.86}$$

то схему (5.3.84) можно записать так:

$$\begin{aligned} \frac{\varphi_{kl}^{j+1/2}-\varphi_{kl}^{j}}{\Delta t_j} + \frac{1}{2}(\Lambda_1\varphi_{kl}^{j+1/2} + \Lambda_2\varphi_{kl}^{j}) = 0, \\ \frac{\varphi_{kl}^{j+1}-\varphi_{kl}^{j+1/2}}{\Delta t_j} + \frac{1}{2}(\Lambda_1\varphi_{kl}^{j+1/2} + \Lambda_2\varphi_{kl}^{j+1}) = 0, \end{aligned} \tag{5.3.87}$$

или

$$\begin{aligned} \left(E+\frac{\Delta t_j}{2}\Lambda_1\right)\varphi^{j+1/2} &= \left(E-\frac{\Delta t_j}{2}\Lambda_2\right)\varphi^{j}, \\ \left(E+\frac{\Delta t_j}{2}\Lambda_2\right)\varphi^{j+1} &= \left(E-\frac{\Delta t_j}{2}\Lambda_1\right)\varphi^{j+1/2}. \end{aligned} \tag{5.3.88}$$

В целых шагах схема (5.3.87) имеет вид

$$\left(E+\frac{\Delta t_j}{2}\Lambda_1\right)\left(E+\frac{\Delta t_j}{2}\Lambda_2\right)\varphi^{j+1} = \left(E-\frac{\Delta t_j}{2}\Lambda_1\right)\left(E-\frac{\Delta t_j}{2}\Lambda_2\right)\varphi^{j}, \tag{5.3.89}$$

или

$$\frac{\varphi^{j+1}-\varphi^{j}}{\Delta t_j} + \frac{1}{2}\Lambda(\varphi^{j+1}+\varphi^{j}) + \frac{\Delta t_j^2}{4}\Lambda_1\Lambda_2\left(\frac{\varphi^{j+1}-\varphi^{j}}{\Delta t_j}\right) = 0. \tag{5.3.90}$$

Из (5.3.90) заключим, что схема (5.3.84) имеет второй порядок точности по τ (при условии наличия достаточной гладкости решения и исходных данных). Если Λ и $\Lambda_1\Lambda_2$ — положительно полуопределенные операторы, то все собственные числа оператора

$$T_j = \left(E+\frac{\Delta t_j}{2}\Lambda+\frac{\Delta t_j^2}{4}\Lambda_1\Lambda_2\right)^{-1}\left(E-\frac{\Delta t_j}{2}\Lambda+\frac{\Delta t_j^2}{4}\Lambda_1\Lambda_2\right)$$

— оператора перехода от φ^j к φ^{j+1} — меньше единицы. Отсюда следует устойчивость и сходимость схемы.

Отметим простоту численной реализации схемы (5.3.84), в которой вычисления на каждом полушаге представляют собой явную схему. Схема (5.3.84) обобщается и на многомерные задачи. При этом расщепление Λ на подходящие операторы Λ_1, Λ_2 (их только два) производится так, чтобы можно было осуществить счет на каждом полушаге по явной схеме. Преимущество схемы повышается с увеличением размерности задачи.

5.4. Многокомпонентное расщепление задач

До сих пор предполагалось, что исходный оператор A представлен в виде суммы двух операторов более простой структуры. При решении сложных задач математической физики зачастую приходится иметь дело с расщеплением операторов на большое число слагаемых. Рассмотрим случай, когда

$$A = \sum_{\alpha=1}^{n} A_\alpha, \tag{5.4.1}$$

причем $A_\alpha \geqslant 0$. Поскольку случай $n = 2$ подробно рассмотрен в 5.3, остановимся только на случае $n > 2$.

Прежде всего можно убедиться, что тривиальное распространение методов расщепления, рассмотренных выше для случая $n = 2$, в общем виде невозможно.

5.4.1. Метод стабилизации

В предположении (5.4.1) метод стабилизации может быть представлен в виде

$$\prod_{\alpha=1}^{n} \left(E + \frac{\tau}{2} A_\alpha\right) \frac{\varphi^{j+1} - \varphi^j}{\tau} + A\varphi^j = f^j, \quad \varphi^0 = g, \tag{5.4.2}$$

где

$$f^j = f(t_{j+1/2}).$$

Схема реализации алгоритма имеет следующий вид:

$$\begin{aligned} F^j &= -A\varphi^j + f^j, \\ \left(E + \frac{\tau}{2}A_1\right)\xi^{j+1/n} &= F^j, \\ \left(E + \frac{\tau}{2}A_2\right)\xi^{j+2/n} &= \xi^{j+1/n}, \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ \left(E + \frac{\tau}{2}A_n\right)\xi^{j+1} &= \xi^{j+\frac{n-1}{n}}, \\ \varphi^{j+1} &= \varphi^j + \tau\xi^{j+1}. \end{aligned} \tag{5.4.3}$$

Легко проверить, что метод стабилизации в случае достаточной гладкости решения имеет второй порядок по точности по τ. Счетная устойчивость будет обеспечена при выполнении условия

$$\|T\| \leqslant 1, \tag{5.4.4}$$

где T — оператор шага, определяемый формулой

$$T = E - \tau \prod_{\alpha=n}^{1} \left(E + \frac{\tau}{2}A_\alpha\right)^{-1} A. \tag{5.4.5}$$

К сожалению, из условия $A_\alpha \geqslant 0$ здесь не следует устойчивость в какой-нибудь норме, как это имело место в случае $n = 2$.

Для установления устойчивости иногда пользуются следующим простым алгоритмическим приемом. Полагая f^j равным нулю, приведем уравнение (5.4.2), разрешенное относительно φ^{j+1}, к следующему виду:

$$\varphi^{j+1} = T\varphi^j. \tag{5.4.6}$$

Поскольку оператор T предполагается не зависящим от времени (т. е. от индекса j), то, решая задачу (5.4.6) при начальном условии

$$\varphi^0 = g \tag{5.4.7}$$

и фиксированном параметре τ, обеспечивающем необходимую аппроксимацию, будем следить только за нормой $\|\varphi\|$. Если эта норма не будет расти, то, по-видимому, $\|T\| < 1$, и, таким образом, можно считать, что условие для счетной устойчивости выполнено. После этого можно переходить к решению неоднородной задачи. Перепишем в следующем виде уравнение (5.4.2):

$$\varphi^{j+1} = T\varphi^j + \tau \prod_{\alpha=n}^{1} \left(E + \frac{\tau}{2}A_\alpha\right)^{-1} f^j. \tag{5.4.8}$$

Отсюда

$$\|\varphi^{j+1}\| \leqslant \|T\|\|\varphi^j\| + \tau \prod_{\alpha=n}^{1} \left\| \left(E + \frac{\tau}{2} A_\alpha\right)^{-1} \right\| \|f^j\|,$$

или, в силу неравенства (5.4.4) и неравенства (1.1.25) из главы 1, получаем

$$\|\varphi^{j+1}\| \leqslant \|\varphi^j\| + \tau\|f^j\|. \tag{5.4.9}$$

С помощью рекуррентной связи приходим к условию устойчивости

$$\|\varphi\| \leqslant \|g\| + \tau j\|f\|, \tag{5.4.10}$$

где

$$\|f\| = \max_j \|f^j\|. \tag{5.4.11}$$

Заметим, что при проверке условия устойчивости $\|T\| \leqslant 1$ мы использовали начальное условие (5.4.7). Это совсем не обязательно. В качестве начального условия можно было выбрать любую функцию из того же класса гладкости, что и функции g и f.

5.4.2. Метод предиктор-корректор

Схема расщепления в этом случае имеет вид

$$\begin{aligned}
\left(E + \frac{\tau}{2} A_1\right) \varphi^{j+\frac{1}{2n}} &= \varphi^j + \frac{\tau}{2} f^j, \\
\left(E + \frac{\tau}{2} A_2\right) \varphi^{j+\frac{1}{2n}} &= \varphi^{j+\frac{1}{2n}}, \\
&\cdots\cdots\cdots\cdots\cdots\cdots \\
\left(E + \frac{\tau}{2} A_n\right) \varphi^{j+1/2} &= \varphi^{j+\frac{n-1}{2n}}, \\
\frac{\varphi^{j+1} - \varphi^j}{\tau} + A\varphi^{j+1/2} &= f^j,
\end{aligned} \tag{5.4.12}$$

где снова предполагается, что $A_\alpha \geqslant 0$ и $f^j = f(t_{j+1/2})$. Система уравнений (5.4.12) сводится к одному уравнению вида

$$\frac{\varphi^{j+1} - \varphi^j}{\tau} + A \prod_{\alpha=n}^{1} \left(E + \frac{\tau}{2} A_\alpha\right)^{-1} \left(\varphi + \frac{\tau}{2} f^j\right) = f^j, \quad \varphi^0 = g. \tag{5.4.13}$$

Метод предиктор-корректор при достаточной гладкости решения имеет второй порядок точности по τ. Уравнение (5.4.13) запишем в виде

$$\varphi^{j+1} = T\varphi^j + \frac{\tau}{2}(E + T) f^j, \tag{5.4.14}$$

где T — оператор шага:

$$T = E - \tau A \prod_{\alpha=n}^{1} \left(E + \frac{\tau}{2} A_\alpha\right)^{-1}. \tag{5.4.15}$$

Требование счетной устойчивости в конечном итоге сводится к оценке нормы оператора T. К сожалению, и в этом случае конструктивное условие $A_\alpha \geqslant 0$ не позволяет доказать устойчивость схемы. Этот вопрос остается открытым.

Чтобы закончить анализ рассмотренных выше двух схем расщепления, остановимся на простейшем случае, когда операторы A_α коммутируют друг с другом и имеют общий базис. Оказывается, этого дополнительного требования достаточно, чтобы из условия $A_\alpha \geqslant 0$ получить устойчивость рассмотренных схем. В самом деле, при условии коммутативности операторы шага T для обеих схем совпадают друг с другом. Рассмотрим для простоты однородную задачу (5.4.6), (5.4.7). Решение этой задачи будем искать в спектральной форме

$$\varphi^j = \sum_k \varphi_k^j u_k, \tag{5.4.16}$$

где u_k — собственные функции задачи (1.1.7) (см. гл. **??**), а $\varphi_k^j = (\varphi^j, u_k^*)$, где u_k^* — собственные функции сопряженной задачи (1.1.7) из главы 1. Поскольку u_k является общим базисом, то

$$Au_k = \lambda_k u_k, \quad A_\alpha u_k = \lambda_k^\alpha u_k, \quad \lambda_k = \sum_{\alpha=1}^{n} \lambda_k^\alpha. \tag{5.4.17}$$

Подставляя разложения (5.4.16) и соответствующие представления для функции g в (5.4.6), (5.4.7), получим для коэффициентов Фурье φ_k^j следующие формулы:

$$\varphi_k^{j+1} = T_k \varphi_k^j, \quad \varphi_k^0 = g_k, \tag{5.4.18}$$

где

$$T_k = 1 - \frac{\tau \lambda_k}{\prod_{\alpha=1}^{n} \left(1 + \frac{\tau}{2}\lambda_k^\alpha\right)}. \tag{5.4.19}$$

Выражение (5.4.19) для T_k преобразуем к виду

$$T_k = \frac{\mu_k - \frac{\tau}{2}\lambda_k}{\mu_k + \frac{\tau}{2}\lambda_k}, \tag{5.4.20}$$

где μ_k — положительные константы при условии $\lambda_k^\alpha \geqslant 0$. Из (5.4.20) вытекает неравенство

$$|T_k| \leqslant 1, \tag{5.4.21}$$

из которого, в соответствии с (5.1.3), следует устойчивость.

Метод стабилизации и метод предиктор-корректор при n-компонентном расщеплении могут быть применены и к случаю, когда оператор A зависит от времени. Однако в данной ситуации априорное установление условия устойчивости оказывается более сложной задачей. Поэтому трудно сказать, насколько оправдано применение рассмотренных двух схем в общих ситуациях. Это стимулировало автора к формулировке более или менее универсального подхода к решению различных сложных и достаточно общих задача на основе идеи расщепления. Таким методом оказался двуциклический метод последовательного расщепления.

5.4.3. Метод покомпонентного расщепления на основе элементарных схем

Попытаемся построить разностный аналог задачи, имеющий второй порядок аппроксимации по τ и абсолютно устойчивый по времени. В соответствии с предположением о многокомпонентном расщеплении будем полагать, что

$$\Lambda^j = \sum_{\alpha=1}^{n} \Lambda_\alpha^j, \tag{5.4.22}$$

где все Λ_α^j — положительно полуопределенные операторы: $\Lambda_\alpha^j \geqslant 0$. Рассмотрим систему уравнений

$$\left(E + \frac{\tau}{2}\Lambda_\alpha^j\right)\varphi^{j+\frac{\alpha}{n}} = \left(E - \frac{\tau}{2}\Lambda_\alpha^j\right)\varphi^{j+\frac{\alpha-1}{n}}, \alpha = 1,\ 2,\ \ldots,\ n. \tag{5.4.23}$$

В случае, когда $\Lambda_\alpha^j \geqslant 0$ коммутативны и $\Lambda_\alpha^j = A_\alpha^{j+1/2}$ или $\Lambda_\alpha^j = (A_\alpha^{j+1} + A_\alpha^j)/2$, схема (5.4.23) является безусловно устойчивой и имеет второй порядок аппроксимации. Этот факт можно установить довольно просто с помощью метода Фурье. Однако для некоммутативных операторов Λ_α^j, как легко заметить, схема (5.4.23) будет, вообще говоря, схемой первого порядка точности по τ и поэтому менее интересна для приложений, чем следующая схема второго порядка аппроксимации:

$$\begin{gathered}\varphi^{j+\frac{\alpha}{2n}} = \left(E - \frac{\tau}{2}\Lambda_\alpha^j\right)\varphi^{j+\frac{\alpha-1}{2n}}, \alpha = 1,\ 2,\ \ldots,\ n, \\ \left(E + \frac{\tau}{2}\Lambda_{2n-\alpha+1}^j\right)\varphi^{j+\frac{\alpha}{2n}} = \varphi^{j+\frac{\alpha-1}{2n}}, \alpha = n+1,\ n+2,\ \ldots,\ 2n.\end{gathered} \tag{5.4.24}$$

Попытаемся определить специальную конструкцию метода полного расщепления на основе (5.4.23), которая дает решение задачи Коши для положительно полуопределенных и некоммутативных операторов Λ_α^j и об-

ладает вторым порядком аппроксимации. Это является в известном смысле полным решением проблемы расщепления.

Заметим, что система уравнений (5.4.23) сводится к одному уравнению вида

$$\varphi^{j+1} = \prod_{\alpha=1}^{n} \left(E + \frac{\tau}{2}\Lambda_\alpha^j\right)^{-1} \left(E - \frac{\tau}{2}\Lambda_\alpha^j\right) \varphi^j. \tag{5.4.25}$$

С помощью (5.4.25) найдем оценку по норме

$$\|\varphi^{j+1}\| \leqslant \prod_{\alpha=1}^{n} \left\| \left(E + \frac{\tau}{2}\Lambda_\alpha^j\right)^{-1} \left(E - \frac{\tau}{2}\Lambda_\alpha^j\right) \right\| \|\varphi^j\|. \tag{5.4.26}$$

На основе леммы Келлога имеем

$$\|\varphi^{j+1}\| \leqslant \|\varphi^j\| \leqslant \cdots \leqslant \|g\|. \tag{5.4.27}$$

Если операторы кососимметричные, т. е. $(\Lambda_\alpha^j\varphi,\ \varphi) = 0$, то

$$\|\varphi^{j+1}\| = \|\varphi^j\| = \cdots = \|g\|. \tag{5.4.28}$$

Таким образом, абсолютная устойчивость этой схемы доказана.

Для того чтобы определить порядок аппроксимации, разложим по степеням малого параметра τ выражение (полагая $\frac{\tau}{2}\|\Lambda_\alpha\| < 1$)

$$T^j = \prod_{\alpha=1}^{n} \left(E + \frac{\tau}{2}\Lambda_\alpha^j\right)^{-1} \left(E - \frac{\tau}{2}\Lambda_\alpha^j\right).$$

Поскольку

$$T^j = \prod_{\alpha=1}^{n} T_\alpha^j,$$

то сначала разложим в ряд операторы T_α^j:

$$T_\alpha^j = E - \tau\Lambda_\alpha^j + \frac{\tau^2}{2}\left(\Lambda_\alpha^j\right)^2 \ldots. \tag{5.4.29}$$

В результате будем иметь

$$T^j = E - \tau\Lambda^j + \frac{\tau^2}{2}\left[(\Lambda^j)^2 + \sum_{\alpha=1}^{n}\sum_{\beta=\alpha+1}^{n} (\Lambda_\alpha^j\Lambda_\beta^j - \Lambda_\beta^j\Lambda_\alpha^j)\right] + O(\tau^3). \tag{5.4.30}$$

В случае, когда операторы Λ_α^j коммутативны, выражение, стоящее под знаком двойной суммы, обращается в нуль, и мы имеем

$$T^j = E - \tau\Lambda^j + \frac{\tau^2}{2}(\Lambda^j)^2 + O(\tau^3). \tag{5.4.31}$$

Сравнивая (5.4.31) с разложением оператора T^j из (5.1.8) в ряд по степеням $\tau\Lambda^j$, где Λ^j определяется с помощью (5.1.22)—(5.1.24), убеждаемся, что в этом частном случае схема (5.4.23) имеет второй порядок аппроксимации по τ. Если операторы Λ_α^j некоммутативны, то схема расщепления оказывается только первого порядка точности по τ. Чтобы построить схему второго порядка точности по τ в некоммутативном случае, необходимо схему (5.4.23) заменить следующей:

$$\varphi = \prod_{\alpha=1}^{n} T_\alpha^j \Phi^{j-1}, \quad \varphi^{j+1} = \prod_{\alpha=1}^{n} T_\alpha^j \Phi^j. \tag{5.4.32}$$

Алгоритмически это означает, что сначала решается система уравнений (5.4.23) на интервале $t_{j-1} \leqslant t \leqslant t_j$ для $\alpha = 1,\ 2,\ \ldots,\ n$, а затем такая же система на интервале $t_j \leqslant t \leqslant t_{j+1}$, но в обратной последовательности $(\alpha = n,\ n-1,\ \ldots,\ 1)$:

$$\begin{aligned} \left(E + \frac{\tau}{2}\Lambda_\alpha^j\right)\varphi^{j+\frac{\alpha}{n}-1} &= \left(E - \frac{\tau}{2}\Lambda_\alpha^j\right)\varphi^{j+\frac{\alpha-1}{n}-1}, \\ &\alpha = 1,\ 2,\ \ldots,\ n, \\ \left(E + \frac{\tau}{2}\Lambda_\alpha^j\right)\varphi^{j+1-\frac{\alpha-1}{n}} &= \left(E - \frac{\tau}{2}\Lambda_\alpha^j\right)\varphi^{j+1-\frac{\alpha}{n}}, \\ &\alpha = n,\ n-1,\ \ldots,\ 1. \end{aligned} \tag{5.4.33}$$

Очевидно, что для полного цикла (5.4.33) имеем

$$\varphi^{j+1} = T^j \varphi^{j-1},$$

где

$$T^j = \prod_{\alpha=1}^{n} T_\alpha^j \prod_{\alpha=n}^{1} T_\alpha^j = E - 2\tau\Lambda^j + \frac{(2\tau)^2}{2}\left(\Lambda^j\right)^2 + O(\tau^3).$$

Таким образом, на интервале $t_{j-1} \leqslant t \leqslant t_{j+1}$ схема (5.4.33) имеет второй порядок точности по τ, если в качестве Λ_α^j взят один из операторов, приведенных в (5.4.22)—(5.4.24).

В заключение заметим, что разностная система (5.4.33) оказывается абсолютно устойчива для $\Lambda_\alpha^j \geqslant 0$.

Для неоднородного уравнения

$$\begin{aligned} &\frac{\partial\varphi}{\partial t} + A\varphi = f, \\ &\varphi = g \text{ при } t = 0, \end{aligned} \tag{5.4.34}$$

где $A(t) \geqslant 0$ и

$$A = \sum_{\alpha=1}^{n} A_\alpha, \quad A_\alpha(t) \geqslant 0$$

на интервале $t_{j-1} \leqslant t \leqslant t_{j+1}$, имеет место следующая схема расщепления:

$$\begin{gathered}
\left(E + \frac{\tau}{2}\Lambda_1^j\right) \Phi^{j-\frac{n-1}{n}} = \left(E - \frac{\tau}{2}\Lambda_1^j\right) \varphi^{j-1}, \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
\left(E + \frac{\tau}{2}\Lambda_n^j\right) (\varphi^j - \tau f^j) = \left(E - \frac{\tau}{2}\Lambda_n^j\right) \varphi^{j-\frac{1}{n}}, \\
\left(E + \frac{\tau}{2}\Lambda_n^j\right) \varphi^{j+\frac{1}{n}} = \left(E - \frac{\tau}{2}\Lambda_n^j\right) (\varphi^j + \tau f^j), \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
\left(E + \frac{\tau}{2}\Lambda_1^j\right) \varphi^{j+1} = \left(E - \frac{\tau}{2}\Lambda_1^j\right) \varphi^{j+\frac{n-1}{n}},
\end{gathered} \tag{5.4.35}$$

где

$$\Lambda_\alpha^j = A_\alpha(t_j).$$

Нетрудно убедиться, что эта схема имеет второй порядок аппроксимации по τ и в предположении необходимой гладкости φ абсолютно устойчива.

Так же как и в случае $n = 2$, n-компонентную систему уравнений (5.4.35) можно записать в эквивалентной форме:

$$\begin{aligned}
\left(E + \frac{\tau}{2}\Lambda_\alpha\right) \varphi^{j-\frac{(n+1)-\alpha}{(n+1)}} &= \left(E - \frac{\tau}{2}\Lambda_\alpha\right) \varphi^{j-\frac{(n+1)-\alpha+1}{(n+1)}}, \\
&\alpha = 1,\ 2,\ \ldots,\ n, \\
\varphi^{j+\frac{1}{(n+1)}} &= \varphi^{j-\frac{1}{(n+1)}} + 2\tau f^j, \\
\left(E + \frac{\tau}{2}\Lambda_{n-\alpha+2}\right) \varphi^{j+\frac{\alpha}{(n+1)}} &= \left(E - \frac{\tau}{2}\Lambda_{n-\alpha+2}\right) \varphi^{j+\frac{\alpha-1}{(n+1)}}, \\
&\alpha = 2,\ 3,\ \ldots,\ n+1.
\end{aligned} \tag{5.4.36}$$

Перейдем к методу расщепления для неявных разностных аппроксимаций. С этой целью рассмотрим задачу

$$\begin{aligned}
&\frac{\partial \varphi}{\partial t} + A\varphi = 0 \quad \text{в } D \times D_t, \\
&\varphi = g^{3)} \quad \text{в } D \text{ при } t = 0.
\end{aligned} \tag{5.4.37}$$

Предположим, что

$$A = \sum_{\alpha=1}^{n} A_\alpha,$$

все $A_\alpha \geqslant 0$ и A_α не зависят от времени. Рассмотрим алгоритм расщепления в следующем виде:

$$\begin{gathered} \frac{\varphi^{j+1/n} - \varphi^j}{\tau} + A_1 \varphi^{j+1/n} = 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \frac{\varphi^{j+1} - \varphi^{j+(n-1)/n}}{\tau} + A_n \varphi^{j+1} = 0. \end{gathered} \tag{5.4.38}$$

Покажем, что такой алгоритм является абсолютно устойчивым. Умножим уравнение

$$\frac{\varphi^{j+\alpha/n} - \varphi^{j+(\alpha-1)/n}}{\tau} + A_\alpha \varphi^{j+\alpha/n} = 0 \tag{5.4.39}$$

скалярно на $\varphi^{j+\alpha/n}$. Получим соотношение

$$(\varphi^{j+\alpha/n} - \varphi^{j+(\alpha-1)/n},\ \varphi^{j+\alpha/n}) + \tau(A_\alpha \varphi^{j+\alpha/n},\ \varphi^{j+\alpha/n}) = 0.$$

Из положительной полуопределенности оператора A_α вытекает, что

$$(\varphi^{j+\alpha/n} - \varphi^{j+(\alpha-1)/n},\ \varphi^{j+\alpha/n}) \leqslant 0$$

или

$$(\varphi^{j+\alpha/n},\ \varphi^{j+\alpha/n}) \leqslant (\varphi^{j+\alpha/n},\ \varphi^{j+(\alpha-1)/n}).$$

Но так как

$$(\varphi^{j+\alpha/n},\ \varphi^{j+(\alpha-1)/n}) \leqslant \frac{1}{2}[(\varphi^{j+\alpha/n},\ \varphi^{j+\alpha/n}) + (\varphi^{j+(\alpha-1)/n},\ \varphi^{j+(\alpha-1)/n})],$$

то

$$\|\varphi^{j+\alpha/n}\| \leqslant \|\varphi^{j+(\alpha-1)/n}\|^2, \quad \alpha = 1,\ 2,\ \ldots,\ n.$$

С помощью этого рекуррентного неравенства получаем

$$\|\varphi^{j+1}\| \leqslant \|\varphi^j\|. \tag{5.4.40}$$

Это означает, что при сделанных предположениях расчет по схеме расщепления (5.4.38) будет абсолютно устойчивым.

Нетрудно убедиться, что система (5.4.38) аппроксимирует исходную задачу с первым порядком точности по τ.

Рассмотрим теперь неоднородную задачу

$$\begin{gathered} \frac{\partial \varphi}{\partial t} + A\varphi = f \text{ в } D \times D_t, \\ \varphi = g \text{ в } D \text{ при } t = 0 \end{gathered} \tag{5.4.41}$$

и схему расщепления для этой задачи

$$\begin{gathered}\frac{\varphi^{j+1/n}-\varphi^{j}}{\tau}+A_1\varphi^{j+1/n}=0,\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ \frac{\varphi^{j+1}-\varphi^{j+(n-1)/n}}{\tau}+A_n\varphi^{j+1}=f^{j}.\end{gathered}\tag{5.4.42}$$

Такая схема расщепления аппроксимирует исходное неоднородное уравнение с точностью до вилечин первого порядка по τ.

Устойчивость схемы (5.4.42) докажем следующим образом. Умножим скалярно каждое из уравнений соответственно на $\varphi^{j+1/n},\ \ldots,\ \varphi^{j+1}$. Тогда аналогично предыдущему получим

$$\|\varphi^{j+\alpha/n}\|\leqslant\|\varphi^{j+(\alpha-1)/n}\|,\quad \alpha=1,\ 2,\ \ldots,\ n-1.\tag{5.4.43}$$

Последнее уравнение, полученное из (5.4.42), рассмотрим более подробно. Имеем

$$(\varphi^{j+1},\ \varphi^{j+1})=(\varphi^{j+(n-1)/n},\ \varphi^{j+1})-\tau(A_n\varphi^{j+1},\ \varphi^{j+1})+\tau(f^{j},\ \varphi^{j+1}).$$

Учитывая, что $A_n\geqslant 0$, получим

$$(\varphi^{j+1},\ \varphi^{j+1})\leqslant(\varphi^{j+(n-1)/n},\ \varphi^{j+1})+\tau(f^{j},\ \varphi^{j+1}).$$

Так как

$$\begin{gathered}|(\varphi^{j+(n-1)/n},\ \varphi^{j+1})|\leqslant\|\varphi^{j+(n-1)/n}\|\|\varphi^{j+1}\|,\\ |(f^{j},\ \varphi^{j+1})\|\leqslant\|f^{j}\|\|\varphi^{j+1}\|,\end{gathered}$$

то

$$\|\varphi^{j+1}\|^2\leqslant\|\varphi^{j+(n-1)/n}\|\|\varphi^{j+1}\|+\tau\|f^{j}\|\|\varphi^{j+1}\|.$$

Сокращая на $\|\varphi^{j+1}\|$, приходим к следующему неравенству:

$$\|\varphi^{j+1}\|\leqslant\|\varphi^{j+(n-1)/n}\|+\tau\|f^{j}\|.$$

Исключая решение с дробными индексами, будем иметь

$$\|\varphi^{j+1}\|\leqslant\|\varphi^{j}\|+\tau\|f^{j}\|.\tag{5.4.44}$$

Учитывая, что

$$\|\varphi^{0}\|=\|g\|,$$

и исключая промежуточные значения решения, получим

$$\|\varphi^{j+1}\| \leqslant \|g\| + \tau j \|f\|, \tag{5.4.45}$$

где

$$\|f\| = \max_j \|f^j\|.$$

Отсюда следует абсолютная устойчивость разностной схемы.

Данный алгоритм расщепления обобщается на случай, когда оператор A зависит от времени. В этом случае в цикле вычислений по схеме расщепления вместо A следует взять подходящую разностную аппроксимацию этого оператора на каждом интервале $t_j \leqslant t \leqslant t_{j+1}$.

5.4.4. Расщепление квазилинейных задач

Рассмотрим эволюционную задачу с оператором A, зависящим от времени и от решения задачи:

$$\begin{gathered} \frac{\partial \varphi}{\partial t} + A(t,\ \varphi)\varphi = 0 \text{ в } D \times D_t, \\ \varphi = g \text{ в } D \text{ при } t = 0. \end{gathered} \tag{5.4.46}$$

Относительно оператора $A(t,\ \varphi)$ предположим, что он неотрицателен, имеет вид

$$A(t,\ \varphi) = \sum_{\alpha=1}^{n} A_\alpha(t,\ \varphi), \tag{5.4.47}$$

$A_\alpha(t,\ \varphi) \geqslant 0$ и обладает достаточной гладкостью. Предположим, далее, что решение φ также является достаточно гладкой функцией времени. Рассмотрим на интервале $t_{j-1} \leqslant t \leqslant t_{j+1}$ схему расщепления

$$\begin{gathered} \frac{\varphi^{j+1/n-1} - \varphi^{j-1}}{\tau} + A_1^j \frac{\varphi^{j+1/n-1} + \varphi^{j-1}}{2} = 0, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \frac{\varphi^{j} - \varphi^{j-1/n}}{\tau} + A_n^j \frac{\varphi^{j} + \varphi^{j-1/n}}{2} = 0, \\ \frac{\varphi^{j+1/n} - \varphi^{j}}{\tau} + A_n^j \frac{\varphi^{j+1/n} + \varphi^{j}}{2} = 0, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \frac{\varphi^{j+1} - \varphi^{j+(n-1)/n}}{\tau} + A_1^j \frac{\varphi^{j+1} + \varphi^{j+(n-1)/n}}{2} = 0, \end{gathered} \tag{5.4.48}$$

где

$$\begin{aligned} A^j =& A(t_j,\ \tilde{\varphi}^j), \\ \tilde{\varphi}^j =& \varphi^{j-1} - \tau A(t_{j-1},\ \varphi^{j-1})\varphi^{j-1}, \\ \tau =& t_j - t_{j-1}. \end{aligned} \tag{5.4.49}$$

Методами, изложенными выше для линейных операторов, зависящих только от времени, несложно доказать, что схема расщепления (5.4.48) при условиях (5.4.49) имеет второй порядок аппроксимации по τ и абсолютно устойчива. Аналогичным образом определяется метод расщепления для неоднородных квазилинейных уравнений. Это открывает широкие возможности применения схем покомпонентного расщепления к решению нестационарных квазилинейных задач гидродинамики, метеорологии, океанологии и других важных областей естествознания.

5.5. Общий подход к покомпонентному расщеплению

При решении многих задач математической физики возникает необходимость расщепления исходных дифференциальных, интегральных или интегро-дифференциальных уравнений на более простые с последующей редукцией их к разностной форме на основе изложенных в настоящей главе алгоритмов. Этот вопрос тесно связан с проблемой слабой аппроксимации исходных уравнений уравнениями более простой структуры.

Пусть имеется некоторая задача математической физики

$$\begin{aligned} \frac{\partial \varphi}{\partial t} + A\varphi = 0 \text{ в } D \times D_t, \\ \varphi = g \text{ в } D \text{ при } t = 0. \end{aligned} \tag{5.5.1}$$

Предположим, что

$$A = \sum_{\alpha=1}^{n} A_\alpha, \tag{5.5.2}$$

причем $A_\alpha \geqslant 0$. Решение φ и функция f предполагаются достаточно гладкими. Задачу (5.5.1) на каждом интервале $\theta_j = \{t_j \leqslant t \leqslant t_{j+1}\}$ представим в следующем виде:

$$\begin{aligned} \frac{\partial \varphi_\alpha}{\partial t} + A_\alpha \varphi_\alpha = 0 \text{ в } D \times \theta_j, \\ \varphi_\alpha^j = \varphi_{\alpha-1}^{j+1} \text{ в } D \text{ при } t = t_j, \\ \alpha = 1,\ 2,\ \ldots,\ n. \end{aligned} \tag{5.5.3}$$

При этом введены следующие обозначения:

$$\varphi_0^{j+1} = \varphi^j, \quad \varphi_n^{j+1} = \varphi^{j+1}. \tag{5.5.4}$$

Ранее было показано, что если к каждому из уравнений применить схему Кранка — Николсона, то придем к система разностных уравнений

$$\frac{\varphi^{j+\alpha/n} - \varphi^{j+(\alpha-1)/n}}{\tau} + A_\alpha \frac{\varphi^{j+\alpha/n} - \varphi^{j+(\alpha-1)/n}}{2} = 0, \qquad \alpha = 1,\ 2,\ \ldots,\ n, \tag{5.5.5}$$

где

$$\varphi^{j+\alpha/n} = \varphi_\alpha^{j+1}, \quad \varphi^{j+1} = \varphi_n^{j+1}. \tag{5.5.6}$$

Предположим, что каждый из операторов A_α в свою очередь представим в виде

$$A_\alpha = \sum_{\beta=1}^{m_\alpha} A_{\alpha\beta}, \tag{5.5.7}$$

где $A_{\alpha\beta} \geqslant 0$. Возникает вопрос: целесообразно ли предварительно оператор A «расщеплять» на A_α, а затем операторы A_α в свою очередь расщеплять на $A_{\alpha\beta}$? Не проще ли сразу A представить через набор операторов $A_{\alpha\beta}$? По этому поводу следует заметить, что, несмотря на внешнюю эквивалентность этих подходов, во многих случаях оказывается целесообразным сначала разложить сложную задачу математической физики на более простые, которые в дальнейшем удобно независимо друг от друга сводить к простейшим задачам.

Рассмотрим систему (5.5.3) и с учетом (5.5.7) расщепим ее на еще более элементарные:

$$\begin{gathered}\frac{\varphi_\alpha^{j+\beta/m_\alpha} - \varphi_\alpha^{j+(\beta-1)/m_\alpha}}{\tau} + A_{\alpha\beta} \frac{\varphi_\alpha^{j+\beta/m_\alpha} + \varphi_\alpha^{j+(\beta-1)/m_\alpha}}{2} = 0, \\ \alpha = 1,\ 2,\ \ldots,\ n, \quad \beta = 1,\ 2,\ \ldots,\ m_\alpha,\end{gathered} \tag{5.5.8}$$

где

$$\varphi_1^j = \varphi^j; \quad \varphi_\alpha^j = \varphi_{\alpha-1}^{j+1}\ (\alpha > 1); \quad \varphi_n^{j+1} = \varphi^{j+1}.$$

Нетрудно видеть, что система расщепленных уравнений (5.5.8) аппроксимирует исходную задачу (5.5.1) с точностью до величин второго порядка по τ, если операторы $A_{\alpha\beta}$ коммутативны. Доказательство этого утверждения основано на том, что с учетом (5.5.2) и (5.5.7) можно изменить упорядочение компонент расщепления, записав

$$A = \sum_{\alpha=1}^{n} \sum_{\beta=1}^{m_\alpha} A_{\alpha\beta} = \sum_{\gamma=1}^{p} A_\gamma.$$

В этом случае мы приходим к задаче

$$\frac{\varphi^{j+\gamma/p}-\varphi^{j+(\gamma-1)/p}}{\tau}+A_\gamma\frac{\varphi^{j+\gamma/p}+\varphi^{j+(\gamma-1)/p}}{2}=0,\qquad \gamma=1,\ 2,\ \ldots,\ p, \tag{5.5.9}$$

которая, как было показано в 4.4, аппроксимирует задачу (5.5.1) со вторым порядком точности по τ. Этот результат остается в силе и тогда, когда операторы $A_{\alpha\beta}$ зависят от времени. В этом случае необходимо на каждом интервале $t_j \leqslant t \leqslant t_{j+1}$ произвести аппроксимацию операторов $A_{\alpha\beta}=\Lambda^j_{\alpha\beta}$ со вторым порядком по τ. Если операторы $\Lambda^j_{\alpha\beta}$ некоммутативны, то методом двуциклической процедуры, описанной в 4.4, приходим к разностной схеме второго порядка точности на каждом интервале $t_{j-1}\leqslant t\leqslant t_{j+1}$.

Резюмируя изложенное, можно утверждать, что если эволюционную задачу вида (5.5.1) при условии $A_\alpha \geqslant 0$ свести к частным задачам эволюционного типа (5.5.3) и затем рассматривать как набор новых эволюционных задач, то, если хотя бы одна из элементарных эволюционных задач редуцируется к разностным схемам первого порядка точности, тогда и аппроксимация исходной задачи (5.5.1) будет первого порядка точности по τ. Если же каждая из таких задач имеет аппроксимацию второго порядка, то в рамках двуциклической процедуры по α и β приходим к аппроксимации второго порядка по τ. Заметим, что если операторы $A_{\alpha\beta}$ некоммутативны, то без двуциклической процедуры мы приходим к аппроксимации задачи (5.5.1) с первым порядком точности. В случае некоммутирующих операторов исходной задачей будет следующая:

$$\frac{\partial\varphi}{\partial t}+\sum_{\alpha=1}^{n}A_\alpha\varphi=0 \text{ в } D\times\theta_j,\qquad \varphi=\varphi^j \text{ в } D \text{ при } t=t_j. \tag{5.5.10}$$

Задачу (5.5.10) редуцируем к системе

$$\frac{\partial\varphi_\alpha}{\partial t}+A_\alpha\varphi_\alpha=0,\ \varphi^j_\alpha=\varphi^{j+1}_{\alpha-1},\quad \alpha=1,\ 2,\ \ldots,\ n. \tag{5.5.11}$$

Пусть $A_\alpha = \sum_{\beta=1}^{m_\alpha} A_{\alpha\beta}$. Тогда каждую из задач (5.5.11) будем решать с помощью двуциклического метода:

$$\begin{gathered}\frac{\varphi_\alpha^{j+\frac{\beta}{2m_\alpha}} - \varphi_\alpha^{j+\frac{\beta-1}{2m_\alpha}}}{\tau/2} + A_{\alpha\beta}\frac{\varphi^{j+\frac{\beta}{2m_\alpha}} + \varphi^{j+\frac{\beta-1}{2m_\alpha}}}{2} = 0,\ \beta = 1,\ 2,\ \ldots,\ m_\alpha,\\ \frac{\varphi_\alpha^{j+\frac{\beta}{2m_\alpha}} - \varphi_\alpha^{j+\frac{\beta-1}{2m_\alpha}}}{\tau/2} + A_{\alpha,\ 2m_\alpha+1-\beta}\frac{\varphi^{j+\frac{\beta}{2m_\alpha}} + \varphi^{j+\frac{\beta-1}{2m_\alpha}}}{2} = 0,\\ \beta = m_\alpha + 1,\ m_\alpha + 2,\ \ldots,\ 2m_\alpha.\end{gathered} \tag{5.5.12}$$

Начальные условия для решения каждой из систем (5.5.12) берутся соответственно в виде

$$\varphi_1^j = \varphi^j, \quad \varphi_\alpha^j = \varphi_{\alpha-1}^{j+1}, \quad (\alpha = 2,\ \ldots,\ n). \tag{5.5.13}$$

Нетрудно проверить, что задача (5.5.12) аппроксимирует на интервале $t_j \leqslant t \leqslant t_{j+1}$ любую из задач (5.5.11) с точностью до τ^2.

Чтобы весь алгоритм приводил к решению задачи (5.5.1) с точностью до τ^2, необходимо, кроме того, чередовать и основные циклы. Так, вместо (5.5.11) на интервале $t_{j-1} \leqslant t \leqslant t_j$ следует иметь систему

$$\begin{gathered}\frac{\partial \varphi_\alpha}{\partial t} + A_\alpha \varphi_\alpha = 0, \quad (\alpha = 1,\ 2,\ \ldots,\ n),\\ \varphi_1^{j-1} = \varphi^{j-1}, \quad \varphi_\alpha^{j-1} = \varphi_{\alpha-1}^j\ (\alpha > 1),\ \varphi^j = \varphi_n^j,\end{gathered} \tag{5.5.14}$$

а на интервале $t_j \leqslant t \leqslant t_{j+1}$ —

$$\begin{gathered}\frac{\partial \varphi_\alpha}{\partial t} + A_{n-\alpha+1}\varphi_\alpha = 0, \quad \alpha = 1,\ 2,\ \ldots,\ n,\\ \varphi_1^j = \varphi^j, \varphi_\alpha^j = \varphi_{\alpha-1}^{j+1}\ (\alpha > 1),\ \varphi^{j+1} = \varphi_n^{j+1}.\end{gathered} \tag{5.5.15}$$

При этом предполагается, что каждая из задач (5.5.14) и (5.5.15) решается с помощью двуциклического метода вида (5.5.12). Заметим, что при условии $A_{\alpha\beta} \geqslant 0$ метод покомпонентного расщепления является абсолютно устойчивым.

В заключение приведем общую схему расщепления для неоднородного уравнения

$$\begin{gathered}\frac{\partial \varphi}{\partial t} + \sum_{\alpha=1}^{n} A_\alpha \varphi = f,\\ \varphi = g \text{ при } t = 0\end{gathered} \tag{5.5.16}$$

на интервале $t_{j-1} \leqslant t \leqslant t_{j+1}$ на основе двуциклического метода. Рассмотрим схемы слабой аппроксимации в дифференциальной форме.

На интервале $t_{j-1} \leqslant t \leqslant t_j$ положим

$$\begin{aligned} &\frac{\partial \varphi_\alpha}{\partial t} + A_\alpha \varphi_\alpha = 0, \quad \alpha = 1,\ 2,\ \ldots,\ n-1, \\ &\frac{\partial \varphi_n}{\partial t} + A_n \varphi_n = f + \frac{\tau}{2} A_n f, \end{aligned} \tag{5.5.17}$$

а на интервале $t_j \leqslant t \leqslant t_{j+1}$ —

$$\begin{aligned} \frac{\partial \varphi_{n+1}}{\partial t} + A_n \varphi_{n+1} &= f - \frac{\tau}{2} A_n f, \\ \frac{\partial \varphi_{n+\alpha}}{\partial t} + A_{n-\alpha+1} \varphi_{n+\alpha} &= 0, \quad \alpha = 2,\ 3,\ \ldots,\ n, \end{aligned} \tag{5.5.18}$$

при условии, что

$$\varphi_1(t_{j-1}) = \varphi(t_{j-1}), \quad \varphi_{\alpha+1}(t_{j-1}) = \varphi_\alpha(t_j), \quad \alpha = 1,\ 2,\ \ldots,\ n-1, \tag{5.5.19}$$

и, соответственно,

$$\varphi_{\alpha+1}(t_j) = \varphi_\alpha(t_{j+1}), \quad \alpha = n,\ n+1,\ n+2,\ \ldots,\ 2n-1. \tag{5.5.20}$$

Если теперь для решения уравнений (5.5.17)—(5.5.19) на интервале $t_{j-1} \leqslant t \leqslant t_{j+1}$ воспользоваться схемами Кранка — Николсона, положив $f = f^j$, то придем к системе (5.4.35) для (5.5.18).

Наряду с системой уравнений (5.5.17), (5.5.18) рассмотрим следующую. На интервале $t_{j-1} \leqslant t \leqslant t_j$

$$\begin{aligned} &\frac{\partial \varphi_1}{\partial t} + A_1 \varphi_1 = 0, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ &\frac{\partial \varphi_n}{\partial t} + A_n \varphi_n = 0, \end{aligned} \tag{5.5.21}$$

на интервале $t_{j-1} \leqslant t \leqslant t_{j+1}$

$$\frac{\partial \varphi_{n+1}}{\partial t} = f \tag{5.5.22}$$

и на интервале $t_j \leqslant t \leqslant t_{j+1}$

$$\begin{aligned} &\frac{\partial \varphi_{n+2}}{\partial t} + A_n \varphi_{n+2} = 0, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ &\frac{\partial \varphi_{2n+1}}{\partial t} + A_1 \varphi_{2n+1} = 0. \end{aligned} \tag{5.5.23}$$

Начальными условиями для системы (5.5.21) будут

$$\varphi_1(t_{j-1}) = \varphi(t_{j-1}), \quad \varphi_\alpha(t_{j-1}) = \varphi_{\alpha-1}(t_j), \quad \alpha = 2,\ 3,\ \ldots,\ n, \tag{5.5.24}$$

для уравнения (5.5.22) —

$$\varphi_{n+1}(t_{j-1}) = \varphi_n(t_j) \tag{5.5.25}$$

и для системы уравнений (5.5.23) —

$$\varphi_\alpha(t_j) = \varphi_{\alpha-1}(t_{j+1}), \quad \alpha = n+2,\ n+3,\ \ldots,\ 2n+1. \tag{5.5.26}$$

Аппроксимация и устойчивость полученных схем гарантирует сходимость (см. 5.1.4).

5.6. Методы решения уравнений гиперболического типа

Широкий класс задач математической физики связан с уравнениями гиперболического типа, численные методы решения которых начали свое развитие с фундаментального исследования Куранта, Фридрихса, Леви. Большой комплекс исследований в дальнейшем был проведен советскими и зарубежными авторами. В настоящее время предложен ряд эффективных алгоритмов решения задач, связанных с уравнениями гиперболического типа в применении к теории колебаний, теории упругости и т. д. в случае многомерных областей. Эти методы, основанные на специальных алгоритмах расщепления, будут рассмотрены в настоящем параграфе.

5.6.1. Метод стабилизации

Рассмотрим задачу

$$\begin{gathered} \frac{\partial^2 \varphi}{\partial t^2} + A\varphi = f \text{ в } D \times D_t, \\ \varphi = p, \frac{\partial \varphi}{\partial t} = q \text{ в } D \text{ при } t = 0. \end{gathered} \tag{5.6.1}$$

Будем считать, что оператор A является конечно-разностным и не зависит от времени, а функции p и q обладают такими свойствами, что допускают достаточную гладкость решения. Предположим, что оператор A положительно определен, т. е.

$$(A\varphi,\ \varphi) \geqslant \gamma^2(\varphi,\ \varphi). \tag{5.6.2}$$

Напомним, что для положительно определенных операторов A справедливо соотношение

$$\gamma^2 = \alpha \left(\frac{A^* + A}{2} \right),$$

где α — минимальное собственное число оператора $(A + A^*)/2$.

Рассмотрим разностную аппроксимацию уравнения (5.6.1) в форме

$$\frac{\varphi^{j+1} - 2\varphi^j + \varphi^{j-1}}{\tau^2} + A\varphi^j = f^j. \tag{5.6.3}$$

Нетрудно показать, что на гладких решениях разностная схема (5.6.3) аппроксимирует исходное уравнение из (5.6.1) со вторым порядком по τ. Присоединим к уравнению (5.6.3) начальные данные. Чтобы не нарушить второго порядка аппроксимации, наряду с уравнением (5.6.3) рассмотрим начальные данные в следующей форме:

$$\varphi^0 = p, \quad \varphi^1 = \left(E - \frac{\tau^2}{2} A \right) p + \tau q + \frac{\tau^2}{2} f^0. \tag{5.6.4}$$

Последнее соотношение (5.6.4) получено разложением решения задачи (5.6.1) в ряд Тейлора в окрестности точки $t = 0$ с последующим исключением производных с помощью уравнения и заданных начальных условий в задаче (5.6.1).

Задача (5.6.3), (5.6.4) поставлена полностью. Теперь нам необходимо исследовать схему (5.6.3) на счетную устойчивость. С этой целью используем спектральный метод.

Пусть u_n и u_n^* — собственные функции, а $\lambda_n = 0$ — собственные числа спектральных задач

$$Au_n = \lambda_n u_n, \quad A^* u_n^* = \lambda_n u_n^*. \tag{5.6.5}$$

Предположим, далее, что $\{u_n\}$ образуют базис. Решение уравнения будем искать в виде

$$\varphi^j = \sum_n \varphi_n^j u_n, \tag{5.6.6}$$

где

$$\varphi_n^j = (\varphi^j,\ u_n^*).$$

Подставив ряд Фурье (5.6.6) в (5.6.3) и умножив результат на u_n^*, получим уравнение для коэффициентов Фурье φ_n^j:

$$\frac{\varphi_n^{j+1} - 2\varphi_n^j + \varphi_n^{j-1}}{\tau^2} + \lambda_n \varphi_n^j = f_n^j. \tag{5.6.7}$$

Общее решение однородного уравнения, соответствующего (5.6.7), будем искать в виде степенной функции

$$\varphi_n^j = \eta_n^j. \tag{5.6.8}$$

Подчеркнем еще раз, что в левой части равенства (5.6.8) j является индексом, а в правой — показателем степени.

Подставляя (5.6.8) в (5.6.7) и полагая $f_n^j = 0$, приходим к характеристическому уравнению для η_n:

$$\eta_n^2 - 2\left(1 - \frac{\tau^2\lambda_n}{2}\right)\eta_n + 1 = 0. \tag{5.6.9}$$

Легко видеть, что при условии

$$\left|1 - \frac{\tau^2\lambda_n}{2}\right| \leqslant 1 \tag{5.6.10}$$

корни уравнения (5.6.9) будут комплексно-сопряженными (в случае строгого неравенства) и по модулю равными единице, т. е.

$$|\eta_n| = 1. \tag{5.6.11}$$

Из условия (5.6.10) получаем

$$\tau^2 \leqslant \frac{4}{\lambda_n}, \quad n = 1,\ 2,\ \ldots. \tag{5.6.12}$$

Очевидно, что (5.6.12) будет выполняться для всех λ_n, если брать значения τ таким, что

$$\tau \leqslant \frac{2}{\sqrt{\beta(A)}}, \tag{5.6.13}$$

где $\beta(A)$ — верхняя граница спектра оператора A. Для случая симметричных операторов $\beta(A) = \|A\|$ и, следовательно,

$$\tau \leqslant \frac{2}{\sqrt{\|A\|}}. \tag{5.6.14}$$

Переходим теперь к рассмотрению неявных разностных схем:

$$\frac{\varphi^{j+1} - 2\varphi^j + \varphi^{j-1}}{\tau^2} + A\frac{\varphi^{j+1} + \varphi^{j-1}}{2} = f^j. \tag{5.6.15}$$

Схема (5.6.15) имеет второй порядок точности по τ и вместе с (5.6.4) со вторым порядком аппроксимирует задачу (5.6.1). Характеристическое

уравнение для (5.6.15) имеет вид

$$\eta_n^2 - \frac{2}{1+\frac{\tau^2\lambda_n}{2}}\eta_n + 1 = 0; \tag{5.6.16}$$

следовательно,

$$\eta_n = \frac{1}{1+\frac{\tau^2\lambda_n}{2}} \pm \sqrt{\left(\frac{1}{1+\frac{\tau^2\lambda_n}{2}}\right)^2 - 1}. \tag{5.6.17}$$

Отсюда видно, что при любых τ и n выполняется равенство

$$|\eta_n| = 1. \tag{5.6.18}$$

Схема (5.6.15) является абсолютно устойчивой (см. Рихтмайер, Мортон [3]).

Рассмотрим теперь случай, когда

$$A = \sum_{\alpha=1}^{n} A_\alpha, \tag{5.6.19}$$

причем все $A_\alpha \geqslant 0$. Для приближенного решения задачи (5.6.1) в этом случае используем разностную аппроксимацию в виде

$$B\frac{\varphi^{j+1} - 2\varphi^j + \varphi^{j-1}}{\tau^2} + A\varphi^j = f^j, \tag{5.6.20}$$

где

$$B = \prod_{\alpha=1}^{n}\left(E + \frac{\tau^2}{2}A_\alpha\right). \tag{5.6.21}$$

Из (5.6.20) и (5.6.21) следует, что уравнение (5.6.20) аппроксимирует исходное уравнение (5.6.1) с точностью до величин второго порядка по τ. Поскольку уравнение (5.6.20) можно привести к виду

$$\frac{\varphi^{j+1} - 2\varphi^j + \varphi^{j-1}}{\tau^2} + B^{-1}A\varphi^j = B^{-1}f^j, \tag{5.6.22}$$

то из анализа Фурье будет следовать необходимое условие устойчивости схемы (5.6.20), (5.6.21):

$$\tau \leqslant \frac{2}{\sqrt{\beta(B^{-1}A)}}. \tag{5.6.23}$$

Таким образом, задача выбора параметра τ, удовлетворяющего условию устойчивости, свелась к вычислению максимального собственного числа задачи

$$Au = \lambda Bu \tag{5.6.24}$$

в предположении, что все собственные числа $B^{-1}A$ положительны. Эта задача решается с помощью итерационного процесса Люстерника.

Запишем схему реализации разностной схемы, соответствующей уравнению (5.6.20):

$$\begin{gathered}\left(E+\frac{\tau^2}{2}A_1\right)\xi^{j+1/n}=-A\varphi^j+f^j,\\ \left(E+\frac{\tau^2}{2}A_2\right)\xi^{j+2/n}=\xi^{j+1/n},\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ \left(E+\frac{\tau^2}{2}A_n\right)\xi^{j+1}=\xi^{j+(n-1)/n},\\ \varphi^{j+1}=2\varphi^j-\varphi^{j-1}+\tau^2\xi^{j+1}.\end{gathered}\tag{5.6.25}$$

Решение этой задачи производится последовательно при $j=2,\ 3,\ \ldots$, причем с использованием начальных данных (5.6.4). Схема (5.6.20) является схемой расщепления.

5.6.2. Сведение уравнения колебаний к эволюционной задаче

Построение абсолютно устойчивых разностных схем для уравнений гиперболического типа, обладающих вторым порядком аппроксимации и эффективно реализуемых на ЭВМ, привело к необходимости создания специальных методов расщепления, аналогичных рассмотренным в случае эволюционных задач.

Основную идею формального сведения гиперболической задачи к эволюционной проиллюстрируем на простейшей задаче о колебании мембраны с периодическими относительно квадрата $D=\{0\leqslant x\leqslant 1, 0\leqslant y\leqslant 1\}$ начальными условиями:

$$\begin{gathered}\frac{\partial^2\varphi}{\partial t^2}=\frac{\partial}{\partial x}a^2\frac{\partial\varphi}{\partial x}+\frac{\partial}{\partial y}a^2\frac{\partial\varphi}{\partial y}\ \text{в}\ D\times D_t,\\ \varphi=p,\quad \frac{\partial\varphi}{\partial t}=q\ \text{в}\ D\ \text{при}\ t=0.\end{gathered}\tag{5.6.26}$$

Здесь $a^2=a^2(x,\ y)$ — квадрат скорости распространения возмущений, $p=p(x,\ y)$, $q=q(x,\ y)$ — заданные функции. Периодическое решение φ обладает достаточной гладкостью для проведения последующих преобразований и построения разностных схем второго порядка точности по всем переменным x, y, t.

Прежде всего уравнение колебаний из (5.6.27) представим в виде системы уравнений

$$\begin{aligned} &\frac{\partial u}{\partial t} - a\frac{\partial \varphi}{\partial x} = 0, \\ &\frac{\partial v}{\partial t} - a\frac{\partial \varphi}{\partial y} = 0 \text{ в } D \times D_t, \\ &\frac{\partial \varphi}{\partial t} - \left(\frac{\partial au}{\partial x} + \frac{\partial av}{\partial y}\right) = 0. \end{aligned} \tag{5.6.27}$$

В качестве начальных данных для функции u, v и φ выберем следующие:

$$u = u^0(x,\ y), \quad v = v^0(x,\ y), \quad \varphi = p(x,\ y) \text{ при } t = 0. \tag{5.6.28}$$

Функции u^0 и v^0 можно выбрать более или менее произвольно, лишь бы они были связаны зависимостью

$$\frac{\partial au^0}{\partial x} + \frac{\partial av^0}{\partial y} = q(x,\ y). \tag{5.6.29}$$

Введем в рассмотрение матрицу A и вектор φ:

$$A = \begin{Vmatrix} 0 & 0 & -a\frac{\partial}{\partial x} \\ 0 & 0 & -a\frac{\partial}{\partial y} \\ -\frac{\partial}{\partial x}a & -\frac{\partial}{\partial y}a & 0 \end{Vmatrix}, \quad \varphi = \begin{Vmatrix} u \\ v \\ \varphi \end{Vmatrix}.$$

Тогда систему уравнений (5.6.27) и начальные данные (5.6.28) можно представить в виде

$$\begin{aligned} &\frac{\partial \varphi}{\partial t} + A\varphi = 0 \text{ в } D \times D_t, \\ &\varphi = \varphi^0 \text{ в } D \text{ при } t = 0, \end{aligned} \tag{5.6.30}$$

где

$$\varphi^0 = \begin{Vmatrix} u^0 \\ v^0 \\ p \end{Vmatrix}.$$

Для исследования определенности оператора A составим функционал

$$(A\varphi,\ \varphi) = -\int\limits_D \left[\frac{\partial}{\partial x}(au\varphi) + \frac{\partial}{\partial y}(au\varphi)\right] dD = -\int\limits_S au_n\varphi dS. \tag{5.6.31}$$

Здесь u_n — нормальная к границе S компонента вектора $\mathbf{u} = u\mathbf{i} + v\mathbf{j}$. Вследствие периодичности a, φ (и производных от решения) в симметричных относительно центра квадрата точках на противоположных сторонах D значения u_n будут равными по абсолютной величине и противоположными по знаку. Поэтому поверхностный интеграл в (5.6.31) обращается в нуль, и

мы приходим к условию

$$(A\varphi,\ \varphi) = 0. \tag{5.6.32}$$

Следует заметить, что если вместо условия периодичности ставится условие жесткого закрепления мембраны $(\varphi = 0)$, то и в этом случае, как видно из (5.6.31), имеет место условие (5.6.32). Это замечание, конечно, справедливо для области D произвольной формы.

Условие (5.6.32) обеспечивает единственность решения задачи. В самом деле, умножая скалярно уравнение (5.6.30) на φ и используя соотношение (5.6.32), получим

$$\frac{d}{dt}\|\varphi\|^2 = 0, \tag{5.6.33}$$

где

$$\|\varphi\| = \left[\int\limits_D (u^2 + v^2 + \varphi^2) dD\right]^{1/2}.$$

Предположим, что

$$u^0 = 0, \quad v^0 = 0, \quad \varphi^0 = 0.$$

Тогда будем иметь

$$\|\varphi^0\| = 0. \tag{5.6.34}$$

Решая уравнение (5.6.33) при условии (5.6.34), получим

$$\|\varphi\| = \|\varphi^0\| = 0.$$

Это означает, что для всех моментов времени имеет место

$$u = 0, \quad v = 0, \quad \varphi = 0,$$

что и доказывает единственность решения задачи.

Перейдем теперь к формулировке метода расщепления для задачи (5.6.30). Введем в рассмотрение матрицы

$$A_1 = \begin{Vmatrix} 0 & 0 & -a\frac{\partial}{\partial x} \\ 0 & 0 & 0 \\ -\frac{\partial}{\partial x}a & 0 & 0 \end{Vmatrix}, \quad A_1 = \begin{Vmatrix} 0 & 0 & 0 \\ 0 & 0 & -a\frac{\partial}{\partial y} \\ 0 & -\frac{\partial}{\partial y}a & 0 \end{Vmatrix}.$$

Очевидно, имеет место соотношение

$$A = A_1 + A_2. \tag{5.6.35}$$

Кроме того, аналогично предыдущему можно показать, что

$$(A_1\varphi,\ \varphi) = 0, \quad (A_2\varphi,\ \varphi) = 0. \tag{5.6.36}$$

Это означает, что задача (5.6.30) на каждом интервале $t_j \leqslant t \leqslant t_{j+1}$ может быть решена с помощью одного из рассмотренных в 4.3 методов расщепления: либо метода стабилизации, либо предиктор-корректора, либо, наконец, метода покомпонентного расщепления. Заметим, что если вместо двумерного уравнения колебаний рассматривается многомерное, то для его решения желательно использовать метод покомпонентного расщепления, который приводит к абсолютно устойчивым разностным схемам второго порядка точности при минимальных требованиях к определенности операторов вида (5.6.36).

Проведем редукцию задачи (5.6.30) на каждом интервале $t_j \leqslant t \leqslant t_{j+1}$, например, на основе метода покомпонентного расщепления. Тогда будем иметь

$$\begin{aligned} \frac{\varphi^{j+1/2} - \varphi^j}{\tau} + A_1\frac{\varphi^{j+1/2} + \varphi^j}{2} = 0, \\ \frac{\varphi^{j+1} - \varphi^{j+1/2}}{\tau} + A_2\frac{\varphi^{j+1} + \varphi^{j+1/2}}{2} = 0. \end{aligned} \tag{5.6.37}$$

В скалярной форме эти уравнения можно представить в виде

$$\begin{aligned} \frac{u^{j+1/2} - u^j}{\tau} &= a\frac{\partial}{\partial x}\left(\frac{\varphi^{j+1/2} + \varphi^j}{2}\right), \\ \frac{v^{j+1/2} - v^j}{\tau} &= 0, \\ \frac{\varphi^{j+1/2} - \varphi^j}{\tau} &= \frac{\partial}{\partial x}\left(a\frac{u^{j+1/2} + u^j}{2}\right) \end{aligned} \tag{5.6.38}$$

и

$$\begin{aligned} \frac{u^{j+1} - u^{j+1/2}}{\tau} &= 0, \\ \frac{v^{j+1} - v^{j+1/2}}{\tau} &= a\frac{\partial}{\partial y}\left(\frac{\varphi^{j+1} + \varphi^{j+1/2}}{2}\right), \\ \frac{\varphi^{j+1} - \varphi^{j+1/2}}{\tau} &= \frac{\partial}{\partial y}\left(a\frac{v^{j+1} + v^{j+1/2}}{2}\right). \end{aligned} \tag{5.6.39}$$

Учитывая, что $v^{j+1/2} = v^j$, а $u^{j+1/2} = u^{j+1}$, упростим системы уравнений (5.6.38), (5.6.39):

$$\begin{aligned} \frac{u^{j+1} - u^j}{\tau} &= a\frac{\partial}{\partial x}\left(\frac{\varphi^{j+1/2} + \varphi^j}{2}\right), \\ \frac{\varphi^{j+1/2} - \varphi^j}{\tau} &= \frac{\partial}{\partial x}\left(a\frac{u^{j+1} + u^j}{2}\right); \end{aligned} \tag{5.6.40}$$

$$\begin{aligned} \frac{v^{j+1}-v^j}{\tau} &= a\frac{\partial}{\partial y}\left(\frac{\varphi^{j+1}+\varphi^{j+1/2}}{2}\right), \\ \frac{\varphi^{j+1}-\varphi^{j+1/2}}{\tau} &= \frac{\partial}{\partial y}\left(a\frac{v^{j+1}+v^j}{2}\right). \end{aligned} \tag{5.6.41}$$

Из системы (5.6.40) находим u^{j+1} и $\varphi^{j+1/2}$, а из (5.6.41) находим v^{j+1} и φ^{j+1}.

Разностные аппроксимации по x и y вводим так, чтобы прийти к абсолютно устойчивым схемам для u, v и φ второго порядка аппроксимации и, с сохранением условий (5.6.36), в конечно-разностном представлении операторов A_1 и A_2. Положим

$$\begin{aligned} \frac{u^{j+1}_{k,\,l}-u^j_{k,\,l}}{\tau} &= \frac{a_{k,\,l}}{h}\left[\left(\frac{\varphi^{j+1/2}+\varphi^j}{2}\right)_{k,\,l} - \left(\frac{\varphi^{j+1/2}+\varphi^j}{2}\right)_{k-1,\,l}\right], \\ \frac{\varphi^{j+1/2}_{k,\,l}-\varphi^j_{k,\,l}}{\tau} &= \frac{1}{h}\left[a_{k+1,\,l}\left(\frac{u^{j+1}+u^j}{2}\right)_{k+1,\,l} - a_{k,\,l}\left(\frac{u^{j+1}+u^j}{2}\right)_{k,\,l}\right]. \end{aligned} \tag{5.6.42}$$

В первом уравнении использована разность «назад», а во втором — разность «вперед». Аналогично для системы (5.6.41) имеем

$$\begin{aligned} \frac{v^{j+1}_{k,\,l}-v^j_{k,\,l}}{\tau} &= \frac{a_{k,\,l}}{h}\left[\left(\frac{\varphi^{j+1}+\varphi^{j+1/2}}{2}\right)_{k,\,l} - \left(\frac{\varphi^{j+1}+\varphi^{j+1/2}}{2}\right)_{k,\,l-1}\right], \\ \frac{\varphi^{j+1}_{k,\,l}-\varphi^{j+1/2}_{k,\,l}}{\tau} &= \frac{1}{h}\left[a_{k,\,l+1}\left(\frac{v^{j+1}+v^j}{2}\right)_{k,\,l+1} - a_{k,\,l}\left(\frac{v^{j+1}+v^j}{2}\right)_{k,\,l}\right]. \end{aligned} \tag{5.6.43}$$

Уравнения (5.6.42) и (5.6.43) записываем для $k = 1,\ 2,\ \ldots,\ N-1$, $l = 1,\ 2,\ \ldots,\ N-1$.

Рассмотрим случай, когда на границе области D задано условие

$$\varphi = 0 \text{ на } \partial D \times D_t. \tag{5.6.44}$$

Тогда при проектировании (5.6.44) на сетку $\partial D_h \times D_\tau$ имеем

$$\varphi^j_{0,\,l} = \varphi^j_{N,\,l} = 0 \quad \text{и} \quad \varphi^j_{k,\,0} = \varphi^j_{k,\,N} = 0 \tag{5.6.45}$$

как для целых, так и для дробных индексов j.

Заметим, что соотношения (5.6.42), (5.6.43) вместе с (5.6.45) дают возможность по известным значениям $\varphi^j_{k,\,l}$, $\varphi^{j+1/2}_{k,\,l}$ и $\varphi^{j+1}_{k,\,l}$ получить $u^{j+1}_{k,\,l}$ и $v^{j+1}_{k,\,l}$ во всех узловых граничных точках.

Уравнения (5.6.42) и (5.6.43) с помощью исключения неизвестных $u^{j+1}_{k,\,l}$ и $v^{j+1}_{k,\,l}$ сведем к разностным уравнениям для величин $\varphi_{k,\,l}$. Для этого введем

в рассмотрение вспомогательные величины

$$\varphi^{j+1/4}_{k,\,l} = \frac{1}{2}(\varphi^{j+1/2}_{k,\,l} + \varphi^{j}_{k,\,l}), \qquad \varphi^{j+3/4}_{k,\,l} = \frac{1}{2}(\varphi^{j+1}_{k,\,l} + \varphi^{j+1/2}_{k,\,l}). \tag{5.6.46}$$

Тогда уравнения (5.6.42) и (5.6.43) запишутся в виде

$$\begin{aligned} &\mu^2_{k+1,\,l}(\varphi^{j+1/4}_{k+1,\,l} - \varphi^{j+1/4}_{k,\,l}) - \mu^2_{k,\,l}(\varphi^{j+1/4}_{k,\,l} - \varphi^{j+1/4}_{k-1,\,l}) - \varphi^{j+1/4}_{k,\,l} = -f^{j+1/4}_{k,\,l}, \\ &\mu^2_{k,\,l+1}(\varphi^{j+3/4}_{k,\,l+1} - \varphi^{j+3/4}_{k,\,l}) - \mu^2_{k,\,l}(\varphi^{j+3/4}_{k,\,l} - \varphi^{j+3/4}_{k,\,l-1}) - \varphi^{j+3/4}_{k,\,l} = -f^{j+3/4}_{k,\,l}, \end{aligned} \tag{5.6.47}$$

где

$$\begin{aligned} \mu_{k,\,l} &= \frac{\tau a_{k,\,l}}{2h}, \\ f^{j+1/4}_{k,\,l} &= 2\varphi^{j}_{k,\,l} + 2(\mu_{k+1,\,l}u^{j}_{k+1,\,l} - \mu_{k,\,l}u^{j}_{k,\,l}), \\ f^{j+3/4}_{k,\,l} &= 2\varphi^{j+1/2}_{k,\,l} + 2(\mu_{k,\,l+1}v^{j}_{k,\,l+1} - \mu_{k,\,l}v^{j}_{k,\,l}). \end{aligned} \tag{5.6.48}$$

Таким образом, приходим к следующему алгоритму численного решения задачи (5.6.30).

Сначала определим начальные поля функций $u^0_{k,\,l}$, $v^0_{k,\,l}$ и $\varphi^0_{k,\,l}$, причем так, чтобы $u^0_{k,\,l}$ и $v^0_{k,\,l}$ удовлетворяли разностному аналогу условия (5.6.29). Затем с помощью первой из формул (5.6.48) находим $f^{j+1/4}_{k,\,l}$ и решаем первое из разностных уравнений (5.6.47) при условии (5.6.45) на границе D_h. По найденным значениям $\varphi^{j+1/4}_{k,\,l}$ с помощью первой из формул (5.6.46) находим

$$\varphi^{j+1/2}_{k,\,l} = 2\varphi^{j+1/4}_{k,\,l} - \varphi^{j}_{k,\,l}.$$

Далее решаем второе из уравнений (5.6.47) с условием (5.6.45). После этого с помощью соотношений (5.6.46) находим

$$\varphi^{j+1}_{k,\,l} = 2\varphi^{j+3/4}_{k,\,l} - \varphi^{j+1/2}_{k,\,l}.$$

Величины $\varphi^{j+1}_{k,\,l}$ затем используются для нахождения $u^{j+1}_{k,\,l}$ и $v^{j+1}_{k,\,l}$ с помощью первых соотношений из (5.6.42) и (5.6.43). Таким образом, алгоритм решения задачи определен полностью.

В заключение следует отметить, что этот метод решения весьма просто обобщается на более сложные уравнения гиперболического типа и обычно приводит к абсолютно устойчивым схемам второго порядка аппроксимации при минимальном требовании к операторам A_α.

5.7. Методы решения многомерного уравнения движения и уравнения переноса

В этом параграфе рассматривается применение методов расщепления к одним из актуальных задач математической физики — задачам для многомерного уравнения движения и нестационарной задаче теории переноса нейтронов. Одновременно с этими методами для построения схем будут использованы подходы, изложенные в предыдущих главах (конечно-разностный, метод интегральных тождеств).

5.7.1. Двумерное уравнение движения с переменными коэффициентами

Рассмотрим на плоскости $(x,\ y)$ задачу о движении ансамбля частиц по заданным траекториям. В рамках методов механики сплошной среды приходим к следующей задаче:

$$\begin{gathered} \frac{\partial \varphi}{\partial t} + u\frac{\partial \varphi}{\partial x} + v\frac{\partial \varphi}{\partial y} = 0 \text{ в } D \times D_t, \\ \varphi(x,\ y,\ 0) = g \text{ в } D. \end{gathered} \tag{5.7.1}$$

Здесь $u = u(x,\ y,\ t)$, $v = v(x,\ y,\ t)$.

Далее предположим, что компоненты вектора скорости u, v в каждый момент времени удовлетворяют уравнению неразрывности

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0. \tag{5.7.2}$$

Пусть областью D является прямоугольник $\{0 \leqslant x \leqslant a,\ 0 \leqslant y \leqslant b\}$ и решение задачи (5.7.1) вместе с коэффициентами u и v является периодическим, принимая на противоположных сторонах прямоугольника одинаковые значения.

Эволюционное уравнение (5.7.1) запишем в операторной форме:

$$\frac{\partial \varphi}{\partial t} + A\varphi = 0, \tag{5.7.3}$$

где

$$\begin{gathered} \varphi = g \text{ при } t = 0, \\ A = u\frac{\partial}{\partial x} + v\frac{\partial}{\partial y}. \end{gathered} \tag{5.7.4}$$

Нетрудно показать, что в исследуемой задаче оператор A удовлетворяет условию $(A\varphi,\ \varphi) = 0$. В самом деле, вводя скалярное произведение в гильбертовом пространстве, запишем

$$(A\varphi,\ \varphi) = \int\limits_0^a dx \int\limits_0^b dy \left(u\frac{\partial \varphi}{\partial x} + v\frac{\partial \varphi}{\partial y} \right) \varphi. \tag{5.7.5}$$

Подынтегральное выражение с учетом (5.4.28) преобразуем к виду

$$\left(u\frac{\partial \varphi}{\partial x} + v\frac{\partial \varphi}{\partial y} \right) \varphi = \frac{\partial u\frac{\varphi^2}{2}}{\partial x} + \frac{\partial v\frac{\varphi^2}{2}}{\partial y}.$$

Поэтому

$$(A\varphi,\ \varphi) = \int\limits_0^a dx \int\limits_0^b dy \left(\frac{\partial u\frac{\varphi^2}{2}}{\partial x} + \frac{\partial v\frac{\varphi^2}{2}}{\partial y} \right). \tag{5.7.6}$$

Отсюда с помощью условия периодичности решения на границах получаем

$$(A\varphi,\ \varphi) = 0. \tag{5.7.7}$$

Попытаемся теперь расщепить оператор A таким образом, чтобы каждый из элементарных операторов $A_\alpha (\alpha = 1,\ 2)$ также удовлетворял условию

$$(A_\alpha \varphi,\ \varphi) = 0. \tag{5.7.8}$$

В этом случае разностная схема покомпонентного расщепления позволит получить абсолютно устойчивую разностную схему второго порядка аппроксимации.

Формальное разложение оператора A на составные части

$$A_1 = u\frac{\partial}{\partial x}, \quad A_2 = v\frac{\partial}{\partial y} \tag{5.7.9}$$

не удовлетворяет условию (5.7.8). Нетрудно убедиться, что имеют место соотношения

$$(A_1\varphi,\ \varphi) = -\frac{1}{2}\int\limits_0^a dx \int\limits_0^b \varphi^2 \frac{\partial u}{\partial x} dy, \quad (A_2\varphi,\ \varphi) = -\frac{1}{2}\int\limits_0^a dx \int\limits_0^b \varphi^2 \frac{\partial v}{\partial y} dy.$$

Это значит, что операторы A_1 и A_2 не могут быть взяты в качестве элементарных для построения схемы последовательного расщепления.

Выберем операторы A_1 и A_2 в более сложной форме:

$$A_1\varphi = u\frac{\partial\varphi}{\partial x} + \frac{\varphi}{2}\frac{\partial u}{\partial x}, \quad A_2\varphi = v\frac{\partial\varphi}{\partial y} + \frac{\varphi}{2}\frac{\partial u}{\partial y}. \tag{5.7.10}$$

Каждый из операторов теперь удовлетворяет требованию (5.7.8), а сумма их в точности равна A. В самом деле, имеем

$$(A_1 + A_2)\varphi = u\frac{\partial\varphi}{\partial x} + v\frac{\partial\varphi}{\partial y} + \frac{\varphi}{2}\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right) = u\frac{\partial\varphi}{\partial x} + v\frac{\partial\varphi}{\partial y} = A\varphi.$$

Здесь мы воспользовались тем, что коэффициенты u и v удовлетворяют уравнению (5.7.3).

Итак, все необходимые условия для применимости метода расщепления теперь выполнены, и мы приходим к схеме расщепления на отрезке $t_{j-1} \leqslant t \leqslant t_{j+1}$:

$$\begin{aligned}
\frac{\varphi^{j-1/2} - \varphi^{j-1}}{\tau} + \left(u^j\frac{\partial}{\partial x} + \frac{1}{2}\frac{\partial u^j}{\partial x}\right)\frac{\varphi^{j-1/2} + \varphi^{j-1}}{2} = 0,\\
\frac{\varphi^{j} - \varphi^{j-1/2}}{\tau} + \left(v^j\frac{\partial}{\partial y} + \frac{1}{2}\frac{\partial v^j}{\partial y}\right)\frac{\varphi^{j} + \varphi^{j-1/2}}{2} = 0,\\
\frac{\varphi^{j+1/2} - \varphi^{j}}{\tau} + \left(v^j\frac{\partial}{\partial y} + \frac{1}{2}\frac{\partial v^j}{\partial y}\right)\frac{\varphi^{j+1/2} + \varphi^{j}}{2} = 0,\\
\frac{\varphi^{j+1} - \varphi^{j+1/2}}{\tau} + \left(u^j\frac{\partial}{\partial x} + \frac{1}{2}\frac{\partial u^j}{\partial x}\right)\frac{\varphi^{j+1} + \varphi^{j+1/2}}{2} = 0.
\end{aligned} \tag{5.7.11}$$

Если функции u и v и решение φ обладают достаточной гладкостью по всем переменным, то схема (5.7.11) имеет второй порядок аппроксимации и будет абсолютно устойчива в том смысле, что

$$\|\varphi^{j+1}\| = \|\varphi^{j-1}\| = \cdots = \|g\|. \tag{5.7.12}$$

Это поучительный пример того, как формальное расщепление на операторы (5.7.9) может скомпрометировать саму идею расщепления, и лишь дополнительные соображения приводят к схемам, теоретически оправданным и эффективным в приложениях.

После такого предварительного рассмотрения перейдем к построению разностных схем решения задачи (5.7.1) по пространственным переменным и времени. Сначала обсудим вопрос о рациональных путях аппроксимации оператора A по пространственным переменным x и y. Удобным методом аппроксимации задач математической физики с сохранением аддитивных свойств оператора и его качественных особенностей является метод покоординатной аппроксимации, которым мы и воспользуемся для построения разностных схем.

Предположим, что коэффициенты u и v достаточно гладкие, и рассмотрим уравнения (5.7.1) в дивергентной форме

$$\begin{gathered}\frac{\partial \varphi}{\partial t}+\frac{\partial u\varphi}{\partial x}+\frac{\partial v\varphi}{\partial y}=0 \text{ в } D\times D_t,\\ \varphi=g \text{ в } D \text{ при } t=0.\end{gathered} \tag{5.7.13}$$

Для построения разностной схемы за основу возьмем оператор A, определяемый выражением

$$A\varphi=\frac{\partial u\varphi}{\partial x}+\frac{\partial v\varphi}{\partial y}-\frac{\varphi}{2}\left(\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}\right). \tag{5.7.14}$$

Разностный аналог этого соотношения рассмотрим в виде

$$\begin{gathered}(A^h\varphi)_{k,l}=\frac{u_{k+1,l}\varphi_{k+1,l}-u_{k-1,l}\varphi_{k-1,l}}{2\Delta x}+\frac{v_{k,l+1}\varphi_{k,l+1}-v_{k,l-1}\varphi_{k,l-1}}{2\Delta y}-\\ -\frac{\varphi_{k,l}}{2}\left(\frac{u_{k+1,l}-u_{k-1,l}}{2\Delta x}+\frac{v_{k,l+1}-v_{k,l-1}}{2\Delta y}\right).\end{gathered} \tag{5.7.15}$$

Очевидно, разностное выражение (5.7.15) аппроксимирует (5.7.14) со вторым порядком относительно Δx и Δy на достаточно гладких функциях u, v и φ. Однако выражение (5.7.15) имеет существенный недостаток, поскольку в такой форме оператор A^h нарушает свою кососимметрическую структуру, т. е. теперь

$$(A^h\varphi,\ \varphi)\neq 0. \tag{5.7.16}$$

Это значит, что привычная нам аппроксимация оказывается неудовлетворительной для конструирования вычислительного алгоритма решения задачи (5.7.1).

Покажем, что аппроксимация выражения (5.7.14) в форме

$$(A^h\varphi)_{k,l}=\frac{u_{k+1/2,l}\varphi_{k+1,l}-u_{k-1/2,l}\varphi_{k-1,l}}{2\Delta x}+\frac{v_{k,l+1/2}\varphi_{k,l+1}-v_{k,l-1/2}\varphi_{k,l-1}}{2\Delta y} \tag{5.7.17}$$

удовлетворяет основному соотношению

$$(A^h\varphi,\ \varphi)=0 \tag{5.7.18}$$

и аппроксимирует выражение (5.7.14) со вторым порядком по Δx и Δy. Воспользуемся следующей аппроксимацией коэффициентов:

$$\begin{aligned} u_{k+1/2,l} &= u_{k+1,l} - \frac{u_{k+1,l} - u_{k,l}}{2}; \\ v_{k,l+1/2} &= v_{k,l+1} - \frac{v_{k,l+1} - v_{k,l}}{2}; \\ u_{k-1/2,l} &= u_{k-1,l} + \frac{u_{k,l} - u_{k-1,l}}{2}; \\ v_{k,l-1/2} &= v_{k,l-1} + \frac{v_{k,l} - v_{k,l-1}}{2}. \end{aligned} \tag{5.7.19}$$

Подставив эти выражения в (5.7.17), после несложных преобразований получим

$$\begin{aligned} (A^h\varphi)_{k,l} = \frac{u_{k+1,l}\varphi_{k+1,l} - u_{k-1,l}\varphi_{k-1,l}}{2\Delta x} + \frac{v_{k,l+1}\varphi_{k,l+1} - v_{k,l-1}\varphi_{k,l-1}}{2\Delta y} - \\ - \frac{\varphi_{k,l}}{2}\left(\frac{u^l_{k+1} - u^l_{k-1}}{2\Delta x} + \frac{v_{k,l+1} - v_{k,l-1}}{2\Delta y}\right) - (\Delta x^2 R_{k,l} + \Delta y^2 Q_{k,l}), \end{aligned} \tag{5.7.20}$$

где величины $R_{k,l}$ и $Q_{k,l}$ при $\Delta x \to 0$ и $\Delta y \to 0$ стремятся к следующим:

$$R_{k,l} \to \frac{1}{4}\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}\frac{\partial \varphi}{\partial x}\right), \quad Q_{k,l} \to \frac{1}{4}\frac{\partial}{\partial y}\left(\frac{\partial v}{\partial y}\frac{\partial \varphi}{\partial y}\right).$$

Предположим теперь, что коэффициенты $u_{k,l}$, $v_{k,l}$ удовлетворяют соотношению

$$\frac{u^l_{k+1} - u^l_{k-1}}{2\Delta x} + \frac{v_{k,l+1} - v_{k,l-1}}{2\Delta y} = O(h^2), \quad h = \max(\Delta x,\ \Delta y). \tag{5.7.21}$$

Если коэффициенты u, v и решение φ имеют ограниченные производные третьего порядка по x и y, то выражение (5.4.45) при условии (5.7.21) отличается от (5.7.15) на тот же второй порядок малости, что и выражение (5.7.15) от (5.7.14). Таким образом, мы показали, что выражение (5.7.17) аппроксимирует (5.7.14) со вторым порядком относительно Δx и Δy.

Покажем, что построенный оператор A^h удовлетворяет условию (5.7.18) и, более того, каждый из операторов A^h_1 и A^h_2, определяемых

$$\begin{aligned} (A^h_1\varphi)_{k,l} &= \frac{u_{k+1/2,l}\varphi_{k+1,l} - u_{k-1,l}\varphi_{k-1/2,l}}{2\Delta x}, \\ (A^h_2\varphi)_{k,l} &= \frac{v_{k,l+1/2}\varphi_{k,l+1} - v_{k,l-1}\varphi_{k,l-1/2}}{2\Delta y}, \end{aligned} \tag{5.7.22}$$

также удовлетворяет условиям

$$(A^h_\alpha\varphi,\ \varphi) = 0. \tag{5.7.23}$$

С этой целью введем в рассмотрение скалярное произведение для векторных величин a, b:

$$(a,\ b) = \sum_k \sum_l a_{k,l} b_{k,l} \Delta x \Delta y.$$

Имеем

$$\begin{aligned}(A_1^h \varphi) &= \frac{1}{2} \sum_k \sum_l \Delta y (u_{k+1/2,l} \varphi_{k+1,l} - u_{k-1/2,l} \varphi_{k-1,l}) \varphi_{k,l}, \\ (A_2^h \varphi) &= \frac{1}{2} \sum_l \sum_k \Delta x (v_{k,l+1/2} \varphi_{k,l+1} - v_{k,l-1/2} \varphi_{k,l-1}) \varphi_{k,l}.\end{aligned} \tag{5.7.24}$$

Приведя подобные члены в (5.7.24), приходим к равенствам (5.4.48). Из условий (5.7.23) немедленно вытекает условие (5.7.18). Итак, необходимые пространственные аппроксимации проведены. Теперь задача состоит во временной редукции системы обыкновенных дифференциальных уравнений

$$\begin{gathered}\frac{d\varphi^n}{dt} + A^h \varphi^h = 0 \text{ в } D_h \times D_t, \\ \varphi^h = g^h \text{ в } D_h \text{ при } t = 0,\end{gathered} \tag{5.7.25}$$

где

$$A^h = A_1^h + A_2^h,$$

φ^h — вектор-функция с компонентами $\varphi_{k,l}$ и A_α^h удовлетворяют условию (5.7.23). Это значит, что задача (5.7.25) может быть решена с помощью метода расщепления. Опуская для удобства индекс h у функций и операторов на отрезке $t_{j-1} \leqslant t \leqslant t_{j+1}$, приходим к системе

$$\begin{aligned}\frac{\varphi^{j-1/2} - \varphi^{j-1}}{\tau} + A_1^j \frac{\varphi^{j-1/2} + \varphi^{j-1}}{2} &= 0, \\ \frac{\varphi^{j} - \varphi^{j-1/2}}{\tau} + A_2^j \frac{\varphi^{j} + \varphi^{j-1/2}}{2} &= 0, \\ \frac{\varphi^{j+1/2} - \varphi^{j}}{\tau} + A_2^j \frac{\varphi^{j+1/2} + \varphi^{j}}{2} &= 0, \\ \frac{\varphi^{j+1} - \varphi^{j+1/2}}{\tau} + A_1^j \frac{\varphi^{j+1} + \varphi^{j+1/2}}{2} &= 0.\end{aligned} \tag{5.7.26}$$

Итак, задача (5.7.1) редуцировалась к системе простейших одномерных разностных уравнений, решение которых возможно с помощью метода факторизации трехточечных разностных уравнений.

5.7.2. Многомерное уравнение движения

Рассмотрим теперь многомерное уравнение гидродинамического переноса субстанции вдоль траектории:

$$\frac{\partial \Phi}{\partial t} + \sum_{\alpha=1}^{n} v_\alpha \frac{\partial \Phi}{\partial x_\alpha} = 0 \text{ в } D \times D_t, \tag{5.7.27}$$

$$\Phi(x,\ 0) = f(x) \text{ в } D.$$

Предположим, что коэффициенты в (5.7.27) удовлетворяют уравнению непрерывности

$$\sum_{\alpha=1}^{n} \frac{\partial v_\alpha}{\partial x_\alpha} = 0 \text{ в } D \times D_t. \tag{5.7.28}$$

Задача (5.7.27) с учетом (5.7.28) может быть приведена к дивергентному виду

$$\frac{\partial \Phi}{\partial t} + \sum_{\alpha=1}^{n} \frac{\partial v_\alpha \Phi}{\partial x_\alpha} = 0 \text{ в } D \times D_t, \tag{5.7.29}$$

$$\Phi(x,\ 0) = f(x) \text{ в } D.$$

Задача (5.7.29) является основной при использовании методов расщепления.

Построим сначала разностное уравнение, соответствующее уравнению из (5.7.29), используя для этой цели схему Кранка — Николсона:

$$\frac{\Phi^{j+1} - \Phi^j}{\tau} + \sum_{\alpha=1}^{n} \frac{\partial u_\alpha^j \Phi^{j+1/2}}{\partial x_\alpha}, = 0 \quad t_j \leqslant t \leqslant t_{j+1}, \tag{5.7.30}$$

где

$$\Phi^{j+1/2} = \frac{1}{2}(\Phi^{j+1} + \Phi^j). \tag{5.7.31}$$

Кроме того, в (5.7.30) мы использовали некоторую аппроксимацию коэффициентов v_α. Естественно выбрать эту аппроксимацию либо первого, либо второго порядка по τ. Например, для получения первого порядка аппроксимации можно принять

$$u_\alpha^j = v_\alpha(x,\ t_j),$$

а для второго порядка —

$$u_\alpha^j = \frac{v_\alpha(x,\ t_{j+1}) + v_\alpha(x,\ t_j)}{2}.$$

Обозначим через

$$A^j\Phi = \sum_{\alpha=1}^{n} A_\alpha^j \Phi, \quad A_\alpha^j \Phi = \frac{\partial u_\alpha^j \Phi}{\partial x_\alpha}.$$

Тогда уравнение (5.7.30) можно записать в виде

$$\frac{\Phi^{j+1} - \Phi^j}{\tau} + A^j \Phi^{j+1/2} = 0, \quad \Phi^{j+1/2} = \frac{1}{2}(\Phi^{j+1} + \Phi^j)$$

или

$$\left(E + \frac{\tau}{2} A^j\right) \Phi^{j+1} = \left(E - \frac{\tau}{2} A^j\right) \Phi^j. \tag{5.7.32}$$

Ради простоты предположим, что решение задачи периодическое относительно n-мерного параллелепипеда D. Если ввести скалярное произведение формулой

$$(a,\ b) = \int_D ab\ dD,$$

то нетрудно проверить, что имеет место равенство

$$(A^j\Phi,\ \Phi) = 0. \tag{5.7.33}$$

Разрешим уравнение (5.7.32) относительно Φ^{j+1}. Тогда получим

$$\Phi^{j+1} = \left(E + \frac{\tau}{2} A^j\right)^{-1} \left(E - \frac{\tau}{2} A^j\right) \Phi^j.$$

Используя лемму Келлога и (5.7.33), находим, что

$$\|\Phi^{j+1}\| = \|\Phi^j\|. \tag{5.7.34}$$

Далее рассмотрим разностную аппроксимацию уравнения (5.7.30) в пространстве геометрических переменных. С этой целью область D спроектируем на D_h и запишем, опуская у u, Φ те индексы, которые не изменяются:

$$\frac{\Phi^{j+1} - \Phi^j}{\tau} + \Lambda^j \Phi^{j+1/2} = 0, \tag{5.7.35}$$

где

$$\Lambda^j = \sum_{\alpha=1}^{n} \Lambda_\alpha^j, \tag{5.7.36}$$

$$\Lambda_\alpha^j \Phi = \frac{1}{2\Delta x_\alpha}(u^j_{\alpha,k_\alpha+1/2}\Phi_{k_\alpha+1} - u^j_{\alpha,k_\alpha-1/2}\Phi_{k_\alpha-1}), \tag{5.7.37}$$

причем k_α — индекс, соответствующий номерам узловых точек переменной x_α. Введем в рассмотрение скалярное произведение

$$(a,\ b) = \sum_{k_1 \dots k_n} a_{k_1 k_2 \dots k_n} b_{k_1 k_2 \dots k_n} \Delta x_1 \Delta x_2 \dots \Delta x_n. \tag{5.7.38}$$

Нетрудно проверить, что

$$(\Lambda^j \Phi,\ \Phi) = 0. \tag{5.7.39}$$

Следовательно, используя (5.7.39), имеем

$$\|\Phi^{j+1}\| = \|\Phi^j\| = \dots = \|f\|,$$

где Φ^j — решение уравнения (5.7.35). Таким образом, при указанной разностной аппроксимации оператора A^j оператором Λ^j из (5.7.35)—(5.7.37) снова приходим к абсолютно устойчивой схеме.

Теперь исследуем аппроксимацию A^j с помощью Λ^j. В связи с этим рассмотрим сначала элементарные операторы A^j_α и Λ^j_α. В дальнейшем используем в выражении $\Lambda^j_\alpha \Phi$ в качестве коэффициентов величины

$$\begin{aligned} u^j_{\alpha, k_\alpha + 1/2} &= u^j_{\alpha, k_\alpha + 1} - \frac{1}{2}(u^j_{\alpha, k_\alpha + 1} - u^j_{\alpha, k_\alpha}), \\ u^j_{\alpha, k_\alpha - 1/2} &= u^j_{\alpha, k_\alpha - 1} + \frac{1}{2}(u^j_{\alpha, k_\alpha} - u^j_{\alpha, k_\alpha - 1}). \end{aligned} \tag{5.7.40}$$

Тогда имеем

$$\begin{aligned} \Lambda^j_\alpha \Phi = \frac{u^j_{\alpha, k_\alpha + 1} \Phi_{\alpha, k_\alpha + 1} - u^j_{\alpha, k_\alpha - 1} \Phi_{\alpha, k_\alpha - 1}}{2 \Delta x_\alpha} - \\ - \frac{\Phi_{k_\alpha}}{2} \frac{u^j_{\alpha, k_\alpha + 1} - u^j_{\alpha, k_\alpha - 1}}{2 \Delta x_\alpha} - \frac{\Delta x^2_\alpha}{4} R^j_\alpha, \end{aligned} \tag{5.7.41}$$

где

$$\begin{aligned} R^j_\alpha = \frac{1}{\Delta x^3_\alpha} [(u^j_{\alpha, k_\alpha + 1} - u^j_{\alpha, k_\alpha})(\Phi_{k_\alpha + 1} - \Phi_{k_\alpha}) - \\ - (u^j_{\alpha, k_\alpha} - u^j_{\alpha, k_\alpha - 1})(\Phi_{k_\alpha} - \Phi_{k_\alpha - 1})]. \end{aligned} \tag{5.7.42}$$

Если $\Delta x_\alpha \to 0$, то, принимая во внимание предположение о достаточной гладкости решения и коэффициентов $u(x,\ t)$, получим

$$R^j_\alpha \to \frac{\partial}{\partial x_\alpha} \left(\frac{\partial u^j_\alpha}{\partial x_\alpha} \frac{\partial \Phi}{\partial x_\alpha} \right).$$

Из (5.7.41) следует, что оператор Λ^j_α, вообще говоря, не аппроксимирует оператор A^j_α.

Рассмотрим полный оператор Λ^j и изучим выражение $\Lambda^j\Phi$. Имеем

$$\Lambda^j\Phi = \sum_{\alpha=1}^{n} \frac{u^j_{\alpha,k_\alpha+1}\Phi_{k_\alpha+1} - u^j_{\alpha,k_\alpha-1}\Phi_{k_\alpha-1}}{2\Delta x_\alpha} - \\ - \sum_{\alpha=1}^{n} \frac{\Phi_\alpha}{2} \frac{u^j_{\alpha,k_\alpha+1} - u^j_{\alpha,k_\alpha-1}}{2\Delta x_\alpha} - \frac{1}{4}\sum_{\alpha=1}^{n} \Delta x_\alpha^2 R^j_\alpha. \tag{5.7.43}$$

Предположим, что разностная запись уравнения неразрывности такова, что

$$\sum_{\alpha=1}^{n} \frac{u^j_{\alpha,k_\alpha+1} - u^j_{\alpha,k_\alpha-1}}{2\Delta x_\alpha} = O(h^2), \tag{5.7.44}$$

где

$$h = \max_\alpha\{\Delta x_\alpha\}.$$

В этом случае вторая сумма в выражении (5.7.43) обращается в нуль, и мы имеем

$$\Lambda^j\Phi = \sum_{\alpha=1}^{n} \frac{u^j_{\alpha,k_\alpha+1}\Phi_{k_\alpha+1} - u^j_{\alpha,k_\alpha-1}\Phi_{k_\alpha-1}}{2\Delta x_\alpha} + O(h^2). \tag{5.7.45}$$

Таким образом, из (5.7.45) следует, что полный оператор Λ^j аппроксимирует A с точностью до величин второго порядка малости по всем геометрическим переменным.

Переходим теперь к расщеплению задачи (5.7.29). С этой целью рассмотрим следующую двуциклическую схему:

$$\left(E + \frac{\tau}{2}\Lambda^j_\alpha\right)\Phi^{j-\frac{n-\alpha}{n}} = \left(E - \frac{\tau}{2}\Lambda^j_\alpha\right)\Phi^{j-\frac{n-\alpha+1}{n}}, \\ \alpha = 1,\ 2,\ \ldots,\ n,$$

$$\left(E + \frac{\tau}{2}\Lambda^j_\alpha\right)\Phi^{j+\frac{n-\alpha+1}{n}} = \left(E - \frac{\tau}{2}\Lambda^j_\alpha\right)\Phi^{j+\frac{n-\alpha}{n}}, \\ \alpha = n,\ n-1,\ \ldots,\ 1. \tag{5.7.46}$$

Коэффициенты в операторе Λ^j_α необходимо выбрать следующим образом: $u^j_\alpha = v_\alpha(x,\ t_j)$. Подходящий выбор обеспечивает второй порядок аппроксимации системы на каждом отрезке

$$t_{j-1} \leqslant t \leqslant t_{j+1}. \tag{5.7.47}$$

Рассмотренный метод расщепления (5.7.46) является абсолютно устойчивым. В самом деле, поскольку аппроксимация операторов Λ^j_α такова, что сохраняется условие

$$(\Lambda^j_\alpha\Phi,\ \Phi) = 0, \tag{5.7.48}$$

то нетрудно доказать с помощью леммы Келлога, что

$$\|T_\alpha^j\| = 1, \quad \alpha = 1,\ 2,\ \ldots,\ n,$$

и, следовательно,

$$\|\Phi^{j-\frac{n-\alpha}{n}}\| = \|\Phi^{j-\frac{n-\alpha+1}{n}}\|, \quad \alpha = 1,\ 2,\ \ldots,\ n,$$
$$\|\Phi^{j+\frac{n-\alpha+1}{n}}\| = \|\Phi^{j+\frac{n-\alpha}{n}}\|, \quad \alpha = n,\ n-1,\ \ldots,\ 1.$$

Исключая промежуточные значения $\|\Phi\|$, получим

$$\|\Phi^{j+1}\| = \|\Phi^{j-1}\|. \tag{5.7.49}$$

Рассмотренный метод решения задач гидродинамического движения легко обобщается на случай квазилинейных уравнений гидродинамики. Требуется лишь схему решения дополнить хорошей схемой экстраполяции коэффициентов u_α на момент времени t_j, зная их в предыдущие моменты.

Нами рассмотрена схема второго порядка аппроксимации по пространственным переменным. Однако возможно распространение алгоритма в случае более точных аппроксимаций. В самом деле, пусть

$$\Lambda_\alpha \Phi = \sum_{m=1}^{p} \beta_m \frac{\frac{1}{2}(u_{k+m} + u_k)\Phi_{k+m} - \frac{1}{2}(u_k + u_{k-m})\Phi_{k-m}}{2(m\Delta x_\alpha)}, \tag{5.7.50}$$

где β_m удовлетворяет следующей системе уравнений:

$$\sum_{m=1}^{p} \beta_m = 1, \quad \sum_{m=1}^{p} m^2\beta_m = 0, \ldots, \quad \sum_{m=1}^{p} m^{2p-2}\beta_m = 0. \tag{5.7.51}$$

Тогда, если

$$x_\alpha = x_{k_\alpha},$$

имеем

$$\sum_{\alpha=1}^{n} \Lambda_\alpha \Phi = \left(\sum_{\alpha=1}^{n} \frac{\partial u_\alpha \Phi}{\partial x_\alpha} \right)_{x_\alpha = x_{k_\alpha}} + O(h^{2p}).$$

При этом предполагается, что коэффициенты удовлетворяют следующему соотношению:

$$\sum_{\alpha=1}^{n} \sum_{m=1}^{p} \beta_m \frac{u_{\alpha,k_\alpha+m} - u_{\alpha,k_\alpha-m}}{2m\Delta x_\alpha} = O(h^{2p}). \tag{5.7.52}$$

Используя далее алгоритм расщепления (5.7.73), получаем решение задачи (5.7.29) с точностью до $O(h^{2p} + \tau^2)$.

Указанный алгоритм обобщается на случай, когда гидродинамическая среда сжимаема. В этом случае основная система уравнений имеет вид

$$\begin{gathered}\frac{\partial \rho \Phi}{\partial t}+\sum_{\alpha=1}^{n} \frac{\partial \rho u_{\alpha} \Phi}{\partial x_{\alpha}}=0, \\ \frac{\partial \rho}{\partial t}+\sum_{\alpha=1}^{n} \frac{\partial \rho u_{\alpha}}{\partial x_{\alpha}}=0.\end{gathered} \tag{5.7.53}$$

Учитывая, что ρ является функцией по существу положительной, систему (5.7.53) можно привести к виду

$$\begin{gathered}\frac{\partial \sqrt{\rho} \Phi}{\partial t}+\sum_{\alpha=1}^{n}\left(\frac{1}{2} u_{\alpha} \frac{\partial \sqrt{\rho} \Phi}{\partial x_{\alpha}}+\frac{1}{2} \frac{\partial \sqrt{\rho} u_{\alpha} \Phi}{\partial x_{\alpha}}\right)=0, \\ \frac{\partial \sqrt{\rho}}{\partial t}+\sum_{\alpha=1}^{n}\left(\frac{1}{2} u_{\alpha} \frac{\partial \sqrt{\rho}}{\partial x_{\alpha}}+\frac{1}{2} \frac{\partial \sqrt{\rho} u_{\alpha}}{\partial x_{\alpha}}\right)=0.\end{gathered} \tag{5.7.54}$$

Обозначая $\sqrt{\rho}\Phi=\psi$, аппроксимируем (5.7.54) выражением

$$\begin{gathered}\frac{\psi^{j+1}-\psi^{j}}{\tau}+\sum_{\alpha=1}^{n} \frac{(u_{\alpha})_{k_{\alpha}+1/2} \psi_{k_{\alpha}+1}^{j+1/2}-(u_{\alpha})_{k_{\alpha}-1/2} \psi_{k_{\alpha}-1}^{j+1/2}}{2 \Delta x_{\alpha}}=0, \\ \frac{\sqrt{\rho^{j+1}}-\sqrt{\psi^{j}}}{\tau}+\sum_{\alpha=1}^{n} \frac{(u_{\alpha})_{k_{\alpha}+1/2} \sqrt{\rho_{k_{\alpha}+1}^{j+1/2}}-(u_{\alpha})_{k_{\alpha}-1/2} \sqrt{\rho_{k_{\alpha}-1}^{j+1/2}}}{2 \Delta x_{\alpha}}=0,\end{gathered} \tag{5.7.55}$$

где

$$\begin{gathered}(u_{\alpha})_{k_{\alpha}+1/2}=\frac{(u_{\alpha})_{k_{\alpha}+1}+(u_{\alpha})_{k_{\alpha}}}{2}, \\ \psi^{j+1/2}=\frac{\psi^{j+1}+\psi^{j}}{2}.\end{gathered}$$

Нетрудно видеть, что (5.7.55) аппроксимирует (5.7.54) со вторым порядком точности по τ и h, кроме того, выполняются соотношения

$$\|\psi^{j+1}\|=\|\psi^{j}\|, \quad \|(\sqrt{\rho^{j+1}})\|=\|(\sqrt{\rho^{j}})\|.$$

Для решения задачи (5.7.55) применим метод, аналогичный (5.7.46).

5.7.3. Нестационарное уравнение переноса нейтронов

Рассмотрим простейшую задачу теории переноса в плоскопараллельной геометрии, которая с учетом начальных данных имеет вид

$$\frac{1}{c}\frac{\partial\varphi}{\partial t}+\mu\frac{\partial\varphi}{\partial z}+\sigma\varphi=\frac{\sigma_s}{2}\int\limits_{-1}^{1}\varphi d\mu+f, \tag{5.7.56}$$

$$\varphi=0 \text{ при } z=0,\ \mu>0, \tag{5.7.57}$$

$$\varphi=0 \text{ при } z=H,\ \mu<0,$$

$$\varphi=\varphi^0 \text{ при } t=0, \tag{5.7.58}$$

где $\varphi(z,\mu,t)$ — плотность частиц в точке z, летящих со скоростью c[1)] под углом υ к оси $0z$ в момент времени t, $\mu=\cos\upsilon$; $f(z,\mu,t)$ — заданные источники излучения; функции $\sigma(z)$, $\sigma_s(z)$ предполагаются кусочно-непрерывными, причем

$$0<\sigma_0\leqslant\sigma\leqslant\sigma_1<\infty,\ 0\leqslant\sigma_s\leqslant\sigma_s^{'}<\infty,$$
$$0<\sigma_{c0}\leqslant\sigma_c=\sigma-\sigma_s.$$

Приведем некоторые удобные для дальнейшего преобразования задачи. Решение (5.7.56) для $\mu>0$ обозначим через φ^+ и для $\mu<0$ обозначим через φ^-. Тогда уравнение переноса можно записать в виде системы двух уравнений при $\mu>0$:

$$\begin{aligned}\frac{\partial\varphi^+}{\partial t}+\mu\frac{\partial\varphi^+}{\partial z}+\sigma\varphi^+&=\frac{\sigma_s}{2}\int\limits_0^1(\varphi^++\varphi^-)d\mu^{'}+f^+,\\ \frac{\partial\varphi^-}{\partial t}-\mu\frac{\partial\varphi^-}{\partial z}+\sigma\varphi^-&=\frac{\sigma_s}{2}\int\limits_0^1(\varphi^++\varphi^-)d\mu^{'}+f^-.\end{aligned} \tag{5.7.59}$$

Граничными условиями для функции φ^+ и φ^- будут следующие:

$$\varphi^+(z,\mu,t)=0 \text{ при } z=0,$$
$$\varphi^-(z,\mu,t)=0 \text{ при } z=H. \tag{5.7.60}$$

[1)] В дальнейшем ради простоты полагаем $c\equiv1$.

Теперь сложим эти два уравнения и вычтем друг из друга. В результате приходим к двум новым уравнениям:

$$\frac{\partial u}{\partial t} + \mu\frac{\partial \upsilon}{\partial z} + \sigma u = \sigma_s \int\limits_0^1 u d\mu' + g, \tag{5.7.61}$$

$$\frac{\partial \upsilon}{\partial t} + \mu\frac{\partial u}{\partial z} + \sigma \upsilon = r, \tag{5.7.62}$$

где

$$u = \frac{1}{2}(\varphi^+ + \varphi^-), \qquad \upsilon = \frac{1}{2}(\varphi^+ - \varphi^-),$$
$$g = \frac{1}{2}(f^+ + f^-), \qquad r = \frac{1}{2}(f^+ - f^-).$$

Нетрудно убедиться, что граничные условия (5.7.60) переходят в следующие:

$$u + \upsilon = 0 \text{ при } z = 0,$$
$$u - \upsilon = 0 \text{ при } z = H, \tag{5.7.63}$$

и начальными данными будут

$$u = u^0,\ \upsilon = \upsilon^0 \text{ при } t = 0. \tag{5.7.64}$$

Задаче (5.7.61)–(5.7.64) придадим операторную форму записи. С этой целью введем в рассмотрение вектор-функции ω, ω^0, F и оператор A:

$$\omega = \left\| \begin{array}{c} u \\ \upsilon \end{array} \right\|, \qquad \omega^0 = \left\| \begin{matrix} u^0 \\ \upsilon^0 \end{matrix} \right\|, \qquad F = \left\| \begin{matrix} g \\ f \end{matrix} \right\|,$$
$$A = \left\| \begin{matrix} \sigma - \sigma_s \int\limits_0^1 d\mu' & \mu\frac{\partial}{\partial z} \\ \mu\frac{\partial}{\partial z} & \sigma \end{matrix} \right\|. \tag{5.7.65}$$

Рассмотрим в $D = [0, H] \times [0, 1]$ гильбертово пространство функции $L_2(D)$ со скалярным произведением

$$(a, b) = \sum_{i=1}^{2} \int\limits_0^1 d\mu \int\limits_0^H a^i b^i dz, \tag{5.7.66}$$

где a^i, b^i — компоненты вектор-функций a и b.

Далее, из этого пространства выделим подпространство вектор-функций Φ, на элементах которого выполняется условие

$$(A\omega, \omega) < +\infty. \tag{5.7.67}$$

Здесь скалярное произведение является функцией времени. Этот факт в дальнейшем специально оговариваться не будет. Потребуем, чтобы компоненты вектор-функций ω были в D непрерывными и имели абсолютно непрерывные первые производные $\partial\omega/\partial z$. Заметим, что гладкость функций u и v в D непосредственно следует из требования гладкости в D функций φ^+ и φ^-. Наконец, из подпространства Φ выделим множество вектор-функций, удовлетворяющих условиям (5.7.63) и имеющих абсолютно непрерывную первую производную по времени. Это подпространство обозначим Φ^0. Очевидно, Φ^0 является областью определения оператора

$$L \equiv \frac{\partial}{\partial t} + A.$$

Тогда приходим к следующей задаче:

$$\begin{gathered} \frac{\partial \omega}{\partial t} + A\omega = F \text{ в } D \times [0,T], \\ \omega = \omega^0 \text{ при } t = 0 \text{ в } D, \end{gathered} \tag{5.7.68}$$

причем

$$F(t) \in L_2\left(D \times [0,T]\right), \quad \omega^0 \in \Phi, \quad \omega(t) \in \Phi^0.$$

Не трудно проверить, что на функциях Φ^0, являющихся также областью определения оператора A, имеет место соотношение

$$(A\omega, \omega) > 0. \tag{5.7.69}$$

Известно, что оператор A — положительно определенный, т. е.

$$(A\omega, \omega) \geqslant \gamma(\omega, \omega), \tag{5.7.70}$$

где γ — положительная константа, связанная с характерным геометрическим размером области.

Переходим к разностной аппроксимации задачи (5.7.61)–(5.7.64) по пространственной переменной z. С этой целью введем две системы узловых точек: основную систему $\{z_k\}_{k=0}^{N}$
$z_0 = 0$, $z_N = H$, а также вспомогательную $\left\{z_{k+1/2}\right\}_{k=0}^{N-1}$. Точки этих двух систем взаимной чередуются, т. е. $z_{k-1/2} < z_k < z_{k+1/2}$.

Проинтегрируем первое из уравнений (5.7.61) по z в пределах $(z_0, z_{1/2})$, $(z_{k-1/2}, z_{k+1/2})(k = 1, 2, \ldots, N-1), (z_{N-1/2}, z_N)$, а второе — в пределах $(z_{k-1}, z_k)(k = 1, 2, \ldots, N)$. Тогда (5.7.61) есть

$$\frac{\partial}{\partial t}\int\limits_{z_0}^{z_{1/2}} u dz + \mu \int\limits_{z_0}^{z_{1/2}} \frac{\partial v}{\partial z} dz + \int\limits_{z_0}^{z_{1/2}} \sigma u dz = \int\limits_{z_0}^{z_{1/2}} dz \int\limits_0^1 \sigma_s u d\mu' = \int\limits_{z_0}^{z_{1/2}} g dz,$$

$$\frac{\partial}{\partial t}\int\limits_{z_0}^{z_1} v dz + \mu \int\limits_{z_0}^{z_1} \frac{\partial u}{\partial z} dz + \int\limits_{z_0}^{z_1} \sigma v dz = \int\limits_{z_0}^{z_1} r dz,$$

. . . . . .

$$\frac{\partial}{\partial t}\int\limits_{z_{k-1/2}}^{z_{k+1/2}} u dz + \mu \int\limits_{z_{k-1/2}}^{z_{k+1/2}} \frac{\partial v}{\partial z} dz + \int\limits_{z_{k-1/2}}^{z_{k+1/2}} \sigma u dz = \int\limits_{z_{k-1/2}}^{z_{k+1/2}} dz \int\limits_{z_0}^{z_1} \sigma_s u d\mu' + \int\limits_{z_{k-1/2}}^{z_{k+1/2}} g dz, \qquad (5.7.71)$$

$$\frac{\partial}{\partial t}\int\limits_{z_k}^{z_{k+1}} v dz + \mu \int\limits_{z_k}^{z_{k+1}} \frac{\partial u}{\partial z} dz + \int\limits_{z_k}^{z_{k+1}} \sigma v dz = \int\limits_{z_k}^{z_{k+1}} r dz,$$

. . . . . .

$$\frac{\partial}{\partial t}\int\limits_{z_{N-1}}^{z_N} v dz + \mu \int\limits_{z_{N-1}}^{z_N} \frac{\partial v}{\partial z} dz + \int\limits_{z_{N-1}}^{z_N} \sigma v dz = \int\limits_{z_{N-1}}^{z_N} r dz,$$

$$\frac{\partial}{\partial t}\int\limits_{z_{N-1/2}}^{z_N} u dz + \mu \int\limits_{z_{N-1/2}}^{z_N} \frac{\partial v}{\partial z} dz + \int\limits_{z_{N-1/2}}^{z_N} \sigma u dz = \int\limits_{z_{N-1/2}}^{z_N} dz \int\limits_0^1 \sigma_s u d\mu' + \int\limits_{z_{N-1/2}}^{z_N} g dz.$$

Введем следующие обозначения:

$$\Delta z_0 = z_{1/2} - z_0, \quad \Delta z_k = z_{k+1/2} - z_{k-1/2}, \quad k = 1, 2, \ldots, N-1,$$
$$\Delta z_N = z_N - z_{N-1/2}, \quad \Delta z_{k-1/2} = z_k - z_{k-1}, \quad k = 1, 2, \ldots, N,$$

$$h = \max(\Delta z_0, \Delta z_N, \Delta z_k, \Delta z_{k+1/2}), \qquad (5.7.72)$$

$$\sigma_k = \frac{1}{\Delta z_k}\int\limits_{z_{k-1/2}}^{z_{k+1/2}} \sigma dz,$$

$$\sigma_{sk} = \frac{1}{\Delta z_k}\int\limits_{z_{k-1/2}}^{z_{k+1/2}} \sigma_s dz, \qquad \sigma_{k+1/2} = \frac{1}{\Delta z_{k+1/2}}\int\limits_{z_k}^{z_{k+1}} \sigma dz, \qquad (5.7.73)$$

$$g_k = \frac{1}{\Delta z_k}\int\limits_{z_{k-1/2}}^{z_{k+1/2}} g dz, \qquad r_{k+1/2} = \frac{1}{\Delta z_{k+1/2}}\int\limits_{z_k}^{z_{k+1}} r dz.$$

Тогда при условии непрерывности функций u и v почти для всех значений z и μ из D и кусочной непрерывности функций σ, σ_s, g и r с возможными разрывами первого рода в точках z_k с помощью методов, изложенных в 2.5, и с учетом граничных условий приходим к следующей разностной аппроксимации уравнений (5.7.61):

$$\frac{\partial u_0}{\partial t} + \mu\frac{v_{1/2} + u_0}{\Delta z_0} + \sigma_0 u_0 = \sigma_{s0}\int\limits_0^1 u_0 d\mu' + g_0,$$

$$\frac{\partial v_{1/2}}{\partial t} + \mu\frac{u_1 - u_0}{\Delta z_{1/2}} + \sigma_{1/2} v_{1/2} = r_{1/2},$$

. . . . . .

$$\begin{gathered}\frac{\partial u_k}{\partial t} + \mu\frac{v_{k+1/2} - v_{k-1/2}}{\Delta z_k} + \sigma_k u_k = \sigma_{sk}\int\limits_0^1 u_k d\mu' + g_k,\\ \frac{\partial v_{k+1/2}}{\partial t} + \mu\frac{u_{k+1} - u_k}{\Delta z_{k+1/2}} + \sigma_{k+1/2} v_{k+1/2} = r_{k+1/2},\end{gathered} \tag{5.7.74}$$

. . . . . .

$$\frac{\partial v_{N-1/2}}{\partial t} + \mu\frac{u_N - u_{N-1}}{\Delta z_{N-1/2}} + \sigma_{N-1/2} v_{N-1/2} = r_{N-1/2},$$

$$\frac{\partial u_N}{\partial t} + \mu\frac{u_N - v_{N-1/2}}{\Delta z_N} + \sigma_N u_N = \sigma_{sN}\int\limits_0^1 u_N d\mu' + g_N.$$

Пусть $M_h(0, 2N)$ есть гильбертово пространство вектор-функций $a = (a_0, a_{1/2}, \ldots, a_{N-1/2}, a_N)$ со скалярным произведением и нормой

$$(a, b) = \sum_{i=0}^{2N}\int\limits_0^1 \Delta z_{i/2} a_{i/2} b_{i/2} d\mu,$$

$$\|a\| = (a, a)^{1/2}, \quad a, b \in M_h(0, 2N).$$

Введем в рассмотрение вектор-функции

$$\varphi = (u_0, v_{1/2}, u_1, \ldots, u_{N-1}, v_{N-1/2}, u_N),$$

$$F = (g_0, r_{1/2}, g_1, \ldots, g_{N-1}, r_{N-1/2}, g_N),$$

$$\varphi^{(0)} = (u_0^{(0)}, v_{1/2}^{(0)}, u_1^{(0)}, \ldots, u_{N-1}^{(0)}, v_{N-1/2}^{(0)}, u_N^{(0)})$$

и матричный оператор $A = L - S$, где L и S имеют вид

$$L = \begin{Vmatrix} \frac{\mu}{\Delta z_0} + \sigma_0 & \frac{\mu}{\Delta z_0} & 0 & \dots & 0 & 0 \\ \frac{-\mu}{\Delta z_{1/2}} & \sigma_{1/2} & \frac{\mu}{\Delta z_{1/2}} & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \sigma_{N-1/2} & \frac{\mu}{\Delta z_{N-1/2}} \\ 0 & 0 & 0 & \dots & \frac{-\mu}{\Delta z_N} & \frac{\mu}{\Delta z_N} + \sigma_N \end{Vmatrix},$$

$$S = \operatorname{diag}\left(\sigma_{s,i/2} \int_0^1 \gamma_{i/2} d\mu \right), \qquad i = 0, 1, \dots, 2N,$$

$$\gamma_{i/2} = \begin{cases} 1, & \text{если } i/2 \text{ — целое число или нуль,} \\ 0, & \text{если } i/2 \text{ — дробное.} \end{cases}$$

С введением указанных выше определений систему (5.7.74) можно записать в операторной форме:

$$\frac{d\varphi}{dt} + A\varphi = F, \qquad t \in [0, T], \tag{5.7.75}$$
$$\varphi = \varphi^{(0)} \text{ при } t = 0.$$

Рассмотрим некоторые свойства операторов задачи (5.7.75). Прежде всего заметим, что оператор S является в $M_h(0, 2N)$ самосопряженным и положительным в силу сделанных предположений относительно исходных данных задачи. Покажем теперь, что операторы A и L положительно определены на M_h. Действительно, пусть $\omega \in M_h$; тогда, рассматривая квадратичную форму $(L\omega, \omega)$, имеем

$$(L\omega, \omega) = \int_0^1 \left(\mu \frac{\omega_{1/2} + \omega_0}{\Delta z_0} + \sigma_0 \omega_0 \right) \Delta z_0 \omega_0 d\mu + \int_0^1 \left(\mu \frac{\omega_1 - \omega_0}{\Delta z_{1/2}} + \sigma_{1/2} \omega_{1/2} \right) \Delta z_{1/2} \omega_{1/2} d\mu +$$
$$+ \sum_{i=1}^{N-1} \int_0^1 \left\{ \Delta z_i \left(\mu \frac{\omega_{i+1/2} - \omega_{i-1/2}}{\Delta z_i} + \sigma_i \omega_i \right) \omega_i + \Delta z_{i+1/2} \left(\mu \frac{\omega_{i+1} - \omega_i}{\Delta z_{i+1/2}} + \sigma_{i+1/2} \omega_{i+1/2} \right) \omega_{i+1/2} \right\} d\mu +$$
$$+ \int_0^1 \Delta z_N \left(\mu \frac{\omega_N - \omega_{N-1/2}}{\Delta z_N} + \sigma_N \omega_N \right) \omega_N d\mu = \sum_{i=0}^{2N} \int_0^1 \Delta z_{i/2} \sigma_{i/2} \omega_{i/2}^2 d\mu + \int_0^1 \mu \left(\omega_0^2 + \omega_N^2 \right) d\mu, \tag{5.7.76}$$

$$(L\omega, \omega) \geqslant \gamma \|\omega\|^2, \qquad \gamma = \text{const} > 0,$$

а также

$$(A\omega,\omega)=\int\limits_0^1\mu(\omega_0^2+\omega_N^2)d\mu+\sum_{i=0}^{2N}\Delta z_{i/2}\int\limits_0^1\left(\sigma_{i/2}\omega_{i/2}^2-\sigma_{si/2}\omega_{i/2}\int\limits_0^1\gamma_{i/2}\omega_{i/2}d\mu\right)d\mu\geqslant\sigma_{c0}\|\omega\|^2.$$

Это и доказывает наши утверждения.

Получим теперь две априорные оценки для решений задачи (5.7.75). Умножим скалярно уравнение (5.7.75) на функцию φ; полученное соотношение проинтегрируем по области $(0,t)$ и получим

$$\begin{gathered}\frac{1}{2}\|\varphi\|^2(t)+\int\limits_0^t(A\varphi,\varphi)dt=\int\limits_0^t(F,\varphi)dt^1+\frac{1}{2}\|\varphi^{(0)}\|^2\leqslant\\ \leqslant\left(\int\limits_0^t\|F\|^2dt^1\right)^{1/2}\left(\int\limits_0^t\|\varphi\|^2dt^1\right)^{1/2}+\frac{1}{2}\|\varphi^{(0)}\|^2\leqslant\\ \leqslant C\left(\int\limits_0^t\|F\|^2dt^1\right)^{1/2}\left(\frac{1}{2}\|\varphi\|^2(t)+\int\limits_0^t(A\varphi,\varphi)dt^1\right)^{1/2}+\frac{1}{2}\|\varphi^{(0)}\|^2,\qquad(5.7.77)\\ C=\text{const}>0.\end{gathered}$$

Воспользовавшись соотношением $|a\cdot b|\leqslant\frac{1}{4\varepsilon}a^2+\varepsilon b^2\quad(\varepsilon>0)$ при соответствующем выборе ε, из последнего неравенства получим априорную оценку для решения φ:

$$\|\varphi\|^2(t)+\int\limits_0^t(A\varphi,\varphi)dt^1\leqslant C\left(\int\limits_0^T\|F\|^2dt^1+\|\varphi^{(0)}\|^2\right),\qquad(5.7.78)$$

где константа $C>0$ и не зависит от t и φ.

С помощью (5.7.78) можно доказать однозначную разрешимость задачи (5.7.75), при этом гладкость по t решения φ будет на порядок выше соответствующей гладкости вектор-функции F. Однако мы не будем здесь останавливаться на этом вопросе, а априори предположим, что (5.7.75) имеет единственное решение φ, обладающее отмеченной выше гладкость по t.

Представим теперь оператор A в виде $A=B+D$, где

$$D=\left\|\begin{array}{cccccc}\frac{\mu}{\Delta z_0}+\sigma_0 & 0 & 0 & \dots & 0 & 0\\ 0 & \sigma_{1/2} & 0 & \dots & 0 & 0\\ \dots & \dots & \dots & \dots & \dots & \dots\\ 0 & 0 & 0 & \dots & \sigma_{N-1/2} & 0\\ 0 & 0 & 0 & \dots & 0 & \frac{\mu}{\Delta z_N}+\sigma_N\end{array}\right\|.$$

Очевидно, что D — положительно определенный оператор в $M_h(0,2N)$. Поэтому уравнение из (5.7.75) эквивалентно следующему:

$$\frac{d}{dt}(D^{-1}\upsilon) + D^{-1/2}AD^{-1/2}\upsilon = D^{-1/2}F, \tag{5.7.79}$$

где $\upsilon = D^{1/2}\varphi$. Умножим (5.7.79) скалярно на υ и результат проинтегрируем по t в интервале $(0,t_1)$. Тогда приходим к соотношению

$$\frac{1}{2}\|\varphi\|^2(t_1) + \int\limits_0^{t_1}(D^{-1/2}AD^{-1/2}\upsilon,\upsilon)dt = \int\limits_0^{t_1}(D^{-1/2}F,\upsilon)dt + \frac{1}{2}\|\varphi^0\|^2,$$

из которого после простых оценок получаем неравенство

$$\begin{aligned}\frac{1}{2}\|\varphi\|^2(t_1) + \int\limits_0^{t_1}(D^{-1/2}\widetilde{D}D^{-1/2}\upsilon,\upsilon)dt \leqslant \\ \leqslant \frac{1}{2}\|\varphi\|^2(t_1) + \int\limits_0^{t_1}(D^{-1/2}AD^{-1/2}\upsilon,\upsilon)dt \leqslant \\ \leqslant \left(\int\limits_0^{t_1}\|D^{-1/2}F\|^2dt\right)^{1/2}\left(\int\limits_0^{t_1}\|\upsilon\|^2dt\right)^{1/2} + \frac{1}{2}\|\varphi^{(0)}\|^2,\end{aligned} \tag{5.7.80}$$

где

$$\widetilde{D} = \left\|\begin{array}{ccccc} \frac{\mu}{\Delta z_0}+\sigma_{c,0} & 0 & \dots & 0 & 0 \\ 0 & \sigma_{c,1/2} & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_{c,N-1/2} & 0 \\ 0 & 0 & \dots & 0 & \frac{\mu}{\Delta z_N}+\sigma_{c,N} \end{array}\right\|.$$

Так как

$$\int\limits_0^{t_1}(D^{-1/2}\widetilde{D}D^{-1/2}\upsilon,\upsilon)dt \geqslant$$

$$\geqslant \int\limits_0^{t_1}dt\int\limits_0^1\left(\Delta z_0\frac{\mu+\Delta z_0\sigma_{c0}}{\mu+\Delta z_0\sigma_0}\upsilon_0^2 + \Delta z_{1/2}\frac{\sigma_{c,1/2}}{\sigma_{1/2}}\upsilon_{1/2}^2 + \dots + \Delta z_N\frac{\mu+\Delta z_N\sigma_{c,N}}{\mu+\Delta z_N\sigma_N}\upsilon_N^2\right)d\mu \geqslant$$

$$\geqslant C_1\int\limits_0^{t_1}\|\upsilon\|^2dt,$$

где постоянная $C_1 > 0$ и не зависит от v и $\Delta z_{i/2}$, то, используя последнее неравенство из (5.7.80), получаем вторую априорную оценку для φ:

$$\frac{1}{2}\|\varphi(t_1)\|^2 + C_1 \int\limits_0^{t_1} \|v\|^2 dt \leqslant \left(\int\limits_0^{t_1} \|D^{-1/2}F\|^2 dt\right)^{1/2} \left(\int\limits_0^{t_1} \|v\|^2 dt\right)^{1/2} + \frac{1}{2}\|\varphi^{(0)}\|^2, \quad (5.7.81)$$

$$\|\varphi(t_1)\|^2 + \int\limits_0^{t_1} \|D^{1/2}\varphi\|^2 dt \leqslant C\left(\int\limits_0^{t_1} \|D^{-1/2}F\|^2 dt + \|\varphi^{(0)}\|^2\right),$$

где константа $C > 0$ и не зависит от φ и $\Delta z_{i/2}$.

Воспользуемся теперь неравенствами (5.7.81) для оценки погрешности, с которой решение задачи (5.7.75) приближает вектор-функцию $\varphi_T = (u(z_0, \mu, t), v(z_{1/2}, \mu, t), \ldots, v(z_{N-1/2}, \mu, t),\ u(z_N, \mu, t))$, составленную по значениям точного решения (5.7.61), предполагая, что решение и исходные данные задачи (5.7.61)–(5.7.64) достаточно гладкие по всем своим переменным, шаг сетки выбран равномерным, т. е. $h = H/N, \quad z_i = ih, \quad z_{i-1/2} = (i - 1/2)h$, и h достаточно мало. Оценки погрешности, которые получаются на этом сравнительно узком классе функций, позволяют нам надеяться, что рассматриваемый здесь метод аппроксимации будет эффективным при решении ряда задач.

Отметим, что ошибки аппроксимации $\varepsilon_i (i = 0, N)$ в первом и последнем уравнениях в (5.7.74) имеют первый порядок относительно h, а остальные $\varepsilon_{i/2}$ — порядка $O(h^2)$. Следовательно, заменяя φ в (5.7.81) на $\varphi_T - \varphi$ и F на

$$\varepsilon = (\varepsilon_0, \varepsilon_{1/2}, \ldots, \varepsilon_{N-1/2}, \varepsilon_N),$$

получаем искомые оценки:

$$\begin{aligned}
&\|\varphi_T - \varphi\|^2(t_1) + \int\limits_0^{t_1} \|D^{1/2}(\varphi_T - \varphi)\|^2 dt \leqslant C \int\limits_0^{t_1} \|D^{-1/2}\varepsilon\|^2 dt \leqslant \\
&\leqslant C \int\limits_0^T dt \int\limits_0^1 \left[\frac{\varepsilon_0^2 h^2}{2(2\mu + h\sigma_0)} + \frac{\varepsilon_{1/2}^2 h}{\sigma_{1/2}} + \cdots + \frac{\varepsilon_{N-1/2}^2 h}{\sigma_{N-1/2}} + \frac{\varepsilon_N^2 h^2}{2(2\mu + h\sigma_N)}\right] d\mu \leqslant \\
&\leqslant O(h^4) + O(h^4) \int\limits_0^1 \frac{d\mu}{(2\mu + h\sigma_0)} + O(h^4) \int\limits_0^1 \frac{d\mu}{2\mu + h\sigma_N} = O\left(h^4 \ln \frac{1}{h}\right), \\
&\max_t \|\varphi_T - \varphi\|(t) + \left(\int\limits_0^T \|D^{1/2}(\varphi_T - \varphi)\|^2 dt\right)^{1/2} = O\left(h^2 \ln^{1/2} \frac{1}{h}\right).
\end{aligned} \quad (5.7.82)$$

Неравенство (5.7.81) может быть использовано для получения оценок погрешности решений (5.7.75) и при более слабых ограничениях относительно гладкости исходных данных и решения задачи.

Переходим к формулировке метода расщепления для решения задачи (5.7.75). Для этого введем операторы $A_1,\ A_2$, действующие в $M_h(0,2N)$:

$$A_1 = \left\| \begin{array}{cccccc} \frac{\mu}{\Delta z_0} & \frac{\mu}{\Delta z_0} & 0 & 0 & 0 & 0 \\ -\frac{\mu}{\Delta z_{1/2}} & 0 & \frac{\mu}{\Delta z_{1/2}} & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0\ \dots & -\frac{\mu}{\Delta z_{N-1/2}} & 0 & \frac{\mu}{\Delta z_{N-1/2}} \\ 0 & 0 & 0\ \dots & 0 & -\frac{\mu}{\Delta z_N} & \frac{\mu}{\Delta z_N} \end{array} \right\|,$$

$$A_2 = \operatorname{diag}\left(\sigma_{i/2} - \sigma_{s,i/2} \int\limits_0^1 d\mu' \gamma_{i/2} \right),$$

при этом

$$A = A_1 + A_2.$$

Легко показать, что оператор A_1 положительно полуопределен, а используя соотношение

$$\int\limits_0^1 \left(\sigma_{i/2} a_{i/2}^2 - \sigma_{s,i/2} a_{i/2} \int\limits_0^1 \gamma_{i/2} a_{i/2} d\mu^1 \right) d\mu \geqslant \sigma_{c,0} \int\limits_0^1 a_{i/2}^2 d\mu,$$

доказать также, что и A_2 положительно определен в $M_h(0,2N)$, т. е.

$$(A_1 a, a) \geqslant 0, \quad (A_2 a, a) \geqslant \gamma \|a\|^2, \quad \gamma = \operatorname{const} > 0.$$

Условия определенности позволяют сформулировать алгоритм решения на основе двуциклического покомпонентного метода расщепления. Сформулируем следующие задачи.

На отрезке $t_{j-1} \leqslant t \leqslant t_j$

$$\begin{aligned} \left(E + \frac{\tau}{2} A_1\right) \varphi^{j-2/3} &= \left(E - \frac{\tau}{2} A_1\right) \varphi^{j-1}, \\ \left(E + \frac{\tau}{2} A_2\right) \varphi^{j-1/3} &= \left(E - \frac{\tau}{2} A_2\right) \varphi^{j-2/3}, \end{aligned} \tag{5.7.83}$$

на отрезке $t_{j-1} \leqslant t \leqslant t_{j+1}$

$$\phi^{j+1/3} = \phi^{j-1/3} + 2\tau F^j \tag{5.7.84}$$

и на отрезке $t_j \leqslant t \leqslant t_{j+1}$

$$\begin{aligned} \left(E + \frac{\tau}{2} A_2\right) \varphi^{j+2/3} &= \left(E - \frac{\tau}{2} A_2\right) \varphi^{j-1/3}, \\ \left(E + \frac{\tau}{2} A_1\right) \varphi^{j+1} &= \left(E - \frac{\tau}{2} A_1\right) \varphi^{j+2/3}, \end{aligned} \tag{5.7.85}$$

где F^j есть вектор с компонентами

$$F_i^j = \frac{1}{t_{j+1} - t_{j-1}} \int\limits_{t_{j-1}}^{t_{j+1}} F_i dt, \quad t_j = j\tau, \quad \tau = \frac{T}{y}.$$

Из свойств $A_1,\ A_2$ следует, что схема (5.7.83)–(5.7.85) (при необходимой гладкости решения) аппроксимирует (5.5.19) с точностью до величины порядка $O(\tau^2)$ и устойчива:

$$\max_j \|\varphi^{(j)}\| \leqslant C(\|\varphi^{(0)}\| + \max_j \|F^{(j)}\|), \tag{5.7.86}$$

где $C = \text{const} > 0$. Из аппроксимации и неравенства (5.7.86) также следует, что

$$\max_j \|\varphi(t_j) - \varphi^{(j)}\| \leqslant O(\tau^2). \tag{5.7.87}$$

Если справедлива оценка (5.7.82), то решения задачи (5.7.83)–(5.7.85) сходятся в метрике пространства $M_h(0, 2N)$ к значениям точного решения исходной задачи при $h,\ \tau \to 0$, причем

$$\max_j \|\varphi_T(t_j) - \varphi^{(j)}\| \leqslant O\left\{h^2 \left|\ln^{1/2} \frac{1}{h}\right| + \tau^2\right\}, \tag{5.7.88}$$

где

$$\varphi_T(t_j) = \left(u(z_0, \mu, t_j), v(z_{1/2}, \mu, t_j), \ldots, v(z_{N-1/2}, \mu, t_j), u(z_N, \mu, t_j)\right).$$

Остановимся на решении системы (5.7.83)–(5.7.85). Так как матрица

$$\left(E + \frac{\tau}{2} A_1\right)$$

является трехдиагональной, то решение первого уравнения из (5.7.83) и второго из (5.7.85) не представляет труда при любом фиксированном μ. Матричный оператор

$$\left(E + \frac{\tau}{2} A_2\right)$$

диагональный, так что решение второго уравнения из (5.7.83) сводится к вычислениям по формулам

$$\varphi_i^{j-1/3} = \frac{1}{1+\dfrac{\tau\sigma_i}{2}}\left[\left(1-\frac{\tau\sigma_i}{2}\right)\varphi_i^{j-2/3} + \frac{\tau\sigma_{s,i}}{1+\dfrac{\tau\sigma_{c,i}}{2}}\int\limits_0^1 \varphi_i^{j-2/3}d\mu\right],$$
$$i = 0, 1, \ldots, N, \tag{5.7.89}$$
$$\varphi_{i-1/2}^{j-1/3} = \frac{1-\dfrac{\tau\sigma_{i-1/2}}{2}}{1+\dfrac{\tau\sigma_{i-1/2}}{2}}\varphi_{i-1/2}^{j-2/3}, \qquad i = 1, 2, \ldots, N.$$

Аналогичные формулы выписываются и при решении первого уравнения из (5.7.85). Итак, численный алгоритм решения системы (5.7.83)–(5.7.85) определен.

Выполним теперь разностную аппроксимацию уравнений по μ. Для этого интервал $0 \leqslant \mu \leqslant 1$ так разобьем на частичные интервалы $\Delta\mu_i$ узловыми точками μ_i, чтобы обеспечить на заданном классе решений наилучшую аппроксимацию интервалов в (5.7.89). Пусть

$$\int\limits_0^1 \psi(\mu)d\mu \cong \sum_{l=1}^{m} s_l\psi_l, \qquad \psi_l = \psi(\mu_l),$$

где s_i — веса выбранной квадратичной формулы. Заменяя интегралы в (5.7.83)–(5.7.85) квадратурными формулами и рассматривая систему при $\mu = \mu_l (l = 1, 2, \ldots, m)$, приходим к системе линейных алгебраических уравнений, аппроксимирующей исходную задачу:

$$\begin{gathered}
\left(E+\frac{\tau}{2}A_{1,l}\right)\varphi_l^{j-2/3} = \left(E-\frac{\tau}{2}A_{1,l}\right)\varphi_l^{j-1},\\
\left(E+\frac{\tau}{2}A_{2,l}\right)\varphi_l^{j-1/3} = \left(E-\frac{\tau}{2}A_{2,l}\right)\varphi_l^{j-2/3},\\
\varphi_l^{j+1/3} = \varphi_l^{j-1/3} + 2\tau F_l^j,\\
\left(E+\frac{\tau}{2}A_{2,l}\right)\varphi_l^{j+2/3} = \left(E-\frac{\tau}{2}A_{2,l}\right)\varphi_l^{j+1/3},\\
\left(E+\frac{\tau}{2}A_{1,l}\right)\varphi_l^{j+1} = \left(E-\frac{\tau}{2}A_{1,l}\right)\varphi_l^{j+2/3},\\
\varphi_l^0 = \varphi_l^{(0)},
\end{gathered} \tag{5.7.90}$$

где $\varphi_l^{j-1},\dots,\varphi_l^{j+1}$ — векторы размерности $(2N+1), l=1,2,\dots,m$,

$$\varphi_l^{(0)} = \varphi^{(0)}(\mu_l),$$
$$A_{1,l} = A_1(\mu_l),$$

а оператор $A_{2,l}$ действует по формуле

$$(A_{2,l}\varphi_l^j)_{i/2} = \sigma_{i/2}\varphi_{l,i/2}^j - \sigma_{s,i/2}\sum_{k=1}^{m} s_k\varphi_{k,i/2}^j, \qquad i=0,1,\dots,2N.$$

Как отмечалось, первое и последнее уравнения в (5.7.90) легко могут быть решены, например с помощью метода прогонки; решение же второго и четвертого уравнений (5.7.90) реализуются по формулам

$$\varphi_{l,i}^{j-1/3} = \frac{1}{1+\dfrac{\tau\sigma_i}{2}}\left[\left(1-\frac{\tau\sigma_i}{2}\right)\varphi_{l,i}^{j-2/3} + \frac{\tau\sigma_{si}}{1+\dfrac{\tau\sigma_{ci}}{2}}\sum_{k=1}^{m} s_k\varphi_{k,i}^{j-2/3}\right],$$
$$i=0,1,\dots,N, \tag{5.7.91}$$
$$\varphi_{l,i-1/2}^{j-1/3} = \frac{1-\dfrac{\tau\sigma_{i-1/2}}{2}}{1+\dfrac{\tau\sigma_{i-1/2}}{2}}\varphi_{l,i-1/2}^{j-2/3}, \qquad i=1,2,\dots,N,$$

для второго уравнения и по тем же формулам (5.7.91) с заменой j на $j+1$ — для четвертого уравнения.

Таким образом, алгоритм определен полностью.

В результате приходим к абсолютно устойчивой схеме порядка точности

$$O(h^2(\ln 1/h)^{1/2} + \tau^2)$$

на гладких решениях. Порядок точности по μ зависит от выбираемой квадратурной формулы. (Отметим, что в рассматриваемой схеме в качестве узловой точки допускается также значение $\mu_l = 0$.)

До сих пор никаких ограничений на выбор параметров схемы не накладывалось. Однако заметим, что если для решения систем в (5.7.90) применяется метод факторизации, то для его устойчивости достаточно потребовать, чтобы $\tau < \min(\Delta z_{i/2})$. При решении практических задач это условие накладывает ограничения на выбор временного шага.

Аналогичным образом могут быть рассмотрены и многомерные уравнения переноса.

5.8. Асимптотический анализ и распараллеливание алгоритмов решения простейшего уравнения диффузии[4)]

В § 4.8 мы рассмотрели асимптотический анализ алгоритмов решения стационарных задач в рамках идеализированной вычислительной системы. В данном параграфе такой анализ мы проведем на примере простейшей модельной нестационарной задачи — смешанной задачи для уравнения диффузии в виде

$$\frac{\partial\varphi}{\partial t} - \frac{\partial}{\partial x}\left(a_1\frac{\partial\varphi}{\partial x}\right) - \frac{\partial}{\partial y}\left(a_2\frac{\partial\varphi}{\partial y}\right) + q\psi = f \text{ в } (0,\ T]\times D,$$

$$\varphi|_{\partial D} = 0, \quad \varphi|_{t=0} = g, \tag{5.8.1}$$

где a_i, q, f, g — заданные достаточно гладкие функции $0 \leqslant q_0 \leqslant q \leqslant q_1$, $0 < \alpha_1 \leqslant a_i \leqslant q_2$, $q_i, \alpha_1 = \text{const} > 0$.

Для численного решения задачи (5.8.1) существует много различных разностных схем. Мы остановимся на трех простейших, наиболее известных схемах. Это — явная и неявная схемы, схема Кранка — Николсона.

Пусть

$$x_i = ih, \quad y_j = jh, \quad i,\ j = 0,\ 1,\ \ldots,\ n+1, \quad h = 1/(n+1),$$
$$N = n^2,$$
$$t_k = k\tau, \quad k = 0,\ 1,\ \ldots,\ M, \quad \tau = T/M.$$

Явная схема для решения задачи (5.8.1) первого порядка аппроксимации по t имеет вид

$$\frac{\varphi^{k+1} - \varphi^k}{\tau} + A\varphi^k = f^k, \quad k = 0,\ 1,\ \ldots,\ M-1, \quad \varphi^0 = \bar{g}, \tag{5.8.2}$$

где A — матрица порядка N, определенная в (4.8.2), f^k и $\bar{g}$ — векторы с компонентами $\{f_{ml}(t_k)\}$, $\{g_{ml}\}$, $m,\ l = 1,\ 2,\ \ldots,\ n$.

Разностная схема (5.8.2) аппроксимирует исходную задачу (5.8.1) с порядком аппроксимации $O(\tau + h^2)$ и условно устойчива при $\tau \leqslant h^2/4$. Имеет место сходимость приближенного решения $\{\varphi^k\}$ к точному решению φ задачи (5.8.1) со скоростью $O(\tau + h^2)$.

[4)] Изложение данного раздела построено на основе работы В. П. Шутяева [4].

Неявная схема для решения задачи (5.8.1) первого порядка аппроксимации по t имеет вид

$$\frac{\varphi^{k+1}-\varphi^k}{\tau}+A\varphi^k=f^{k+1},\quad k=0,\ 1,\ \ldots,\ M-1,\quad \varphi^0=\bar{g}. \tag{5.8.3}$$

Разностная схема (5.8.3) аппроксимирует исходную задачу (5.8.1) с порядком аппроксимации $O(\tau+h^2)$ и абсолютно устойчива. Имеет место сходимость приближенного решения $\{\varphi^k\}$, полученного по схеме (5.8.3), к точному решению φ задачи (5.8.1) со скоростью $O(\tau+h^2)$.

Схема Кранка — Николсона для решения задачи (5.8.1) имеет вид

$$\begin{gathered}\frac{\varphi^{k+1}-\varphi^k}{\tau}+A\frac{\varphi^{k+1}+\varphi^k}{\tau}=f^{k+1/2},\quad k=1,\ 2,\ \ldots,\ M-1,\\ \varphi^0=\bar{g},\end{gathered} \tag{5.8.4}$$

где $f^{k+1/2}$ — вектор с компонентами $\{f_{m,\,l}(\tau(k+1/2))\}$, $m,\ l=1,\ 2,\ \ldots,\ n$.

Разностная схема (5.8.4) аппроксимирует исходную задачу (5.8.1) с порядком аппроксимации $O(\tau^2+h^2)$ и абсолютно устойчива. Имеет место сходимость приближенного решения $\{\varphi^k\}$ к точному решению φ задачи (5.8.1) со скоростью $O(\tau^2+h^2)$. Если нам нужно найти решение задачи (5.8.1) с точностью $O(\tau+h^2)$, то можно использовать явную или неявную схемы с шагами h и τ либо схему Кранка — Николсона с шагами h и $\sqrt{\tau}$ по пространству и по времени соответственно:

$$\begin{gathered}\left(E+\frac{\sqrt{\tau}}{2}A\right)\varphi^{k+1}=\left(E-\frac{\sqrt{\tau}}{2}A\right)\varphi^k+\sqrt{\tau}f^{k+1/2},\quad k=0,\ 1,\ \ldots,\ M'-1,\\ \varphi^0=\bar{g},\end{gathered} \tag{5.8.5}$$

Здесь $M'=T/\sqrt{\tau}=\sqrt{T}\sqrt{M}$ (если $\tau=T/M$).

Будем предполагать, что используется идеализированная модель многопроцессорной ЭВМ, описанная в § 4.8 при $1\leqslant p\leqslant N$. Через T_p будем обозначать по-прежнему число тактов, необходимых для реализации алгоритма с помощью p процессоров.

Решение задачи (5.8.1) будем искать с точностью $\varepsilon=O(\tau+h^2)$, где $h=1/(n+1)$ — шаг сетки по пространственным переменным $x,\ y$, а $\tau=T/M$ — шаг сетки по времени.

Явная схема. Явную схему рассмотрим в виде

$$\begin{gathered}\varphi^{k+1}=\varphi^k+\tau(-A\varphi^k+f^k),\quad k=0,\ 1,\ \ldots,\ M'-1,\\ \varphi^0=\bar{g},\end{gathered} \tag{5.8.6}$$

причем при условии $\tau \leqslant h^2/4$. В этом случае, как легко видеть,

$$T_1 = O(NM).$$

При наличии p процессоров каждый шаг метода (5.8.6) можно реализовать за $O(N/p)$ тактов; в итоге

$$T_p = O\left(\frac{N}{p}M\right). \tag{5.8.7}$$

Неявная схема. Неявную схему запишем в виде

$$\begin{gathered}(E+\tau A)\varphi^{k+1} = \varphi^k + \tau f^{k+1}, \quad k = 0,\ 1,\ \ldots,\ M-1,\\ \varphi^0 = \bar{g}.\end{gathered} \tag{5.8.8}$$

На каждом шаге схемы (5.8.8) нужно находить решение системы $(E+\tau A)\varphi^{k+1} = \varphi^k + \tau f^{k+1}$. Это можно сделать с помощью одного из методов, описанных в 4.8.2.

Если мы будем применять для решения системы $(E+\tau A)\varphi^{k+1} = \varphi^k + \tau f^{k+1}$ нетрадиционные методы, то мы должны учесть, что при различных соотношениях τ и h будет различным число обусловленности матриц $E+\tau A$, что повлияет на скорость сходимости соответствующих итерационных методов. Параллельные же алгоритмы для реализации итерационных методов останутся прежними, теми же, что и в 4.8.2, поскольку матрица A из 4.8.2 и матрица $E+\tau A$ имеют одну и ту же структуру.

В связи с этим мы должны лишь найти k_0 — число итераций в каждом методе, необходимых для достижения заданной точности по h: $\epsilon = O(N^{-1})$ в зависимости от соотношения τ и h.

Мы будем рассматривать три случая: 1) $\tau >> h^2$; 2) $\tau/h = \text{const}$; 3) $\tau/h^2 = \text{const}$, а также семь итерационных методов для решения системы $(E+\tau A)\varphi^{k+1} = \varphi^k + \tau f^{k+1}$: метод простой итерации; метод Гаусса — Зейделя; метод чебышевского ускорения; метод последовательной верхней релаксации с оптимальным выбором параметра (SOR); метод последовательной релаксации по линиям (SLOR); метод симметричной верхней релаксации (SSOR); метод переменных направлений (МПН).

Нетрудно пересчитать, каким будет число k_0 в каждом из этих итерационных методов при решении системы $(E+\tau A)\varphi^{k+1} = \varphi^k + \tau f^{k+1}$ с точностью $\varepsilon = O(N^{-1})$. Эти значения k_0 приведены в таблице 5.1.

Теперь для параллельной реализации методов решения системы $(E+\tau A)\varphi^{k+1} = \varphi^k + \tau f^{k+1}$ нам необходимо пересмотреть таблицу 4.3 из 4.8.2 с учетом изменения k_0 при различных соотношениях τ и h и с учетом того, что число процессоров фиксировано и равно p. В итоге мы приходим к таблице

Таблица 5.1. Число итераций k_0 при различных соотношениях τ и h

Методы решения системы $(E+\tau A)u=f$	Число итераций k_0, необходимых для достижения точности $\varepsilon = O(N^{-1})$		
	$\tau >> h^2$	τ/h = const	τ/h^2 = const
Метод простой итерации	$O(N\ln N)$	$O(N^{1/2}\ln N)$	$O(\ln N)$
Метод Гаусса — Зейделя	$O(N\ln N)$	$O(N^{1/2}\ln N)$	$O(\ln N)$
Метод чебышевского ускорения	$O(N^{1/2}\ln N)$	$O(N^{1/4}\ln N)$	$O(\ln N)$
SOR	$O(N^{1/2}\ln N)$	$O(N^{1/4}\ln N)$	$O(\ln N)$
SLOR	$O(N^{1/2}\ln N)$	$O(N^{1/4}\ln N)$	$O(\ln N)$
SSOR	$O(N^{1/4}\ln N)$	$O(N^{1/8}\ln N)$	$O(\ln N)$
МПН	$O(\ln^2 N)$	$O(\ln^2 N)$	$O(\ln N)$

значений $T_p^{(1)}$. Проделав M шагов по формуле (5.8.8), мы получим число тактов, необходимых для реализации алгоритма (5.8.8):

$$T_p = MT_p^{(1)}.$$

Эти значения T_p при различных соотношениях τ и h приведены в таблице 5.2.

Таблица 5.2. Реализация неявной схемы на p процессорах, $1 \leqslant p \leqslant N$

Методы решения системы $(E+\tau A)u=f$	Число тактов T_p, $1 \leqslant p \leqslant N$		
	$\tau >> h^2$	τ/h = const	τ/h^2 = const
Метод простой итерации	$M\frac{N^2}{p}\ln N$	$M\frac{N^{3/2}}{p}\ln N$	$M\frac{N}{p}\ln N$
Метод Гаусса — Зейделя	$M\frac{N^2}{p}\ln N$	$M\frac{N^{3/2}}{p}\ln N$	$M\frac{N}{p}\ln N$
Метод чебышевского ускорения	$M\frac{N^{3/2}}{p}\ln N)$	$M\frac{N^{5/4}}{p}\ln N$	$M\frac{N}{p}\ln N$
SOR	$M\frac{N^{3/2}}{p}\ln N$	$M\frac{N^{5/4}}{p}\ln N$	$M\frac{N}{p}\ln N$
SLOR	$M\sqrt{\frac{N}{p}}(\ln p + +\sqrt{\frac{N}{p}})\times N^{1/2}\ln N$	$M\sqrt{\frac{N}{p}}(\ln p + +\sqrt{\frac{N}{p}})\times N^{1/4}\ln N$	$M\sqrt{\frac{N}{p}}(\ln p + +\sqrt{\frac{N}{p}})\times \ln N$
SSOR	$M\frac{N^{5/4}}{p}\ln N, 1 \leqslant \leqslant p \leqslant \sqrt{n}$	$M\frac{N^{9/8}}{p}\ln N, 1 \leqslant \leqslant p \leqslant \sqrt{n}$	$M\frac{N}{p}\ln N, 1 \leqslant p \leqslant \leqslant \sqrt{n}$
МПН	$M\sqrt{\frac{N}{p}}(\ln p + +\sqrt{\frac{N}{p}})\times \ln^2 N$	$M\sqrt{\frac{N}{p}}(\ln p + +\sqrt{\frac{N}{p}})\times \ln^2 N$	$M\sqrt{\frac{N}{p}}(\ln p + +\sqrt{\frac{N}{p}})\times \ln N$
ДБПФ	$M\frac{N}{p}\ln N$	$M\frac{N}{p}\ln N$	$M\frac{N}{p}\ln N$

Из таблицы 5.2 следует, что лучше всего применять неявную схему для соотношения шагов τ/h^2 = const. В этом случае почти все методы дают одинаковую оценку для T_p, совпадающую с оценкой числа тактов для дискретного быстрого преобразования Фурье (ДБПФ):

$$T_p = O\left(\frac{N}{p}\ln N\right). \tag{5.8.9}$$

Схема Кранка — Николсона. Как мы указывали выше, чтобы найти решение задачи (5.8.1) с точностью $O(\tau + h^2)$, достаточно воспользоваться схемой Кранка — Николсона в виде (5.8.5):

$$\left(E+\frac{\sqrt{\tau}}{2}A\right)\varphi^{k+1} = \left(E-\frac{\sqrt{\tau}}{2}A\right)\varphi^k + \sqrt{\tau}f^{k+1/2}, \quad k=0,\ 1,\ \ldots,\ M'-1, \qquad \varphi^0 = \bar{g}. \tag{5.8.10}$$

Аналогично неявной схеме (5.8.5) на каждом шаге методу (5.8.10) нужно решать систему вида $(E + \frac{\sqrt{\tau}}{2}A)u = f$. Как мы видели выше, все здесь зависит от соотношения шагов τ и h. Мы будем рассматривать два случая:

1) $\tau >> h^2$ и 2) τ/h^2 = const. (В случае τ/h = const результаты по числу тактов T_p получаются хуже, чем для случая τ/h^2 = const, и мы его не рассматриваем.) Эти случаи соответствуют:

1) $\sqrt{\tau} >> h^2$ и 2) $\sqrt{\tau}/h$ = const. Значит, мы можем воспользоваться таблицами 5.1 и 5.2, вместо τ рассматривая $\sqrt{\tau}$. Из этих таблиц мы получаем число тактов $T_p^{(1)}$, необходимое для реализации одного шага метода (5.8.10). Тогда для реализации $M' = \sqrt{T}\sqrt{M} = O(\sqrt{M})$ шагов алгоритма (5.8.10) мы получим значение числа тактов $T_p = M'T_p^{(1)}$, необходимых для решения задачи (5.8.1) с помощью схемы Кранка — Николсона с точностью $\varepsilon = O(\tau + h^2) = O(M^{-1} + N^{-1})$ на p процессорах. Эти значения приведены в таблице 5.3.

Таблица 5.3. Реализация схемы Кранка — Николсона на p процессорах, $1 \leqslant p \leqslant N$

Методы решения системы $(E + \frac{\sqrt{\tau}}{2}A)u = f$	Число тактов T_p, $1 \leqslant p \leqslant N$	
	$\tau >> h^2$	τ/h^2 = const
Метод простой итерации	$\sqrt{M}\frac{N^2}{p}\ln N$	$\sqrt{M}\frac{N^{3/2}}{p}\ln N$
Метод Гаусса — Зейделя	$\sqrt{M}\frac{N^2}{p}\ln N$	$\sqrt{M}\frac{N^{3/2}}{p}\ln N$
Метод чебышевского ускорения	$\sqrt{M}\frac{N^{3/2}}{p}\ln N$	$\sqrt{M}\frac{N^{5/4}}{p}\ln N$
SOR	$\sqrt{M}\frac{N^{3/2}}{p}\ln N$	$\sqrt{M}\frac{N^{5/4}}{p}\ln N$
SLOR	$\sqrt{M}\sqrt{\frac{N}{p}}\left(\ln p + \sqrt{\frac{N}{p}}\right)\sqrt{N}\ln N$	$\sqrt{M}\sqrt{\frac{N}{p}}\left(\ln p + \sqrt{\frac{N}{p}}\right)N^{1/4}\ln N$
SSOR	$\sqrt{M}\frac{N^{5/4}}{p}\ln N, 1 \leqslant p \leqslant N$	$\sqrt{M}\frac{N^{9/8}}{p}\ln N, 1 \leqslant p \leqslant N$
МПН	$\sqrt{M}\sqrt{\frac{N}{p}}\left(\ln p + \sqrt{\frac{N}{p}}\right)\ln^2 N$	$\sqrt{M}\sqrt{\frac{N}{p}}\left(\ln p + \sqrt{\frac{N}{p}}\right)\ln^2 N$
ДБПФ	$\sqrt{M}\frac{N}{p}\ln N$	$\sqrt{M}\frac{N}{p}\ln N$

Сравнение трех схем. Для сравнения всех трех схем рассмотрим случай τ/h^2 = const. Это означает, что N/M = const, и мы решаем задачу (5.8.1) с точностью $\varepsilon = O(h^2) = O(N^{-1})$. Из таблиц 5.2 и 5.3 при $M = N$ получаем значения числа тактов T_p в случае τ/h^2 = const. Эти значения T_p приведены в таблице 5.4.

Таблица 5.4. Сравнение трех схем в случае τ/h^2 = const по числу тактов T_p, $1 \leqslant p \leqslant N$

Явная схема	Неявная схема	Схема Кранка — Николсон	Метод решения системы линейных уравнений
$\frac{N^2}{p}$	$\frac{N^2}{p}\ln N$	$\frac{N^2}{p}\ln N$	Метод простой итерации
	$\frac{N^2}{p}\ln N,\ 1 \leqslant p \leqslant \sqrt{N}$	$\frac{N^2}{p}\ln N$	Метод Гаусса — Зейделя
	$\frac{N^2}{p}\ln N$	$\frac{N^{7/4}}{p}\ln N$	Метод чебышевского ускорения
	$\frac{N^2}{p}\ln N$	$\frac{N^{7/4}}{p}\ln N$	SOR
	$\sqrt{\frac{N}{p}}\left(\ln p + \sqrt{\frac{N}{p}}\right) N \ln N$	$\sqrt{\frac{N}{p}}\left(\ln p + \sqrt{\frac{N}{p}}\right) N^{3/4}\ln N$	SLOR
	$\frac{N^2}{p}\ln N, 1 \leqslant p \leqslant \sqrt{N}$	$\frac{N^{13/8}}{p}\ln N, 1 \leqslant p \leqslant \sqrt{N}$	SSOR
	$\sqrt{\frac{N}{p}}\left(\ln p + \sqrt{\frac{N}{p}}\right) N \ln N$	$\sqrt{\frac{N}{p}}\left(\ln p + \sqrt{\frac{N}{p}}\right) \sqrt{N}\ln N$	МПН
	$\frac{N^2}{p}\ln N$	$\frac{N^{3/2}}{p}\ln N$	ДБПФ

Хотя наиболее естественен параллелизм в случае применения явных схем, однако, как следует из таблицы 5.4, наилучший результат в данном случае получается при использовании схемы Кранка — Николсона. Число тактов, необходимых для решения исходной задачи (5.8.1) с точностью $\varepsilon = O(N^{-1})$ с помощью схемы Кранка — Николсона на p процессорах при применении ДБПФ, равно

$$T_p = O\left(\frac{N^{3/2}}{p}\ln N\right). \tag{5.8.11}$$

Здесь $R_p = T_1/T_p \sim p,\ E_p \sim 1$. Если использовать явную схему, то

$$T_p = O\left(\frac{N^2}{p}\right), \tag{5.8.12}$$

причем

$$R_p = \frac{T_1}{T_p} = p, \quad E_p \sim 1.$$

При решении исходной задачи (5.8.1) с помощью неявной схемы получаем в большинстве случаев

$$T_p = O\left(\frac{N^2}{p} \ln N\right), \tag{5.8.13}$$

а при использовании методов SLOR и МПН —

$$T_p = O\left(\frac{N^{3/2}}{\sqrt{p}} \ln p \ln N + \frac{N^2}{p} \ln N\right). \tag{5.8.14}$$

В последнем случае при $1 \leqslant p < N$ $T_p = O\left(\frac{N^2}{p} \ln N\right)$, а при $p = N$ результат ухудшается (см. другие схемы):

$$T_p = O(N \ln^2 N). \tag{5.8.15}$$

В заключение данной главы отметим, что анализ алгоритмов, описанный в § 4.8 и 5.8, может служить одним из критериев выбора той или иной архитектуры электронно-вычислительной машины, разрабатываемой для решения задач в той или иной прикладной области. Он может также использоваться и при выборе алгоритмов для уже существующих вычислительных систем.

Глава 6.

Повышение точности приближенных решений по Ричардсону

В предыдущих главах рассматривались, как правило, сходящиеся разностные схемы. В принципе с их помощью можно найти решение аппроксимируемой дифференциальной задачи со сколь угодно высокой точностью, выбирая достаточно малые шаги разностной сетки. Но тогда увеличиваются как размерность приближенной задачи, так и затраты на ее решение, особенно для многомерных задач. В связи с этим объем оперативной памяти и быстродействие ЭВМ во многих случаях существенно ограничивают достигаемую точность приближенного решения.

Поэтому построение приближенных решений высокой точности является весьма актуальной задачей вычислительной математики. Известны различные подходы к построению таких решений. Наиболее широкую сферу приложений нашли разностные и вариационно-разностные схемы повышенного порядка точности и повышение точности приближенных решений на последовательности сеток.

Основное внимание в настоящей главе будет уделено второму подходу, который идейно восходит к Ричардсону [4] и назван им *экстраполяцией к пределу*. Метод состоит в использовании последовательностей сеток и соответствующих им однотипных аппроксимаций для построения приближенных решений заданного порядка точности. Применение такого подхода позволяет использовать в расчетах только стандартные разностные аппроксимации задач первого и второго порядков точности. В настоящее время усилиями советских и зарубежных математиков обосновано применение экстраполяции к пределу для самых разных задач, включая нелиней-

ные. Из всего многообразия этих задач автором отобраны и далее излагаются наиболее простые.

6.1. Обыкновенное дифференциальное уравнение первого порядка

В этом параграфе на примере простейшего линейного уравнения мы детально рассмотрим эффект от применения экстраполяции Ричардсона по шагу разностной сетки.

Пусть $u(t)$ — решение дифференциального уравнения

$$u' + a(t)u = f(t), \quad t \in (0,1), \tag{6.1.1}$$

с начальным условием

$$u(0) = u_0. \tag{6.1.2}$$

Относительно задачи (6.1.1), (6.1.2) предполагается, что

$$a(t) \geqslant 0, \quad t \in (0,1), \tag{6.1.3}$$

и гладкость всех функций достаточна для проведения дальнейших выкладок.

Построим равномерное разбиение отрезка $[0,1]$ «целыми» узлами

$$t_j = j\tau, \quad j = 0,1,\ldots,M, \tag{6.1.4}$$

где M целое, с шагом $\tau = 1/M$ и с «промежуточными» узлами

$$t_{j+1/2} = (j+1/2)\tau,\ j = 0,1,\ldots,M-1. \tag{6.1.5}$$

В соответствии со схемой Кранка — Николсона (см § 5.2) заменим исходное дифференциальное уравнение (6.1.1) в промежуточных узлах приближенной системой алгебраических уравнений

$$\frac{u^{j+1}-u^j}{\tau} + a^{j+1/2}\frac{u^{j+1}+u^j}{2} = f^{j+1/2}, \quad j = 0,1,\ldots,M-1. \tag{6.1.6}$$

Если к этим формулам добавить начальное условие

$$u^0 = u_0, \tag{6.1.7}$$

то все u^j могут быть определены рекуррентно:

$$u^j = \left(1 + \frac{\tau}{2} a^{j-1/2}\right)^{-1} \left[\left(1 - \frac{\tau}{2} a^{j-1/2}\right) u^{j-1} - \frac{\tau}{2} f^{j-1/2}\right],$$
$$j = 1, 2, \ldots, M. \tag{6.1.8}$$

Как уже не раз отмечалось, разностная задача (6.1.6), (6.1.7) имеет второй порядок аппроксимации. Из рассуждений § 5.2 с заменой положительно определенной матрицы Λ^j на положительное число $a^{j+1/2}$ вытекает оценка устойчивости

$$|u^j| \leqslant |u_0| + \max_{0 \leqslant j \leqslant M-1} |f^{j+1/2}|. \tag{6.1.9}$$

Поэтому, согласно § 1.4, решение разностной задачи приближает решение дифференциальной задачи со вторым порядком точности

$$\max_{0 \leqslant j \leqslant M} |u^j - (u)^j| \leqslant C_1 \tau^2, \tag{6.1.10}$$

где C_1 — постоянная, не зависящая от τ.

Проиллюстрируем эти теоретические оценки практическими расчетами на ЭВМ для уравнения

$$u' + tu = t, \quad t \in (0, 1), \tag{6.1.11}$$

с начальным условием

$$u(0) = 2. \tag{6.1.12}$$

Легко проверяется, что точным решением задачи (6.1.11), (6.1.12) является функция

$$u(t) = e^{-t^2/2} + 1. \tag{6.1.13}$$

Численный эксперимент состоял в построении приближенного решения задачи (6.1.4)—(6.1.8) для M, равного последовательно 10, 20, 50, 100, 200, и в определении величины

$$\xi = \max_{0 \leqslant j \leqslant M} |u^j - (u)^j|. \tag{6.1.14}$$

Для наглядности изображения зависимости величины ξ от M результаты расчетов приведены на рис. 6.1 в логарифмических координатах.

Как легко видеть из рисунка, численные эксперименты подтверждают оценку (6.1.10). Дальнейшее сопоставление результатов счета и наклона теоретической прямой указывает на неулучшаемость оценки (6.1.10) по порядку, т. е.

$$\xi \geqslant C_2 \tau^2 \tag{6.1.15}$$

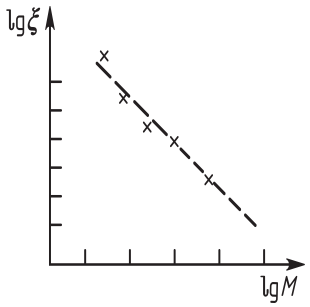


Рис. 6.1.

с некоторой положительной константой C_2, не зависящей от τ. Таким образом, максимальная погрешность приближенного решения (6.1.14) оказалась величиной порядка τ^γ с показателем $\gamma = 2$.

Детальные наблюдения не максимальной погрешности, а погрешности приближенного решения в каждой точке разностной сетки позволили Ричардсону в 1910 г. выдвинуть гипотезу о следующем поведении погрешности приближенного решения при $\tau \to 0$:

$$u^j - (u)^j = \tau^2 (v)^j + \eta^j, \quad j = 0, 1, \ldots, M. \tag{6.1.16}$$

При этом главная часть погрешности $\tau^2 (v)^j$ является произведением τ^2 на значения в узлах разностной сетки функции $v(t)$, не зависящей от τ, а остаточный член η^j для всех $j = 0, 1, \ldots, M$ является величиной $O(\tau^4)$.

В этом разделе мы ставим целью обоснование этого факта для задачи (6.1.1)—(6.1.3).

Обоснование обычно проводится в три приема. Сначала строятся некоторые необходимые условия для выполнения условий (6.1.16), затем из них конструируется краевая задача, решением которой будет функция $v(t)$, а после этого уже доказывается ограниченность сеточной функции η^j/τ^4.

Итак, если соотношения (6.1.16) выполнены, то можно записать

$$u^j = (u)^j + \tau^2 (v)^j + \ldots, \; j = 0, 1, \ldots, M, \tag{6.1.17}$$

где предположено, что остаточный член является величиной $O(\tau^4)$. Зафиксируем j и, подставляя значения u^j в уравнение (6.1.6), получим

$$\frac{(u)^{j+1} - (u)^j}{\tau} + a^{j+1/2} \frac{(u)^{j+1} + (u)^j}{2} + \\ + \tau^2 \left(\frac{(v)^{j+1} - (v)^j}{\tau} + a^{j+1/2} \frac{(v)^{j+1} + (v)^j}{2} \right) + \cdots = f^{j+1/2}.$$

Используя разложение Тейлора для функций u и v:

$$(u)^{j+1/2\pm1/2} = \left(u \pm \frac{\tau}{2}u' + \frac{\tau^2}{8}u'' \pm \frac{\tau^3}{48}u'''\right)^{j+1/2} + \ldots,$$
$$(v)^{j+1/2\pm1/2} = \left(v \pm \frac{\tau}{2}v'\right)^{j+1/2} + \ldots, \tag{6.1.18}$$

приведем каждое слагаемое к значению функции в узле $j+1/2$:

$$\left(u' + au + \tau^2\left(\frac{1}{24}u''' + \frac{1}{8}au''\right) + \tau^2(v' + av)\right) + \cdots = f^{j+1/2}.$$

Величины порядка $O(\tau^0)$ исключаются ввиду выполнения уравнения (6.1.1) в узле $j+1/2$, поэтому для достижения равенства при $\tau \to 0$ необходимо выполнение соотношения

$$(v' + av)^{j+1/2} = \left(-\frac{1}{24}u''' - \frac{1}{8}au''\right)^{j+1/2}, \quad j = 0, 1, \ldots M-1. \tag{6.1.19}$$

Отдельно рассмотрим равенство (6.1.17) для $j = 0$:

$$u^0 = u_0 + \tau(v)^0 + \ldots. \tag{6.1.20}$$

Из начальных условий (6.1.2) и (6.1.7) следует равенство $u^0 = u_0$; отсюда при $\tau \to 0$ приходим к условию для v^0:

$$(v)^0 = 0. \tag{6.1.21}$$

На этом построение необходимых условий, вытекающих из равенства (6.1.17), заканчивается. Теперь для определения функции $v(t)$ потребуем большего: пусть (6.1.19) выполняется не только в узлах сетки, но и на всем интервале $(0,1)$, т. е. справедливо равенство

$$v'(t) + a(t)v(t) = -\frac{1}{24}u'''(t) - \frac{1}{8}au''(t), \quad t \in (0,1). \tag{6.1.22}$$

Если это уравнение (относительно функции $v(t)$) дополнить начальным условием

$$v(0) = 0, \tag{6.1.23}$$

вытекающим из (6.1.21), то получается дифференциальная задача, имеющая единственное достаточно гладкое решение $v(t)$, причем независимость функции v от τ очевидна.

Докажем теперь, что сеточная функция η, определяемая $M+1$ соотношением

$$\eta^j = u^j - (u)^j - \tau^2(v)^j, \tag{6.1.24}$$

будет иметь своими значениями величины порядка $O(\tau^4)$.

Для этого зафиксируем некоторое j и подставим функцию η в разностный оператор задачи (6.1.6):

$$\frac{\eta^{j+1}-\eta^j}{\tau}+a^{j+1/2}\frac{\eta^{j+1}+\eta^j}{2}=\frac{u^{j+1}-u^j}{\tau}+a^{j+1/2}\frac{u^{j+1}+u^j}{2}-$$
$$-\left(\frac{(u)^{j+1}-(u)^j}{\tau}+a^{j+1/2}\frac{(u)^{j+1}+(u)^j}{2}\right)-$$
$$-\tau^2\left(\frac{(v)^{j+1}-(v)^j}{\tau}+a^{j+1/2}\frac{(v)^{j+1}+(v)^j}{2}\right). \quad (6.1.25)$$

Преобразуем правую часть этого соотношения. Из формулы (6.1.6) следует, что два первых слагаемых правой части равны числу $f^{j+1/2}$. Все функции в последующих слагаемых приведем к значениям в узле $j+1/2$, используя разложения (6.1.18) с остаточными членами в форме Лагранжа:

$$\frac{\eta^{j+1}-\eta^j}{\tau}+a^{j+1/2}\frac{\eta^{j+1}+\eta^j}{2}=$$
$$=\left(f-u'-au-\tau^2\left(\frac{1}{24}u'''+\frac{1}{8}au''\right)-\tau^2(v'+av)\right)^{j+1/2}-$$
$$-\tau^4\left(\frac{1}{16\cdot 5!}u^{\mathrm{V}}(\xi_j^1)+\frac{1}{16\cdot 4!}u^{\mathrm{IV}}(\xi_j^2)\right)-\tau^4\left(\frac{1}{24}v'''(\xi_j^3)+\frac{1}{8}a^{j+1/2}v''(\xi_j^4)\right),$$

где ξ_j^i — некоторые точки интервала (t_j,t_{j+1}).

Вспоминая, что функция u является решением уравнения (6.1.1), а v — решением уравнения (6.1.22), упростим последнее равенство:

$$\frac{\eta^{j+1}-\eta^j}{\tau}+a^{j+1/2}\frac{\eta^{j+1}+\eta^j}{2}=$$
$$=-\tau^4\left(\frac{1}{16\cdot 5!}u^{\mathrm{V}}(\xi_j^1)+\frac{1}{16\cdot 4!}u^{\mathrm{IV}}(\xi_j^2)+\frac{1}{24}v'''(\xi_j^3)+\frac{1}{8}a^{j+1/2}v''(\xi_j^4)\right). \quad (6.1.26)$$

Если функции u и v имеют на отрезке $[0,1]$ непрерывные производные, участвующие в (6.1.26), то правая часть представляет собой величину, модуль которой ограничен числом $C_3\tau^4$ с константой C_3, не зависящей от τ. Используя этот факт, справедливый для всех $j=0,1,\ldots,M-1$, и оценку (6.1.9), получаем

$$|\eta^j|\leqslant|\eta^0|+C_3\tau^4.$$

Вследствие (6.1.2), (6.1.7) и (6.1.23) из определения η^j вытекает, что $\eta^0=0$. Поэтому

$$|\eta^j|\leqslant C_3\tau^4,\quad j=0,1,\ldots,M, \quad (6.1.27)$$

и гипотеза (6.1.16) доказана.

Теперь изложим метод повышения точности, основанный на этом разложении.

Пусть u_τ и $u_{\tau/2}$ — решение двух приближенных задач (6.1.6), (6.1.7) с шагами τ и $\tau/2$ соответственно. Несмотря на то, что точность каждого из решений является величиной порядка $O(\tau^2)$, мы довольно просто построим из них решение с четвертым порядком точности на сетке с шагом τ. Пусть $t_j = j\tau$ — произвольная точка разностной сетки с шагом τ; тогда линейная комбинация

$$\bar{u}^j = \frac{4}{3}u^{2j}_{\tau/2} - \frac{1}{3}u^j_\tau \tag{6.1.28}$$

приближает точное решение u с четвертым порядком точности по τ:

$$|\bar{u}^j - u(j\tau)| \leqslant C_4\tau^4, \quad j = 0, 1, \ldots, M, \tag{6.1.29}$$

с константой C_4, не зависящей от τ.

Докажем это. В точке $j\tau$ справедливы разложения (6.1.16) для обоих приближенных решений

$$u^j_\tau = u(j\tau) + \tau^2 v(j\tau) + \eta^j_\tau,$$
$$u^{2j}_{\tau/2} = u(j\tau) + \frac{\tau^2}{4}v(j\tau) + \eta^{2j}_{\tau/2}.$$

Поэтому из (6.1.28) имеем

$$\bar{u}^j = \left(\frac{4}{3} - \frac{1}{3}\right)u(j\tau) + \tau^2\left(\frac{4}{3}\cdot\frac{1}{4} - \frac{1}{3}\right)v(j\tau) + \left(\frac{4}{3}\eta^{2j}_{\tau/2} - \frac{1}{3}\eta^j_\tau\right).$$

Отсюда вытекает, что

$$\bar{u}^j = u(j\tau) + \frac{4}{3}\eta^{2j}_{\tau/2} - \frac{1}{3}\eta^j_\tau.$$

Учитывая оценку (6.1.27) для функции η^j, получаем неравенство

$$\left|\frac{4}{3}\eta^{2j}_{\tau/2} - \frac{1}{3}\eta^j_\tau\right| \leqslant \frac{4}{3}\cdot C_3\frac{\tau^4}{16} + C_3\frac{\tau^4}{3} = \frac{5}{12}C_3\tau^4,$$

откуда следует (6.1.29) при условии, что $C_4 = 5C_3/12$.

Следует отметить, что сформулированный подход может быть распространен на случай неравномерных сеток. Однако в этом случае требуется более сложный анализ.

Возможно дальнейшее обобщение в сторону увеличения числа слагаемых разложения (6.1.16), что приводит к построению приближенного

решения с порядком точности выше четвертого. Это вытекает из общих теорем, приведенных в следующем параграфе.

6.2. Общие результаты

В предыдущем параграфе построен пример повышения точности приближенного решения для конкретной задачи. Здесь мы изложим в абстрактной форме достаточные условия повышения точности разностных решений для широкого класса задач.

6.2.1. Теорема о разложении

Пусть D — ограниченная область в n-мерном евклидовом пространстве $\mathbf{R}^n$; $\bar{D}$ — ее замыкание; ∂D — граница области либо ее часть. Рассмотрим задачу математической физики

$$\begin{aligned} A\varphi &= f \quad \text{в} \quad D, \\ a\varphi &= g \quad \text{на} \quad \partial D, \end{aligned} \tag{6.2.1}$$

где A, a — линейные операторы; f, g, φ — функции с областями определения D, ∂D, $\bar{D}$ соответственно. Естественно предположить, что эта задача имеет единственное решение в классе достаточно гладких функций. Для многих задач математической физики найдены условия (налагаемые на правые части), выполнение которых влечет существование, единственность и некоторую гладкость решения. Обычно они формулируются следующим образом.

Пусть F^k, G^k, Φ^k (k — некоторый целый индекс) — классы функций, определенных соответственно на D, ∂D, $\bar{D}$. Тогда предположим выполненным условие, гарантирующее разрешимость задачи для гладких правых частей.

Условие А. Для любого $k = 0, 1, \ldots, m$ и любой пары функций $f \in F^k$, $g \in G^k$ существует единственное решение $\varphi \in \Phi^k$ задачи (6.2.1).

Например, для задачи (6.1.1), (6.1.2) эти обозначения принимают следующий вид:

$$\begin{aligned} &D = (0,1), \quad \partial D = \{0\}, \quad \bar{D} = [0,1], \\ &F^k = C^{2k+2}[0,1], \quad \Phi^k = C^{2k+3}[0,1], \end{aligned}$$

а G^k для любого k совпадает с множеством вещественных чисел $(-\infty, \infty)$. Тогда для достаточно гладкого коэффициента $a(t)$ (например, $a \in C^{2m+2}[0,1]$) выполнение условия А очевидно.

Для численного решения задачи (6.2.1) введем разностную сетку $\bar{D}_h \in \bar{D}$ с переменным параметром h, который в принципе может быть сколь угодно малым. Дифференциальную задачу заменим конечно-разностной (алгебраической) системой уравнений, определенных в узлах конечных подмножеств $D_h \in D$ и $\partial D_h \in \partial D$, а приближенное решение будем искать в пространстве сеточных функций, определенных в узлах сетки $D_h \in \bar{D}$. Имеем

$$\begin{aligned} A^h\varphi^h &= f \quad \text{в} \quad D_h, \\ a^h\varphi^h &= g \quad \text{на} \quad \partial D_h. \end{aligned} \tag{6.2.2}$$

Здесь A^h, a^h — линейные алгебраические операторы, а φ^h — сеточная функция, аппроксимирующая в узлах $\bar{D}_h$ решение φ исходной дифференциальной задачи. В линейных пространствах φ_h, F_h, G_h сеточных функций, определенных на $\bar{D}_h$, D_h, ∂D_h, введем нормы $\|\cdot\|_{\Phi_h}$, $\|\cdot\|_{F_h}$, $\|\cdot\|_{G_h}$ соответственно и потребуем, чтобы задача (6.2.2) была устойчивой.

Условие В. Если сеточная функция $u^h \in \Phi_h$ является решением задачи

$$\begin{aligned} A^h\varphi^h &= f^h \quad \text{в} \quad D_h, \\ a^h\varphi^h &= g^h \quad \text{на} \quad \partial D_h, \end{aligned} \tag{6.2.3}$$

где $f^h \in F_h$, $g^h \in G_h$, то справедлива оценка

$$\|u^h\|_{\Phi_h} \leqslant c_1\|f^h\|_{F_h} + c_2\|g^h\|_{G_h} \tag{6.2.4}$$

с константами, не зависящими от h, f^h, g^h.

Проиллюстрируем это условие на примере задачи из § 6.1. Для этого введем нормы

$$\begin{gathered} \|u\|_{\Phi_h} = \max_{0\leqslant j\leqslant M} |u^j|, \\ \|f\|_{F_h} = \max_{0\leqslant j\leqslant M-1} |f^{j+1/2}|, \\ \|g\|_{G_h} = |g^0|. \end{gathered}$$

Тогда неравенство (6.1.9) записывается в виде (6.2.4), где $c_1 = c_2 = 1$, и, следовательно, условие В выполнено, т. е. задачи (6.1.6), (6.1.7) устойчивы.

Еще одно условие касается способа аппроксимации дифференциальных операторов разностными соотношениями.

Условие С. Для любой функции $u \in \Phi^k$, где $0 \leqslant k \leqslant m$, справедливы разложения

$$\begin{aligned} A^h\varphi^h &= Au + \sum_{j=1}^{k} h^j B_j + \sigma^h \quad \text{на} \quad D_h, \\ a^h\varphi^h &= au + \sum_{j=1}^{k} h^j b_j + \rho^h \quad \text{на} \quad \partial D_h, \end{aligned} \tag{6.2.5}$$

причем функции B_j, b_j не зависят от h, $B_j \in F^{k-j}$, $b_j \in G^{k-j}$, а для остаточных членов справедливы оценки

$$\|\sigma^h\|_{F_h} \leqslant c_3 h^{k+\beta}, \quad \|\rho^h\|_{G_h} \leqslant c_4 h^{k+\beta}, \tag{6.2.6}$$

где константы c_3, c_4 не зависят от h, а $\beta > 0$ и не зависит от h, k, u.

В случае $k = 0$, когда у сумм в (6.2.5) верхний предел меньше нижнего, они считаются равными нулю. Поэтому (6.2.5) превращается в условие аппроксимации операторов A, a разностными операторами A^h, a^h с порядком β.

Отметим, что условие С для стандартных разностных схем, как правило, выполняется. Это легко проверить с помощью формулы Тейлора. Например, пусть $u \in C^{k+2}[0,1]$. Тогда для первых разностей имеют место разложения

$$\begin{aligned} \frac{u(x+h)-u(x)}{h} &= u'(x) + \sum_{j=1}^{k} h^j \frac{u^{(j+1)}(x)}{(j+1)!} + \sigma_1^h(x), \\ \frac{u(x)-u(x-h)}{h} &= u'(x) + \sum_{j=1}^{k} h^j \frac{(-1)^j u^{(j+1)}(x)}{(j+1)!} + \sigma_2^h(x), \end{aligned} \tag{6.2.7}$$

где

$$|\sigma_i(x)| \leqslant \frac{h^{k+1}}{(k+2)!} \max_{x\in[0,1]} |u^{(k+2)}(x)|. \tag{6.2.8}$$

Для центральной разности аналогичное разложение содержит только четные степени h:

$$\frac{u(x+h/2)-u(x-h/2)}{h} = u'(x) + \sum_{j=1}^{[k/2]} h^{2j} \frac{u^{(2j+1)}(x)}{(2j+1)!4^j} + \sigma_3^h(x), \tag{6.2.9}$$

где $|\sigma_3^h(x)| \leqslant \frac{h^{k+1}}{(k+2)!2^{k+1}} \max_{x\in[0,1]} |u^{(k+2)}(x)|$.

Докажем следующую теорему.

Пусть для задач (6.2.1), (6.2.2) справедливы условия А, В, С и $f \in F^m$, $g \in G^m$. Тогда для разностного решения φ^h справедливо разложение

$$\varphi^h = \varphi + \sum_{k=1}^{m} h^k v_k + \eta^h \quad \text{на} \quad \bar{D}_h. \tag{6.2.10}$$

Здесь функции $v_k \in \Phi^{m-k}$ не зависят от h, а для остаточного члена η^h справедлива оценка

$$\|\eta^h\|_{\Phi_h} \leqslant c_5 h^{m+\beta}, \tag{6.2.11}$$

где константа c_5 не зависит от h.

Рассмотрим произвольный набор не зависящих от h функций $v_j \in \Phi^{m-j}$, где $j = 1, 2, \ldots, m$. По этим функциям и двум решениям φ и φ^h определим сеточную функцию

$$\eta^h = \varphi^h - \varphi - \sum_{j=1}^{m} h^j v_j \quad \text{на} \quad \bar{D}_h. \tag{6.2.12}$$

Выразим отсюда φ^h и подставим в уравнения (6.2.2). Получим

$$\begin{aligned} A^h\varphi + \sum_{j=1}^{m} h^j A^h v_j + A^h\eta^h = f \quad \text{на} \quad D_h, \\ a^h\varphi + \sum_{j=1}^{m} h^j a^h v_j + a^h\eta^h = g \quad \text{на} \quad \partial D_h. \end{aligned} \tag{6.2.13}$$

В соответствии с условием С мы можем записать разложения

$$\begin{aligned} A^h\varphi = f + \sum_{i=1}^{m} h^i B_{0i} + \sigma_0^h \quad \text{на} \quad D_h, \\ a^h\varphi = g + \sum_{i=1}^{m} h^i b_{0i} + \rho_0^h \quad \text{на} \quad \partial D_h \end{aligned} \tag{6.2.14}$$

и

$$\begin{aligned} A^h v_j = A v_j + \sum_{i=1}^{m-j} h^i B_{ji} + \sigma_j^h \quad \text{на} \quad D_h, \\ a^h v_j = a v_j + \sum_{i=1}^{m-j} h^i b_{ji} + \rho_j^h \quad \text{на} \quad \partial D_h. \end{aligned} \tag{6.2.15}$$

Здесь

$$B_{ji} \in F^{m-j-i},\ b_{ji} \in G^{m-j-i}, \tag{6.2.16}$$

причем B_{ji}, b_{ji} не зависят от h, а для остаточных членов выполняются неравенства

$$\|\sigma_j^h\|_{F_h} \leqslant c_{j1} h^{m-j+\beta},\ \|\rho_j^h\|_{G_h} \leqslant c_{j2} h^{m-h+\beta} \tag{6.2.17}$$

с константами c_{j1} и c_{j2}, не зависящими от h. Используя разложения (6.2.14), (6.2.15), приведем равенства (6.2.13) к следующему виду:

$$\begin{aligned} f + \sum_{j=1}^{m} h^j A v_j + \sum_{j=0}^{m} h^j \sum_{i=1}^{m-j} h^i B_{ji} + \sum_{j=0}^{m} h^j \sigma_j^h + A^h \eta^h = f \quad \text{на} \quad D_h, \\ g + \sum_{j=1}^{m} h^j a v_j + \sum_{j=0}^{m} h^j \sum_{i=1}^{m-j} h^i b_{ji} + \sum_{j=0}^{m} h^j \rho_j^h + a^h \eta^h = g \quad \text{на} \quad \partial D_h. \end{aligned} \tag{6.2.18}$$

Полагая

$$\xi^h = \sum_{j=0}^{m} h^j \sigma_j^h, \quad \zeta^h = \sum_{j=0}^{m} h^j \rho_j^h$$

и используя оценки (6.2.17), получим, что

$$\|\xi^h\|_{F_h} \leqslant h^{m+\beta} \tilde{c}_1, \quad \|\zeta^h\|_{G_h} \leqslant h^{m+\beta} \tilde{c}_2, \tag{6.2.19}$$

где

$$\tilde{c}_1 = \sum_{j=0}^{m} c_{j1}, \quad \tilde{c}_2 = \sum_{j=0}^{m} c_{j2}.$$

Соотношения (6.2.18) путем несложных преобразований с учетом введенных обозначений можно привести к виду

$$\begin{aligned} \sum_{j=1}^{m} h^j \left(A v_j + \sum_{i=1}^{j} B_{j-i,i} \right) + \xi^h + A^h \eta^h = 0 \quad \text{на} \quad D_h, \\ \sum_{j=1}^{m} h^j \left(a v_j + \sum_{i=1}^{j} b_{j-i,i} \right) + \zeta^h + a^h \eta^h = 0 \quad \text{на} \quad \partial D_h. \end{aligned} \tag{6.2.20}$$

Итак, для произвольного набора функций $v_j \in \Phi^{m-j}$ и функции η^h, определяемой равенством (6.2.12), получены равенства (6.2.20) с остаточными членами ξ^h и ζ^h, удовлетворяющими оценкам (6.2.19).

Выберем теперь функции v_j ($j = 1, 2, \ldots, m$) как решения дифференциальных задач

$$\begin{aligned} A v_j = -\sum_{i=1}^{j} B_{j-i,i} \quad \text{в} \quad D, \\ a v_j = -\sum_{i=1}^{j} b_{j-i,i} \quad \text{на} \quad \partial D. \end{aligned} \tag{6.2.21}$$

Например, функция v_1 находится в результате решения задачи

$$\begin{aligned} A v_1 = -B_{0,1} \quad \text{в} \quad D, \\ a v_j = -b_{0,1} \quad \text{на} \quad \partial D. \end{aligned}$$

Из условия C применительно к разложению (6.2.14) следует, что $B_{01} \in F^{m-1}$ и $b_{01} \in G^{m-1}$. Поэтому функция v_1 определяется единственным образом и $v_1 \in \Phi^{m-1}$ (см. условие A). Предположим, что для $j = 1, 2, \ldots, k$, где $1 \leqslant k \leqslant m$, уже найдены функции $v_j \in \Phi^{m-j}$. Тогда в силу условия C справедливы k разложений (6.2.15) для $j = 1, 2, \ldots, k$, удовлетворяющих условиям (6.2.16). Выпишем задачу (6.2.21) для $j = k + 1$:

$$\begin{aligned} Av_{k+1} &= -\sum_{i=1}^{k+1} B_{k-i+1,i} \quad \text{в} \quad D, \\ av_{k+1} &= -\sum_{i=1}^{k+1} b_{k-i+1,i} \quad \text{на} \quad \partial D. \end{aligned} \tag{6.2.22}$$

Правые части по свойству (6.2.16) принадлежат F^{m-k-1} и G^{m-k-1} соответственно. Поэтому из условия A следует, что задача (6.2.22) имеет единственное решение $v_{k+1} \in \Phi^{m-k-1}$; независимость v_{k+1} от h очевидна.

Итак, мы указали способ построения не зависящих от h функций $v_j \in \Phi^{m-j}$, где $j = 1, 2, \ldots, m$. Для этого конкретного набора функций v_j также справедливы тождества (6.2.20) с оценками (6.2.19), причем в силу (6.2.21) соотношения (6.2.20) принимают вид

$$\begin{aligned} A^h \eta^h &= -\xi^h \quad \text{в} \quad D_h, \\ a^h \eta^h &= -\zeta^h \quad \text{на} \quad \partial D_h. \end{aligned}$$

Из условия B следует неравенство

$$\|\eta^h\|_{\Phi_h} \leqslant c(\|\xi^h\|_{F_h} + \|\zeta^h\|_{G_h}).$$

Используя оценки (6.2.19), получаем (6.2.11), где $c_5 = c(\tilde{c}_1 + \tilde{c}_2)$. Выражая φ^h из равенства (6.2.12), мы получаем разложение (6.2.10) с требуемыми свойствами. Теорема доказана.

Если в разностных аппроксимациях используются только центральные разности, то, как правило, коэффициенты B_j, b_j в разложениях (6.2.5) обращаются в нуль для нечетных j. Такая ситуация наблюдалась, например, в предыдущем параграфе, а также в разложении (6.2.9). Как следствие, разложение (6.2.10) также не будет содержать нечетных степеней. Это позволит значительно сократить количество вычислений на ЭВМ. Сформулируем аналог условия C для этого важного случая.

Условие D. Для любой функции $u \in \Phi^k$ $(0 \leqslant k \leqslant m)$ справедливы разложения

$$\begin{aligned} A^h u &= Au + \sum_{i=1}^{k} h^{2j} B_j + \sigma^h \quad \text{на} \quad D_h, \\ a^h u &= au + \sum_{i=1}^{k} h^{2j} b_j + \rho^h \quad \text{на} \quad \partial D_h \end{aligned} \tag{6.2.23}$$

с оценками остаточных членов

$$\|\sigma^h\|_{F_h} \leqslant c_6 h^{2k+\beta}, \ \|\rho^h\|_{G_h} \leqslant c_7 h^{2k+\beta}. \tag{6.2.24}$$

Соответствующая формулировка теоремы о разложении выглядит следующим образом.

Пусть для задачи (6.2.1), (6.2.2) справедливы условия A, B, D. Тогда выполняется разложение

$$\varphi^h = \varphi + \sum_{k=1}^{m} h^{2k} v_k + \eta^h \text{ на } \bar{D}_h \tag{6.2.25}$$

с оценкой остаточного члена

$$\|\eta^h\|_{\Phi_h} \leqslant c_8 h^{2m+\beta}. \tag{6.2.26}$$

Причем, как и ранее, $B_j \in F^{k-j}$, $b_j \in G^{k-j}$, $v_k \in \Phi^{m-k}$, притом B_j, b_j, v_k, c_6, c_7, c_8 не зависят от h, а $\beta > 0$ и не зависит от h, k, u.

Эта теорема доказывается аналогично предыдущей.

Проверим выполнение условий теоремы для задачи (6.1.1), (6.1.2). В ходе изложения этого раздела мы уже выясняли справедливость условий A, B. Остается проверить условие D. Из формулы Тейлора для функции $u \in C^{2k+3}[0,1]$ вытекает разложение

$$\begin{aligned} &\frac{u(x+h/2) - u(x-h/2)}{h} + a(x)\frac{u(x+h/2) + u(x-h/2)}{2} = \\ &= u'(x) + a(x)u(x) + \sum_{j=1}^{k} h^{2j} \left(\frac{u^{(2j+1)}(x)}{(2j+1)!2^{2j}} + \frac{a(x)u^{(2j)}(x)}{(2j)!2^{2j}} \right) + \sigma^h(x), \end{aligned}$$

где

$$\begin{aligned} |\sigma^h(x)| \leqslant h^{2k+2} \Bigg(&\frac{1}{(2k+3)!2^{2k+2}} \max_{x\in[0,1]} |u^{(2k+3)}(x)| + \\ &+ \frac{1}{(2k+2)!2^{2k+2}} \max_{x\in[0,1]} |a(x)| \cdot \max_{x\in[0,1]} |u^{(2k+2)}(x)| \Bigg). \end{aligned}$$

Таким образом, если положить $F^k = C^{2k+2}[0,1]$, $\Phi^k = C^{2k+3}[0,1]$, то первое из разложений (6.2.23) будет выполнено с константой $\beta = 2$ и функциями

$$B_j(x) = \frac{u^{(2j+1)}(x)}{(2j+1)!2^{2j}} + \frac{a(x)u^{(2j)}(x)}{(2j)!2^{2j}}, \quad j = 1, 2, \ldots, k.$$

Поскольку начальное условие (6.1.2) аппроксимируется точно, т. е. $a^h u(0) = au(0) = u(0)$, то второе разложение в (6.2.23) также будет иметь место, если положить $b_j = 0$ ($j = 1, \ldots, k$), $\rho^h = 0$. Итак, условие D выполнено и, следовательно, справедливы разложение (6.2.25) и оценка (6.2.26).

6.2.2. Ускорение сходимости

Рассмотрим некоторые применения разложений, полученных в предыдущем разделе. Будем считать, что в пространстве Φ_h введена равномерная норма

$$\|v\|_{\Phi_h} = \max_{x \in \bar{D}_h} |v(x)|. \tag{6.2.27}$$

Пусть для задачи (6.2.1) выполнено условие A с целым $m \geqslant 1$, и предположим, что x — общая точка $(m+1)$ разностных сеток $\bar{D}_{h_i}$ с шагами h_1, h_2, …, h_{m+1}. На каждой из этих сеток построим приближенную задачу

$$\begin{aligned} &A^{h_i}\varphi^{h_i} = f \text{ на } D_{h_i}, \\ &a^{h_i}\varphi^{h_i} = g \text{ на } \partial D_{h_i}, \\ &i = 1, 2, \ldots, m+1. \end{aligned} \tag{6.2.28}$$

При выполнении условия B решение φ^{h_i} этой задачи существует и единственно. Таким образом, в точке x имеется $(m+1)$ приближенных решений φ^{h_i}. Причем, если выполнено условие C, то для каждого из них справедливо разложение

$$\varphi^{h_i} = \varphi + \sum_{k=1}^{m} h_i^k v_k + \eta^{h_i}, \quad x \in \bigcap_{i=1}^{m+1} \bar{D}_{h_i}, \tag{6.2.29}$$

и оценка

$$|\eta^{h_i}(x)| \leqslant c_5 h_i^{m+\beta}. \tag{6.2.30}$$

Рассмотрим систему

$$\begin{aligned} \sum_{i=1}^{m+1} \gamma_i &= 1, \\ \sum_{i=1}^{m+1} \gamma_i h_i^k &= 0, \quad k = 1, 2, \ldots, m. \end{aligned} \tag{6.2.31}$$

Определитель матрицы этой системы является определителем Вандермонда:

$$V(h_1, h_2, \ldots, h_{m+1}) = \begin{vmatrix} 1 & 1 & \ldots & 1 \\ h_1 & h_2 & \ldots & h_{m+1} \\ \ldots & \ldots & \ldots & \ldots \\ h_1^m & h_2^m & \ldots & h_{m+1}^m \end{vmatrix}.$$

Как известно,

$$V(h_1, h_2, \ldots, h_{m+1}) = \prod_{1 \leqslant i < j \leqslant m+1} (h_j - h_i).$$

Отсюда видно, что система (6.2.31) невырождена, если все h_i попарно различны. Предположим теперь, что h_i занумерованы по возрастанию и

$$\frac{h_{i+1}}{h_i} \geqslant 1 + c_9 \tag{6.2.32}$$

с константой $c_9 > 0$. Применим метод Крамера для решения системы (6.2.31). Имеем

$$\gamma_i = \frac{V(h_1, \ldots, h_{i-1}, 0, h_{i+1}, \ldots, h_{m+1})}{V(h_1, \ldots, h_{i-1}, h_i, h_{i+1}, \ldots, h_{m+1})},$$

откуда

$$\gamma_i = \prod_{1 \leqslant k < i} \frac{-h_k}{h_i - h_k} \cdot \prod_{i < k \leqslant m+1} \frac{h_k}{h_k - h_i}. \tag{6.2.33}$$

Из условия (6.2.32) следует, что

$$\frac{h_k}{h_i - h_k} = \frac{1}{\dfrac{h_i}{h_k} - 1} \leqslant \frac{1}{c_9}, \qquad \text{если } k < i,$$

$$\frac{h_k}{h_k - h_i} = \frac{1}{1 - \dfrac{h_i}{h_k}} \leqslant \frac{1}{1 - \dfrac{1}{1 + c_9}} \leqslant 1 + \frac{1}{c_9}, \quad \text{если } i < k.$$

Используя эти равенства, из соотношения (6.2.33) получаем

$$|\gamma_i| \leqslant \left(1 + \frac{1}{c_9}\right)^m. \tag{6.2.34}$$

Эта оценка указывает на ограниченность величин $|\gamma_i|$.

Составим линейную комбинацию функций φ^{h_i} с полученными весами γ_i

$$\bar{\varphi}(x) = \sum_{i=1}^{m+1} \gamma_i \varphi^{h_i}(x) \tag{6.2.35}$$

и докажем, что полученное решение $\bar{\varphi}(x)$ имеет точность порядка $h_{m+1}^{m+\beta}$. Подставим разложения (6.2.29) в правую часть формулы (6.2.35). Получим

$$\bar{\varphi}(x) = \varphi(x) \sum_{i=1}^{m+1} \gamma_i + \sum_{k=1}^{m} \left(v_k(x) \sum_{i=1}^{m+1} h_i^k \gamma_i \right) + \sum_{i=1}^{m+1} \gamma_i \eta^{h_i}(x).$$

Ввиду того, что γ_i являются решением системы (6.2.31), это равенство упрощается:

$$\bar{\varphi}(x) = \varphi(x) + \sum_{i=1}^{m+1} \gamma_i \eta^{h_i}(x).$$

Привлекая оценку (6.2.30) для η^{h_i} и оценку (6.2.34) для γ_i, получим, что

$$|\bar{\varphi}(x) - \varphi(x)| \leqslant \sum_{i=1}^{m+1} c_5 h_i^{m+\beta} \left(1 + \frac{1}{c_9}\right)^m \leqslant c_5(m+1) \left(1 + \frac{1}{c_9}\right)^m h_{m+1}^{m+\beta}. \qquad (6.2.36)$$

Таким образом, при выполнении условий A, B, C и неравенства (6.2.32) линейная комбинация $\bar{\varphi}(x)$ приближает точное решение $\varphi(x)$ с порядком $h_{m+1}^{m+\beta}$.

Остановимся теперь на выборе последовательности параметров h_i. Чаще всего используются два способа. Один из них заключается в том, что последовательность сеток $\bar{D}_{h_i}$ берется с набором параметров $h_k = h/k$, где $h > 0$, $k = 1, 2, \ldots, m+1$. В этом случае условие (6.2.32) выполняется с константой $c_9 = 1/m$ для любого $h > 0$, причем из формулы (6.2.33) следует, что

$$\gamma_k = \frac{(-1)^{m-k+1} k^{m+1}}{k!(m-k+1)!}, \quad k = 1, 2, \ldots, m+1. \qquad (6.2.37)$$

При втором способе параметры выбираются по правилу $h_k = h/2^{k-1}$, где $h > 0$, $k = 1, 2, \ldots, m+1$. В этом случае условие (6.2.32) выполняется для любого $h > 0$ с константой $c_9 = 1$. Формула (6.2.33) дает явное выражение для весов γ_k, но их вычисление затруднительно для больших m. Поэтому для вычисления суммы

$$\sum_{i=1}^{m+1} \gamma_i \varphi^{h_i}(x),$$

стоящей в правой части равенства (6.2.35), используется правило Ромберга. Оно заключается в последовательном вычислении величин

$$T_i^{(1)} = 2\varphi^{h_{i+1}(x)} - \varphi^{h_i}(x), \quad i = 1, 2, \ldots, m.$$

Затем осуществляется рекуррентный процесс

$$T_i^{(k)} = \frac{2^k T_{i+1}^{(k-1)} - T_i^{(k-1)}}{2^k - 1}, \quad i = 1, 2, \ldots, m-k+1,\ k = 2, \ldots, m.$$

В результате этих вычислений мы приходим к следующей величине:

$$T_1^{(m)} = \sum_{i=1}^{m+1} \gamma_i \varphi^{h_i}(x).$$

Сформулируем теперь аналогичные результаты для случая, когда регулярная часть разложения содержит лишь четные степени h, т. е. вместо условия C выполняется условие D. Тогда для решений φ^{h_i} задачи (6.2.28) выполняется разложение

$$\varphi^{h_i} = \varphi + \sum_{k=1}^{m} h_i^{2k} v_k + \eta^{h_i}, \quad x \in \bigcup_{i=1}^{m+1} \bar{D}_{h_i}, \tag{6.2.38}$$

и оценка

$$|\eta^{h_i}(x)| \leqslant c_8 h_i^{2m+\beta}. \tag{6.2.39}$$

Веса μ_i найдем из системы

$$\begin{gathered} \sum_{i=1}^{m+1} \mu_i = 1, \\ \sum_{i=1}^{m+1} \mu_i h_i^{2k} = 0, \quad k = 1, 2, \ldots, m. \end{gathered} \tag{6.2.40}$$

Определитель матрицы этой системы связан с определителем Вандермонда следующей формулой:

$$V(h_1^2, h_2^2, \ldots, h_{m+1}^2) = \prod_{1 \leqslant i < j \leqslant m+1} (h_j^2 - h_i^2).$$

Поэтому система (6.2.40) невырождена, если все h_i попарно различны. Ее решение находится по методу Крамера:

$$\mu_i = \frac{V(h_1^2, \ldots, h_{i-1}^2, 0, h_{i+1}^2, \ldots, h_{m+1}^2)}{V(h_1^2, \ldots, h_{i-1}^2, h_i^2, h_{i+1}^2, \ldots, h_{m+1}^2)},$$

откуда следует, что

$$\mu_i = \prod_{1 \leqslant k < i} \frac{-h_k^2}{h_i^2 - h_k^2} \prod_{i < k \leqslant m+1} \frac{h_k^2}{h_k^2 - h_i^2}. \tag{6.2.41}$$

Если, кроме того, выполняется неравенство (6.2.35), то для μ_i справедлива оценка

$$|\mu_i| \leqslant \left(1 + \frac{1}{2c_9}\right)^m, \tag{6.2.42}$$

показывающая ограниченность величин $|\mu_i\|$.

Составим линейную комбинацию с найденными весами

$$\bar{\varphi}(x) = \sum_{i=1}^{m+1} \mu_i \varphi^{h_i}(x) \tag{6.2.43}$$

и докажем, что полученное решение $\bar{\varphi}(x)$ имеет точность порядка $h_{m+1}^{2m+\beta}$. Подставим разложение (6.2.38) в (6.2.43). Используя равенства (6.2.40), придем к соотношению

$$\bar{\varphi}(x) = \varphi(x) + \sum_{i=1}^{m+1} \mu_i \eta^{h_i}(x).$$

Привлекая оценки (6.2.39), (6.2.42), получаем неравенство

$$|\bar{\varphi}(x) - \varphi(x)| \leqslant c_8(m+1)\left(1 + \frac{1}{2c_9}\right)^m h_{m+1}^{2m+\beta}. \tag{6.2.44}$$

Следовательно, при выполнении A, B, D и неравенства (6.2.32) линейная комбинация $\bar{\varphi}(x)$ приближает точное решение $\varphi(x)$ с порядком $h_{m+1}^{2m+\beta}$.

Рассмотрим подробнее правило вычисления весов μ_i для двух специальных способов выбора параметров h_i. В том случае, когда $h_i = h/i$, где $h_i > 0$, $i = 1, 2, \ldots, m+1$, формула (6.2.41) упрощается и

$$\mu_i = 2\frac{(-1)^{m-i+1}\, i^{2m+2}}{(m+i+1)!(m-i+1)!}, \quad i = 1, 2, \ldots, m+1. \tag{6.2.45}$$

В частности, для задачи из § 6.1 мы брали веса $\mu_1 = -1/3$, $\mu_2 = 4/3$, поскольку m было равно единице.

При втором способе выбора $h_i = h/2^{i-1}$, где $h > 0$, $i = 1, 2, \ldots, m+1$, можно использовать как формулу (6.2.30), так и правило Ромберга. Для вычисления суммы

$$\sum_{i=1}^{m+1} \mu_i \varphi^{h_i}(x)$$

оно выглядит следующим образом. Вычислим сначала величины

$$K_i^{(1)} = \frac{4}{3}\varphi^{h_{i+1}}(x) - \frac{1}{3}\varphi^{h_i}(x),\ i = 1, 2, \ldots, m.$$

Затем подсчитаем последовательность значений

$$K_i^{(j)} = \frac{4^j K_{i+1}^{(j-1)} - K_i^{(j-1)}}{4^j - 1}, \quad i = 1, 2, \ldots, m - j + 1, \ k = 2, \ldots, m.$$

Итогом вычислений является следующая величина:

$$K_1^{(m)} = \sum_{i=1}^{m+1} \mu_i \varphi^{h_i}(x).$$

Вновь вернемся к задаче (6.1.1)—(6.1.3) и соответствующей ей разностной задаче (6.1.6), (6.1.7). Мы уже выяснили, что условие $f \in C^{2m+2}[0,1]$ гарантирует существование разложения вида (6.2.38), (6.2.39). Для решения u^{τ_i} задачи (6.1.6), (6.1.7) на сетке с шагом τ_i оно записывается в следующей форме:

$$u^{\tau_i}(j\tau) = u(j\tau) + \sum_{k=1}^{m} \tau_i^{2k} v_k(j\tau) + \eta^{\tau_i}(j\tau), \tag{6.2.46}$$

$$j = 1, 2, \ldots, M_i = 1/\tau_i,$$

причем

$$|\eta^{\tau_i}(j\tau)| \leqslant c_8 \tau_i^{2m+2}. \tag{6.2.47}$$

Следовательно, к этой задаче применимо правило уточнения, аналогичное формулам (6.2.40), (6.2.43).

Рассмотрим конкретную ситуацию, когда параметры разностной сетки выбираются равными $\tau, \tau/2, \ldots, \tau/m + 1$. Тогда правило уточнения имеет следующий вид:

$$\bar{u}(j\tau) = \sum_{i=1}^{m+1} \mu_i u^{\tau/i}(j\tau), \quad j = 0, 1, \ldots, M, \tag{6.2.48}$$

где μ_i определяется формулой (6.2.45). В итоге $\bar{u}(j\tau)$ приближает значение $u(j\tau)$ с точностью $O(\tau^{2m+2})$.

Легко заметить, что уточненное решение получается на сравнительно редкой сетке; в то же время у решений $u^{\tau/i}$ используется лишь небольшая часть значений. При другом выборе последовательности τ_i общих узлов нескольких сеток может оказаться меньше. Поэтому для вычисления приближенных значений с высокой точностью в произвольной точке области определения решения исходной задачи следует использовать интерполяцию. Рассмотрим простейшую ситуацию с применением интерполяционных многочленов Ньютона.

Пусть t — произвольная точка отрезка $[0,1]$. Зафиксируем произвольный номер i и рассмотрим сетку с шагом τ_i. На этой сетке возьмем $(2m+2)$

узлов, ближайших к t. Построим по значениям приближенного решения u^{τ_i} в этих узлах интерполяционный многочлен Ньютона степени $2m+1$, который обозначим через $P_i(x;u^{\tau_i})$. Из разложения (6.2.46) следует, что

$$P_i(x;u^{\tau_i}) = P_i(x;u) + \sum_{k=1}^{m} \tau_i^{2k} P_i(x,v_k) + P_i(x;\eta^{\tau_i}).$$

Поскольку гладкие функции интерполируются многочленом Ньютона с точностью $O(\tau_i^{2m+2})$, а его коэффициенты ограничены, то

$$P_i(t;u^{\tau_i}) = u(t) + \sum_{k=1}^{m} \tau_i^{2k} v_k(t) + O(\tau_i^{2m+2}).$$

Это разложение справедливо для $i = 1,2,\ldots,m+1$, поэтому оно дает возможность применить способ уточнения разностных решений

$$\bar{u}(t) = \sum_{i=1}^{m+1} \mu_i P_i(t;u^{\tau_i}) \tag{6.2.49}$$

с теми же весами μ_i из (6.2.40). В этом случае

$$|\bar{u}(t) - u(t)| \leqslant \sum_{i=1}^{m+1} c_{10}|\gamma_i|\tau_i^{2m+2}$$

с константой c_{10}, в которую входят константы интерполяционной формулы и погрешности решения в узлах сетки.

6.3. Простейшие интегральные уравнения

Излагаемые далее интегральные уравнения являются наиболее простым объектом для иллюстрации общих теорем § 6.2 ввиду отсутствия граничных условий.

6.3.1. Уравнение Фредгольма второго рода

Рассмотрим уравнение

$$\varphi(x) = \int_0^1 K(x,t)\varphi(t)\,dt + f(x), \quad x \in [0,1]. \tag{6.3.1}$$

Относительно функции f и ядра K предположим выполнение следующих условий:

$$f \in C^{2m+2}[0,1], \quad K \in C^{2m+2}([0,1]\times[0,1]), \tag{6.3.2}$$

$$x = \max_{x\in[0,1],\, t\in[0,1]} |K(x,t)| < 1 \tag{6.3.3}$$

с целым $m \geqslant 1$. Тогда решение φ этой задачи существует и единственно в классе $C^{2m+2}[0,1]$. Более того, справедливо условие А из 6.2.1 с классами $\Phi^k = F^k = C^{2k+2}[0,1]$.

Для приближенного решения уравнения (6.3.1) введем разностную сетку

$$x_{i+1/2} = (i+1/2)h, \quad i = 0,1,\ldots,N-1, \tag{6.3.4}$$

с шагом $h = 1/N$ и заменим интеграл в (6.3.1) квадратурной формулой средних прямоугольников. В итоге в узлах сетки получим систему линейных уравнений

$$\varphi^h(x_{i+1/2}) = \sum_{j=0}^{N-1} hK(x_{i+1/2}, x_{j+1/2})\varphi^h(x_{j+1/2}) + f(x_{i+1/2}), \tag{6.3.5}$$

$$i = 0,1,\ldots,N-1.$$

Предполагая, что существует хотя бы одно решение этой системы, выведем априорную оценку. Пусть максимальня по модулю компонента вектора φ^h имеет номер k. Тогда из (6.3.5) и (6.3.3) вытекает, что

$$|\varphi^h(x_{k+1/2})| \leqslant \sum_{j=0}^{N-1} h|K(x_{k+1/2}, x_{j+1/2})|\,|\varphi^h(x_{j+1/2})| + |f(x_{k+1/2})| \leqslant$$

$$\leqslant \kappa|\varphi^h(x_{k+1/2})| + \max_{0\leqslant j\leqslant N-1} |f(x_{j+1/2})|.$$

Отсюда следует, что

$$\max_{0\leqslant i\leqslant N-1} |\varphi^h(x_{i+1/2})| \leqslant \frac{1}{1-\kappa} \max_{0\leqslant i\leqslant N-1} |f(x_{i+1/2})|. \tag{6.3.6}$$

Эта оценка гарантирует как однозначную разрешимость системы (6.3.5), так и устойчивость ее решения. Тем самым мы установили справедливость условия В из 6.2.1, где

$$\|u\|_{\Phi_h} = \|u\|_{F_h} = \max_{0\leqslant i\leqslant N-1} |u(x_{i+1/2})|.$$

Для квадратурной формулы средних прямоугольников хорошо известен следующий результат. Если $u \in C^{2k}[0,1]$, то

$$\sum_{i=0}^{N-1} hu(x_{i+1/2}) = \int_0^1 u(x)\,dx - \sum_{j=1}^{k} h^{2j} \frac{1-2^{-2j+1}}{(2j)!} B_{2j} u^{(2j-1)}(x)|_{x=0}^{x=1} + h^{2k} \frac{B_{2k}}{(2k)!} u^{(2k)}(\xi), \tag{6.3.7}$$

где $\xi \in [0,1]$, $f(x)|_{x=0}^{x=1} = f(1) - f(0)$, B_j — числа Бернулли: $B_0 = 1$, $B_2 = 1/6$, $B_4 = -1/30$, $B_6 = -1/42, \ldots$. Это разложение можно получить, например, как сумму аналогичных разложений по отрезкам $[ih, (i+1)h]$, на которых применяется формула Тейлора для приведения всех значений функции u и ее производных к сумме значений в точке $x_{i+1/2}$. Обозначим через A и A^h операторы, действующие по формулам

$$Au(x) = u(x) - \int_0^1 K(x,t)u(t)\,dt, \quad x \in [0,1],$$

$$A^h u(x_{i+1/2}) = u(x_{i+1/2}) - \sum_{j=0}^{N-1} K(x_{i+1/2}, x_{j+1/2}) u(x_{j+1/2}) h,$$

$$i = 0, 1, \ldots, N-1.$$

Тогда для любой функции $u \in C^{2k+2}[0,1]$ из (6.3.7) вытекает разложение вида (6.2.12):

$$A^h u(x_{i+1/2}) = Au(x_{i+1/2}) + \sum_{j=1}^{K} h^{2j} g_j(x_{i+1/2}) + \sigma^h(x_{i+1/2}), \tag{6.3.8}$$

$$i = 0, 1, \ldots, N-1, \tag{6.3.9}$$

где

$$g_j(x) = \frac{1-2^{-2j+1}}{(2j)!} B_{2j} \frac{\partial^{2j-1}}{\partial t^{2j-1}} (K(x,t)u(t))|_{t=0}^{t=1}, \; x \in [0,1], \tag{6.3.10}$$

$$|\sigma^h(x_{i+1/2})| \leqslant c_1 h^{2m+2}. \tag{6.3.11}$$

Выполнение этих условий достаточно для справедливости разложения

$$\varphi^h(x_{i+1/2}) = \varphi(x_{i+1/2}) + \sum_{k=1}^{m} h^{2k} v_k(x_{i+1/2}) + \eta^h(x_{i+1/2}), \tag{6.3.12}$$

$$i = 0, 1, \ldots, N-1. \tag{6.3.13}$$

На его основе возможно использование метода уточнения из 6.2.2 для разложения с четными степенями параметра. Для этого построим разностные сетки с шагом h_i, равным $h, h/3, \ldots, h/(2m+1)$, и на каждой из них решим систему линейных уравнений (6.3.5). Все полученные решения φ^{h_i} ($i = = 1, 2, \ldots, m+1$) определены на сетке (6.3.4) с шагом h. Выберем μ_i как решение системы

$$\sum_{i=1}^{m+1} \mu_i = 1,$$
$$\sum_{i=1}^{m+1} \mu_i h_i^{2k} = 0, \quad k = 1, 2, \ldots, m+1.$$

Составим линейную комбинацию с полученными коэффициентами μ_i:

$$\bar{\varphi}(x_{i+1/2}) = \sum_{k=1}^{m+1} \mu_k \varphi^{h_k}(x_{i+1/2}), \quad i = 0, 1, \ldots, N-1.$$

На основании 6.2.2 имеет место оценка

$$|\varphi(x_{i+1/2}) - \bar{\varphi}(x_{i+1/2})| \leqslant c_2 h^{2m+2}, \quad i = 0, 1, \ldots, N-1,$$

где c_2 не зависит от h.

Отметим, что обычные способы деления шага $h, h/2, h/3, \ldots$, а также $h, h/2, h/4, \ldots$, вообще говоря, не дают ни одного общего узла у последовательности сеток вида (6.3.4).

6.3.2. Уравнение Вольтерра первого рода

Рассмотрим простейшее интегральное уравнение Вольтерра первого рода

$$\int_0^x K(x,t)\varphi(t)\,dt = f(x), \quad x \in [0,1]. \tag{6.3.14}$$

Относительно правой части f и ядра K предположим, что выполнены условия

$$f \in C^{2m+3}[0,1], \quad \frac{\partial K}{\partial x} \in C^{2m+3}(\bar{Q}), \tag{6.3.15}$$

где целое $m \geqslant 1$, Q — треугольник $0 \leqslant t \leqslant x \leqslant 1$. Предположим также, что

$$f(0) = 0, \ \min_{x \in [0,1]} |K(x,x)| = k_1 \neq 0. \tag{6.3.16}$$

Тогда существует единственное решение φ уравнения (6.3.14) из класса $C^{2m+2}[0,1]$.

Для приближенного решения уравнения (6.3.14) введем разностные сетки с целыми и средними узлами

$$x_i = ih, \qquad i = 0, 1, \ldots, N,$$
$$x_{i+1/2} = (i + 1/2)h, \qquad i = 0, 1, \ldots, N - 1,$$

и воспользуемся квадратурной формулой средних прямоугольников. В итоге приходим к равенствам

$$\sum_{j=0}^{i} hK(x_{i+1}, x_{j+1/2})\varphi^h(x_{j+1/2}) = f(x_{i+1}), \quad i = 0, 1, \ldots, N - 1. \tag{6.3.17}$$

Равенства (6.3.17) представляют собой систему линейных алгебраических уравнений относительно φ^h с треугольной матрицей. Из условия (6.3.16) следует, что для достаточно малых h диагональные элементы этой матрицы будут превышать по модулю некоторое положительное число, например

$$|hK(x_{i+1}, x_{i+1/2})| \geqslant hk_1/2, \tag{6.3.18}$$

и система (6.3.17) имеет единственное решение. Выведем для него априорную оценку. С этой целью найдем разность двух уравнений (6.3.17) с номерами i, $i - 1$ и преобразуем ее.

Имеем

$$K(x_{i+1}, x_{i+1/2})\varphi^h(x_{i+1/2}) =$$
$$= -\sum_{j=0}^{i-1}\{K(x_{i+1}, x_{j+1/2}) - K(x_i, x_{j+1/2})\}\varphi^h(x_{j+1/2}) + \frac{f(x_{i+1}) - f(x_i)}{h}.$$

Учитывая условия (6.3.16), (6.3.18), приходим к неравенству

$$|\varphi^h(x_{i+1/2})| \leqslant \frac{2}{k_1}\left(hk_2\sum_{j=0}^{i-1}|\varphi^h(x_{j+1/2})| + f_1\right), \tag{6.3.19}$$

где

$$k_2 = \max_{(x,t)\in\bar{Q}}\left|\frac{\partial K}{\partial x}(x, t)\right|, \; f_1 = \max_{0\leqslant j\leqslant N-1}\frac{|f(x_{i+1}) - f(x_i)|}{h}.$$

Методом математической индукции из (6.3.19) нетрудно показать, что

$$|\varphi^h(x_{i+1/2})| \leqslant \frac{2f_1}{k_1}\left(1 + \frac{2hk_2}{k_1}\right)^i.$$

Используя неравенства

$$1+\frac{2hk_2}{k_1}\leqslant e^{2hk_2/k_1},\quad ih\leqslant 1,$$

приходим к оценке

$$|\varphi^h(x_{i+1/2})|\leqslant\frac{2f_1}{k_1}e^{2k_2/k_1},\quad i=0,1,\ldots,N-1. \tag{6.3.20}$$

Если положить

$$\|\varphi\|_{\Phi_h}=\max_{0\leqslant i\leqslant N-1}|\varphi(x_{i+1/2})|,\ \|f\|_{F_h}=\max_{0\leqslant i\leqslant N-1}\frac{|f(x_{i+1})-f(x_i)|}{h},$$

то оценку (6.3.20) можно записать в виде

$$\|\varphi\|_{\Phi_h}\leqslant c_3\|f\|_{F_h}.$$

Осталось проверить выполнение условия C из § 6.2. Для этого возьмем классы гладкости

$$\Phi^k=C^{2k+2}[0,1],\ F^k=C^{2k+3}[0,1]$$

и положим

$$Au(x)=\int_0^x K(x,t)u(t)\,dt,\quad x\in[0,1],$$

$$A^hu(x_i)=\sum_{j=0}^{i}K(x_i,x_{j+1/2})u(x_{j+1/2})h,\quad i=1,2,\ldots,N.$$

Используя разложение вида (6.3.7) на отрезке $[0,ih]$ для функции $u\in C^{2k+2}[0,1]$, приходим к разложению

$$A^hu(x_i)=Au(x_i)+\sum_{j=1}^{k}h^{2j}g_j(x_i)+\sigma^h(x_i),\quad i=1,2,\ldots,N,$$

причем $g_j\in C^{2k-2j+3}[0,1]$ и

$$|\sigma^h(x_i)|\leqslant c_4h^{2k+2}.$$

Полагая $\sigma^h(0)=0$, получаем, что

$$\|\sigma^h\|_{F_h}\leqslant 2c_4h^{2k+1}.$$

Поэтому справедлива теорема из 6.2.1 о разложении решения системы (6.3.17) по четным степеням h с параметром $\beta=1$. Это разложение используется так же, как и в предыдущем разделе.

Выберем параметр h так, чтобы выполнялось условие (6.3.18), и построим разностные сетки с шагами h_i, равными $h, h/3, \ldots, h/(2m+1)$. На каждой из них найдем решения φ^{h_i} системы (6.3.17). В узлах сетки с шагом h составим линейную комбинацию

$$\bar{\varphi}(x_{i+1/2}) = \sum_{k=1}^{m+1} \mu_k \varphi^{h_k}(x_{i+1/2}), \quad i = 0, 1, \ldots, N-1, \tag{6.3.21}$$

с весами, определяемыми из системы

$$\sum_{i=1}^{m+1} \mu_i = 1,$$
$$\sum_{i=1}^{m+1} \mu_i h_i^{2k} = 0, \quad k = 1, 2, \ldots, m+1.$$

Тогда на основании теоремы из 6.2.2 имеет место оценка

$$|\varphi(x_{i+1/2}) - \bar{\varphi}(x_{i+1/2})| \leqslant c_5 h^{2m+1}, \quad i = 0, 1, \ldots, N-1.$$

Заметим, что достигнутый порядок точности оказался меньше, чем для уравнения Фредгольма второго рода при одинаковых требованиях к гладкости правой части.

6.4. Одномерное уравнение диффузии

В этом параграфе будет изучаться задача Дирихле для уравнения диффузии простейшего вида

$$-\frac{d^2u}{dx^2} + q(x)u = f(x), \quad x \in (0,1), \tag{6.4.1}$$
$$u(0) = 0, \quad u(1) = 0, \tag{6.4.2}$$

причем предполагается, что

$$q(x) \geqslant 0, \tag{6.4.3}$$

а гладкость функций q, f и, следовательно, u достаточна для проведения последующих выкладок.

Сначала для разностной схемы на последовательности сеток на основе результатов § 6.2 мы построим приближенные решения с точностью $O(h^{2k})$ для $k \geqslant 2$. Далее, на примере задачи (6.4.1), (6.4.2) рассмотрим модификацию экстраполяции к пределу, разработанную применительно к вариационно-разностному методу Галеркина.

6.4.1. Разностный метод

Построим разностную сетку

$$x_k = kh,\ k = 0, 1, \ldots, N, \tag{6.4.4}$$

с равномерным шагом $h = 1/N$ и перейдем от задачи (6.4.1), (6.4.2) к приближенной задаче

$$\begin{gathered}\frac{-u^h_{k-1} + 2u^h_k - u^h_{k+1}}{h^2} + q_k u^h_k = f_k, \quad k = 1, 2, \ldots, N-1, \\ u^h_0 = u^h_N = 0\end{gathered} \tag{6.4.5}$$

со вторым порядком аппроксимации.

Проверим для задач (6.4.1), (6.4.2) и (6.4.5) выполнение условий A, B, D из § 6.2. Условие A, характеризующее однозначную разрешимость задачи (6.4.1), (6.4.2), связано с гладкостью коэффициента q. Потребуем, чтобы q принадлежало классу $C^{2m+2}[0,1]$, где целое $m \geqslant 1$. Тогда для любой правой части $f \in C^{2m+2}[0,1]$ существует единственное решение $u \in C^{2m+4}[0,1]$, удовлетворяющее условию (6.4.2). Итак, для выполнения условия A можно положить $F^k = C^{2k+2}[0,1]$, а Φ^k — подмножество функций из $C^{2k+4}[0,1]$, обращающихся в нуль на концах отрезка $[0,1]$.

Для проверки условия B выведем априорную оценку. Умножим каждое из уравнений (6.4.5) на hu^h_k и сложим:

$$\frac{1}{h}\sum_{k=1}^{N}\{(u^h_k - u^h_{k-1})^2 + q_k(u^h_k)^2\} = \sum_{k=1}^{N-1} f_k u^h_k h.$$

Отбросив в левой части положительные слагаемые $q_k(u^h_k)^2$ и заменив f_k и u^h_k на большие, получим, что

$$\frac{1}{h}\sum_{k=1}^{N}(u^h_k - u^h_{k-1})^2 \leqslant \max_{1\leqslant k\leqslant N-1}|f_k| \cdot \max_{1\leqslant k\leqslant N-1}|u^h_k|. \tag{6.4.6}$$

Для оценки снизу привлечем соотношение, вытекающее из неравенства Коши — Буняковского для сеточной функции y, равной нулю в точке $x = 0$:

$$|y_k| = \left|\sum_{j=1}^{k}(y_j - y_{j-1})\right| \leqslant k^{1/2}\left(\sum_{j=1}^{k}(y_j - y_{j-1})^2\right)^{1/2} \leqslant N^{1/2}\left(\sum_{j=1}^{k}(y_j - y_{j-1})^2\right)^{1/2}.$$

Поскольку $u_0^h = 0$, то, полагая $y_k = u_k^h$, получим

$$\max_{1 \leqslant k \leqslant N-1} |u_k^h|^2 \leqslant \frac{1}{h} \sum_{k=1}^{N} (u_k^h - u_{k-1}^h)^2.$$

Объединим это неравенство с (6.4.6). Имеем

$$\max_{1 \leqslant k \leqslant N-1} |u_k^h|^2 \leqslant \max_{1 \leqslant k \leqslant N-1} |f_k| \cdot \max_{1 \leqslant k \leqslant N-1} |u_k^h|.$$

Отсюда следует оценка

$$\|u^h\|_{\Phi^h} \leqslant \|f\|_{F^h}, \tag{6.4.7}$$

где

$$\|u^h\|_{\Phi^h} = \max_{1 \leqslant k \leqslant N-1} |u_k^h|, \ \|f\|_{F_h} = \max_{1 \leqslant k \leqslant N-1} |f_k|.$$

Оценка (6.4.7) означает выполнение условия B. Покажем, что условие D также выполняется. В самом деле, пусть $u \in C^{2m+4}[0,1]$. Обозначим

$$Au(x) = -u''(x) + q(x)u(x), \quad x \in (0,1),$$
$$A^h u(x_k) = \frac{-u(x_{k-1}) + 2u(x_k) - u(x_{k+1})}{h^2} + q(x_k)u(x_k),$$
$$k = 1, 2, \ldots, N-1.$$

Используя формулу Тейлора

$$u(x_{k \pm 1}) = \sum_{r=0}^{2m+1} \frac{(\pm h)^r}{r!} u^{(r)}(x_k) + O(h^{2m+2}),$$

приходим к разложению

$$A^h u(x_k) = Au(x_k) + \sum_{j=0}^{m} h^{2j} g_j(x_k) + \sigma^h(x_k), \quad k = 1, 2, \ldots, N-1,$$

где

$$\|\sigma^h\|_{F_h} \leqslant c_2 h^{2m+2}.$$

Рассмотрим теперь метод повышения точности разностных решений задачи (6.4.5). Пусть x — общая точка $(m+1)$ разностных сеток с шагами $h_1 > h_2 > \cdots > h_{m+1}$, причем

$$\frac{h_j}{h_j + 1} \geqslant 1 + c_3, \quad c_3 > 0.$$

Построим на каждой сетке приближенную задачу (6.4.5) и найдем ее решение u^{h_j}, например, методом прогонки. Таким образом, в точке x имеется $(m+1)$ приближенных значений $u^{h_j}(x)$. Пусть коэффициенты μ_j являются решением системы

$$\sum_{j=1}^{m+1} \mu_j = 1, \quad \sum_{j=1}^{m+1} h_j^{2k} \mu_j = 0, \quad k = 1, 2, \ldots, m+1.$$

Составим линейную комбинацию

$$\tilde{u}(x) = \sum_{j=1}^{m+1} \mu_j u^{h_j}(x).$$

На основании теоремы из 6.2.2 для этого решения справедлива оценка

$$|u(x) - \bar{u}(x)| \leqslant c_4 h^{2m+2}.$$

Изложенная здесь методика легко обобщается на третью краевую задачу, а также на уравнения с переменным коэффициентом

$$-(ku')' + qu = f,$$

включая случай разрывного коэффициента k. В последней ситуации необходимо проделать кусочно-линейное преобразование независимой переменной так, чтобы впоследствии точки разрыва были целыми узлами каждой из разностных сеток с шагами $h_1, h_2, \ldots, h_{m+1}$.

6.4.2. Метод Галеркина

В отличие от вариационно-разностных методов высокой точности, основанных на применении большего, чем обычно, числа базисных функций высокой гладкости (пример таких базисных функций приведен в § 2.4), здесь излагается метод повышения точности приближенных решений, опирающийся на линейное комбинирование решений вариационно-разностных задач с разными шагами сетки при использовании только кусочно-линейных функций.

В целях построения приближенной задачи разобьем интервал на $[0, 1]$ на N равных частей длины $h = 1/N$ точками

$$x_k = kh, \ k = 0, 1, \ldots, N, \tag{6.4.8}$$

и построим базисные функции ω_k из 2.3.2, которые имеют вид

$$\omega_k(x) = \begin{cases} 0, & \text{если } x \in (-\infty, x_{k-1}], \\ 1 + \dfrac{x - x_k}{h}, & \text{если } x \in (x_{k-1}, x_k], \\ 1 - \dfrac{x - x_k}{h}, & \text{если } x \in (x_k, x_{k+1}), \\ 0, & \text{если } x \in [x_{k+1}, +\infty), \end{cases}$$
$$k = 1, 2, \ldots, N-1.$$

Умножим уравнение (6.4.1) на каждую из базисных функций и проинтегрируем полученные равенства. Если функция $u(x)$ — решение задачи (6.4.1), (6.4.2), то в результате получим тождества

$$-\int\limits_{x_{k-1}}^{x_{k+1}} \left(\frac{d^2u}{dx^2}\omega_k + qu\omega_k \right) dx = \int\limits_{x_{k-1}}^{x_{k+1}} f\omega_k \, dx, \quad k = 1, 2, \ldots, N-1. \tag{6.4.9}$$

Преобразуем их, проводя интегрирование по частям и используя условие $\omega_k(x_{k\pm1}) = 0$:

$$\int\limits_{x_{k-1}}^{x_{k+1}} \left(\frac{du}{dx}\frac{d\omega_k}{dx} + qu\omega_k \right) dx = \int\limits_{x_{k-1}}^{x_{k+1}} f\omega_k \, dx.$$

Введем некоторые обозначения, упрощающие дальнейшие выкладки:

$$(v, w) = \int\limits_0^1 v(x)w(x) \, dx, \quad [v, w] = \int\limits_0^1 \left(\frac{dv}{dx}\frac{dw}{dx} + qvw \right) dx.$$

С их помощью предыдущие равенства записываются в виде

$$[u, \omega_k] = (f, \omega_k), \quad i = 1, 2, \ldots, N-1. \tag{6.4.10}$$

Основываясь на этих тождествах, будем искать приближенное решение $u^h(x)$ в виде

$$u^h(x) = \sum_{l=1}^{N-1} \alpha_i^h \omega_l(x) \tag{6.4.11}$$

с некоторым набором констант α_l^h, определяемых из равенств, полученных подстановкой $u^h(x)$ в (6.4.10) вместо $u(x)$:

$$\sum_{l=1}^{N-1} \alpha_l^h [\omega_l, \omega_k] = (f, \omega_k), \quad k = 1, 2, \ldots, N-1. \tag{6.4.12}$$

Система (6.4.12) для наглядности может быть записана в матричном виде

$$A\alpha \equiv \begin{Vmatrix} b_1 & c_1 & 0 & \dots & 0 & 0 \\ a_2 & b_2 & c_2 & \dots & 0 & 0 \\ 0 & a_3 & b_3 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & b_{N-2} & c_{N-2} \\ 0 & 0 & 0 & \dots & a_{N-1} & b_{N-1} \end{Vmatrix} \begin{Vmatrix} \alpha_1^h \\ \alpha_2^h \\ \alpha_3^h \\ \dots \\ \alpha_{N-2}^h \\ \alpha_{N-1}^h \end{Vmatrix} = \begin{Vmatrix} g_1 \\ g_2 \\ g_3 \\ \dots \\ g_{N-2} \\ g_{N-1} \end{Vmatrix} \tag{6.4.13}$$

с элементами

$$a_k = -\frac{1}{h} + \int\limits_{x_{k-1}}^{x_k} \left(1 + \frac{x - x_k}{h}\right) \frac{x_k - x}{h} q\, dx = [\omega_{k-1}, \omega_k],$$

$$b_k = \frac{2}{h} + \int\limits_{x_{k-1}}^{x_k} \left(1 + \frac{x - x_k}{h}\right)^2 q\, dx + \int\limits_{x_k}^{x_{k+1}} \left(1 - \frac{x - x_k}{h}\right)^2 q\, dx = [\omega_k, \omega_k],$$

$$c_k = a_{k+1} = [\omega_{k+1}, \omega_k],$$

$$g_k = \int\limits_{x_{k-1}}^{x_k} \left(1 + \frac{x - x_k}{h}\right) f\, dx + \int\limits_{x_k}^{x_{k+1}} \left(1 - \frac{x - x_k}{h}\right) f\, dx = (f, \omega_k).$$

Очевидно, что функция, не являющаяся полиномом степени меньше 2, не может быть приближена равномерно на отрезке $[0, 1]$ с точностью более $O(h^2)$ кусочно-линейными функциями с базисом из $\omega_i^h(x)$. Поэтому трансформируем гипотезу Ричардсона в следующую: для коэффициентов α_k^h при $h \to 0$ имеет место разложение

$$\alpha_k^h = u(kh) + h^2 v(kh) + \eta_k^h, \quad k = 0, 1, \dots, N, \tag{6.4.14}$$

где $u(x)$ — решение задачи (6.4.1), (6.4.2); $v(x)$ — некоторая гладкая, не зависящая от h функция; η_k^h — сеточная функция, модуль которой равномерно по k ограничен величиной порядка $O(h^4)$:

$$|\eta_k^h| \leqslant C_1 h^4 \tag{6.4.15}$$

с константой C, не зависящей от h и k.

Использование общих результатов § 6.2 без учета специфики метода Галеркина приводит к значительно завышенным требованиям гладкости функций q, f. Поэтому мы приведем далее обоснование разложения (6.4.14), специально ориентированное на метод конечных элементов и дающее четвертый порядок точности с более слабыми требованиями гладкости коэффициентов уравнения (6.4.1), чем в разностном методе.

Обоснование разложения (6.4.14) проведем, как в 6.1, в три этапа.

Сначала мы приведем некоторые рассуждения, поясняющие выбор функции $v(x)$, найдем эту функцию, а затем докажем оценку (6.4.15) для остаточного члена.

Итак, если соотношение (6.4.14) имеет место, то, выделяя во всех выкладках только два главных члена, получим

$$\alpha_k^h = u(x_k) + h^2 v(x_k) + O(h^4). \tag{6.4.16}$$

Подставим это выражение в систему (6.4.13):

$$\begin{gathered} a_k u(x_{k-1}) + b_k u(x_k) + c_k u(x_{k+1}) + \\ + h^2(a_k v(x_{k-1}) + b_k v(x_k) + c_k v(x_{k+1})) + O(h^4) = g_k, \\ k = 1, 2, \ldots, N-1, \end{gathered} \tag{6.4.17}$$

где для общности положено $a_1 = c_{N-1} = 0$. Величина остаточного члена $O(h^4)$ не столь очевидна. Однако дальше мы покажем, что на самом деле эта величина еще меньше.

По значениям функций $u(x)$ и $v(x)$ в узлах разностной сетки построим непрерывные функции

$$\tilde{u}(x) = \sum_{l=1}^{N-1} u(x_l)\omega_l(x) \text{ и } \tilde{v}(x) = \sum_{l=1}^{N-1} v(x_l)\omega_l(x),$$

называемые далее кусочно-линейными восполнениями $u(x)$ и $v(x)$ соответственно. Вспоминая вид коэффициентов a_k, b_k, c_k, g_k, приходим вместо (6.4.17) к равенствам

$$[\tilde{u}, \omega_k] + h^2[\tilde{v}, \omega_k] + O(h^4) = (f, \omega_k), \ k = 1, 2, \ldots, N-1. \tag{6.4.18}$$

Интегрированием по частям легко выводятся равенства

$$\begin{gathered} -\frac{1}{h}u(x_{k-1}) + \frac{2}{h}u(x_k) - \frac{1}{h}u(x_{k+1}) = \int_0^1 \frac{du}{dx}\frac{d\omega_k}{dx}\,dx = \int_0^1 \frac{d\tilde{u}}{dx}\frac{d\omega_k}{dx}\,dx, \\ k = 1, 2, \ldots, N-1. \end{gathered} \tag{6.4.19}$$

Используем теперь разложение Тейлора для функций $u(x)$ и $q(x)$:

$$\begin{aligned} u(x) &= u(x_k) + (x - x_k)\frac{du}{dx}(x_k) + \frac{(x-x_k)^2}{2}\frac{d^2u}{dx^2}(x_k) + O(h^3), \\ q(x) &= q(x_k) + (x - x_k)\frac{dq}{dx}(x_k) + \frac{(x-x_k)^2}{2}\frac{d^2q}{dx^2}(x_k) + O(h^3) \end{aligned} \tag{6.4.20}$$

на каждом интервале (x_{k-1}, x_{k+1}) для доказательства равенства

$$(q\tilde{u}, \omega_k) = (qu, \omega_k) + h^2(w, \omega_k) + O(h^4), \tag{6.4.21}$$

где

$$w(x) = -\frac{1}{12}q(x)\frac{d^2u}{dx^2}(x).$$

Так как

$$\tilde{u}(x) = \begin{cases} \frac{1}{h}(u(x_k)(x - x_{k-1}) + u(x_{k-1})(x_k - x)), & x \in (x_{k-1}, x_k), \\ \frac{1}{h}(u(x_k)(x_{k+1} - x) + u(x_{k+1})(x - x_k)), & x \in (x_k, x_{k+1}), \end{cases}$$

то

$$\tilde{u}(x) = \begin{cases} u(x_k) + (x_k - x)u'(x_k) + (x_k - x)\frac{h}{2}u''(x_k) + O(h^3) \text{ для } x \in (x_{k-1}, x_k), \\ u(x_k) + (x - x_k)u'(x_k) + (x - x_k)\frac{h}{2}u''(x_k) + O(h^3) \text{ для } x \in (x_k, x_{k+1}). \end{cases}$$

Поэтому, используя последнее равенство и разложения (6.4.20) и проводя интегрирование полиномов под интегралами, получим

$$(q\tilde{u}, \omega_k) = hu(x_k)q(x_k) - h^3\left(\frac{1}{12}u(x_k)q''(x_k) + \frac{1}{12}u'(x_k)q(x_k) + \frac{1}{12}u''(x_k)q(x_k)\right) + O(h^4),$$

$$(qu, \omega_k) = hu(x_k)q(x_k) - h^3\frac{1}{12}(u(x_k)q(x_k))'' + O(h^4),$$

$$h^2(w, \omega_k) = h^3w(x_k) + O(h^4).$$

Сопоставление коэффициентов при членах порядка h и h^2 и доказывает равенство (6.4.21).

Объединяя равенства (6.4.2) и (6.4.21), получим

$$[\tilde{u}, \omega_k] = [u, \omega_k] + h^2(w, \omega_k) + O(h^4). \tag{6.4.22}$$

Проведение этих выкладок для функции $v(x)$ с меньшим числом членов разложения (6.4.20) дает соотношение

$$h^2(\tilde{v}, \omega_k) = h^2(v, \omega_k) + O(h^4).$$

Привлекая для преобразования этого равенства тождество (6.4.2), справедливое для $v(x)$, имеем

$$h^2[\tilde{v}, \omega_k] = h^2[v, \omega_k] + O(h^4). \tag{6.4.23}$$

Наконец, замена слагаемых в левой части (6.4.18) с помощью (6.4.22) и (6.4.23) имеет своим следствием равенства

$$[u,\omega_k]+h^2(w,\omega_k)+h^2[v,\omega_k]+O(h^4)=(f,\omega_k), \tag{6.4.24}$$

$$k=1,2,\ldots,N-1. \tag{6.4.25}$$

Так как мы хотим выполнение тождества для всех h, то потребуем совпадения коэффициентов при главных степенях.

Сравнение коэффициентов для степеней h^0 и h^2 дает равенства

$$[u,\omega_k]=(f,\omega_k), \tag{6.4.26}$$

$$[v,\omega_k]=-(w,\omega_k),\quad k=1,2,\ldots,N-1. \tag{6.4.27}$$

Этими равенствами построение необходимых условий заканчивается. Первая группа этих равенств совпадает с (6.4.10), являющимся следствием уравнения (6.4.1). Для выполнения последующих равенств достаточно взять в качестве $v(x)$ решение краевой задачи

$$-\frac{d^2v}{dx^2}+qv=-w(x), \tag{6.4.28}$$

$$v(0)=v(1)=0. \tag{6.4.29}$$

Таким образом, построение уравнения, определяющего функцию $v(x)$, можно считать законченным. Остается доказать ограниченность (6.4.15) сеточной функции, определяемой равенствами

$$\eta_k^h=\alpha_k^h-u(x_k)-h^2v(x_k),\quad k=0,1,\ldots,N. \tag{6.4.30}$$

Для доказательства построим по значениям η_l^h в узлах разностной сетки кусочно-линейное восполнение

$$\tilde{\eta}^h(x)=\sum_{l=1}^{N-1}\eta_l^h\omega_l(x).$$

Тогда (6.4.30) перепишется в виде

$$\tilde{\eta}^h(x)=u^h(x)-\tilde{u}(x)-h^2\tilde{v}(x),$$

что влечет равенство интегралов

$$[\tilde{\eta}(x),\omega_k]=[u^h,\omega_k]-[\tilde{u},\omega_k]-h^2[\tilde{v},\omega_k]. \tag{6.4.31}$$

Полагая, что функция u имеет непрерывные производные на $[0,1]$ вплоть до четвертого порядка, получим разложение Тейлора для функции u:

$$u(x) = \sum_{l=0}^{3} \frac{d^l u}{dx^l}(x_k)\frac{(x-x_k)^l}{l!} + \frac{(x-x_k)^4}{4!}\frac{d^4 u}{dx^4}(\xi_k^x), \quad x \in (x_{k-1}, x_{k+1}),$$

и аналогичные разложения для функций q и v, которые короче на два члена:

$$v(x) = v(x_k) + (x-x_k)\frac{dv}{dx}(x_k) + \frac{(x-x_k)^2}{2}\frac{d^2 v}{dx^2}(\rho_k^x), \quad x \in (x_{k-1}, x_{k+1}),$$

где точки ξ_k^x и ρ_k^x лежат в интервале (x_{k-1}, x_{k+1}). С указанными разложениями повторим выкладки, проведенные ранее для получения необходимых условий. Тогда равенство (6.4.31) приводится к следующему виду:

$$[\tilde{\eta}, \omega_k] = \sigma_k, \tag{6.4.32}$$

причем $|\sigma_k| \leqslant C_2 h^5$; последнее обусловлено сокращением всех членов порядка h и h^3 ввиду выбора функции $v(x)$. Напомним, что $v(x)$ является решением задачи (6.4.28), (6.4.29).

Умножим каждое из равенств (6.4.32) на η_k^h и сложим по всем k:

$$[\tilde{\eta}^h, \tilde{\eta}^h] = \sum_{k=1}^{N-1} \sigma_k \eta_k^h.$$

Заменим, далее, значения σ_k и η_k^h на максимальные. Тогда будем иметь

$$[\tilde{\eta}^h, \tilde{\eta}^h] \leqslant N C_2 h^5 \max_{1 \leqslant k \leqslant N-1} |\eta_k^h| = C_2 h^4 \max_{0 \leqslant x \leqslant 1} |\tilde{\eta}^h(x)|. \tag{6.4.33}$$

Теперь получим оценку снизу. Для этого привлечем неравенство, являющееся следствием неравенства Коши — Буняковского для функции $y(x)$, равной нулю в точке $x = 0$:

$$|y(x)| = \Big| \int_0^x y'(t)\,dt \Big| \leqslant \left(\int_0^x 1\,dt \right)^{1/2} \left(\int_0^x (y'(t))^2\,dt \right)^{1/2} \leqslant$$
$$\leqslant \sqrt{x} \left(\int_0^x (y'(t))^2\,dt \right)^{1/2} \leqslant \left(\int_0^x \{(y'(t))^2 + q(y(t))^2\}\,dt \right)^{1/2}.$$

Так как $\tilde{\eta}^h(x)$ равна нулю в точке $x = 0$, то

$$|\tilde{\eta}^h(x)|^2 \leqslant [\tilde{\eta}^h, \tilde{\eta}^h], \quad x \in [0,1],$$

и из оценки (6.4.33) получаем

$$\max_{0\leqslant x\leqslant 1}|\tilde{\eta}^h(x)|^2 \leqslant [\tilde{\eta}^h,\tilde{\eta}^h] \leqslant C_2h^4 \max_{0\leqslant x\leqslant 1}|\tilde{\eta}^h(x)|.$$

Таким образом,

$$\max_{0\leqslant x\leqslant 1}|\tilde{\eta}^h(x)| = \max_{1\leqslant l\leqslant N}|\tilde{\eta}^h_l(x)| \leqslant C_2h^4, \tag{6.4.34}$$

что и заключает обоснование разложения (6.4.14). Дальнейшие рассуждения аналогичны тем, которые были приведены в § 6.1, а именно: построим две равномерные сетки с шагами h и $h/2$ и найдем решения α^h и $\alpha^{h/2}$ двух задач (6.4.12), соответствующих этим сеткам. Если $x_k = kh$ — точка сетки, то решение с четвертым порядком точности строится в виде линейной комбинации

$$\bar{u}_k = \frac{4}{3}\alpha^{h/2}_{2k} - \frac{1}{3}\alpha^h_k.$$

Доказательство практически ничем не отличается от того, которое применяется в случае обыкновенных дифференциальных уравнений первого порядка, и дает оценку

$$|\bar{u}_k - u(x_k)| \leqslant \frac{5}{12}C_2h^4.$$

Описанный метод обобщается на случай других краевых задач. Кроме того, он может быть применен также и для решения краевых задач для квазилинейных уравнений и уравнений с переменными коэффициентами.

6.5. Нестационарные задачи

В этом параграфе мы сначала рассмотрим повышение точности разностного решения для уравнения теплопроводности с одной пространственной переменной и на примере этого уравнения изучим учет условий согласования, возникающих в нестационарных задачах при сопоставлении начальных и краевых условий. Во второй части мы применим общие результаты о повышении точности разностных решений на последовательности сеток к методу расщепления для системы дифференциальных уравнений, в частности, для системы, полученной в результате замены пространственных производных разностными операторами.

6.5.1. Уравнение теплопроводности

Рассмотрим задачу об остывании нагретого стержня

$$\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial x^2}, \quad x \in [0,1], \quad t \in [0,T], \tag{6.5.1}$$

имеющего известное начальное распределение температуры

$$u(x,0) = u_0(x), \quad x \in [0,1]; \tag{6.5.2}$$

на концах стрежня поддерживается нулевая температура:

$$u(0,t) = u(1,t) = 0, \quad t \in (0,T]. \tag{6.5.3}$$

Пусть

$$u_0 \in C^\infty[0,1]. \tag{6.5.4}$$

Тогда в каждой точке области $Q = (0,1) \times (0,T)$ будут существовать непрерывные частные производные решения u любого порядка. Однако для непрерывности на замыкании $\bar{Q}$ необходимо потребовать выполнение дополнительных условий. Если решение $u(x,t)$ непрерывно на $\bar{Q}$, то из сопоставления равенств (6.5.2), (6.5.3) в углах $(0,0)$, $(0,1)$ прямоугольника $\bar{Q}$ следует, что

$$u_0(0) = 0, \quad u_0(1) = 0.$$

Эти равенства мы будем называть условиями согласования порядка 0. Если на $\bar{Q}$ непрерывны производные $\partial u/\partial t$, $\partial^2 u/\partial x^2$, то из уравнения (6.5.1)—(6.5.3) в углах $(0,0)$, $(0,1)$ прямоугольника вытекают условия согласования порядка 1:

$$\frac{d^2 u_0}{dx^2}(0) = 0, \quad \frac{d^2 u_0}{dx^2}(1) = 0.$$

Дифференцируя уравнения (6.5.1) в предположении непрерывности на $\bar{Q}$ всех получающихся производных, получим равенства

$$\frac{\partial^k u}{\partial t^k} = a^{2k} \frac{\partial^{2k} u}{\partial x^{2k}} \text{ на } \bar{Q}, \quad k = 1, 2, \ldots,$$

из которых в силу условий (6.5.2), (6.5.3) в углах $(0,0)$, $(0,1)$ вытекают соотношения

$$\frac{d^{2k} u_0}{dx^{2k}}(0) = 0, \quad \frac{d^{2k} u_0}{dx^{2k}}(1) = 0, \tag{6.5.5}$$

называемые условиями согласования порядка k. Заметим, что выполнения этих условий и достаточно для непрерывности на $\bar{Q}$ соответствующих производных.

Потребуем от исходной задачи (6.5.1)—(6.5.3) выполнения условий согласования порядка 0, 1, 2, 3. Вместе с (6.5.4) это гарантирует непрерывность на $\bar{Q}$ (и, значит, ограниченность) всех производных вида

$$\frac{\partial^{k+l} u}{\partial t^k \partial x^l}, \quad \text{где } 2k + l \leqslant 6. \tag{6.5.6}$$

Построим для приближенного решения задачи (6.5.1)—(6.5.3) разностную сетку, полученную пересечением прямых

$$x_k = kh, \quad k = 0, 1, \ldots, N,$$
$$t_j = j\tau, \quad j = 0, 1, \ldots, M.$$

Здесь $h = 1/N$, $\tau = T/M$ — шаги разностной сетки по x, t. Заменим уравнение (6.5.1) неявной разностной схемой

$$\frac{v_k^j - v_k^{j-1}}{\tau} = a^2 \frac{v_{k-1}^j - 2v_k^j + v_{k+1}^j}{h^2}, \tag{6.5.7}$$
$$k = 1, 2, \ldots, N-1; \quad j = 1, 2, \ldots, M,$$

а начальные и краевые условия — системой неравенств

$$v_k^0 = u_0(x_k), \quad k = 0, 1, \ldots, N, \tag{6.5.8}$$
$$v_0^j = v_N^j = 0, \quad j = 1, 2, \ldots, M. \tag{6.5.9}$$

Далее нам потребуется априорная оценка для этой задачи с неоднородной правой частью и нулевыми граничными значениями. Пусть η_k^j — решение задачи

$$\frac{\eta_k^j - \eta_k^{j-1}}{\tau} = a^2 \frac{\eta_{k-1}^j - 2\eta_k^j + \eta_{k+1}^j}{h^2} + \sigma_k^j, \tag{6.5.10}$$
$$k = 1, 2, \ldots, N-1; \quad j = 1, 2, \ldots, M,$$
$$\eta_k^0 = 0, \quad k = 0, 1, \ldots, N, \tag{6.5.11}$$
$$\eta_0^j = \eta_N^j = 0, \quad j = 1, 2, \ldots, M. \tag{6.5.12}$$

Методом математической индукции доказывается, что

$$\max_{0 \leqslant k \leqslant N} |\eta_k^j| \leqslant \tau j \max_{\substack{1 \leqslant k \leqslant N-1 \\ 1 \leqslant j' \leqslant j}} |\sigma_k^{j'}|. \tag{6.5.13}$$

В самом деле, для $j = 0$ утверждение (6.5.13) вытекает из начального условия (6.5.11). Пусть оно выполнено на слое j. Рассмотрим максимальную по модулю компоненту η_k^{j+1} на слое $j + 1$. Пусть она имеет номер k_0. Ясно,

что $k_0 \neq 0$ и $k_0 \neq N$. Тогда из уравнения (6.5.10) следует, что

$$\left(\frac{1}{\tau}+\frac{2a^2}{h^2}\right)|\eta_{k_0}^{j+1}| = \left|\frac{1}{\tau}\eta_{k_0}^{j}+\frac{a^2}{h^2}\eta_{k_0-1}^{j+1}+\frac{a^2}{h^2}\eta_{k_0+1}^{j+1}+\sigma_{k_0}^{j+1}\right| \leqslant$$
$$\leqslant \frac{2a^2}{h^2}|\eta_{k_0}^{j+1}| + (j+1)\max_{\substack{1\leqslant k\leqslant N-1\\ 1\leqslant j'\leqslant j}}|\sigma_k^{j'}|.$$

Отсюда

$$\max_{0\leqslant k\leqslant N}|\eta_k^{j+1}| = |\eta_{k_0}^{j+1}| \leqslant (j+1)\tau \max_{\substack{1\leqslant k\leqslant N-1\\ 1\leqslant j'\leqslant j+1}}|\sigma_k^{j'}|,$$

а следовательно, верна оценка (6.5.13).

Из (6.5.13) вытекает, что задача (6.5.7)—(6.5.9) однозначно разрешима, поскольку однородная задача $u_0(x_k) = 0$ $(k = 0, 1, \ldots, N)$ не может иметь других решений, кроме тривиального.

Докажем, что для решения разностной задачи (6.5.7)—(6.5.9) имеет место разложение

$$v_k^j = u(x_k, t_j) + h^2 w(x_k, t_j) + \tau z(x_k, t_j) + \eta_k^j,$$
$$k = 0, 1, \ldots, N; \quad j = 0, 1, \ldots, M, \tag{6.5.14}$$

где функции ω и z не зависят от τ, h, а сеточная функция η_k^j достаточно мала:

$$\max_{\substack{0\leqslant k\leqslant N\\ 0\leqslant j\leqslant N}}|\eta_k^j| \leqslant c_1(h^4+\tau^2). \tag{6.5.15}$$

Опустим нестрогие предварительные рассуждения, приводящие к формированию задач для w, z, и будем сразу искать эти функции как решение задач

$$\frac{\partial w}{\partial t} = a^2\frac{\partial^2 w}{\partial x^2} + \frac{a^2}{12}\frac{\partial^4 u}{\partial x^4} \text{ на } Q, \tag{6.5.16}$$
$$w(x,0) = 0, \quad x \in [0,1],$$
$$w(1,t) = w(0,t) = 0, \quad t \in (0,T],$$
$$\frac{\partial z}{\partial t} = a^2\frac{\partial^2 z}{\partial x^2} - \frac{1}{2}\frac{\partial^2 u}{\partial t^2} \text{ на } Q,$$
$$z(x,0) = 0, \quad x \in [0,1], \tag{6.5.17}$$
$$z(1,t) = z(0,t) = 0, \quad t \in (0,T].$$

Поскольку функция u имеет на $\bar{Q}$ непрерывные производные вида 6.5.6, то решение задачи 6.5.16 существует и единственно. Более того, для него выполняются условия согласования порядка 0, 1, 2. В самом деле, условия согласования, вытекающее из непрерывности w на $\bar{Q}$, выполняет-

ся автоматически. Условие согласования порядка 1, необходимое и достаточное для непрерывности на $\bar{Q}$ производных $\partial w/\partial t$, $\partial^2 w/\partial x^2$, вытекает из условия 6.5.5 с номером $k = 2$ и имеет вид

$$\frac{\partial^4 u}{\partial x^4}(0,0) = 0, \quad \frac{\partial^4 u}{\partial x^4}(1,0) = 0.$$

Для вывода условия согласования порядка 2 продифференцируем по t уравнение из (6.5.16). Имеем

$$\frac{\partial^2 w}{\partial t^2} = a^2 \frac{\partial^2}{\partial x^2}\left(\frac{\partial w}{\partial t}\right) + \frac{a^2}{12}\frac{\partial^4}{\partial x^4}\left(\frac{\partial u}{\partial t}\right) =$$
$$= a^2 \frac{\partial^2}{\partial x^2}\left(a^2 \frac{\partial^2 w}{\partial x^2} + \frac{a^2}{12}\frac{\partial^4 u}{\partial x^4}\right) + \frac{a^2}{12}\frac{\partial^4}{\partial x^4} a^2 \frac{\partial^2 u}{\partial x^2} = a^4 \frac{\partial^4 w}{\partial x^4} + \frac{a^4}{6}\frac{\partial^6 u}{\partial x^6}.$$

Учитывая нулевые начальные и краевые значения, приходим к условию согласования порядка 2:

$$\frac{a^4}{6}\frac{\partial^6 u}{\partial x^6}(0,0) = 0, \quad \frac{a^4}{6}\frac{\partial^6 u}{\partial x^6}(1,0) = 0.$$

Оно, очевидно, справедливо на основании (6.5.5) для $k = 3$. Таким образом, для задачи (6.5.16) выполнены условия согласования порядка 0, 1, 2. Они достаточны для непрерывности на $\bar{Q}$ производных вида

$$\frac{\partial^{k+l} w}{\partial t^k \partial x^l}, \quad \text{где} \quad 2k + l \leqslant 4. \tag{6.5.18}$$

Аналогичное утверждение справедливо для решения задачи (6.5.17).

Итак, функции v_k^j, u, w и z однозначно определены в узлах сетки (x_k, t_j). Введем сеточную функцию

$$\eta_k^j = v_k^j - u_k^j - h^2 w_k^j - \tau z_k^j, \tag{6.5.19}$$
$$k = 0, 1, \ldots, N; \quad j = 0, 1, \ldots, M,$$

и докажем, что ее значения имеют величину порядка $\tau^2 + h^4$. Для этого выразим из (6.5.19) v_k^j и подставим разностное уравнение (6.5.7). Имеем

$$\frac{(u + h^2 w + \tau z)_k^j - (u + h^2 w + \tau z)_k^{j-1}}{\tau} + \frac{\eta_k^j - \eta_k^{j-1}}{\tau} =$$
$$= a^2 \frac{(u + h^2 w + \tau z)_{k-1}^j - 2(u + h^2 w + \tau z)_k^j + (u + h^2 w + \tau z)_{k+1}^j}{h^2} + a^2 \frac{\eta_{k-1}^j - 2\eta_k^j + \eta_{k+1}^j}{h^2}.$$

Используя разложения в ряд Тейлора для функций u, w и z, приходим к равенству

$$\left(\frac{\partial u}{\partial t}+h^2\frac{\partial w}{\partial t}+\tau\frac{\partial z}{\partial t}+\frac{\tau}{2}\frac{\partial^2 u}{\partial t^2}\right)_k^j+\rho_k^j+\frac{\eta_k^j-\eta_k^{j-1}}{\tau}= \\ a^2\left(\frac{\partial^2 u}{\partial x^2}+h^2\frac{\partial^2 w}{\partial x^2}+\tau\frac{\partial^2 z}{\partial x^2}+\frac{1}{12}\frac{\partial^4 u}{\partial x^4}\right)_k^j+E_k^j+a^2\frac{\eta_{k-1}^j-2\eta_k^j+\eta_{k+1}^j}{h^2}. \tag{6.5.20}$$

Здесь ρ_k^j, E_k^j — остаточные члены формулы Тейлора, для которых на основании ограниченности соответствующих производных функций u, w и z справедливы оценки

$$|\rho_k^j|\leqslant c_2\tau^2+c_3\tau h^2\leqslant\left(c_2+\frac{c_3}{2}\right)\tau^2+\frac{c_3}{2}h^4, \\ |E_k^j|\leqslant c_4h^4+c_5\tau h^2\leqslant\frac{c_5}{2}\tau^2+\left(c_4+\frac{c_5}{2}\right)h^4. \tag{6.5.21}$$

Заметим, что с учетом (6.5.1), (6.5.16), (6.5.17) соотношение (6.5.21) приводится к виду

$$\frac{\eta_k^j-\eta_k^{j-1}}{\tau}=a^2\frac{\eta_{k-1}^j-2\eta_{k-1}^j-2\eta_k^j+\eta_{k+1}^j}{h^2}+\sigma_k^j, \\ k=1,2,\ldots,N-1;\quad j=1,2,\ldots,M, \tag{6.5.22}$$

причем для $\sigma_k^j=E_k^j-\rho_k^j$ получается оценка

$$|\sigma_k^j|\leqslant c_6(\tau^2+h^4),\quad k=1,2,\ldots,N;\quad j=1,2,\ldots,M. \tag{6.5.23}$$

Кроме того, из (6.5.19) и граничных условий для функций u, w, z, v_k^j следует, что

$$\eta_k^0=0,\qquad k=0,1,\ldots,N, \\ \eta_0^j=\eta_N^j=0,\quad j=1,2,\ldots,M.$$

Поэтому справедлива априорная оценка (6.5.13), из которой вытекает неравенство

$$|\eta_k^j|\leqslant Tc_6(\tau^2+h^4),\quad k=0,1,\ldots,N,\quad j=0,1,\ldots,M. \tag{6.5.24}$$

Следовательно, оценка (6.5.15) и разложение (6.5.14) нужными нам свойствами обладают.

Порядок гладкости решения u по x и по t различен. Поэтому разложение (6.5.14) вполне естественно содержит τ^2 наравне с h^4.

Изложим теперь метод повышения точности, основанный на разложении (6.5.14). Выберем целые $M\geqslant 2$ и $N\geqslant 2$ и построим разностную сетку

с временным шагом $\tau = T/M$ и пространственным шагом $h = 1/N$. Найдем решение задачи (6.5.7)—(6.5.9) и обозначим его через u_h^τ. Затем построим разностную сетку с шагом $\tau/4$ по времени и шагом $h/2$ по пространству. Вновь решим задачу (6.5.7)—(6.5.9) и новое решение обозначим через $u_{h/2}^{\tau/4}$. Теперь в узлах сетки (x_k, t_j) с шагами τ, h имеется два приближенных решения: u_h^τ и $u_{h/2}^{\tau/4}$. Составим линейную комбинацию

$$\bar{u}(x_k, t_j) = \frac{4}{3} u_{h/2}^{\tau/4}(x_k, t_j) - \frac{1}{3} u_h^\tau(x_k, t_j), \tag{6.5.25}$$

$$k = 0, 1, \ldots, N; \quad j = 0, 1, \ldots, M, \tag{6.5.26}$$

и покажем, что уточненное разностное решение $\bar{u}$ приближает точное решение с порядком точности $\tau^2 + h^4$. В самом деле, в каждом узле (x_k, t_j) справедливы разложения

$$u_h^\tau(x_k, t_j) = u(x_k, t_j) + h^2 w(x_k, t_j) + \tau z(x_k, t_j) + O(\tau^2 + h^4),$$
$$u_{h/2}^{\tau/4}(x_k, t_j) = u(x_k, t_j) + \frac{h^2}{4} w(x_k, t_j) + \frac{\tau}{4} z(x_k, t_j) + O(\tau^2 + h^4).$$

Поскольку функции u, w, z не зависят от τ и h, то сложение с весами — 1/3, 4/3 дает равенство

$$\bar{u}(x_k, t_j) = u(x_k, t_j) + O(\tau^2 + h^4).$$

Таким образом разностное решение (6.5.25), полученное линейной комбинацией приближенных решений с точностью порядка $\tau + h^2$, приближает решение u с точностью порядка $\tau^2 + h^4$.

6.5.2. Метод расщепления для эволюционной задачи

В этом параграфе изучается дифференциальная задача

$$\frac{du}{dt} + A(t)u = f(t), \quad t \in (0, 1), \tag{6.5.27}$$
$$u(0) = u_0,$$

где $A(t)$ — матрица с $n \times n$ элементами — функциями, а $u(t)$ и $f(t)$ — вектор-функции с n компонентами. Предположим, что матрица A представима в виде суммы

$$A(t) = A_1(t) + A_2(t),$$

где $A_i(t)$ — положительно определенные матрицы для каждого $t \in [0, 1]$.

Разобьем отрезок $[0,1]$ на M равных частей точками

$$t_j = j\tau, \quad j = 0, 1, \ldots, M, \tag{6.5.28}$$

с шагом $\tau = 1/M$. Для численного решения задачи (6.5.27) рассмотрим неявную схему расщепления

$$(E + \tau A_1(t_j))u^{j-1/2} = u^{j-1} + \tau f^j,$$
$$(E + \tau A_2(t_j))u^j = u^{j-1/2}, \tag{6.5.29}$$
$$j = 1, 2, \ldots, M, \quad u^0 = u_0, \tag{6.5.30}$$

где E — единичная матрица.

Для n-мерных векторов введем норму

$$\|v\| = \left(\sum_{i=1}^{n} v_i^2\right)^{1/2}.$$

Из 4.4.3 вытекает оценка, характеризующая устойчивость этой схемы:

$$\max_{0 \leqslant j \leqslant M} \|u^j\| \leqslant \|u_0\| + \max_{1 \leqslant j \leqslant M} \|f^j\|. \tag{6.5.31}$$

Поставим себе цель — доказать разложение

$$u_j = \sum_{k=0}^{l-1} \tau^k v_k(t_j) + \eta_\tau^j, \quad j = 0, 1, \ldots, M, \tag{6.5.32}$$

с гладкими, не зависящими от τ вектор-функциями v_k и сеточной вектор-функцией η_τ, удовлетворяющей неравенству

$$\|\eta_\tau^j\| \leqslant C_1 \tau^l \tag{6.5.33}$$

с не зависящей от τ и k константой C_1.

Для этого запишем систему (6.5.29), исключив промежуточные значения $u^{j-1/2}$:

$$\frac{1}{\tau}(E + \tau A_1(t_j))(E + \tau A_2(t_j))u^j - \frac{1}{\tau}u^{j-1} = f^j, \tag{6.5.34}$$
$$j = 1, 2, \ldots, M.$$

Проверим выполнение условий A, B, C из § 6.2. Пусть F^k — множество n-мерных вектор-функций с компонентами из класса $C^{k+1}[0,1]$, Φ^k — множество n-мерных вектор-функций с компонентами из $C^{k+2}[0,1]$, а G^k (k — любое) — множество n-мерных векторов с вещественными коэффициентами. Тогда

из предположения, что элементы матрицы A принадлежат $C^{m+1}[0,1]$, следует выполнение первого из условий. Условие B эквивалентно неравенству (6.5.31) в следующих обозначениях:

$$\|u\|_{\Phi_h} = \max_{0 \leqslant j \leqslant M} \|u^j\|, \quad \|f\|_{F_h} = \max_{1 \leqslant j \leqslant M} \|f^j\|, \quad \|u\|_{G_h} = \|u^0\|.$$

Осталось проверить условие C. Используем для этого, как и раньше, разложение в ряд Тейлора. Простые выкладки приводят к разложениям (6.2.5) из 6.2, в которых $h = \tau$:

$$B_1 = A_1 A_2 u + \frac{1}{2}\frac{d^2 u}{dt^2},$$
$$B_l = \frac{(-1)^l}{(l+1)!}\frac{d^{l+1}u}{dt^{l+1}}, \quad l = 2, \ldots, k,$$
$$b_l = 0, \quad i = 1, 2, \ldots, k.$$

Таким образом, условия A, B, C выполнены, а также справедлива теорема о разложении разностного решения по степеням параметра h (здесь τ) с коэффициентом $\beta = 1$. Полученное разложение используется обычным образом. На разностных сетках с шагами $\tau_i = \tau/i$ ($i = 1, 2, \ldots, m+1$, $\tau = 1/M$) решаем систему (6.5.29). Полученные решения u^{τ_i} определены на сетке с шагом τ. Выберем γ_i как решение системы

$$\sum_{i=1}^{m+1} \gamma_i = 1, \quad \sum_{i=1}^{m+1} \gamma_i h_i^k = 0, \quad k = 1, 2, \ldots, m+1.$$

Составим линейную комбинацию с полученными коэффициентами γ_i:

$$\bar{u}(t_j) = \sum_{k=1}^{m+1} \gamma_k u^{\tau_k}(t_j), \quad j = 0, 1, \ldots, M.$$

Тогда на основании теоремы из 6.2.2 имеет место оценка

$$\|u(t_j) - \bar{u}(t_j)\| \leqslant c_2 \tau^{m+1}, \quad j = 0, 1, \ldots, M.$$

6.6. Экстраполяция Ричардсона для многомерных задач

В многомерных задачах применение экстраполяции к простейшим разностным схемам оказывается делом более трудным по ряду обстоя-

тельств. Мы перечислим главные из них и укажем на некоторые пути, позволяющие обойти эти трудности.

В качестве модели для применения метода в многомерном случае рассмотрим двумерное уравнение Пуассона

$$\Delta u = f \text{ в } D \tag{6.6.1}$$

с краевым условием

$$u = 0 \text{ на } \partial D. \tag{6.6.2}$$

В случае, когда область D имеет негладкую криволинейную границу ∂D, простейшие разностные аналоги вблизи границы не позволяют разложить ошибку аппроксимации по шагу сетки аналогично тому, как это делается при разложении ошибки аппроксимации внутри области. В самом деле, в этом случае представление

$$u^h(x) = u(x) + h^2 v_1(x) + h^4 v_2(x) + \dots, \tag{6.6.3}$$

естественное для рассмотренных ранее одномерных задач, уже не имеет смысла для негладких границ ∂D. В этом случае величины v_i в разложении (6.6.3) имеют разрывные данные и не приводят к гладким решениям, которые позволили бы рекуррентно определить все необходимые функции v_k, $(k < i)$.

Поэтому вблизи границы приходится отказаться от простейших разностных аналогов и применять более сложные, которые обычно приводят к большому количеству ненулевых коэффициентов уравнений вблизи границы. В результате этого удается получить вид погрешности аппроксимации, совпадающей внутри и в приграничной полосе области. Что касается метода конечных элементов, то требование непрерывности вплоть до границы коэффициентов разложения погрешности аппроксимации приводит к выбору в приграничной полосе более сложных базисных функций, которые также приводят к большему количеству неизвестных в приграничных уравнениях.

В том случае, когда граница ∂D состоит из отрезков прямых и возможны простые регулярные построения разностной сетки и конечных элементов вплоть до границы, применение метода экстраполяции становится затруднительным из-за неограниченного роста производных от решения вблизи углов области. В некоторых задачах производные от решения настолько быстро возрастают вблизи этих особых точек, что не удается получить даже минимальную точность. Аналогично обстоит дело с гладкостью решения вблизи пересечения границы и линий разрыва первого рода ко-

эффициентов уравнения. На рис. 6.2 приведены четыре типа особых точек, причем первые три могут не дать обычной точности решения порядка h.

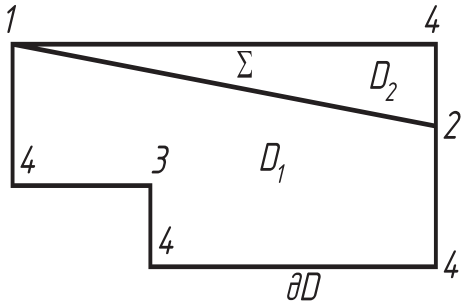


Рис. 6.2.

Обойти указанные трудности можно следующим образом. При построении разностных уравнений на основе метода Галеркина целесообразно заменить обычные скалярные произведения на взвешенные, с весами, стремящимися к нулю при приближении к особым точкам. Пусть, например, особенность имеет место в начале координат. Тогда вместо нормы

$$\|u\|_{\overset{\circ}{W}{}_2^1} = \left[\int\limits_\Omega \left[\left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 \right] d\Omega \right]^{1/2}, \tag{6.6.4}$$

естественной для задачи (6.6.1), (6.6.2) и функций, обращающихся в нуль на границе области, используем следующую:

$$\|u\|^0_{\alpha,\overset{\circ}{W}{}_2^1} = \left[\int\limits_\Omega \left[\left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 \right] (x^2+y^2)^\alpha \, d\Omega \right]^{1/2}. \tag{6.6.5}$$

Здесь α — степень сингулярности решения, которая априори предполагается известной. Проводя обычные построения в методе Галеркина со скалярным произведением с весом $(x^2+y^2)^\alpha$, мы приходим к приближенному решению u^h, удовлетворяющему оценке

$$\|u-u^h\|_{\alpha,\overset{\circ}{W}{}_2^1} \leqslant C\|u-\tilde{u}\|_{\alpha,\overset{\circ}{W}{}_2^1}, \tag{6.6.6}$$

где $\tilde{u}$ — интерполянт функции u в пространстве базисных функций.

Необходимо отметить, что возможности этого метода ограничены тем, что параметр α нельзя выбирать более 1.

Другой возможностью является сгущение разностной сетки или конечных элементов с тем, чтобы сделать погрешность аппроксимации задачи приемлемой. Эта погрешность слагается из интегралов по элементар-

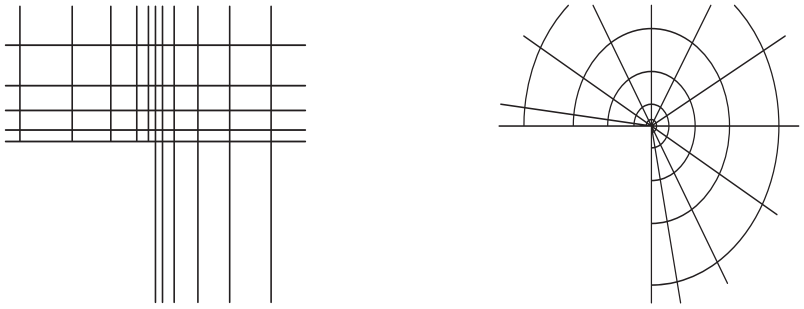
ным областям, которые оцениваются величинами вида

$$Ch_i^{\beta}\|u\|_{k,i}, \tag{6.6.7}$$

где h_i — диаметр i-й элементарной области или ячейки сетки, $\|u\|_{k,i}$ — норма функции на этой области, содержащая k-ю производную, а C — константа, не зависящая от этих множителей. Тогда разумно потребовать, чтобы величины (6.6.7) оказались одинаковыми в D. С этой целью можно уменьшать шаги h_i обратно пропорционально величинам $\|u\|_{k,i}$ при подходе к особым точкам. Однако сгущение сеток нельзя делать чрезмерным, так как число обусловленности алгебраических систем методов Ритца и Галеркина существенно зависит от величины

$$h_{\max}/h_{\min}\,(h_{\min} = \min_i h_i),$$

которая оказывается тем большей, чем сильнее сгущается сетка. Исключения составляют простейшие кусочно-линейные базисные функции на треугольниках, для которых удается нормировать алгебраические системы диагональной матрицей так, что эта величина не оказывает существенного влияния на число обусловленности.



(a) (b)

Рис. 6.3.

Для сгущения сеток вблизи особых точек используются подходы, схематически изображенные на рис. 6.3. При этом обычно способ 6.3a оказывается более простым в алгоритмическом смысле, так как он влечет за собой незначительное усложнение программы на ЭВМ, в отличие от способа 6.3b, который оказывается несколько сложнее, так как предполагает «склеивание» полярной сетки с прямоугольной, либо с другими полярными сетками в случае двух и более особенностей. Однако способ 6.3b экономичнее с точки зрения общего количества узлов, так как в этом случае сетка

сгущается только около особой точки, в то время как в способе 6.3а сгущение происходит по линиям, в том числе и вдали от особой точки.

Рассмотрим теперь еще один метод повышения точности, суть которого состоит в том, что к обычным базисным функциям главы 2 добавляются дополнительные, детально описывающие поведение производных от решения в окрестности особой точки. Такой подход продиктован тем, что для правой части задачи (6.6.1), (6.6.2) из $W_2^k(D)$ решение u представимо в виде суммы гладкой функции $v \in W_2^{k+2}(D)$ и конечного числа слагаемых, которые в полярных координатах с центром в особой точке записываются в виде

$$w_i = \mu_i(\varphi) r^{\gamma_i} \ln^{p_i} r,$$

где $\mu_i(\varphi)$ — аналитическая функция угловой переменной φ; γ_i — вещественная положительная константа; p_i — целое неотрицательное число. Проведение доказательства для каждого конкретного случая обычно выявляет эти функции с точностью до постоянного множителя. Таким образом, в методе Галеркина отыскиваются не только веса в разложении гладкой составляющей по базисным функциям метода конечных элементов, но и веса особых функций. Основная трудность применения этого метода состоит в том, что функции w_i усложняют структуру расположения ненулевых элементов алгебраической системы, так как носители этих функций захватывают большое количество элементарных областей.

Использование особых базисных функций, детально описывающих поведение решения вблизи особых точек, делает возможным применение внутри области простейших элементов с последующей экстраполяцией по шагу сетки для достижения высокой точности. Для задачи (6.6.1), (6.6.2) укажем только общую схему алгоритма и его обоснования, так как двумерный случай предоставляет большое число различных модификаций, которые здесь не рассматриваются.

Необходимо отметить, что исследование повышения точности привлекает обширную информацию о гладкости решений эллиптических уравнений на областях различной формы.

Пусть для простоты область является прямоугольником со сторонами длиной a и b. Двумя семействами параллельных линий, содержащих стороны прямоугольника, построим равномерную прямоугольную сетку с шагами $h_x = a/N$ и $h_y = b/N$ соответственно (целое $N > 2$). Для окончания триангуляции области в каждом прямоугольнике сетки проведем диагональ, например, слева вверх. Для удобства введем параметр $h = 1/N$, характеризующий размеры сетки и связанный с h_x и h_y равенствами $h_x = ah$, $h_y = bh$.

Для каждого внутреннего узла X_i введем базисную функцию φ_i из главы 2, линейную на каждом треугольнике. Кроме этих $(N-1)^2$ функций, введем дополнительные базисные функции ψ_k, равные функциям w_i в некоторой области особых точек (углов прямоугольника и пересечений границы с возможными линиями разрыва коэффициентов, являющихся линиями разностной сетки) и гладко продолженные на остальную часть области так, чтобы удовлетворить краевому условию (6.6.2). Число таких дополнительных функций в каждой конкретной задаче зависит от степени точности, которую мы хотим получить.

В соответствии с методом Ритца или Галеркина мы получаем решение

$$u^h(X) = \sum_{l=1}^{(N-1)^2} \alpha_I^h \varphi_l(X) + \sum_{k=1}^{M} \beta_k^h \psi_k(X), \tag{6.6.8}$$

которое в узлах разностной сетки (вершинах триангуляции) разлагается по параметру h:

$$u^h(X) = u(X) + \sum_{l=1}^{m-1} h^{2l} v_l(X) + h^{2m} \xi_h(X) \tag{6.6.9}$$

с функциями $v_l(X)$, не зависящими от h, и ограниченной сеточной функцией $\xi_h(X)$. В том случае, когда это разложение доказано, необходимо избавиться от членов порядка $h^2, \ldots, h^{2m-2}$, что и приводит к решению $u(x)$ с точностью порядка h^{2m}.

Обоснование этого разложения основано на возможности разложения весов α_i^h и β_k^h в формуле (6.6.8) также по степеням h:

$$\begin{aligned} \alpha_i^h =& v(X_i) + \sum_{l=1}^{m-1} h^{2l} Z_l(X_i) + h^{2m} \eta_h(X_i), \\ \beta_k^h =& \gamma_k + \sum_{l=1}^{m-1} h^{2l} \eta_{k,l} + h^{2m} \xi_{k,h}, \end{aligned} \tag{6.6.10}$$

где $v(X)$ — гладкая составляющая решения $u(X)$; функции $Z_l(X)$ не зависят от h; сеточная функция η_h ограничена при $h \to 0$; а $\xi_{k,h}$ — числа, причем γ_k и $\eta_{k,l}$ не зависят от h, а ограничены при $h \to 0$.

Проводя выкладки, вытекающие из разложений (6.6.10), можно сформулировать условия (вспомогательные дифференциальные задачи, которые определяют функции $Z_l(X)$ и константы $\eta_{k,l}$, являющиеся коэффициентами при функциях $\varphi_k(X)$ в разложениях решений вспомогательных задач) на гладкие составляющие $v_l(X)$ и особые функции $\psi_k(X)$. Такая схема реализации, конечно, дает только указание к возможному подходу, который требует тщательного обоснования в каждом конкретном случае.

Глава 7.

Метод Шварца и разделения области[1)]

Трудности построения схем для краевой задачи и составления соответствующих программ для численной реализации схем на ЭВМ во многом зависят от геометрии области D, в которой ищется решение задачи. Поэтому привлекательными являются методы, которые позволяют свести процесс решения исходной задачи к последовательности задач, рассматриваемых в областях, имеющих более простую форму по сравнению с D. Одним из таких методов является *альтернирующий метод Шварца*. В этом методе область D представляется в виде объединения конечного числа подобластей D_M ($m = 1, 2, \ldots, M$), причем каждая из подобластей имеет ненулевое пересечение с другими. Предельный случай этого метода, когда подобласти D_m ($m = 1, 2, \ldots, M$) могут иметь лишь общую границу, естественно назвать *методом разделения области*. В последние годы интерес к этому методу значительно возрос. Одновременно было выявлено, что некоторые случаи метода разделения области могут приводить к несходящимся итерационным процессам, если только в алгоритм не вводить числовые параметры и не выбирать их так, чтобы сделать норму некоторого оператора перехода меньше единицы. Необходимость введения числовых параметров естественным образом влечет интересный для исследования вопрос о выборе оптимальных параметров, при которых итерационный процесс был бы наиболее быстро сходящимся в подходящих нормах. Решение проблемы выбора оптимальных параметров весьма важно с точки зрения построения экономичных алгоритмов.

Проблему сведения процесса решения исходной задачи в области D со сложной границей можно также попытаться осуществить путем прибли-

[1)] Данная глава по просьбе автора написана В. И. Агошковым, Ю. А. Кузнецовым, А. М. Мацокиным

женной замены исходной задачи на задачу, заданную в более простой области $\tilde{D} \supset D$, например в параллелепипеде. Такой подход получил название *метода фиктивных областей*.

Как легко заметить, упомянутые методы объединяются идеей перехода от решения задач в сложных областях к решению задач или последовательности задач в областях, форма границ которых была бы более простой. Некоторые из алгоритмов этих методов и будут рассмотрены в данной главе.

7.1. Метод Шварца

7.1.1. Формулировка метода

Рассмотрим задачу Дирихле

$$Lu \equiv -\sum_{i,j=1}^{2} \frac{\partial}{\partial x_i} a_{ij}(x) \frac{\partial}{\partial x_j} + c(x)u = f(x), \tag{7.1.1}$$

$$x = (x_1, x_2) \in D,$$

$$u(x) = 0, \quad x \in \partial D, \tag{7.1.2}$$

где D — ограниченная область из $\mathbf{R}^2$ с кусочно-гладкой границей ∂D, $a_{ij} = a_{ji}$, $c(x) \geqslant 0$ — кусочно-гладкие вещественные коэффициенты, причем выполнены условия

$$\mu_0 \sum_{i=1}^{2} \xi_i^2 \leqslant \sum_{i,j=1}^{2} a_{ij}(x)\xi_i\xi_j \leqslant \mu_1 \sum_{i=1}^{2} \xi_i^2, \tag{7.1.3}$$

$$\mu_0, \mu_1 = \text{const} > 0, \quad x \in \bar{D}.$$

Функция $f(x)$ предполагается принадлежащей гильбертову пространству вещественных функций $L_2(D)$.

Представим область D в виде объединения конечного числа своих подобластей $D_1, \ldots, D_M$ (с липшицевыми границами $\partial D_1, \ldots, \partial D_M$):

$$D = \bigcup_{m=1}^{M} D_m, \tag{7.1.4}$$

а также выберем некоторую функцию $u^0(x)$, удовлетворяющую граничному условию (7.1.2). Тогда альтернирующий метод Шварца состоит в построении последовательности функций $\{u^\alpha(x)\}$, определяемых путем решения

следующих задач:

$$\begin{aligned}
&Lu^{n+\frac{1}{M}} = f(x), && x \in D_1,\\
&u^{n+\frac{1}{M}}(x) = u^n(x), && x \in \bar{D} \setminus D_1,\\
&\ldots\ldots\ldots\ldots\ldots\ldots && \ldots\ldots\ldots\\
&Lu^{n+\frac{m}{M}} = f(x), && x \in D_m,\\
&u^{n+\frac{m}{M}}(x) = u^{n+\frac{m-1}{M}}(x), && x \in \bar{D} \setminus D_m,\\
&\ldots\ldots\ldots\ldots\ldots\ldots && \ldots\ldots\ldots\\
&Lu^{n+1} = f(x), && x \in D_M,\\
&u^{n+1}(x) = u^{n+\frac{M-1}{M}}(x), && x \in \bar{D} \setminus D_M,\\
&n = 0, 1, \ldots.
\end{aligned} \tag{7.1.5}$$

Таким образом, чтобы определить функцию $u^{n+m/M}$, достаточно в подобласти D_m решить задачу Дирихле

$$\begin{aligned}
&Lu^{n+\frac{m}{M}} = f(x) && \text{в} && D_m,\\
&u^{n+\frac{m}{M}}(x) = u^{n+\frac{m-1}{M}}(x) && \text{на} && \partial D_m,
\end{aligned} \tag{7.1.6}$$

а после решение этой задачи продолжить в $\bar{D} \setminus \bar{D}_m$ уже найденной функцией $u^{n+(m-1)/M}(x)$. В результате получим функцию $u^{n+m/M}(x)$, определенную на $\bar{D}$. Очевидно, что если подобласти D_m имеют границы сравнительно простой формы, то отыскание решений подзадач (7.1.6) может быть значительно проще, чем решение исходной задачи в области D.

Отметим, что при сделанных ограничениях на коэффициенты и границы ∂D, ∂D_m все рассмотренные задачи надо понимать в обобщенной форме записи. Приведем эту форму. Рассмотрим пространство $\overset{\circ}{W}{}_2^1(D) \subset L_2(D)$ (которое определено в главе **??**) и введем в нем новое скалярное произведение и норму вида

$$[u, v] = \int\limits_D \left(\sum_{i,j=1}^{2} a_{ij}(x) \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} + c(x) uv \right) dx, \quad [u] = [u, u]^{1/2},$$

эквивалентные соответственно скалярному произведению $(u, v)_{W_2^1(D)}$ и норме $\|u\|_{W_2^1(D)}$ при $u, v \in \overset{\circ}{W}{}_2^1(D)$. Тогда обобщенная форма записи задачи (7.1.1), (7.1.2) имеет вид

$$[u, v] = f(v), \tag{7.1.7}$$

где $u(x) \in \overset{\circ}{W}{}_2^1(D)$, $v(x)$ — произвольная функция из $\overset{\circ}{W}{}_2^1(D)$, а также

$$f(v) = \int\limits_D f(x)v(x)\,dx$$

— линейный непрерывный функционал над $\overset{\circ}{W}{}_2^1(D)$.

Выделим в $\overset{\circ}{W}{}_2^1(D)$ замкнутые подпространства

$$W_m = \{v \in \overset{\circ}{W}{}_2^1(D) :\ v(x) = 0,\ x \in \bar{D} \setminus D_m\}, \quad m = 1, 2, \ldots, M, \tag{7.1.8}$$

и обозначим разность $(u^{n+m/M} - u^{n+(m-1)/M})$ через $z_m^n(x)$. Легко заметить, что $z_m^n \in W_m$. Теперь в терминах обобщенных постановок задач алгоритм (7.1.5) можно переписать следующим образом: при заданной функции $u^0(x) \in \overset{\circ}{W}{}_2^1(D)$ требуется определить функции $\{u^\alpha\}$ из итерационного алгоритма вида

$$u^{n+1/M} = u^n + z_1^n,$$
$$[z_1^n, v] = f(v) - [u^n, v]$$

(v — любая функция из W_1, $z_1^n \in W_1$);

. . . . . .

$$u^{n+\frac{m}{M}} = u^{n+\frac{m-1}{M}} + z_m^n,$$
$$[z_m^n, v] = f(v) - [u^{n+\frac{m-1}{M}}, v] \tag{7.1.9}$$

(v — любая функция из W_m, $z_m^n \in W_m$);

. . . . . .

$$u^{n+1} = u^{n+\frac{M-1}{M}} + z_M^n,$$
$$[z_M^n, v] = f(v) - [u^{n+\frac{M-1}{M}}, v]$$

(v — любая функция из W_M, $z_M^n \in W_M$);

. . . . . .

$$n = 0, 1, \ldots.$$

Здесь подзадача

$$[z_m^n, v] = f_m(v), \tag{7.1.10}$$

где

$$f_m(v) \equiv f(v) - [u^{n+\frac{m-1}{M}}, v],$$

v — любая функция из W_m, представляет собой известную обобщенную форму записи задачи Дирихле в подобласти D_m для функции $z_m^n \in W_m$. В силу известных результатов теории краевых задач данная подзадача при введенных выше ограничениях на коэффициенты и границу ∂D_m имеет единственное решение z_m^n. Численное решение задач такого типа достаточно просто осуществить, например, проекционными алгоритмами с привлечением базисных функций с финитными носителями (см. гл. 2).

7.1.2. Сходимость метода

Исследуем проблему сходимости метода, сформулированного в 7.1.1. Введем в пространстве $\overset{\circ}{W}{}_2^1(D)$ операторы ортогонального проектирования на подпространства W_m:

$$R_m : \overset{\circ}{W}{}_2^1(D) \to W_m, \quad m = 1, 2, \ldots, M,$$

и пусть

$$Q = E - R_m, \quad m = 1, 2, \ldots, M,$$

где E — тождественный оператор. Поскольку для любой функции $v \in W_m$ для точного решения задачи (7.1.7) имеем

$$f(v) = [u, v] = [Q_m u, v] + [R_m u, v] = [R_m u, v],$$

тогда

$$[z_m^n, v] = [R_m u, v] - [u^{n+\frac{m-1}{M}}, v].$$

Если здесь принять функцию v равной $R_m w$, где w — произвольная функция из $\overset{\circ}{W}{}_2^1(D)$, то, привлекая свойства R_m, из последнего равенства имеем

$$[R_m z_m^n, w] = [R_m u, w] - [R_m u^{n+\frac{m-1}{M}}, w],$$

т. е.

$$z_m^n = R_m z_m^n = R_m u - R_m u^{n+\frac{m-1}{M}}.$$

А тогда

$$u^{n+\frac{m}{M}} = u^{n+\frac{m-1}{M}} + z_m^n$$

можно переписать в виде

$$u^{n+\frac{m}{M}} = (R_m + Q_m) u^{n+\frac{m-1}{M}} + z_m^n = Q_m u^{n+\frac{m-1}{M}} + R_m u.$$

Учитывая последнее соотношение, итерационный процесс (7.1.9) можно переписать в следующем виде:

$$\begin{aligned} &u^{n+\frac{1}{M}} = Q_1 u^n + R_1 u, \\ &u^{n+\frac{m}{M}} = Q_m u^{n+\frac{m-1}{M}} + R_m u, \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &u^{n+1} = Q_M u^{n+\frac{M-1}{M}} + R_M u, \\ &n = 0, 1, \ldots, \quad u^0 \in \overset{\circ}{W}{}_2^1(D), \end{aligned} \tag{7.1.11}$$

а для ошибки $\psi^n = u^n - u$ получим уравнение

$$\psi^{n+1} = Q_M Q_{M-1} \ldots Q_1 \psi^n, \quad n = 0, 1, \ldots. \tag{7.1.12}$$

Очевидно, что итерационный процесс (7.1.11) будет сходиться со скоростью геометрической прогрессии только в том случае, если норма $\|T\|$ оператора шага $T = Q_M Q_{M-1} \ldots Q_1$ в $\overset{\circ}{W}{}_2^1(D)$ меньше единицы. В следующем утверждении формулируется необходимое условие такой сходимости: если норма оператора шага T итерационного процесса (7.1.11) меньше единицы, то

$$\overset{\circ}{W}{}_2^1(D) = W_1 + W_2 + \cdots + W_M \tag{7.1.13}$$

и для любого элемента $v \in \overset{\circ}{W}{}_2^1(D)$ существует представление

$$v = v_1 + v_2 + \cdots + v_M, \quad v_m \in W_m, \quad m = 1, 2, \ldots, M, \tag{7.1.14}$$

такое, что

$$[v_m] \leqslant \gamma [v], \quad m = 1, 2, \ldots, M, \tag{7.1.15}$$

с постоянной $\gamma \leqslant (1 - \|T\|)^{-1}$.

Докажем это утверждение. Легко проверить, что

$$E - T = R_1 + R_2 Q_1 + R_3 Q_2 Q_1 + \cdots + R_M Q_{M-1} \ldots Q_1.$$

Так как по условию теоремы $\|T\| < 1$, то оператор $E - T$ обратим к $\|(E - T)^{-1}\| \leqslant (1 - \|T\|)^{-1}$. Тогда для любого элемента $v \in \overset{\circ}{W}{}_2^1(D)$ справедливо представление

$$v = (R_1 + R_2 Q_1 + R_3 Q_2 Q_1 + \cdots + R_M Q_{M-1} \ldots Q_1)(E - T)^{-1} v = v_1 + v_2 + \cdots + v_M,$$

где $v_m = R_m Q_{m-1} \dots Q_1 (E-T)^{-1} v \in W_m$, и так как нормы ортогональных проекторов равны единице, то

$$[v_m] \leqslant \|(E-T)^{-1}\|[v] \leqslant (1-\|T\|)^{-1}[v],\ m = 1, 2, \dots, M,$$

и утверждение доказано.

Необходимое условие сходимости в $\overset{\circ}{W}{}_2^1(D)$ итерационного процесса (7.1.11) со скоростью геометрической прогрессии является и достаточным условием. Но прежде чем доказать этот факт, покажем, что если выполняется необходимое условие скорости сходимости итерационного процесса (7.1.11), сформулированного выше, то оператор

$$R = R_1 + R_2 + \dots + R_M,$$

являющийся суммой ортопроекторов, обратим оператор R^{-1} ограничен.

Действительно, поскольку R самосопряжен в $\overset{\circ}{W}{}_2^1(D)$, то нам достаточно установить неравенства

$$[Rv, v] \geqslant c[v, v]$$

для любого элемента $v \in \overset{\circ}{W}{}_2^1(D)$ с некоторой положительной постоянной c. Для любого элемента $u \in \overset{\circ}{W}{}_2^1(D)$ имеем

$$[u] = \sup_{v \neq 0} \frac{|[u, v]|}{[v]} = \sup_{v \neq 0} \frac{|[u_1 v_1 + \dots + u_M v_M]|}{[v]},$$

где разложение $v = v_1 + v_2 + \dots + v_M$ таково, что $v_m \in W_m$, $[v_m] \leqslant \gamma[v]$ ($m = 1, 2, \dots, M$).

Тогда

$$\begin{aligned}
[u] &\leqslant \sup_{v \neq 0} \sum_{m=1}^{M} \frac{|[u, v_m]|}{[v]} = \sup_{v \neq 0} \sum_{m=1}^{M} \frac{|[R_m u, v_m]|}{[v]} \leqslant \\
&\leqslant \sup_{v \neq 0} \sum_{m=1}^{M} \frac{|[R_m u][v_m]|}{[v]} \leqslant \gamma \sum_{m=1}^{M} [R_m u] \leqslant \gamma \sqrt{M}\, ([R_m u, R_m u])^{1/2} = \\
&= \gamma \sqrt{M}\, ([R_m u, u])^{1/2} = \gamma \sqrt{M} [Ru, u]^{1/2},
\end{aligned}$$

т. е. $[Ru, u] \geqslant (M\gamma^2)^{-1}[u, u]$, что и требовалось доказать.

Докажем теперь, что необходимое условие скорости сходимости, сформулированное ранее, является и достаточным условием сходимости итерационного процесса (7.1.11) со скоростью геометрической прогрессии. Кроме того, существует такая постоянная $q = q(M, \gamma)$, что

$$\|T\| \leqslant q(M, \gamma) < 1. \tag{7.1.16}$$

Предположим, что необходимое условие такой скорости сходимости процесса (7.1.11) выполняется, а $\|T\| = \|QMQ_{M-1}\dots Q_1\| = 1$. Тогда для любого достаточно малого $\varepsilon > 0$ существует такой элемент $v_\varepsilon \in \overset{\circ}{W}{}_2^1(D)$, что

$$[v_\varepsilon] = 1, \quad 1 - \varepsilon \leqslant [Tv_\varepsilon]^2 \leqslant 1.$$

Отсюда легко получить, что

$$1 - \varepsilon \leqslant [Q_1 v_\varepsilon]^2 \leqslant 1, \quad [R_1 v_\varepsilon]^2 \leqslant \varepsilon_1 = \varepsilon.$$

Поскольку

$$[Q_M Q_{M-1}\dots Q_2 v_\varepsilon] \geqslant [Tv_\varepsilon] - [Q_M Q_{M-1}\dots Q_2 R_1 v_\varepsilon] \geqslant \sqrt{1-\varepsilon_1} - \sqrt{\varepsilon_1} = \sqrt{1-\varepsilon_2},$$

то для $T_2 = Q_M Q_{M-1}\dots Q_2$ имеем

$$1 - \varepsilon_2 \leqslant [T_2 v_\varepsilon]^2 \leqslant 1.$$

Очевидно, что, повторив предыдущие рассуждения, придем к оценкам

$$[R_2 v_\varepsilon]^2 \leqslant \varepsilon_2, \quad 1 - \varepsilon_3 \leqslant [T_3 v_\varepsilon]^2 \leqslant 1,$$

где $\varepsilon_3 = 1 - (\sqrt{1-\varepsilon_2} - \sqrt{\varepsilon_2})^2$, $T_3 = Q_M Q_{M-1}\dots Q_3$. Повторяя эти выкладки еще $M-2$ раза, будем иметь оценки

$$[R_m v_\varepsilon]^2 \leqslant \varepsilon_m, \quad \varepsilon_m = 1 - (\sqrt{1-\varepsilon_{m-1}} - \sqrt{\varepsilon_{m-1}})^2, \ m = 3, \dots, M.$$

Из предыдущего изложения следует существование ограниченного оператора $R^{-1} = (R_1 + \dots + R_M)^{-1}$. Тогда

$$1 = [v_\varepsilon] = [R^{-1} R v_\varepsilon] \leqslant \|R^{-1}\| \sum_{m=1}^{M} [R_m v_\varepsilon] \leqslant \|R^{-1}\| \sum_{m=1}^{M} \sqrt{\varepsilon_m}.$$

Легко видеть, что при достаточно малом ε неравенство

$$1 \leqslant \|R^{-1}\| \sum_{m=1}^{M} \sqrt{\varepsilon_m}$$

противоречиво и, следовательно, предположение $\|T\| = 1$ неверно, т. е. $\|T\| < 1$ и итерационный процесс (7.1.11) сходится со скоростью геометрической прогрессии.

Если предположить, что норма оператора T приближается к единице при изменении подпространств $W_1, \dots, W_M$, а постоянная γ при этом не из-

меняется, то, повторив практически без изменения вышеизложенные рассуждения, придем к заключению о справедливости последнего из доказываемых утверждений.

В заключение данного параграфа заметим, что, несмотря на достаточно общий характер условий (7.1.13)—(7.1.15), их можно в ряде случаев использовать для выбора областей D_m ($m = 1, 2, \ldots, M$). Так, например, пусть $M = 2$ и Δ_m есть часть границы ∂D_m, принадлежащая D. Легко заметить, что если Δ_1, Δ_2 имеют точки касания или если точки, в которых Δ_1, Δ_2 соприкасаются, принадлежат ∂D и являются вершинами «нулевых» углов, то в этих случаях условие (7.1.13) заведомо не выполнено. С другой стороны, если подобласть $\Delta D = D_1 \cap D_2$ имеет липшицеву границу и точки соприкосновения Δ_1, Δ_2 с ∂D расположены друг от друга на положительном расстоянии, то условия (7.1.13)—(7.1.15) оказываются выполненными. А значит, метод Шварца в этом случае будет сходиться со скоростью геометрической прогрессии.

7.2. Метод разделения области

В данном параграфе мы рассмотрим лишь некоторые из алгоритмов метода разделения области. Ради определенности они будут изложены применительно к конкретной задаче — задаче Дирихле для эллиптического уравнения второго порядка.

7.2.1. Алгоритмы метода разделения области

Пусть в $\mathbf{R}^n$ задана ограниченная область D с кусочно-гладкой границей ∂D, локально удовлетворяющей условию Липшица. Рассмотрим задачу

$$Lu \equiv -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i} a_{ij}(x) \frac{\partial u}{\partial x_j} + c(x)u = f(x),, \tag{7.2.1}$$

$$x = (x_1, \ldots, x_n) \in D,$$

$$u(x) = 0, \quad x \in \partial D, \tag{7.2.2}$$

где $a_{ij}(x) = a_{ji}(x)$, $c(x) \geqslant 0$ — кусочно-гладкие вещественные функции, причем

$$\mu_0 \sum_{i=1}^{2} \xi_i^2 \leqslant \sum_{i,j=1}^{2} a_{ij}(x)\xi_i\xi_j \leqslant \mu_1 \sum_{i=1}^{2} \xi_i^2$$

$$\mu_0, \mu_1 = \text{const} > 0, \quad x \in \bar{D} = D \cup \partial D,$$

$$f(x) \in L_2(D).$$

Обобщенная форма записи задачи (7.2.1), (7.2.2) имеет вид

$$[u, v] = (f, v), \tag{7.2.3}$$

где v — произвольная функция из $\overset{\circ}{W}{}_2^1(D)$, u — искомое решение из класса $\overset{\circ}{W}{}_2^1(D)$, $(\cdot,\cdot) \equiv (\cdot,\cdot)_{L_2(D)}$, а также

$$[u, v] = \int\limits_D \left(\sum_{i,j=1}^{2} a_{ij}(x) \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} + c(x)uv \right) dx.$$

Несложно доказать, что при сделанных ограничениях задача (7.2.1), (7.2.2) имеет единственное обобщенное решение $u(x) \in \overset{\circ}{W}{}_2^1(D)$, удовлетворяющее (7.2.3). При необходимости гладкости коэффициентов и дополнительных ограничениях на ∂D это решение будет удовлетворять почти всюду в D уравнению (7.2.1). Поэтому в дальнейшем, когда мы для упрощения изложения будем записывать задачи в классической форме (т. е. в форме записи, подобной (7.2.1), (7.2.2)), будем считать, что решения и исходные данные этих задач обладают необходимой гладкостью.

Пусть поверхность γ разбивает область D на две подобласти D_1 и D_2. Границу подобласти обозначим через $\Gamma_1 = \gamma \cup \partial D_1$, а границу D_2 — через $\Gamma_2 = \gamma \cup \partial D_2$, где ∂D_i — часть границы Γ_i, принадлежащая ∂D, т. е. $\partial D_i = \partial D \cap \Gamma_i$. Предполагается, что границы Γ_1, Γ_2 также локально удовлетворяют условию Липшица, кроме того, пусть $\text{mes}(\partial D_i) > 0$ $(i = 1, 2)$.

Для решения задачи (7.2.1), (7.2.2) используем следующий алгоритм. Пусть функция $u_i^0(x)$ $(i = 1, 2)$ определена на D_i, при этом $u_i^k(x) \in W_2^1(D_i)$, $u_i^k = 0$ на ∂D_i. Предположим, что на γ определена (может быть, в слабом смысле) внешняя по отношению к D_i конормальная производная

$$\frac{\partial u_i^k}{\partial \nu_i} = \sum_{l,j=1}^{n} a_{lj}(x) \frac{\partial u_i^k}{\partial x_j} \cos(n_i, x_l),$$

где n_i — внешняя по отношению к D_i нормаль. Пусть $p_k \geqslant 0$, $q_k \geqslant 0$, $\{\alpha_k\}$, $\{\beta_k\}$ есть заданные постоянные. Определим по функциям u_1^k, u_2^k приближения u_1^{k+1}, u_2^{k+1} следующим образом. Полагаем

$$v_2^k = p_k u_2^k - \frac{\partial u_2^k}{\partial \nu_2} \quad \text{на} \quad \gamma. \tag{7.2.4}$$

Ищем в D_1 решение задачи

$$\begin{aligned} Lu_1^{k+1/2} &= f, \quad x \in D_1, \\ u_1^{k+1/2} &= 0 \quad \text{на} \quad \partial D_1, \\ \frac{\partial u_1^{k+1/2}}{\partial \nu_1} + p_k u_1^{k+1/2} &= v_2^k \quad \text{на} \quad \gamma. \end{aligned} \tag{7.2.5}$$

Вычисляем $u_1^{k+1}(x)$ по формуле

$$u_1^{k+1} = u_1^k + \alpha_{k+1}(u_1^{k+1/2} - u_1^k), \quad x \in \bar{D}_1. \tag{7.2.6}$$

Находим

$$v_1^{k+1} = -q_k \frac{\partial u_1^{k+1}}{\partial \nu_1} + u_1^{k+1} \quad \text{на} \quad \gamma \tag{7.2.7}$$

Решаем задачу

$$\begin{aligned} Lu_2^{k+1/2} &= f, \quad x \in D_2, \\ u_2^{k+1/2} &= 0 \quad \text{на} \quad \partial D_2, \\ q_k \frac{\partial u_2^{k+1/2}}{\partial \nu_2} + u_2^{k+1/2} &= v_1^k \quad \text{на} \quad \gamma. \end{aligned} \tag{7.2.8}$$

Вычисляем $u_2^{k+1}(x)$ по формуле

$$u_2^{k+1} = u_2^k + \beta_{k+1}(u_2^{k+1/2} - u_2^k), \quad x \in \bar{D}_2. \tag{7.2.9}$$

Формулами (7.2.4)—(7.2.9) задан ряд алгоритмов метода разделения области. Выбирая так или иначе постоянные p_k, q_k, α_k, β_k, мы получим соответствующий алгоритм. Приведем некоторые частные случаи итерационного метода (7.2.4)—(7.2.9).

Если

$$p_k = q_k = 0, \quad \alpha_k = \beta_k = \alpha,$$

то итерационный процесс принимает вид

$$\begin{aligned}
Lu_1^{k+1/2} &= f, \quad x \in D_1,\\
u_1^{k+1/2} &= 0 \quad \text{на} \quad \partial D_1,\\
\frac{\partial u_1^{k+1/2}}{\partial \nu_1} &= -\frac{\partial u_2^k}{\partial \nu_2} \quad \text{на} \quad \gamma,\\
u_1^{k+1} &= u_1^k + \alpha(u_1^{k+1/2} - u_1^k), \quad x \in \bar{D}_1,\\
Lu_2^{k+1/2} &= f, \quad x \in D_2,\\
u_2^{k+1/2} &= 0 \quad \text{на} \quad \partial D_2,\\
u_2^{k+1/2} &= u_1^{k+1} \quad \text{на} \quad \gamma,\\
u_2^{k+1} &= u_2^k + \alpha(u_2^{k+1/2} - u_2^k), \quad x \in \bar{D}_2.
\end{aligned} \tag{7.2.10}$$

При

$$p_k = q_k = q, \quad \alpha_k = \beta_k = 1$$

имеем

$$\begin{aligned}
v_2^k &= qu_2^k - \frac{\partial u_2^k}{\partial \nu_2} \quad \text{на} \quad \gamma,\\
Lu_1^{k+1} &= f, \quad x \in D_1,\\
u_1^{k+1} &= 0 \quad \text{на} \quad \partial D_1,\\
\frac{\partial u_1^{k+1}}{\partial \nu_1} + qu_1^{k+1} &= v_2^k \quad \text{на} \quad \gamma,\\
v_1^{k+1} &= -q\frac{\partial u_1^{k+1}}{\partial \nu_1} + u_1^{k+1} \quad \text{на} \quad \gamma,\\
Lu_2^{k+1} &= f, \quad x \in D_2,\\
u_2^{k+1} &= 0 \quad \text{на} \quad \partial D_2,\\
q\frac{\partial u_2^{k+1}}{\partial \nu_2} + u_2^{k+1} &= v_1^{k+1} \quad \text{на} \quad \gamma.
\end{aligned} \tag{7.2.11}$$

Если же

$$p_k = q_k = 0, \quad \alpha_{k+1} = \beta_{k+1} = 1,$$

то из (7.2.4)—(7.2.9) получаем метод простой итерации:

$$\begin{aligned}
Lu_1^{k+1} &= f, \quad x \in D_1,\\
u_1^{k+1} &= 0 \quad \text{на} \quad \partial D_1,\\
\frac{\partial u_1^{k+1}}{\partial \nu_1} &= -\frac{\partial u_2^k}{\partial \nu_2} \quad \text{на} \quad \gamma,\\
Lu_2^{k+1} &= f, \quad x \in D_2,\\
u_2^{k+1} &= 0 \quad \text{на} \quad \partial D_2,\\
u_2^{k+1} &= u_1^{k+1} \quad \text{на} \quad \gamma.
\end{aligned} \tag{7.2.12}$$

Приняв

$$p_k = q_k = 0, \quad \alpha_{k+1} = 1,$$

приходим к следующему алгоритму:

$$\begin{aligned}
Lu_1^{k+1} &= f, \quad x \in D_1, \\
u_1^{k+1} &= 0 \quad \text{на} \quad \partial D_1, \\
\frac{\partial u_1^{k+1}}{\partial \nu_1} &= -\frac{\partial u_2^k}{\partial \nu_2} \quad \text{на} \quad \gamma, \\
Lu_2^{k+1/2} &= f, \quad x \in D_2, \\
u_2^{k+1/2} &= 0 \quad \text{на} \quad \partial D_2, \\
u_2^{k+1/2} &= u_1^{k+1} \quad \text{на} \quad \gamma, \\
u_2^{k+1} &= u_2^k + \beta_{k+1}(u_2^{k+1/2} - u_2^k), \quad x \in \bar{D}_2.
\end{aligned} \tag{7.2.13}$$

Если в (7.2.4)—(7.2.9) положить

$$\alpha_k = \beta_k = 1,$$

то получаем алгоритм вида

$$\begin{aligned}
v_2^k &= p_k u_2^k - \frac{\partial u_2^k}{\partial \nu_2} \quad \text{на} \quad \gamma, \\
Lu_1^{k+1} &= f, \quad x \in D_1, \\
u_1^{k+1} &= 0 \quad \text{на} \quad \partial D_1, \\
\frac{\partial u_1^{k+1}}{\partial \nu_1} + p_k u_1^{k+1} &= v_2^k \quad \text{на} \quad \gamma, \\
v_1^{k+1} &= -q_k \frac{\partial u_1^{k+1}}{\partial \nu_1} + u_1^{k+1} \quad \text{на} \quad \gamma, \\
Lu_2^{k+1} &= f, \quad x \in D_2, \\
u_2^{k+1} &= 0 \quad \text{на} \quad \partial D_2, \\
q_k \frac{\partial u_2^{k+1}}{\partial \nu_2} + u_2^{k+1} &= v_1^{k+1} \quad \text{на} \quad \gamma.
\end{aligned} \tag{7.2.14}$$

Как мы видим, в приведенных алгоритмах, как правило, присутствует один или несколько параметров $p_k, \ldots, \beta_k$. Выбирая их подходящим образом, мы можем добиться не только сходимости соответствующего алгоритма, но и, возможно, того, чтобы его скорость сходимости была наиболее быстрой. Так, используя теорию чебышевских итерационных процессов, путем специального выбора параметров $\{\beta_k\}$ в (7.2.13) можно прийти к алгоритму с оптимальной скоростью сходимости. При рассмотрении алгоритма (7.2.14) можно показать, что здесь оператор перехода является оператором пере-

хода метода переменных направлений. Следовательно, в этом алгоритме можно применить известные приемы выбора $\{p_k\}$, $\{q_k\}$ с целью ускорения итерационного процесса.

Отметим важность введения в итерационные алгоритмы метода разделения области параметров p_k, q_k, α_k, β_k. Так, например, рассмотрим метод простой итерации (7.2.12), т. е. процесс без каких-либо параметров. Оказывается, что он, вообще говоря, не будет сходиться. Действительно, пусть $D \subset \mathbf{R}^2$, $D_1 = \{0 < x_1 < 1, 0 < x_2 < 1\}$, $D_2 = \{-1 < x_1 < 1, 0 < x_2 < 1\}$, а оператор L есть $Lu = -\Delta u$. Тогда алгоритм (7.2.12) имеет вид

$$\begin{aligned}
-\Delta u_1^{k+1} &= f, \quad x \in D_1, \\
u_1^{k+1} &= 0 \quad \text{на} \quad \partial D_1, \\
\frac{\partial u_1^{k+1}}{\partial n_1} &= -\frac{\partial u_2^k}{\partial n_2} \quad \text{на} \quad \gamma, \\
-\Delta u_2^{k+1} &= f, \quad x \in D_2, \\
u_2^{k+1} &= 0 \quad \text{на} \quad \partial D_2, \\
u_2^{k+1} &= u_1^{k+1} \quad \text{на} \quad \gamma,
\end{aligned}$$

где u_2^0 — заданная функция.

Рассматривая этот процесс для ошибок $\varepsilon_i^\alpha = u_i - u_i^\alpha$, легко заметить, что $\varepsilon_2^k = \varepsilon_2^0$, $\varepsilon_1^k = \varepsilon_1^1$ и $\|\nabla \varepsilon_1^{k+1}\|^2_{L_2(D_1)} = \|\nabla \varepsilon_2^k\|^2_{L_2(D_2)} = \|\nabla \varepsilon_1^k\|^2_{L_2(D_1)} = \dots$.

Таким образом, нормы $\|\nabla \varepsilon_i^\alpha\|_{L_2(D_i)}$ для каждого i не изменяются в результате итерационного процесса, который в данном случае оказывается не сходящимся в метриках $\|\cdot\|_{W_2^1(D_i)}$ ($i = 1, 2$).

7.2.2. Сходимость алгоритмов

Проиллюстрируем теперь некоторые подходы к изучению проблемы сходимости сформулированных выше алгоритмов разделения области. Для простоты мы будем рассматривать лишь алгоритм (7.2.13) при $\beta_k \equiv \beta$ ($k = 1, 2, \dots$):

$$\begin{aligned}
Lu_1^{k+1} &= f, \quad x \in D_1, \\
u_1^{k+1} &= 0 \quad \text{на} \quad \partial D_1, \\
\frac{\partial u_1^{k+1}}{\partial \nu_1} &= -\frac{\partial u_2^k}{\partial \nu_2} \quad \text{на} \quad \gamma, \\
Lu_2^{k+1/2} &= f, \quad x \in D_2, \\
u_2^{k+1/2} &= 0 \quad \text{на} \quad \partial D_2, \\
u_2^{k+1/2} &= u_1^{k+1} \quad \text{на} \quad \gamma, \\
u_2^{k+1} &= u_2^k + \beta(u_2^{k+1/2} - u_2^k), \quad x \in \bar{D}_2.
\end{aligned} \tag{7.2.15}$$

В дальнейшем нам потребуется несколько вспомогательных утверждений, касающихся L-гармонических в D_i функций $u(x) \in W_2^1(D_i)$, равных нулю на ∂D_i. Здесь функцию $u(x) \in W_2^1(D_i)$ мы называем L-гармонической, если она почти всюду в D_i удовлетворяет уравнению $Lu = 0$.

Пусть для определенности рассматриваются подобласть D_1 и L-гармонические в D_1 функции $u(x) \in W_2^1(D_1)$, такие, что $u = 0$ на ∂D_1. Множество таких функций обозначим через F_1. Зададим на F_1 оператор вида

$$B_1 u = \frac{\partial}{\partial \nu_1} u|_\gamma, \tag{7.2.16}$$

т. е. оператор конормальной производной на γ. В силу ограничений на коэффициенты уравнения (7.2.1) оператор симметричен и положительно определен в $L_2(\gamma)$. Действительно, если $u, v \in F_1$, то, пользуясь формулой Грина, имеем

$$0 = (Lu, v)_{L_2(D_1)} = [u, v]_1 - \left(\frac{\partial u}{\partial \nu_1}, v\right)_{L_2(\gamma)},$$
$$\left(\frac{\partial u}{\partial \nu_1}, v\right)_{L_2(\gamma)} = [u, v]_1,$$

где

$$[u, v]_1 = \int\limits_{D_1} \left(\sum_{i,j=1}^{2} a_{ij}(x) \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} + c(x) uv \right) dx.$$

А тогда

$$\left(\frac{\partial u}{\partial \nu_1}, v\right)_{L_2(\gamma)} = [u, v]_1 = \left(u, \frac{\partial v}{\partial \nu_1}\right)_{L_2(\gamma)},$$

т. е. оператор B_1 симметричен в $L_2(\gamma)$. Он является также положительно определенным. Действительно,

$$\left(\frac{\partial u}{\partial \nu_1}, u\right)_{L_2(\gamma)} = [u, u]_1 \geqslant \mu_0 \|\nabla u\|^2_{L_2(D_1)}.$$

А поскольку $u = 0$ на ∂D_1, то

$$\|\nabla u\|^2_{L_2(D_1)} \geqslant c_1 \|u\|^2_{L_2(D_1)}, \quad c_1 = \text{const} > 0,$$
$$\|\nabla u\|^2_{L_2(D_1)} \geqslant c_2 \|u\|^2_{W_2^1(D_1)} \geqslant c_3 \|u\|^2_{L_2(\gamma)},$$

где $c_i = \text{const} > 0$. Следовательно,

$$\left(\frac{\partial u}{\partial \nu_1}, v\right)_{L_2(\gamma)} \geqslant \mu_0 c_3 \|u\|^2_{L_2(\gamma)},$$

т. е. оператор B_1 положительно определен в $L_2(\gamma)$. Отмеченные свойства B_1 позволяют на F_1 ввести скалярное произведение и норму вида

$$(u,v)_{F_1} = \left(\frac{\partial u}{\partial \nu_1}, v\right)_{L_2(\gamma)} = [u,v]_1,$$
$$\|u\|_{F_1} = \left(\frac{\partial u}{\partial \nu_1}, v\right)^{1/2}_{L_2(\gamma)} = [u,u]_1^{1/2}.$$

Заметим также, что если функция $g \in L_2(\gamma)$ принадлежит области значений оператора B_1, то уравнение $B_1 v = g$ имеет единственное решение $v = B_1^{-1} g$, т. е. B_1 имеет обратный оператор.

Аналогичными свойствами обладает оператор

$$B_2 u = \left.\frac{\partial u}{\partial \nu_2}\right|_{\gamma},$$

определенный на L-гармонических в D_2 функциях $u \in W_2^1(D_2)$, равных нулю на ∂D_2. Множество этих функций обозначаем через F_2, а скалярное произведение на F_2 и норма, порождаемые оператором B_2, имеют вид

$$(u,v)_{F_2} = \left(\frac{\partial u}{\partial \nu_2}, v\right)_{L_2(\gamma)} = [u,v]_2, \quad \|u\|_{F_2} = (u,u)_{F_2}^{1/2},$$

где

$$[u,v]_2 = \int\limits_{D_2} \left(\sum_{i,j=1}^{2} a_{ij}(x) \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} + c(x) uv \right) dx.$$

Рассмотрим такие две функции $U_1 \in F_1$, $U_2 \in F_2$, что $U_1|_\gamma = U_2|_\gamma = g$. Очевидно, что они являются решениями задач

$$\begin{gathered} LU_k = 0, \qquad x \in D_k, \\ U_k = 0 \quad \text{на} \quad \partial D_k, \\ U_k = g \quad \text{на} \quad \gamma,\ k = 1,2. \end{gathered} \tag{7.2.17}$$

Предположим, что коэффициенты уравнения (7.2.1) и разбиение D на D_1, D_2 таковы, что для отношения

$$d = \frac{[U_2,U_2]_2}{[U_1,U_1]_1} = \frac{(B_2U_2,U_2)_{L_2(\gamma)}}{(B_1U_1,U_1)_{L_2(\gamma)}} \tag{7.2.18}$$

справедливы оценки

$$m \leqslant d \leqslant M, \tag{7.2.19}$$

где постоянные m, M не зависят от функции g. Легко заметить, что при $g \not\equiv 0$ имеем $m > 0$, $M < \infty$. Конкретные значения m, M целесообразно получать для каждого случая разбиения D на D_1, D_2 и вида оператора L.

Пример 1. Пусть $Lu = -\Delta u$ в D_k $(k = 1, 2)$, а D, D_1, D_2 таковы, что D_1, D_2 симметричны относительно γ (т. е. D является зеркальным отражением D_2 относительно γ). Замечаем, что в этом случае $m = M = 1$.

Исследуем теперь проблему сходимости процесса (7.2.15). Для этого запишем его для ошибок $\varepsilon_k^\alpha(x) = u(x) - u_k^\alpha(x)$, $x \in \bar{D}_k$ $(k = 1, 2)$:

$$\begin{aligned}
L\varepsilon_1^{k+1} &= f, \quad x \in D_1, \\
\varepsilon_1^{k+1} &= 0 \quad \text{на} \quad \partial D_1, \\
\frac{\partial \varepsilon_1^{k+1}}{\partial \nu_1} &= -\frac{\partial \varepsilon_2^k}{\partial \nu_2} \quad \text{на} \quad \gamma, \\
L\varepsilon_2^{k+1/2} &= f, \quad x \in D_2, \\
\varepsilon_2^{k+1/2} &= 0 \quad \text{на} \quad \partial D_2, \\
\varepsilon_2^{k+1/2} &= \varepsilon_1^{k+1} \quad \text{на} \quad \gamma, \\
\varepsilon_2^{k+1} &= \varepsilon_2^k + \beta(\varepsilon_2^{k+1/2} - \varepsilon_2^k), \quad x \in \bar{D}_2.
\end{aligned} \tag{7.2.20}$$

Поскольку $\varepsilon_1^{k+1} \in F_1$, $\varepsilon_2^{k+1} \in F_2$, то из уравнений (7.2.20) можно оставить лишь уравнения на γ:

$$\begin{aligned}
\frac{\partial \varepsilon_1^{k+1}}{\partial \nu_1} &= -\frac{\partial \varepsilon_2^k}{\partial \nu_2} \quad \text{на} \quad \gamma, \\
\varepsilon_2^{k+1/2} &= \varepsilon_1^{k+1} \quad \text{на} \quad \gamma, \\
\varepsilon_2^{k+1} &= \varepsilon_2^k + \beta(\varepsilon_2^{k+1/2} - \varepsilon_2^k) \quad \text{на} \quad \gamma.
\end{aligned} \tag{7.2.21}$$

Из первого из них находим $\varepsilon_1^{k+1} = -B_1^{-1} B_2 \varepsilon_2^k$, а затем из остальных двух имеем

$$\varepsilon_2^{k+1} = \varepsilon_2^k - \beta(B_1^{-1} B_2 + E)\varepsilon_2^k \quad \text{на} \quad \gamma, \tag{7.2.22}$$

где E — тождественный оператор. Рассмотрим это уравнение в пространстве F_1, считая функции ε_2^α продолженными на D_1 как решения задач (7.2.17) при $k = 1$, $g = \varepsilon_2^\alpha$ на γ. Замечаем, что в силу свойств операторов B_1, B_2 оператор

$$A = B_1^{-1} B_2 + E$$

в пространстве F_1 симметричен и положительно определен. Действительно,

$$(Au, v)_{F_1} = (B_1 Au, v)_{L_2(\gamma)} = (B_2 u, v)_{L_2(\gamma)} + (B_1 u, v)_{L_2(\gamma)} = (u, Av)_{F_1},$$
$$(Au, u)_{F_1} \geqslant \|u\|_{F_1}^2.$$

Кроме того, из (7.2.19) следуют оценки для верхней и нижней границ оператора A:

$$\sup_{\|u\|_{F_1}=1}(Au,u)_{F_1}\leqslant 1+M\equiv\tilde{\beta},$$
$$\inf_{\|u\|_{F_1}=1}(Au,u)_{F_1}\geqslant 1+m\equiv\tilde{\alpha}.$$

А значит, если мы выберем значение параметров β равным (см. 4.2.1)

$$\beta=\frac{2}{\tilde{\alpha}+\tilde{\beta}}=\frac{2}{2+m+M},$$

то получим итерационный процесс (7.2.15) с оптимальной скоростью сходимости. Оценка скорости сходимости в этом случае будет иметь вид

$$\|\varepsilon_2^{k+1}\|_{F_1}\leqslant\frac{M-m}{2+m+M}\|\varepsilon_2^k\|_{F_1}\leqslant\cdots\leqslant c\left(\frac{M-m}{2+m+M}\right)^{k+1},$$

где постоянная c не зависит от k. Как следствие этой оценки имеем

$$\sum_{j=1}^{2}\|u-u_j^{k+1}\|_{W_2^1(D_j)}\leqslant c\sum_{j=1}^{2}[u-u_j^{k+1},u-u_j^{k+1}]_j^{1/2}\leqslant c\left(\frac{M-m}{2+m+M}\right)^{k+1}\to 0,\quad k\to\infty. \tag{7.2.23}$$

Итак, сходимость алгоритма установлена, а также доказана оценка скорости сходимости. Причем в данном алгоритме параметр β выбран оптимальным образом. В этом случае, если имеем лишь оценки вида $m\lesssim d\lesssim M$, можно надеяться, что при значении $\beta=2/(2+m+M)$ рассматриваемый алгоритм будет сходящимся, а скорость его сходимости будет близка к оптимальной.

Пример 2. Пусть оператор L и области D, D_1, D_2 те же самые, что и в примере 1. Тогда здесь при $\beta=1/2$ процесс (7.2.15) сойдется на одну итерацию (что легко заметить из (7.2.23)).

Аналогичным образом можно исследовать и другие алгоритмы метода разделения области.

7.2.3. Распараллеливание процесса решения задач[2)]

Пусть $D\subset R_2$ ($x_1\equiv x, x_2\equiv y$) есть прямоугольник со сторонами A (по оси x) и b (по оси y) и в D рассматривается задача (7.2.1), (7.2.2). Как уже отмечалось, обобщенное решение этой задачи существует, и оно единственно.

[2)] См. В. И. Агошков, В. И. Лебедев [23].

Требуется отыскать его численное приближение. Предположим, что предлагаемый метод дискретизации этой задачи (разностный, проекционно-сеточный и т. п.) приводит к системе уравнений, «трудно» поддающейся решению с помощью имеющейся у пользователя ЭВМ. Поэтому в данном случае задачу можно попытаться решить с помощью метода разделения областей, который сводит процесс решения всей задачи к последовательному решению однотипных задач по подобластям. Такой итерационный алгоритм потребует ЭВМ уже с меньшими ресурсами памяти. С другой стороны, если у пользователя имеется доступ к системе из нескольких ЭВМ, то возникает возможность применения к решению всей задачи параллельных алгоритмов, что позволит значительно ускорить процесс построения решения во всей области. Таким образом, здесь возникает вопрос создания алгоритма, пригодного для распараллеливания вычислений. Одним из таких алгоритмов является нестационарный алгоритм метода разделения областей с переменными параметрами, задаваемый формулами (7.2.13).

Разобьем D на подобласти прямыми $x = x_i$, параллельными оси $(0, y)$, $i = 1, 2, \ldots, N$, $x_0 = 0$, $x_{N+1} = A$. Границу между подобластями, лежащую на прямой $x = x_i$, обозначим $\gamma_{i+1/2}$, и пусть

$$\gamma = \bigcup_{j=1}^{N} \gamma_{j+1/2}, \quad a_j = x_j - x_{j-1}, \quad j = 1, 2, \ldots, N,$$

$$a_{\min} = \min_i a_i, \quad a_{\max} = \max_i a_i.$$

Нечетные по порядку (нумерация ведется слева направо) подобласти обозначим $D^{(1)}_{2j-1}$ ($j = 1, \ldots, J_1$), а нечетные — через $D^{(2)}_{2j}$ ($j = 1, 2, \ldots, J_2$). Пусть также $D_1 = \bigcup_{j=1}^{J_1} D^{(1)}_{2j-1}$, $D_2 = \bigcup_{j=1}^{J_2} D^{(2)}_{2j}$, Γ_j — граница области D_j ($j = 1, 2$), $\partial D_j = \Gamma_j \setminus \gamma$ (рис. 7.1). Через ∂D_i обозначается часть границы области D_i, совпадающая с ∂D, а через γ — граница между D_1 и D_2, $\gamma \subset D$. Итерационный алгоритм решения задачи записывается в виде (7.2.13).

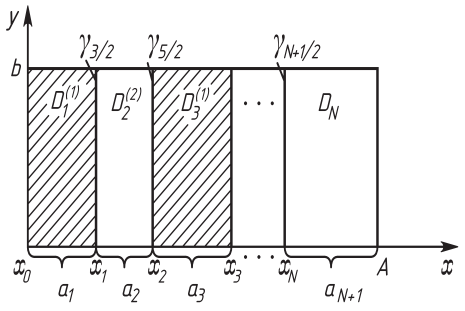


Рис. 7.1.

Остановимся на некоторых вопросах реализации описанного алгоритма. В силу введенных разбиений областей D, D_1, D_2 замечаем, что каждая из задач в (7.2.13) допускает разбиение на ряд подзадач в областях $\{D_j^{(1)}\}$, $\{D_j^{(2)}\}$. Введем обозначения

$$u_1^\alpha = u_{1,j}^\alpha \quad \text{в} \quad D_j^{(1)} \subset D_1,$$
$$u_2^\alpha = u_{2,j}^\alpha \quad \text{в} \quad D_j^{(2)} \subset D_2.$$

Тогда алгоритм (7.2.13) можно записывать так:

$$\begin{aligned} Lu_{1,2j-1}^{k+1} &= f, \quad x \in D_{2j-1}^{(1)}, \\ u_{1,2j-1}^{k+1} &= 0 \quad \text{на} \quad (\partial D_1 \cap \Gamma_{2j-1}^{(1)}), \\ \frac{\partial u_{1,2j-1}^{k+1}}{\partial \nu_1} &= -\frac{\partial u_2^k}{\partial \nu_2} \quad \text{на} \quad (\gamma \cap \Gamma_{2j-1}^{(1)}), \\ j &= 1, 2, \ldots, J_1; \end{aligned} \tag{7.2.24}$$

$$\begin{aligned} Lu_{2,2j}^{k+1/2} &= f, \quad x \in D_{2j}^{(2)}, \\ u_{2,2j}^{k+1/2} &= 0 \quad \text{на} \quad (\partial D_2 \cap \Gamma_{2j}^{(2)}), \\ u_{2,2j}^{k+1/2} &= u_1^{k+1} \quad \text{на} \quad (\gamma \cap \Gamma_{2j}^{(2)}), \\ u_{2,2j}^{k+1} &= u_{2,2j}^k + \beta_{k+1}(u_{2,2j}^{k+1/2} - u_{2,2j}^k), \quad x \in \bar{D}_{2j}^{(2)}. \\ j &= 1, 2, \ldots, J_2. \end{aligned} \tag{7.2.25}$$

Здесь $\Gamma_i^{(k)}$ есть граница $D_i^{(k)}$ ($k = 1, 2$).

Из (7.2.23), (7.2.24) заключаем, что при наличии $J_{\max} = \max(J_1, J_2)$ параллельно действующих ЭВМ или специализированных процессоров решения задач (7.2.25) и J_2 задач из (7.2.24) можно осуществлять параллельно. Следовательно, при наличии системы из $J_{\max}$ параллельно работающих ЭВМ применение для решения задачи (7.2.1), (7.2.2) алгоритма (7.2.23), (7.2.24) может позволить ускорить процесс получения приближенного решения исходной задачи.

С другой стороны, если у пользователя имеется доступ к ЭВМ с небольшим быстродействием и оперативной памятью, то алгоритм (7.2.23), (7.2.24) также допускает реализацию на данной ЭВМ путем последовательного решения подзадач на каждой из областей $D_{2j-1}^{(1)}$, $D_{2j}^{(2)}$. Как легко заметить, запоминать здесь необходимо лишь информацию о приближенном решении задачи в окрестностях γ, на что требуется значительно меньший объем памяти ЭВМ, чем для запоминания информации о решении во всей области D.

Пусть введены обозначения

$$M = -\frac{\mu_0 \left(\text{ch}\frac{\pi a_{\min}}{b} - 1\right)\pi^2}{(\pi^2\mu_1 + a_1 b^2)\left(\text{ch}\frac{\pi a_{\min}}{b} + 1\right)},$$
$$m = -\frac{(\pi^2\mu_1 + a_1 b^2)\left(\text{ch}\frac{\pi a_{\min}}{b} + 1\right)}{\mu_0\pi^2\left(\text{ch}\frac{\pi a_{\min}}{b} - 1\right)},$$

а $\{\omega_i\}_{i=1}^N$ есть заданный набор параметров, выбор которых обеспечивает устойчивость итерационного процесса (см. 4.2.4). Сформулируем без доказательства следующее утверждение: если области D, D_1, D_2 имеют вид, изображенный на рис. 7.1, а параметры $\{\beta_i\}$ выбираются по формуле

$$\beta_i = \left[1 - \frac{1}{2}(M + m - (M - m)\cos\omega_i\pi)\right]^{-1}, \tag{7.2.26}$$

то алгоритм (7.2.23), (7.2.24) обладает оптимальной скоростью сходимости, а после $l \cdot N$ итераций алгоритма, в котором параметры β_i циклически повторяются с периодом N ($\beta_i = \beta_{i+N}$), справедливы оценки погрешностей

$$\|u_i - u_i^{l\cdot N}\|_{W_2^1(D_i)} \leqslant cq^l, \quad i = 1, 2, \tag{7.2.27}$$

где $q = \left|T_N\left(\frac{2-M-m}{M-m}\right)\right|^{-1} < 1$, $T_N(t)$ — полином Чебышева порядка N, при этом постоянная c не зависит от номера итерации и функций u_i, $u_i^{(k)}$.

Итак, в данном параграфе сформулирован ряд алгоритмов метода разделения области в применении к задачам без предварительной аппроксимации их каким-либо методом. Однако, можно сначала осуществить данную аппроксимацию, а затем уже рассматривать методы решения полученных схем и получать алгоритмы, которые можно интерпретировать как методы разделения области (либо как специфические алгоритмы линейной алгебры). Данное направление вычислительной математики в настоящее время также интенсивно развивается.

7.3. Методы разделения области в нестационарных задачах

Алгоритмы, изложенные в предыдущих параграфах данной главы, можно применить и для решения нестационарных задач. Один из простейших путей их использования состоит в следующем. Сначала нестационарные задачи аппроксимируются по временной переменной подходящей раз-

ностной схемой так, чтобы на каждом шаге получались уравнения с операторами (конечно, если это возможно сделать), допускающими применение рассмотренных ранее алгоритмов. А затем осуществляют решение последовательности стационарных задач данными алгоритмами.

Однако возможны и другие пути применения метода разделения области в нестационарных задачах. При этом они могут быть такими, чтобы учитывать специфику задачи, т. е. тот факт, что она является нестационарной. Одно из таких применений метода разделения мы изложим в данном параграфе, и его можно рассматривать как применение метода разделения для реализации неявных разностных схем для параболических задач.

Пусть D — ограниченная область m-мерного евклидова пространства $(m = 1, 2, 3)$ с кусочно-линейной границей ∂D; Γ_0 — часть ∂D, состоящая из конечного числа отрезков прямых, и Γ_1 — такое открытое подмножество ∂D, что $\Gamma_0 \cap \Gamma_1 = \varnothing$ и $\Gamma_0 \cup \bar{\Gamma}_1 = \partial D$.

Рассмотрим параболическую дифференциальную задачу

$$\begin{aligned} \frac{\partial u}{\partial t} + Lu = f \quad &\text{в} \quad D \times (0, T], \\ u = 0 \quad &\text{на} \quad \Gamma_0 \times (0, T], \\ lu = 0 \quad &\text{на} \quad \Gamma_1 \times (0, T], \\ u(0) = g \quad &\text{в} \quad D. \end{aligned}$$

Здесь $lu \equiv \partial u / \partial \nu + \sigma u$, где ν — вектор внешней нормали к Γ_1 и σ — неотрицательная кусочно-гладкая функция, а

$$Lu \equiv -\sum_{i,j=1}^{m} \frac{\partial}{\partial x_i} a_{ij} \frac{\partial u}{\partial x_j} + au \tag{7.3.1}$$

— симметричный эллиптический оператор с кусочно-гладкими ограниченными (по модулю) коэффициентами a_{ij} и a, которые зависят только от переменной $x = (x_1, \ldots, x_m)$.

Построим на отрезке $[0, T]$ равномерную сетку с шагом $\tau = T/M$, где M — некоторое положительное целое число. Затем построим в D регулярную прямоугольную сетку D_h с числом узлов N, где $h = N^{-m}$, и определим сеточные граничные множества $\Gamma_{0,h}$ и $\Gamma_{1,h}$.

Для построения сеток определим сеточные операторы L_h и l_h и предположим, что на множестве всех сеточных функций $\{v_h\}$, удовлетворяющих однородным краевым условиям $v_h = 0$ на $\Gamma_{0,h}$ и $l_h v_h = 0$ на $\Gamma_{1,h}$, оператор L_h является симметричным.

В результате проделанных построений для решения сходной дифференциальной системы мы можем применить либо неявную схему точности

$O(\tau)$:

$$\begin{aligned} \frac{u_h^k - u_h^{k-1}}{\tau} + L_h u_h^k &= f_h^k \quad \text{в} \quad D_h, \\ u_h^k &= 0 \quad \text{на} \quad \Gamma_{0,h}, \\ l_h u_h^k &= 0 \quad \text{на} \quad \Gamma_{1,h}, \\ u_h^0 &= g_h \quad \text{в} \quad D_h, \\ k &= 1, 2, \ldots, M, \end{aligned} \tag{7.3.2}$$

либо неявную схему точности $O(\tau^2)$, называемую часто схемой Кранка — Николсона:

$$\begin{aligned} \frac{u_h^k - u_h^{k-1}}{\tau} + L_h \frac{u_h^k + u_h^{k-1}}{2} &= f_h^{k-1/2} \quad \text{в} \quad D_h, \\ u_h^k &= 0 \quad \text{на} \quad \Gamma_{0,h}, \\ l_h u_h^k &= 0 \quad \text{на} \quad \Gamma_{1,h}, \\ u_h^0 &= g_h \quad \text{в} \quad D_h, \\ k &= 1, 2, \ldots, M. \end{aligned} \tag{7.3.3}$$

Далее мы ограничимся рассмотрением только схемы (7.3.2), поскольку (7.3.3) может быть рассмотрена аналогично.

Предметом нашего изучения является реализация одного шага схемы (7.3.2) с некоторой заданной точностью ε. Это означает, что для заданной сеточной функции v_h мы должны вычислить с точностью ε в некоторой сеточной норме $\|\cdot\|_h$ решение u_h сеточной системы

$$\begin{gathered} (E_h + \tau L_h) u_h = v_h + \tau f_h \quad \text{в} \quad D_h, \\ u_h = 0 \quad \text{на} \quad \Gamma_{0,h}, \quad l_h u_h = 0 \quad \text{на} \quad \Gamma_{1,h}, \end{gathered} \tag{7.3.4}$$

где E_h — единичный сеточный оператор, т. е. построить сеточную функцию $\hat{u}_h$, удовлетворяющую неравенству

$$\|\hat{u}_h - u_h\|_h \leqslant \varepsilon. \tag{7.3.5}$$

Перепишем (7.3.4) в виде сеточной системы

$$\begin{gathered} (E_h + \tau L_h) w_h = \tau \xi_h \quad \text{в} \quad D_h, \\ w_h = \psi_{0,h} \quad \text{на} \quad \Gamma_{0,h}, \quad l_h w_h = \psi_{1,h} \quad \text{на} \quad \Gamma_{1,h}, \end{gathered} \tag{7.3.6}$$

где $w_h = u_h - v_h$, $\psi_{0,h} = -v_h$, $\psi_{1,h} = -l_h v_h$ и $\xi_h = -L_h v_h + f_h$. Заметим, что в случае, когда v_h удовлетворяет на $\Gamma_{0,h}$ и $\Gamma_{1,h}$ соответствующим однородным сеточным краевым условиям, решение w_h системы (7.3.6) тоже удовлетворяет этим однородным краевым условиям. Сначала мы рассмотрим только этот случай.

Зададим для некоторого, вообще говоря, произвольного узла $y = (y_1, \ldots, y_m)$ сетки D_h сеточную функцию источник

$$\xi_h(x, y) \equiv \xi_h(x) = \begin{cases} h^{-m}, & x = y, \\ 0, & x \neq y. \end{cases}$$

Тогда при выполнении сделанных выше предположений имеет место следующее утверждение: существует такая не зависящая от h и τ положительная константа c, что для всех узлов x сетки $D_h \cup \Gamma_{1,h}$, удовлетворяющих неравенству[3)]

$$|x - y| \equiv \left[\sum_{i=1}^{m} (x_i - y_i)^2\right]^{1/2} \geqslant c\sqrt{\tau} \ln \left(\frac{\varepsilon h^m}{\tau}\right)^{-1}, \tag{7.3.7}$$

и для любого $\varepsilon = (0, 1)$ выполняется оценка

$$|w_h(x)| < \varepsilon. \tag{7.3.8}$$

Проиллюстрируем последнее утверждение рассмотрением трех простейших примеров.

Пример 1. Применим для решения простейшего уравнения теплопроводности

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = f, \quad (x, t) \in (0, 1) \times (0, T] \tag{7.3.9}$$

с однородными граничными условиями Дирихле $u(0, T) = u(1, t) = 0$ и начальным условием $u(x, 0) = g(x)$ неявную разностную схему

$$\begin{gathered} \frac{u_i^k - u_i^{k-1}}{\tau} - \frac{u_{i-1}^k - 2u_i^k + u_{i+1}^k}{h^2} = f_i^k, \quad i = 1, 2, \ldots, N, \\ u_0^k = u_{N+1}^k = 0, \quad k = 1, 2, \ldots, M, \end{gathered} \tag{7.3.10}$$

где $h = 1/(N + 1)$ и $u_i^0 = g_i$ ($i = 1, 2, \ldots, N$).

Для этой схемы система (7.3.6) имеет вид

$$\begin{gathered} w_i - \frac{\tau}{h^2}(w_{i-1} - 2w_i + w_{i+1}) = \tau\xi_i, \quad i = 1, 2, \ldots, N, \\ w_0 = w_{N+1} = 0. \end{gathered} \tag{7.3.11}$$

Выберем некоторое целое j ($1 \leqslant j \leqslant N$) и положим

$$\xi_i = \begin{cases} h^{-l}, & i = j, \\ 0, & i \neq j. \end{cases}$$

[3)] Ю. А. Кузнецов [23], Г. И. Марчук, Ю. А. Кузнецов [23].

Тогда с помощью несложных вычислений можно показать, что

$$0 < w_i < \frac{\tau}{h}\exp\left\{-\frac{|x_i - x_j|}{2\sqrt{\tau}}\right\}. \tag{7.3.12}$$

Отсюда следует, что для всех узлов $x_i = ih$, удовлетворяющих неравенству

$$|x_i - x_j| > 2\sqrt{\tau}\ln\left(\frac{\varepsilon h}{\tau}\right)^{-1}, \tag{7.3.13}$$

выполняется оценка

$$|w_i| < \varepsilon. \tag{7.3.14}$$

Таким образом, в данном случае константа c из (7.3.7) может быть выбрана равной двум.

Пример 2. Рассмотрим простейшее двумерное уравнение теплопроводности

$$\frac{\partial u}{\partial t} - \Delta u = f, \quad (x,t) \in D \times (0,T], \tag{7.3.15}$$

где $D = (0,1) \times (0,1)$, с однородными краевыми условиями Дирихле. Для решения этого уравнения применим неявную разностную схему

$$\begin{gathered}\frac{u_{ij}^k - u_{ij}^{k-1}}{\tau} - (\Delta_h u^k)_{ij} = f_{ij}^k, \quad 1 \leqslant i,j \leqslant N, \\ u_{ij}^k = 0 \quad \text{на} \quad \partial D_h,\end{gathered} \tag{7.3.16}$$

с пятиточечным разностным оператором Лапласа Δ_h на равномерной сетке D_h с шагом $h = 1/(N+1)$. Для этой схемы система (7.3.6) имеет вид

$$\begin{gathered}w_{ij} - \tau(\Delta_h w)_{ij} = \tau\xi_{ij},, \quad 1 \leqslant i,j \leqslant N, \\ w_{ij} = 0 \quad \text{на} \quad \partial D_h.\end{gathered} \tag{7.3.17}$$

Выберем некоторый узел $\hat{y} = (\hat{y}_1, \hat{y}_2)$ сетки D_h и положим

$$\xi_{ij} = \begin{cases} h^{-2}, & x_{ij} \equiv (x_{1,i}, x_{2,j}) = \hat{y}, \\ 0, & x_{ij} \neq \hat{y}. \end{cases} \tag{7.3.18}$$

С помощью дискретного аналога метода разделения переменных (например, по переменной x_2) и результата из примера 1 легко показать, что для любого узла $x = x_{ij}$ сетки D_h, удовлетворяющего неравенству

$$|x - y| > 2\sqrt{\tau}\ln\left(\frac{\varepsilon h^2}{\tau}\right)^{-1}, \tag{7.3.19}$$

выполняется оценка

$$|w_{ij}| < \varepsilon. \tag{7.3.20}$$

Пример 3 (двумерная задача для полигональной области). Рассмотрим двумерное уравнение теплопроводности

$$\frac{\partial u}{\partial t} - \Delta u = f, \quad (x,t) \in D \times (0,T], \tag{7.3.21}$$

с однородными краевыми условиями Дирихле. Мы предположим, что $D \subset \Pi = (0;1) \times (0;1)$ и ее граница ∂D пересекает линии квадратной сетки Π_h с шагом $h = 1/(1+\tilde{N})$, где $\tilde{N}$ — некоторое положительное целое только в узлах этой сетки. Тогда для решения (7.3.21) может быть применена неявная разностная схема (7.3.16), которая приводит к системе (7.3.17), (7.3.18) с тем лишь отличием, что уравнения рассматриваются только для узлов сетки, принадлежащих D.

Используя теорию монотонных матриц (теория M-матриц), нетрудно показать, что решение системы (7.3.17), (7.3.18) мажорирует сверху решение аналогичной системы для рассматриваемого примера. Следовательно, утверждение (7.3.19), (7.3.20) сохраняет силу и для данного случая.

Основываясь на сформулированном выше утверждении и рассмотренных примерах, применим для приближенного решения системы (7.3.6) с однородными сеточными краевыми условиями метод разделения области. С этой целью сначала покроем иходную прямоугольную сетку D_h некоторой более крупной прямоугольной сеткой G_h, как это показано на рис. 7.2. Мы будем предполагать, что линии сетки G_h принадлежат множеству линий сетки D_h. Таким образом, мы осуществим разбиение сетки D_h на некоторые сеточные подобласти $D_h^{(k,l)}$.

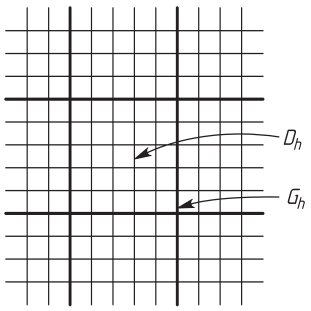


Рис. 7.2.

Поставим задачу вычислить решение системы (7.3.6) для узлов сетки $D_h^{(k,l)}$ для некоторых выбранных значений k и l с заданной точностью $\varepsilon > 0$. Для простоты изложения предположим, что правая часть системы (7.3.6)

удовлетворяет неравенству $\|\xi_h\| < 1$. Погрузим сеточную область $D_h^{(k,l)}$ в сеточную подобласть $\hat{D}_h^{(k,l)}$ области D^h, как показано, например, на рис. 7.3, при дополнительном предположении, что для любого узла x сеточной границы $\partial\hat{D}_h^{(k,l)}$ выполняется неравенство

$$\rho(x, \partial D_h^{(k,l)}) > c\sqrt{\tau}\ln\left(\frac{\varepsilon h^m}{\tau}\right)^{-1}, \tag{7.3.22}$$

где c — константа из неравенства (7.3.7). Здесь через ρ обозначается минимальное геометрическое расстояние между точкой x и точками границы $\partial D_h^{(k,l)}$.

Рис. 7.3.

Рассмотрим теперь следующую систему сеточных уравнений:

$$\begin{aligned}
(E_h + \tau L_h)\hat{w}_k^{(k,l)} &= \tau\xi_h \quad \text{в} \quad \hat{D}_k^{(k,l)} \\
\hat{w}_k^{(k,l)} &= 0 \quad \text{на} \quad \partial\hat{D}_k^{(k,l)} \cap D_h, \\
\hat{w}_k^{(k,l)} &= \psi_{0,h} \quad \text{на} \quad \Gamma_{0,h}, \\
l_h\hat{w}_k^{(k,l)} &= \psi_{1,h} \quad \text{на} \quad \Gamma_{1,h}.
\end{aligned} \tag{7.3.23}$$

Опираясь на предыдущие построения и сделанные предположения, нетрудно показать, что для любого узла x сетки $D_h^{(k,l)} \cap (\Gamma_{1,h} \cap \partial D_h^{(k,l)})$ выполняется неравенство

$$|w_h(x) - \hat{w}_h^{(k,l)}| < \varepsilon. \tag{7.3.24}$$

Таким образом, неявная разностная схема (7.3.6) может быть реализована с заданной точностью ε независимо для каждой из подобластей $D_h^{(k,l)}$, что особенно важно для многопроцессорных ЭВМ. Конечно, при использовании рассмотренного варианта метода разделения области возникает ряд вопросов: как разбивать сеточные области на подобласти? как выбирать $\hat{D}_h^{(k,l)}$ в зависимости от исходных данных задачи и параметров сеток? какие методы использовать для решения подзадач? как регуляризировать потоки данных при реализации метода на ЭВМ различной архитектуры? и т. д. Некоторые из этих вопросов, а также другие возможные примене-

ния оценок (7.3.7), (7.3.8) уже обсуждались в публикациях. В заключение отметим, что рассмотренный подход легко обобщается на задачи с неоднородными краевыми условиями, а также на неявные разностные схемы, когда для аппроксимации задачи по пространственным переменным используется проекционно-сеточный метод. Представляет большой интерес использование установленных свойств сеточного оператора системы (7.3.6) для конструирования явно-неявных разностных схем (с перекрывающимися подобластями) и конструирования методов, сочетающих в себе идеи метода разделения области и метода расщепления.

7.4. Метод фиктивных областей

Рассмотрим эллиптическое дифференциальное уравнение

$$Lu \equiv -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial u}{\partial x_j}\right) + c(x)u = f(x), \tag{7.4.1}$$

$$x = (x_1, \ldots, x_n) \in D_1,$$

заданное в ограниченной области D_1 с краевым условием (первая краевая задача) вида

$$u(x) = 0,\ x \in \partial D_1. \tag{7.4.2}$$

Предположим, что коэффициенты и решение задачи (7.4.1), (7.4.2) достаточно гладки, $a_{ij}(x) = a_{ji}(x)$, $c(x) \geqslant 0$ и выполняется условие эллиптичности

$$\inf_{x \in D_1} \sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geqslant \mu \sum_{i=1}^{n} \xi_i^2 \tag{7.4.3}$$

с положительной постоянной μ, не зависящей от произвольного вектора $\xi = (\xi_1, \ldots, \xi_n)$.

Трудности составления программ для численного решения краевой задачи (7.4.1), (7.4.2) разностными или вариационно-разностными методами во многом зависят от геометрии области D_1. Поэтому целесообразно строить программы не для конкретных областей, а для более или менее широкого класса областей. Одним из возможных путей решения этой проблемы является замена краевой задачи (7.4.1), (7.4.2) на задачу, в определенном смысле близкую к ней, но заданную в более простой области, например в параллелепипеде. Такой подход получил название *метода фиктивных областей*.

Метод, о котором пойдет речь, основан на хорошо известном физикам факте, что в среде с относительно большими значениями коэффициента

диффузии изменение плотности диффундирующего вещества относительно мало. В связи с этим естественно попытаться пополнить начальную область D_1 до параллелепипеда и доопределить коэффициенты $a_{ij}(x)$ уравнения (7.4.1) достаточно большими значениями в добавленной области. Задав на границе параллелепипеда краевое условие вида (7.4.2), можно ожидать, что решение полученной задачи мало отличаться от нуля в добавленной области и будет почти совпадать с решением задачи (7.4.1), (7.4.2) в исходной области D_1.

Поясним эту идею на простом примере одномерной краевой задачи

$$\begin{gathered} \frac{d^2u}{dx^2} = -2, \quad 0 < x < 0,5, \\ u(0) = u(0,5) = 0, \end{gathered} \tag{7.4.4}$$

точное решение которой $u(x) = x(0,5 - x)$. Заменим (7.4.4) задачей

$$\begin{gathered} \frac{d}{dx}\left(a(x)\frac{dv}{dx}\right) = f(x), \quad 0 < x < 1, \\ v(0) = v(1) = 0, \\ \left(a\frac{dv}{dx}\right)(0,5-0) = \left(a\frac{dv}{dx}\right)(0,5+0), \end{gathered} \tag{7.4.5}$$

$$a(x) = \begin{cases} 1, & 0 < x < 0,5, \\ \dfrac{1}{\varepsilon^2}, & 0,5 < x < 1, \end{cases}$$

$$f(x) = \begin{cases} -2, & 0 < x < 0,5, \\ 0, & 0,5 < x < 1. \end{cases}$$

Точное решение задачи (7.4.5) имеет вид

$$v(x) = \begin{cases} x\left(\dfrac{1+2\varepsilon^2}{1+\varepsilon^2}0,5 - x\right), & 0 < x < 0,5, \\ \dfrac{\varepsilon^2}{2(1+\varepsilon^2)}(1-x), & 0,5 < x < 1. \end{cases}$$

Очевидно, что

$$\begin{gathered} \lim_{\varepsilon\to 0} v(x) = u(x) \quad \text{для} \quad x \in [0;0,5], \\ \lim_{\varepsilon\to 0} v(x) = 0 \quad \text{для} \quad x \in (0,5;1). \end{gathered}$$

Следовательно, при достаточно малых ε область $0,5 < x \leqslant 1$ можно рассматривать как фиктивную. Нечто подобное имеет место и в многомерном случае.

Сформулируем теперь метод фиктивных областей для решения первой краевой задачи (7.4.1), (7.4.2). Обозначим через D_2 дополнение области D_1 до параллелепипеда D и через S — общую часть границ D_1 и D_2. В области D рассмотрим уравнение

$$L_\varepsilon v \equiv -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(A_{ij}(x)\frac{\partial v}{\partial x_j}\right) + Cv = F(x), \tag{7.4.6}$$

где

$$A_{ij}(x) = \begin{cases} a_{ij}(x), & x \in D_1, \\ 0, & x \in D_2, \quad i \neq j, \\ \varepsilon^{-2}, & x \in D_2, \quad i = j, \end{cases}$$

$$C(x) = \begin{cases} c(x), & x \in D_1, \\ 0, & x \in D_2, \end{cases} \qquad F(x) = \begin{cases} f(x), & x \in D_1, \\ 0, & x \in D_2, \end{cases}$$

Заметим, что уравнение (7.4.6) не обязательно выбирать однородным в области D_2.

Для уравнения (7.4.6) поставим краевую задачу

$$v(x) = 0, \ x \in \partial D, \tag{7.4.7}$$

$$[v(x)]|_S = 0, \quad \left[\sum_{i,j=1}^{n} A_{ij}(x)\cos(\nu, x_i)\frac{\partial v}{\partial x_j}\right]\Bigg|_S = 0. \tag{7.4.8}$$

Здесь ν — нормаль к границе S; $[\ ]|_S$ обозначает скачок функции на поверхности S.

Условимся считать решение $u(x)$ задачи (7.4.1), (7.4.2) равным нулю в области D_2; оценим разность $w(x) = u(x) - v(x)$. Нетрудно видеть, что функция $w(x)$ удовлетворяет следующей краевой задаче:

$$L_\varepsilon w(x) = 0, \ x \in D, \ x \notin S, \tag{7.4.9}$$

$$w(x) = 0, \ x \in \partial D, \tag{7.4.10}$$

$$[w(x)]|_S = 0, \quad \left[\sum_{i,j=1}^{n} A_{ij}(x)\cos(\nu, x_i)\frac{\partial w}{\partial x_j}\right]\Bigg|_S = \varphi(x), \tag{7.4.11}$$

где $\varphi(x) = \sum_{i,j=1}^{n} a_{ij}\cos(\nu, x_i)\frac{\partial u}{\partial x_i}$, $x \in S$.

Умножая уравнение (7.4.9) на $w(x)$ и интегрируя по D с учетом условий (7.4.10) и (7.4.11), получим тождество

$$\int\limits_{D_1}\left(\sum_{i,j=1}^{n} a_{ij}(x)\frac{\partial w}{\partial x_i}\frac{\partial w}{\partial x_j}+Cw^2\right)dx+\frac{1}{\varepsilon^2}\int\limits_{D_2}\sum_{i=1}^{n}\left(\frac{\partial w}{\partial x_i}\right)^2 dx=-\int\limits_{S}\varphi(x)w(x)\,ds. \quad (7.4.12)$$

Отбрасывая первое слагаемое левой части (см. (7.4.3)) и применяя неравенство Коши — Шварца к правой части равенства (7.4.12), получим

$$\frac{1}{\varepsilon^2}\int\limits_{D_2}\sum_{i=1}^{n}\left(\frac{\partial w}{\partial x_i}\right)^2 dx\leqslant\sqrt{\int\limits_{S}\varphi^2(x)\,ds}\sqrt{\int\limits_{S}w^2(x)\,ds}. \quad (7.4.13)$$

Для того чтобы оценить правую часть полученного неравенства, воспользуемся соотношением

$$\int\limits_{S}w^2(x)\,ds\leqslant C_1\left(\delta\int\limits_{\omega_\delta}w^2(x)\,ds+\frac{1}{\delta}\int\limits_{\omega_\delta}\sum_{i=1}^{n}\left(\frac{\partial w}{\partial x_i}\right)^2 dx\right), \quad (7.4.14)$$

которое справедливо для пограничной полосы ω_δ ширины $0<\delta\leqslant\delta_0$, где δ_0 не зависит от функции $w(x)$, и неравенством Фридрихса

$$\int\limits_{D}w^2(x)\,dx\leqslant C_2\int\limits_{D}\sum_{i=1}^{n}\left(\frac{\partial w}{\partial x_i}\right)^2 dx, \quad (7.4.15)$$

которое справедливо для функций $w(x)$, равных нулю на части границы области D.

Выбирая в (7.4.14) $\delta=\delta_0$ и расширяя интегрирование в правой части неравенства, получим

$$\int\limits_{S}w^2(x)\,ds\leqslant C_3\left(\int\limits_{D_2}w^2(x)\,ds+\int\limits_{D_2}\sum_{i=1}^{n}\left(\frac{\partial w}{\partial x_i}\right)^2 dx\right). \quad (7.4.16)$$

Так как $w(x)=0$ при $x\in\partial D$, то к первому слагаемому в правой части (7.4.21) можно применить неравенство (7.4.15). Получим

$$\int\limits_{S}w^2(x)\,ds\leqslant C_4\int\limits_{D_2}\sum_{i=1}^{n}\left(\frac{\partial w}{\partial x_i}\right)^2 dx. \quad (7.4.17)$$

Учитывая неравенство (7.4.17), получим из (7.4.13) следующую оценку:

$$\left(\int\limits_{D_2}\sum_{i=1}^{n}\left(\frac{\partial w}{\partial x_i}\right)^2 dx\right)^{1/2} \leqslant C_4\varepsilon^2, \tag{7.4.18}$$

а из (7.4.15) —

$$\left(\int\limits_{D_2} w^2(x)\,dx\right)^{1/2} \leqslant C_5\varepsilon^2. \tag{7.4.19}$$

Аналогичным образом из тождества (7.4.12), используя условие эллиптичности (7.4.3), получим неравенство

$$\mu\int\limits_{D_1}\sum_{i=1}^{n}\left(\frac{\partial w}{\partial x_i}\right)^2 dx \leqslant \sqrt{\int\limits_S \varphi^2(x)\,ds}\sqrt{\int\limits_S w^2(x)\,ds} \leqslant C_4\varepsilon^2. \tag{7.4.20}$$

Кроме того, используя обобщенное неравенство Фридрихса

$$\int\limits_{D_1} w^2(x)\,ds \leqslant C_6\left(\int\limits_S w^2(x)\,ds + \int\limits_{D_1}\sum_{i=1}^{n}\left(\frac{\partial w}{\partial x_i}\right)^2 dx\right) \tag{7.4.21}$$

и оценки (7.4.17), (7.4.18) и (7.4.20), приходим к неравенству

$$\int\limits_{D_1} w^2(x)\,dx \leqslant C_7\varepsilon^2. \tag{7.4.22}$$

Таким образом, мы доказали, что решение $v(x)$ задачи (7.4.6)—(7.4.8) приближает решение $u(x)$ задачи (7.4.1), (7.4.2) в $W_2^1(D_1)$ с точностью ε, т. е. (см. (7.4.20) и (7.4.22))

$$\|u-v\|_{W_2^1(D_1)} \leqslant C_8\varepsilon, \tag{7.4.23}$$

где постоянная C_8 не зависит от ε.

Более тонкий анализ[4)] приводит к следующей оптимальной по ε оценке:

$$\|u-v\|_{C(D_1)} \leqslant C_9\varepsilon^2. \tag{7.4.24}$$

Окончательно можно дать следующую схему приближенного решения первой краевой задачи (7.4.1), (7.4.2) в произвольной ограниченной области D_1. Заключим область D_1 в наименьший параллелепипед D. Выберем ε так, чтобы решение задачи (7.4.6)—(7.4.8) приближало решение исходной задачи с нужной точностью. Затем решим задачу (7.4.6)—(7.4.8) разностным методом с необходимой точностью.

[4)] См. В. Д. Копченов [4].

В заключение параграфа сформулируем метод фиктивных областей применительно к решению третьей краевой задачи:

$$-\sum_{i,j=1}^{n}\frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial u}{\partial x_j}\right)+\sum_{i=1}^{n}b_i(x)\frac{\partial u}{\partial x_i}+c(x)u=f(x), \tag{7.4.25}$$

$$x=(x_1,\ldots,x_n)\in D_1,$$

$$\sum_{i,j=1}^{n}a_{ij}(x)\cos(\nu,x_i)\frac{\partial u}{\partial x_j}+\sigma(x)u=0,\quad x\in\partial D_1. \tag{7.4.26}$$

Здесь $\sigma(x)$ — достаточно гладкая и неотрицательная на ∂D_1 функция.

Расширим область D_1 до параллелепипеда $D\supset\bar{D}_1$. Предположим, что существует сфера радиуса $\rho>0$, лежащая внутри области D_1 и такая, что ею можно коснуться изнутри области D_1 любой точки поверхности ∂D_1. В приграничной полосе области D_1 ширины $\varepsilon\leqslant\rho$ определим функцию $\sigma_\varepsilon(x)$ по формуле

$$\sigma_\varepsilon(x)=\frac{2}{\varepsilon}\sigma(x_\tau)\left(1+\frac{\tau}{\varepsilon}\right),$$

$$-\varepsilon\leqslant\tau\leqslant 0,\ x_{\tau i}=x_i-\tau\cos(\nu,x_i),$$

где $|\tau|$ — расстояние по нормали от точки x до ∂D_1. В оставшейся части области D положим функцию $\sigma_\varepsilon(x)$ равной нулю.

Рассмотрим в области D уравнение

$$-\sum_{i,j}^{n}\frac{\partial}{\partial x_i}\left(A_{ij}(x)\frac{\partial v}{\partial x_j}\right)+\sum_{i=1}^{n}B_i(x)\frac{\partial v}{\partial x_i}+C(x)v+\sigma_\varepsilon(x)v=F(x), \tag{7.4.27}$$

где

$$A_{ij}(x)=\begin{cases}a_{ij}(x), & x\in D_1,\\ 0, & x\in D_2,\quad i\neq j,\\ \varepsilon, & x\in D_2,\quad i=j;\end{cases}\qquad B_i(x)=\begin{cases}b_i(x), & x\in D_1,\\ 0, & x\in D_2;\end{cases}$$

$$C(x)=\begin{cases}c(x), & x\in D_1,\\ 0, & x\in D_2;\end{cases}\quad F(x)=\begin{cases}f(x), & x\in D_1,\\ 0, & x\in D_2.\end{cases}$$

Для уравнения (7.4.27) поставим краевую задачу:

$$v(x)=0,\ x\in\partial D, \tag{7.4.28}$$

и условия согласования

$$[v(x)]|_{D_1} = 0, \quad \left[\sum_{i,j=1}^{n} A_{ij}(x)\cos(\nu, x_i)\frac{\partial v}{\partial x_j}\right]\Bigg|_{\partial D_1} = 0. \tag{7.4.29}$$

Для достаточно гладких коэффициентов задачи (7.4.25)—(7.4.26) можно доказать следующую оценку

$$\|u - v\|^2_{W_2^1(D_1)} \leqslant \varepsilon C_{10}\|f\|^2_{L_2(D_1)}, \tag{7.4.30}$$

где постоянная C_{10} не зависит от выбора $\varepsilon > 0$.

Переход от третьей краевой задачи (7.4.25), (7.4.26) к первой краевой задаче (7.4.27)—(7.4.29) можно осуществить следующим образом. Предположим, что решение задачи (7.4.27)—(7.4.29) при $\varepsilon \to 0$ сходится к некоторой функции $\tilde{u}(x)$ в норме пространства $W_2^1(D)$. Решение $v(x)$ задачи (7.4.27)—(7.4.29) удовлетворяет интегральному тождеству

$$\int\limits_{D_1} \sum a_{ij}\frac{\partial v}{\partial x_i}\frac{\partial \varphi}{\partial x_j}\,dx + \varepsilon\int\limits_{D_2}\sum \frac{\partial v}{\partial x_i}\frac{\partial \varphi}{\partial x_i}\,dx + \int\limits_{D_1} C(x)v\varphi\,dx + \int\limits_{D_1}\sigma_\varepsilon v\varphi\,dx = \int\limits_{D_1} f\varphi\,dx \tag{7.4.31}$$

при любой функции $\varphi \in \overset{\circ}{W}{}_2^1(D_1)$ (для простоты мы положили $b_i(x) = 0$). Устремляя в (7.4.31) ε к нулю и учитывая вид функции $\sigma_\varepsilon(x)$, получим тождество[5)]

$$\int\limits_{D_1} \sum a_{ij}\frac{\partial \tilde{u}}{\partial x_i}\frac{\partial \varphi}{\partial x_j}\,dx + \int\limits_{D_1} C\tilde{u}\varphi\,dx + \int \sigma\tilde{u}\varphi\,ds = \int\limits_{D_1} f\varphi\,dx, \tag{7.4.32}$$

которое справедливо для всех $\varphi \in W_2^1(D_1)$. Но так как решение задачи (7.4.25), (7.4.26) удовлетворяет интегральному тождеству (7.4.32), то функции $u(x)$ и $\tilde{u}(x)$ в области D_1 должны совпадать.

[5)] См. Л. А. Руховец [4].

Глава 8.

Сопряженные уравнения и методы возмущений

В последние годы появляется возможность решения все более широкого круга важных прикладных задач. Однако на практике часто возникают столь сложные задачи, что и современный уровень вычислительной техники оказывается недостаточным для их успешного решения. В силу этого такие задачи приходится заменять упрощенными математическими моделями, полученными некоторыми достаточно строгими методами. И одним из основных таких методов в настоящее время является метод возмущений, который здесь уже выступает как мощный метод математического моделирования. Метод возмущений эффективно используется и для получения приближенного решения исходной задачи. Сейчас, в эпоху бурного развития вычислительной техники и вычислительной математики, алгоритмы возмущений все больше и больше находят применение как в теореитических исследованиях, так и в практических расчетах.

В настоящей главе будут изложены основные положения метода регулярных возмущений применительно к неоднородным задачам, задачам на собственные значения, к вычислению функционалов. Одновременно будет показана та значительная роль, которую играют в данных методах сопряженные уравнения и их решения.

8.1. Основные и сопряженные уравнения. Алгоритмы возмущений

Пусть F есть гильбертово пространство со скалярным произведением $(\cdot,\cdot)_F \equiv (\cdot,\cdot)$ и нормой $\|\cdot\|_F \equiv \|\cdot\| = (\cdot,\cdot)^{1/2}$. Считаем, что F является самосо-

пряженным. Пусть A — линейный замкнутый оператор, действующий из F в F. Область его определения $D(A)$ предполагается плотной в F. Рассмотрим следующее неоднородное уравнение:

$$A\varphi = f, \tag{8.1.1}$$

которое будем предполагать однозначно разрешимым при любом элементе f из области значений $R(A)$ оператора A. Другими словами, предполагается существование обратного оператора A^{-1}, определенного на $R(A)$. Уравнение (8.1.1) будем называть *невозмущенным основным уравнением*. Наряду с (8.1.1) введем *возмущенное уравнение*

$$A_\varepsilon \varphi_\varepsilon = f_\varepsilon, \tag{8.1.2}$$

где $A_\varepsilon = A + \varepsilon\delta A$; δA — возмущающий оператор из F в F с областью определения, совпадающей с $D(A)$ (т. е. $D(A_\varepsilon) = D(A)$) и областью значений $R(\delta A) \subset R(A)$; ε — числовой параметр, $f_\varepsilon = f + \varepsilon\delta f$, $\delta f \in R(A)$. Отметим, что часто параметр ε вводится в возмущенное уравнение формально, а после проведения исследований он полагается равным необходимому значению. Так, например, если оператор возмущенного уравнения есть $A + \delta A$, то его можно заменить оператором $A_\varepsilon = A + \varepsilon\delta A$ с параметром ε. Если положить ε равным нулю, A_ε становится равным A; если же $\varepsilon = 1$, то A_ε совпадает с интересующим нас оператором $A+\delta A$. Отмеченным обстоятельством нередко пользуются в алгоритмах теории возмущений. Однако во многих задачах в качестве параметра ε можно выбрать тот или иной параметр, который фактически присутствует в уравнении и имеет физический смысл.

Итак, ставится задача отыскания решения (8.1.20) с помощью возмущений. Его формальная схема состоит в следующем. Предположим, что решение уравнения (8.1.20) существует и представимо в виде ряда по ε:

$$\varphi_\varepsilon = \varphi_0 + \varepsilon\varphi_1 + \varepsilon^2\varphi_2 + \dots, \tag{8.1.3}$$

где $\{\varphi_i\}_{i=0}^{\infty} \subset D(A)$. Подставляя (8.1.3) в (8.1.20), имеем

$$(A + \delta A)(\varphi_0 + \varepsilon\varphi_1 + \varepsilon^2\varphi_2 + \dots) = f + \varepsilon\delta f.$$

Приравнивая в обеих частях этого равенства коэффициенты при одинаковых степенях ε, получаем следующие уравнения для φ_i ($i = 0, 1, 2, \dots$):

$$\begin{aligned} &A\varphi_0 = f, \\ &Au_1 = \delta f - \delta A u_0, \\ &Au_i = -\delta A u_{i-1}, \quad i = 2, 3, \dots . \end{aligned} \tag{8.1.4}$$

Если предположить, что эти уравнения разрешимы и их уравнения — функции $\{\varphi_i\}$ — найдены, то мы можем построить функцию по формуле (8.1.3), которую назовем формальным решением уравнения (8.1.20). Если мы ограничимся вычислением $\varphi_0, \varphi_1, \dots, \varphi_N$, то функцию

$$\varphi_\varepsilon^{(N)} = \varphi_0 + \varepsilon\varphi_1 + \dots + \varepsilon^N \varphi_N \tag{8.1.5}$$

назовем приближением N-го порядка к u_ε (конечно, пока также формальным приближением).

Уже из изложенного выше следует ряд вопросов, которые требуют ответа в алгоритмах возмущений. В число этих вопросов входят следующие:

1. Разрешимо ли уравнение (8.1.20)?
2. Представимо ли решение возмущенной задачи в виде ряда (8.1.3)?
3. Разрешима ли система (8.1.4)? (Здесь вопрос не только в обратимости оператора A, но и в принадлежности к $R(A)$ правых частей уравнений системы.)
4. Сходится ли ряд (8.1.3), и если сходится, то будет ли функция φ_ε решением уравнения (8.1.20)?
5. Какова скорость сходимости ряда (8.1.3) и какова оценка погрешности $\|\varphi_\varepsilon - \varphi_\varepsilon^N\|$?
6. Если требуется построить приближенное решение уравнения (8.1.20) с заданной точностью, то как это проще сделать: решая непосредственно уравнение (8.1.20) или решая последовательно необходимое число уравнений системы (8.1.4)? Какова при этом эффективность второго пути приближенного решения задачи («эффективность алгоритмов возмущений»)?

Если ответы на поставленные вопросы (и, возможно, на ряд других) не даются, то изложение и применение в практических расчетах алгоритмов возмущений в той или иной степени следует считать формальным.

Опишем два подхода, позволяющих получить ответы на некоторые из поставленных вопросов, в частности на первые четыре из них, которые можно считать принципиальными в теории возмущений. Эти подходы в настоящее время считаются классическими. Первый из них основан на исследовании разрешимости уравнения (8.1.20) (при соответствующих огра-

ничениях на ε и δA) и установлении аналитичности решения φ_ε по ε. После этого, как правило, ответы на вопросы 3 и 4 являются уже следствием полученных утверждений. Во втором подходе исходят из факта обратимости оператора A и исследования разрешимости системы уравнений (8.1.4). После доказательства существования решений уравнений данной системы образуют ряд (8.1.3), доказывают его сходимость, а затем необходимо показать, что φ_ε удовлетворяет уравнению (8.1.20).

Рассмотрим схему рассуждений, относящуюся к первому подходу обоснования алгоритма возмущений. При этом мы ограничимся случаем, когда уравнение (8.1.1) корректно разрешимо, т. е. для оператора A выполнено соотношение

$$\|\varphi\| \leqslant m\|A\varphi\|, \quad m = \text{const} > 0 \tag{8.1.6}$$

при любом элементе $\varphi \in D(A)$. При наличии неравенства (8.1.6) для нормы оператора A^{-1} справедлива оценка

$$\|A^{-1}\| \leqslant m,$$

и при малых значениях $|\varepsilon|\,\|\delta A\|$ уравнение (8.1.20) будет также корректно разрешимым. Достаточным условием для этого является выполнение следующего неравенства:

$$m\,|\varepsilon|\,\|\delta A\| \equiv q < 1. \tag{8.1.7}$$

Действительно, в этом случае имеем

$$\begin{aligned}\|(A+\varepsilon\delta A)\varphi\| = \|A(I+A^{-1}\varepsilon\delta A)\varphi\| &\geqslant \frac{1}{m}\|(I+\varepsilon A^{-1}\delta A)\varphi\| \geqslant \\ \geqslant \frac{1}{m}(\|\varphi\| - |\varepsilon|\,\|A^{-1}\delta A\varphi\|) &\geqslant \frac{1}{m}(1-|\varepsilon|m\|\delta A\|)\|\varphi\| = \\ &= \frac{1}{m}(1-q)\|\varphi\|,\end{aligned}$$

т. е. уравнение (8.1.20) корректно разрешимо. Теперь для решения φ_ε возмущенного уравнения получаем следующие представления:

$$\varphi_\varepsilon = (A+\varepsilon\delta A)^{-1}f_\varepsilon = A^{-1}(I+\varepsilon\delta A\,A^{-1})^{-1}f_\varepsilon.$$

Однако для $T_\varepsilon \equiv -\varepsilon\delta A\,A^{-1}$ справедливы соотношения

$$\|T_\varepsilon\| \leqslant q < 1,$$

$$(I-T_\varepsilon)^{-1} = \sum_{i=0}^{\infty} T_\varepsilon^i = \sum_{i=0}^{\infty} \varepsilon^i(-1)^i(\delta A\,A^{-1})^i$$

(причем рассматриваемый здесь ряд операторов является сходящимся!). Поэтому

$$\varphi_\varepsilon = \sum_{i=0}^{\infty} \varepsilon^i (-1)^i A^{-1} (\delta A\, A^{-1})^i (f + \varepsilon \delta f) \equiv \sum_{i=0}^{\infty} \varepsilon^i \varphi_i, \tag{8.1.8}$$

где

$$\varphi_0 = A^{-1} f, \qquad \varphi_1 = A^{-1}(\delta f - \delta A \varphi_0), \tag{8.1.9}$$

$$\varphi_i = (-1)^{i-1} (A^{-1} \delta A)^{i-1} \varphi, \ i \geqslant 2.$$

Таким образом, разрешимость уравнения (8.1.20) установлена, а также доказана представимость φ_ε в виде сходящегося ряда (8.1.3). Уравнения (8.1.4) являются лишь другой формой записи соотношений (8.1.9), а разрешимость их вытекает из корректной разрешимости уравнения (8.1.1) и сходимости ряда $\sum\limits_{i=0}^{\infty} \varepsilon^i (-1)^i (\delta A\, A^{-1})^i$ при выполнении неравенства (8.1.7).

Приведенную выше схему рассуждений часто рассматривают как известную схему метода последовательных приближений в применении к уравнению (8.1.20). Действительно, возмущенное уравнение эквивалентно следующему:

$$\varphi_\varepsilon = -\varepsilon A^{-1} \delta A \varphi_\varepsilon + f'_\varepsilon, \tag{8.1.10}$$

где $f'_\varepsilon = A^{-1}(f + \varepsilon \delta f)$, а оператор $T_\varepsilon = -\varepsilon A^{-1} \delta A$ при выполнении (8.1.7) является сжимающим. Поэтому процесс последовательных приближений

$$\psi_\varepsilon^{(n)} = -\varepsilon A^{-1} \delta A \psi_\varepsilon^{(n-1)} + f'_\varepsilon, \quad n = 1, 2, \ldots,$$

будет сходящимся в F при $n \to \infty$, и в результате мы вновь получим представление $\varphi_\varepsilon = \lim_{n\to\infty} \psi_\varepsilon^{(n)}$ в виде ряда (8.1.8) и придем к уравнениям (8.1.9) (или (8.1.4)).

Легко заметить, что если принять предложение (8.1.7), то нетрудно оценить и скорость сходимости функций (8.1.5) к φ_ε:

$$\begin{aligned} \|\varphi_\varepsilon - \varphi_\varepsilon^{(N)}\| &\leqslant \sum_{i=N+1}^{\infty} |\varepsilon|^i m^{i-1} \|\delta A\|^{i-1} \|\varphi_1\| \leqslant \\ &\leqslant \sum_{i=N+1}^{\infty} |\varepsilon|^i m^i \|\delta A\|^{i-1} \|\delta f - \delta A \varphi_0\| = \\ &= |\varepsilon|^{N+1} m^{N+1} \|\delta f - \delta A \varphi_0\| \, \|\delta A\|^N / (1 - q). \end{aligned} \tag{8.1.11}$$

Во втором подходе к исследованию алгоритма возмущений после формального получения системы уравнений (8.1.4) необходимо прежде всего установить разрешимость этих уравнений. Так, первое уравнение имеет правую часть из $R(A)$. Тогда при наличии соотношения (8.1.6) будем иметь

корректную разрешимость данного уравнения. Во втором уравнении имеем $\delta f \in R(A)$, $\delta A\varphi_0 \in R(A)$, а значит, и второе уравнение корректно разрешимо. Аналогичный вывод можно сделать и об остальных уравнениях системы (8.1.4). Таким образом, все решения этих уравнений существуют и принадлежат $D(A)$. Кроме того, для данных решений будут справедливыми представления (8.1.9).

Образуем теперь функции $\varphi_\varepsilon^{(N)} = \varphi_0 + \varepsilon\varphi_1 + \cdots + \varepsilon^N\varphi_N$ $(N = 1, 2, \dots)$. Предполагая выполнение неравенства (8.1.7), имеем

$$\|\varphi_\varepsilon^{(N)} - \varphi_\varepsilon^{(N+M)}\| = \|\sum_{i=N+1}^{N+M} \varepsilon^i\varphi_i\| \leqslant \|\varphi_1\| \sum_{i=N+1}^{N+M} \varepsilon q^{i-1} \to 0, \quad N, M \to \infty.$$

Следовательно, последовательность $\{\varphi_\varepsilon^{(N)}\}$ сходится: $\varphi_\varepsilon = \lim_{N\to\infty} \varphi_\varepsilon^{(N)} \in F$. После этого нетрудно показать, что φ_ε удовлетворяет уравнению (8.1.10), а также что $u_\varepsilon \in D(A)$ и $A_\varepsilon\varphi_\varepsilon = f_\varepsilon$. Оценка скорости сходимости (8.1.11) здесь также остается справедливой.

Итак, из изложенного выше следует, что и в первом, и во втором подходах к исследованию алгоритмов возмущений существенную роль играют наличие соотношений (8.1.6) (*априорные оценки в возмущенной задаче*) и выполнение ограничений типа (8.1.7) (*ограничения на величину возмущения*). Именно при наличии этих факторов часто удается достаточно полно изучить тот или иной алгоритм возмущений и дать его обоснование. Отсутствие одного из них, как правило, в значительной степени усложняет исследование.

В дальнейшем уравнения (8.1.1), (8.1.20) мы будем называть также *основными уравнениями*. Наряду с ними будут рассматриваться уравнения с операторами A^*, A_ε^*, являющиеся сопряженными соответственно к A, A_ε. Операторы A^*, A_ε^* вводятся с помощью соотношений (*тождество Лагранжа*)

$$\begin{aligned} (A\varphi, \varphi^*) &= (\varphi, A^*\varphi^*),\ \varphi \in D(A),\ \varphi^* \in D(A^*), \\ (A_\varepsilon\varphi, \varphi^*) &= (\varphi, A_\varepsilon^*\varphi^*),\ \varphi \in D(A_\varepsilon),\ \varphi^* \in D(A_\varepsilon^*). \end{aligned} \tag{8.1.12}$$

В дальнейшем мы предполагаем, что в области определения $D(A^*)$, $D(A_\varepsilon^*)$ операторов A^*, A_ε^* совпадают.

Рассмотрим *сопряженное уравнение* вида

$$A^*\varphi^* = g, \tag{8.1.13}$$

где $g \in F$. При изучении алгоритмов возмущений для решения такого типа уравнений предполагаем, что исходное основное уравнение (8.1.1) везде разрешимо в F, т. е. область значений $R(A)$ оператора совпадает со всем

пространством F. Это ограничение позволяет нам гарантировать корректную разрешимость сопряженного уравнения (8.1.13), т. е. мы будем иметь соотношение

$$\|\varphi^*\| \leqslant m\|A^*\varphi^*\|, \quad \varphi^* \in D(A^*). \quad m = \text{const} > 0. \tag{8.1.14}$$

Тогда возмущенное сопряженное уравнение

$$(A^* + \varepsilon\delta A^*)\varphi_\varepsilon^* = g + \varepsilon\delta g, \tag{8.1.15}$$

где $D(\delta A^*) = D(A^*)$, $R(\delta A^*) \subseteq R(A^*)$, $\delta g \in R(A^*)$, при выполнении ограничения

$$m\,|\varepsilon|\,\|\delta A^*\| \equiv q < 1 \tag{8.1.16}$$

будет также корректно разрешимым. Повторяя теперь рассуждения, проведенные при рассмотрении основных задач, можно доказать, что решение уравнения (8.1.15) представимо сходящимся в F рядом

$$\varphi_\varepsilon^* = \varphi_0^* + \varepsilon\varphi_1^* + \varepsilon^2\varphi_2^* + \ldots, \tag{8.1.17}$$

где функции $\{\varphi_i^*\}$ принадлежат $D(A^*)$, и они могут быть определены путем последовательного решения уравнений

$$\begin{aligned} &A^*\varphi_0^* = g, \\ &A^*\varphi_1^* = \delta g - \delta A^*\varphi_0^*, \\ &A^*\varphi_i^* = -\delta A^*\varphi_{i-1}^*, \quad i = 2, 3, \ldots. \end{aligned} \tag{8.1.18}$$

Если найти лишь функцию $\{\varphi_i^*\}_{i=0}^N$, то функцию

$$\varphi_*^{(N)} = \varphi_0^* + \varepsilon\varphi_1^* + \cdots + \varepsilon^N\varphi_N^* \tag{8.1.19}$$

можно принять за приближение N-го порядка к решению уравнения (8.1.15).

Обоснование алгоритма возмущений в сопряженных задачах можно осуществить так же, как и в основных.

Замечание. Если в (8.1.20) и (8.1.15) принять, что

$$\delta A^* = (\delta A)^* \tag{8.1.20}$$

(хотя это необязательно), то в этом случае будем иметь $(A+\varepsilon\delta A)^* = A^*+\varepsilon\delta A^*$, т. е. уравнение (8.1.15) является сопряженным по отношению к основному

невозмущенному уравнению (8.1.20). Именно этот случай часто интересен в прикладных задачах.

Теперь отметим несколько фактов, подчеркивающих важное значение решений сопряженных уравнений в прикладных задачах. Пусть в процессе решения основного уравнения (8.1.1) нам необходимо вычислить некоторый функционал

$$J_p = (\varphi, p), \tag{8.1.21}$$

где $p \in F$. Сделать это можно по приведенной (заданной) формуле после отыскания φ. Однако поскольку для решений уравнений (8.1.1), (8.1.13) имеем равенство (*соотношение сопряженности*)

$$(f, \varphi^*) = (\varphi, g), \tag{8.1.22}$$

то, принимая в (8.1.13) элемент g равным p, получаем следующее представление для интересующего нас функционала:

$$J_p = (f, \varphi_p^*), \tag{8.1.23}$$

где φ_p^* есть решение сопряженного уравнения (8.1.13) со специальной правой частью $g = p$. Представление (8.1.23) обладает рядом преимуществ и широко используется при получении экономичных алгоритмов возмущений для отыскания приближенных значений линейных функционалов (об этом будет идти речь в дальнейшем). Теперь с целью выявления физического смысла решений сопряженных уравнений и для упрощения изложения предположим, что J_p имеет вид

$$J_p = \int\limits_D p(x)\varphi(x)\,dx, \tag{8.1.24}$$

где D — ограниченная область евклидова пространства, а элементы рассматриваемых пространств являются вещественными функциями от переменной $x \in D$ и $F = F^* = L_2(D)$. Поскольку

$$J_p = \int\limits_D \varphi_p^*(x) f(x)\,dx, \tag{8.1.25}$$

то отсюда вытекает следующая физическая интерпретация функции $\varphi_p^*(x)$: она описывает «отклик» функционала J_p на изменение функции $f(x)$ в правой части уравнения (8.1.1). Легко заметить, что изменение f на множестве $\delta D \subset D$, на котором $\varphi_p^*(x)$ мало, внесет малый вклад в J_p по сравнению с вкладом, который дало бы аналогичное изменение f в подобласти из D, где

φ_p^* велико. В силу сказанного, решения u_p^* сопряженных уравнений часто называют *функциями ценности* по отношению к функционалу J_p. Как мы увидим далее, это свойство решений сопряженных уравнений было положено во многих работах в основу решения алгоритмами возмущений неоднородных задач, задач оптимального управления, обратных задач, задач планирования эксперимента и др.

Пример. Пусть в пространстве $F \equiv L_2(D)$, $D = \{0 < x, y < 1\} \subset \mathbf{R}^2$, действует оператор A, определяемый дифференциальным выражением

$$A\varphi = -\mu\Delta\varphi + u\frac{\partial\varphi}{\partial x} + v\frac{\partial\varphi}{\partial y},$$
$$\mu = \text{const} > 0, \quad \Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}, \quad u(x,y),\ v(x,y) \in C^{(1)}(D)$$

и областью определения $D(A) \equiv W_2^2(D) \cap \overset{\circ}{W}{}_2^1(D)$, где $W_2^2(D)$, $\overset{\circ}{W}{}_2^1(D)$ есть пространства Соболева (см. гл. 1). Оператор A является замкнутым. Он обладает дискретным спектром. Предположим, что функции $u(x,y)$, $v(x,y)$ таковы, что нуль не принадлежит спектру оператора A. В этом случае задача

$$\begin{gathered} -\mu\Delta\varphi + u\frac{\partial\varphi}{\partial x} + v\frac{\partial\varphi}{\partial y} = f, \quad f \in L_2(D), \\ \varphi\big|_{\partial D} = 0, \end{gathered} \tag{8.1.26}$$

или в операторной форме

$$A\varphi = f, \tag{8.1.27}$$

однозначно разрешима при любой функции $f \in L_2(D)$, т. е. $R(A) = L_2(D)$, а уравнение везде разрешимо. Тогда сопряженное уравнение

$$A^*\varphi^* = g, \tag{8.1.28}$$

где

$$A^*\varphi^* = -\mu\Delta\varphi^* - \frac{\partial(u\varphi^*)}{\partial x} - \frac{\partial(v\varphi^*)}{\partial y},$$
$$D(A^*) = D(A),$$

корректно разрешимо, а также существует такая положительная постоянная m, что

$$\|\varphi^*\|_{L_2} \leqslant m\|A^*\varphi^*\|_{L_2}, \quad \varphi^* \in D(A^*). \tag{8.1.29}$$

Если наряду с (8.1.27), (8.1.28) рассмотреть возмущенные уравнения

$$A_\varepsilon\varphi_\varepsilon = f, \tag{8.1.30}$$

$$A_\varepsilon^* \varphi_\varepsilon^* = g, \tag{8.1.31}$$

где

$$A_\varepsilon \varphi_\varepsilon = -\mu\Delta\varphi_\varepsilon + (u + \varepsilon\delta u)\frac{\partial\varphi_\varepsilon}{\partial x} + (v + \varepsilon\delta v)\frac{\partial\varphi_\varepsilon}{\partial y},$$
$$A_\varepsilon^* \varphi_\varepsilon^* = -\mu\Delta\varphi_\varepsilon - \frac{\partial((u + \varepsilon\delta u)\varphi_\varepsilon^*)}{\partial x} - \frac{\partial((v + \varepsilon\delta v)\varphi_\varepsilon^*)}{\partial y},$$
$$D(A_\varepsilon) = D(A_\varepsilon^*) = D(A),$$
$$\delta u, \delta v \in C^{(1)}(D),\ 0 \leqslant \varepsilon \leqslant 1,$$

при достаточно малых ε (или малых δu, δv в метрике $C^{(1)}$) алгоритмы возмущений для решения (8.1.30), (8.1.31) будут сходящимися. И после решения, например, задач

$$\begin{aligned}
-\mu\Delta\varphi_0 + u\frac{\partial\varphi_0}{\partial x} + v\frac{\partial\varphi_0}{\partial y} &= f, & \varphi_0|_{\partial D} &= 0,\\
-\mu\Delta\varphi_1 + u\frac{\partial\varphi_1}{\partial x} + v\frac{\partial\varphi_1}{\partial y} &= -\delta u\frac{\partial\varphi_0}{\partial x} - \delta v\frac{\partial\varphi_0}{\partial y}, & \varphi_1|_{\partial D} &= 0,\\
-\mu\Delta\varphi_0^* - \frac{\partial(u\varphi_0^*)}{\partial x} - \frac{\partial(v\varphi_0^*)}{\partial y} &= g, & \varphi_0^*|_{\partial D} &= 0,\\
-\mu\Delta\varphi_1^* - \frac{\partial(u\varphi_1^*)}{\partial x} - \frac{\partial(v\varphi_1^*)}{\partial y} &= \frac{\partial(u\varphi_0^*)}{\partial x} + \frac{\partial(v\varphi_0^*)}{\partial y}, & \varphi_1^*|_{\partial D} &= 0,
\end{aligned}$$

можно построить приближения первого порядка

$$\varphi_\varepsilon^{(1)} = \varphi_0 + \varepsilon\varphi_1, \quad \varphi_\varepsilon^{*(1)} = \varphi_0^* + \varepsilon\varphi_1^*$$

к φ_ε, φ_ε^* соответственно.

8.2. Метод теории возмущений для задач на собственные значения

Пусть A — линейный оператор, действующий в гильбертовом пространстве F с областью определения $D(A) \subset F$ и областью значений $R(A) \subset F$. Предполагаем, что область определения $D(A)$ плотна в F.

Для оператора A рассмотрим задачу на собственные значения

$$A\varphi = \lambda\varphi. \tag{8.2.1}$$

Предположим, что нам известны некоторое собственное значение λ_0 оператора A и отвечающий ему собственный элемент $\varphi_0 \in D(A)$, являющиеся решением задачи (8.2.1).

Пусть теперь свойства оператора A задачи (8.2.1) как-то изменились и вместо оператора A мы имеем дело с некоторым оператором $\tilde{A}$. Тогда приходим к возмущенной задаче

$$\tilde{A}\tilde{\varphi} = \tilde{\lambda}\tilde{\varphi}. \tag{8.2.2}$$

Возникает вопрос: как найти решение $\tilde{\lambda}, \tilde{\varphi}$ задачи (8.2.2), используя информацию о возмущении оператора A и о решении невозмущенной задачи (8.2.1)? Другими словами, как изменятся решения λ_0, φ_0 задачи (8.2.1) при внесении возмущения в оператор A? Этот вопрос в общем случае сложный и требует тщательного анализа спектра как исходного, так и возмущенного операторов A и $\tilde{A}$.

Мы приведем здесь формальную схему алгоритма возмущений для отыскания $\tilde{\lambda}$, $\tilde{\varphi}$ в случае аналитических возмущений коэффициентов задачи.

Будем рассматривать случай возмущенного изолированного собственного значения λ_0 кратности 1. Пусть $\bar{\lambda}_0$ является однократным собственным значением сопряженного оператора A^* и уравнение $A\varphi = \lambda_0\varphi + f$ разрешимо для тех и только тех правых частей f, которые ортогональны φ_0^* — решению однородной сопряженной задачи $A^*\varphi_0^* = \bar{\lambda}_0\varphi_0^*$.

В отличие от предыдущего параграфа, мы будем рассматривать более общий случай возмущения. Предположим, что возмущение оператора A определяется некоторым параметром ε (вещественным или комплексным) так, что $\tilde{A} = A(\varepsilon)$ представляет собой аналитическую функцию, регулярную в окрестности точки $\varepsilon = 0$:

$$\tilde{A} = \sum_{i=0}^{\infty} \varepsilon^i A^{(i)}, \tag{8.2.3}$$

где $A^{(0)} = A$, а $A^{(i)}$ ($i = 1, 2, \dots$) — некоторые линейные операторы, действующие в X с областью определения $D(A)$.

Допустим, что возмущенный оператор $\tilde{A} = A(\varepsilon)$ имеет собственное значение $\tilde{\lambda} = \lambda(\varepsilon)$, которое вместе с собственным вектором $\tilde{\varphi} = \varphi(\varepsilon)$ зависит от ε аналитически, т. е. $\tilde{\lambda}$ и $\tilde{\varphi}$ представляются в виде рядов по степеням ε:

$$\begin{aligned} \tilde{\lambda} \equiv \tilde{\lambda}(\varepsilon) &= \sum_{i=0}^{\infty} \varepsilon^i \lambda^{(i)}, \quad \lambda^{(0)} = \lambda_0, \\ \tilde{\varphi} \equiv \tilde{\varphi}(\varepsilon) &= \sum_{i=0}^{\infty} \varepsilon^i \varphi^{(i)}, \quad \varphi^{(0)} = \varphi_0, \end{aligned} \tag{8.2.4}$$

сходящихся в некоторой окрестности точки $\varepsilon = 0$.

Подставляя ряды (8.2.3), (8.2.4) в уравнение (8.2.2) и приравнивая члены при одинаковых степенях ε, получаем систему

$$\begin{aligned}
&(A-\lambda_0 E)\varphi^{(0)}=0,\\
&(A-\lambda_0 E)\varphi^{(1)}=\lambda^{(1)}\varphi^{(0)}-A^{(1)}\varphi^{(0)},\\
&(A-\lambda_0 E)\varphi^{(2)}=\lambda^{(1)}\varphi^{(1)}+\lambda^{(2)}\varphi^{(0)}-A^{(1)}\varphi^{(1)}-A^{(2)}\varphi^{(0)},\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&(A-\lambda_0 E)\varphi^{(n)}=\sum_{i=1}^{n}\lambda^{(i)}\varphi^{(n-i)}-\sum_{i=1}^{n}A^{(i)}\varphi^{(n-i)},
\end{aligned}\tag{8.2.5}$$

где E — тождественный оператор, действующий в F.

Перепишем (8.2.5) в виде

$$\begin{aligned}
&(A-\lambda_0 E)\varphi^{(0)}=0,\\
&(A-\lambda_0 E)\varphi^{(1)}=f_1,\\
&\ldots\ldots\ldots\ldots\ldots\\
&(A-\lambda_0 E)\varphi^{(n)}=f_n,\\
&\ldots\ldots\ldots\ldots\ldots
\end{aligned}\tag{8.2.6}$$

где правые части f_n ($n=1,2,\ldots$) определяются по формулам

$$f_n=\sum_{i=1}^{n}\lambda^{(i)}\varphi^{(n-i)}-\sum_{i=1}^{n}A^{(i)}\varphi^{(n-i)}.$$

Первое уравнение системы (8.2.6) заведомо выполняется, поскольку это есть не что иное, как невозмущенная задача (8.2.1) при $\lambda=\lambda_0$, $\varphi=\varphi^{(0)}=\varphi_0$. Для разрешимости второго и последующих уравнений системы (8.2.6) необходимо потребовать выполнения условия ортогональности правых частей f_n элементу φ_0^* — решению сопряженной однородной задачи $A^*\varphi_n^*=\bar{\lambda}_0\varphi_0^*$:

$$(f_n,\varphi_0^*)=0,\quad n=1,2,\ldots.\tag{8.2.7}$$

(Здесь символ $(\cdot,\cdot)$ обозначает скалярное произведение в F.)

Конечно, элемент $\varphi(\varepsilon)$ определяется из (8.2.6) не однозначно, а с точностью до множителя, который может быть любой функцией от ε. Чтобы избежать этого, обычно вводят дополнительные условия (условия нормировки) на $\varphi(\varepsilon)$. Мы будем рассматривать следующее условие:

$$(\varphi(\varepsilon),\varphi_0^*)=1,\tag{8.2.8}$$

из которого следует, что

$$(\varphi_0, \varphi_0^*) = 1,\ (\varphi^{(1)}, \varphi_0^*) = (\varphi^{(2)}, \varphi_0^*) = \cdots = 0. \tag{8.2.9}$$

Соотношения (8.2.7) и (8.2.9) позволяют определить коэффициенты разложения собственного значения $\lambda(\varepsilon)$ (так называемые поправки $\lambda^{(n)}$):

$$\begin{aligned}
&\lambda^{(1)} = (A^{(1)}\varphi^{(0)}, \varphi_0^*),\\
&\lambda^{(2)} = (A^{(1)}\varphi^{(1)}, \varphi_0^*) + (A^{(2)}\varphi^{(0)}, \varphi_0^*),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&\lambda^{(n)} = \sum_{i=1}^{n} (A^{(i)}\varphi^{(n-i)}, \varphi_0^*),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned} \tag{8.2.10}$$

Зная поправку $\lambda^{(1)}$, из второго уравнения системы (8.2.5) при условиях $(f_1, \varphi_0^*) = 0$ и $(\varphi^{(1)}, \varphi_0^*) = 0$ мы однозначно определим $\varphi^{(1)}$, а по $\varphi^{(0)}$ и $\varphi^{(1)}$ мы построим поправку $\lambda^{(2)}$ по формуле (8.2.10). Зная поправки $\lambda^{(1)}$, $\lambda^{(2)}$ из (8.2.5) при условиях (f_2, φ_0^*) и $(\varphi^{(2)}, \varphi_0^*) = 0$, мы найдем $\varphi^{(2)}$ и т. д.: по поправкам $\lambda^{(1)}$, $\lambda^{(2)}$, $\ldots$, $\lambda^{(n)}$ из $(n+1)$-го уравнения системы (8.2.5) при условиях $(f_n, \varphi_0^*) = 0$ и $(\varphi^{(n)}, \varphi_0^*) = 0$ мы однозначно определим $\varphi^{(n)}$. Таким образом, из системы (8.2.5) мы можем последовательно находить $\varphi^{(1)}, \varphi^{(2)}, \ldots, \varphi^{(n)}, \ldots$. Если мы обрываем решение системы (8.2.5) при некотором $n = N$, то алгоритм (8.2.5) обычно называют алгоритмом теории возмущений N-го порядка.

Аналогично изложенному формулируется алгоритм возмущений и для задач на собственные значения с сопряженными операторами.

Сделаем некоторые замечания, касающиеся изложенного выше метода. Хотя он известен давно, однако его обоснование для многих задач нетривиально. Отметим основные проблемы, возникающие здесь.

Одной из трудностей данного метода является доказательство того, что $\tilde{\lambda}$ и $\tilde{\varphi}$ допускают представления в виде сходящихся рядов (8.2.4). Это связано со свойствами исходного оператора A и возмущенного оператора $A(\varepsilon)$.

Другая трудность теории описанного алгоритма — оценка скорости сходимости приближений

$$\lambda^{(N)} = \sum_{i=0}^{N} \varepsilon^i \lambda^{(i)} \ \text{ и } \ \varphi^{(N)} = \sum_{i=0}^{N} \varepsilon^i \varphi^{(i)}$$

при $N \to \infty$ к $\tilde{\lambda}$ и $\tilde{\varphi}$ соответственно. Эта проблема особенно важна при проведении практических расчетов, она тесно связана с получением границ изменения ε, при которых алгоритм будет сходящимся.

Особой проблемой при использовании сформулированного алгоритма является решение неоднородных вырожденных уравнений из системы (8.2.5).

Кроме того, в каждом конкретном случае возникает проблема целесообразности применения алгоритма теории возмущений (8.2.5).

Все эти проблемы сложны в общем случае и до сих пор являются основными проблемами математической теории возмущений операторов в гильбертовом пространстве.

До сих пор мы говорили о возмущениях собственных значений оператора A, т. е. точечного спектра оператора A (здесь мы считаем, что спектр состоит из трех подмножеств: точечного, непрерывного и остаточного спектров). Однако при возмущении оператора A изменяются не только собственные значения, но и весь его спектр, и, таким образом, возникает задача о возмущении всего спектра оператора A. Эта задача несравненно сложнее задачи о возмущении отдельных собственных значений. Дело в том, что спектр оператора довольно чувствителен даже к незначительным изменениям оператора: собственные значения могут расщепляться или сливаться, а непрерывная часть спектра может целиком перейти в чисто точечный спектр. К задаче возмущения непрерывного спектра нельзя подходить так же, как к задаче возмущения точечного спектра, т. е. нельзя ограничиваться исследованием отдельных точек или отдельных отрезков спектра. В этом случае приходится рассматривать все части спектра одновременно. Указанные обстоятельства потребовали разработки новых методик проведения исследований и алгоритмов. Описание их выходит за рамки данной книги, и найти его можно в оригинальных работах.

Пример. Пусть $F = L_2(0,1)$ — вещественное пространство. В качестве A рассмотрим оператор

$$A\varphi = -p_0 \frac{d^2\varphi}{dx^2} + q_0\varphi, \quad \varphi \in D(A), \quad p_0, q_0 = \text{const} = 0,$$

с областью определения

$$D(A) = \{u : u \in W_2^2(0,1),\, u(0) = u(1) = 0\},$$

где $W_2^2(0,1)$ — гильбертово пространство функций со скалярным произведением и нормой

$$(u,v)_{W_2^2(0,1)} = \int_0^2 \left(\frac{d^2u}{dx^2}\frac{d^2v}{dx^2} + \frac{du}{dx}\frac{dv}{dx} + uv \right) dx,$$
$$\|u\|_{W_2^2(0,1)} = (u,u)^{1/2}_{W_2^2(0,1)}.$$

Нетрудно показать, что $D(A)$ плотно в F и оператор A самосопряжен.

Рассмотрим для A задачу на собственные значения

$$A\varphi = \lambda\varphi,$$

которую можно переписать в виде

$$-p_0\frac{d^2\varphi}{dx^2} + q_0\varphi = \lambda\varphi, \quad x \in (0,1),$$
$$\varphi(0) = \varphi(1) = 0.$$

Это есть не что иное, как простейшая задача Штурма — Лиувилля. Известно, что спектр оператора A состоит из счетного числа простых собственных значений

$$\lambda_k = p_0\pi^2k^2 + q_0,\ k = 1,2,\dots.$$

Каждому значению λ_k соответствует собственная функция $\varphi_k = \sqrt{2}\sin(k\pi x)$, причем функции φ_k ($k = 1,2,\dots$) образуют полную ортонормированную систему в $F = L_2$.

Вместе с A рассмотрим возмущенный оператор $A(\varepsilon) = A + \varepsilon\delta A$ с областью определения $D(A)$, где

$$\delta A\varphi = -\frac{d}{dx}p(x)\frac{d\varphi}{dx} + q(x)\varphi, \quad \varphi \in D(A),$$

где

$$0 < p_1 \leqslant p(x) \leqslant p_2, \quad 0 \leqslant q(x) \leqslant q_1, \quad p_1,p_2,q_1 = \text{const} > 0,$$
$$p(x), q(x) \in C^{(1)}(0,1).$$

Для возмущенного оператора $A(\varepsilon)$ также рассмотрим задачу на собственные значения

$$A(\varepsilon)\tilde{\varphi} = \tilde{\lambda}\tilde{\varphi},$$

которую можно записать также в виде

$$-\frac{d}{dx}(p_0+\varepsilon p(x))\frac{d\tilde{\varphi}}{dx}+(q_0+\varepsilon q(x))\tilde{\varphi}(x)=\tilde{\lambda}\tilde{\varphi},\quad x\in(0,1),$$
$$\tilde{\varphi}(0)=\tilde{\varphi}(1)=0.$$

Известно, что эта задача имеет счетное число простых собственных значений $\tilde{\lambda}_1,\tilde{\lambda}_2,\ldots$, явный вид которых в общем случае неизвестен. Для их приближенного отыскания можно воспользоваться алгоритмом возмущений, который позволяет найти $\tilde{\lambda}_1,\tilde{\lambda}_1,\ldots$ с любой наперед заданной точностью.

Так, доказано, что для каждого собственного значения λ_k существуют простое собственное значение $\tilde{\lambda}_k$ и соответствующий ему собственный элемент $\tilde{\varphi}_k$ возмущенного оператора $A(\varepsilon)=A+\varepsilon\delta A$, представляющихся в виде рядов по ε:

$$\tilde{\lambda}_k=\lambda_k+\varepsilon\lambda_{k,1}+\varepsilon^2\lambda_{k,2}+\ldots$$
$$\tilde{\varphi}_k=\varphi_k+\varepsilon\varphi_{k,1}+\varepsilon^2\varphi_{k,2}+\ldots,$$

сходящихся в некоторой окрестности точки $\varepsilon=0$.

Для отыскания поправок $\lambda_{k,1},\lambda_{k,2},\ldots,\varphi_{k,0},\varphi_{k,1},\ldots$ воспользуемся алгоритмом возмущений. Так, зная $\varphi_{k,0}\equiv\varphi_k$, находим $\lambda_{k,1}$:

$$\lambda_{k,1}=(A_1\varphi_k,\varphi_k)=\int\limits_0^1\left(p(x)\left(\frac{d\varphi_k}{dx}\right)^2+q(x)(\varphi_k)^2\right)dx.$$

После чего имеем

$$\tilde{\lambda}_k=p_0\pi^2k^2+q_0+\varepsilon\lambda_{k,1}+O(\varepsilon^2).$$

Для вычисления поправки $\varphi_{k,1}$ решаем задачу

$$-p_0\frac{d^2\varphi_{k,1}}{dx^2}+q_0\varphi_{k,1}-\lambda_{k,1}\varphi_{k,1}=f_{k,1},\quad x\in(0,1),$$
$$\varphi_{k,1}(0)=\varphi_{k,1}(1)=0$$

с правой частью

$$f_{k,1}=\lambda_{k,1}\varphi_k+\frac{d}{dx}p(x)\frac{d\varphi_k}{dx}-q(x)\varphi_k$$

и при условии

$$(\varphi_{k,1},\varphi_k)=0.$$

Поскольку функция $f_{k,1}$ ортогональна f_k, то эта задача имеет единственное решение. Вычислив подходящим методом $\varphi_{k,1}$, получаем

$$\tilde{\varphi}_k \cong \varphi_k + \varepsilon\varphi_{k,1}.$$

Аналогичным образом можно найти значения $\lambda_{k,1}, \ldots, \lambda_{k,N}$ и функции $\varphi_{k,2}, \ldots, \varphi_{k,N}$ при $N > 1$.

8.3. Сопряженные уравнения и теория возмущений для линейных функционалов

В настоящее время математические модели в различных отраслях науки становятся все более сложными и содержательными. Поэтому они требуют предварительного глубокого анализа и, в первую очередь, исследования чувствительности решения и его функционалов в зависимости от вариаций параметров задачи и входных данных. И именно здесь мы сталкиваемся с необходимостью привлечения в исследованиях сопряженных уравнений в зависимости от тех или иных функционалов задач. Сопряженные уравнения в этом случае позволяют оценить вариации функционала в зависимости от вариаций входных параметров исходнвх задач. В данном параграфе мы изложим некоторые из методов теории возмущений для линейных функционалов. Причем базироваться они будут как на решениях основных уравнений, так и на решениях уравнений с сопряженными операторами.

Пусть по-прежнему, как и в предыдущем параграфе, A — линейный оператор, действующий в гильбертовом пространстве F с областью определения $D(A) \subset F$, плотной в F. Предполагаем, что в F заданы некоторое скалярное произведение $(\cdot, \cdot)$ и норма $\|\cdot\|$:

$$\|\varphi\| = (\varphi, \varphi)^{1/2}, \;\; \varphi \in F.$$

Рассмотрим линейное уравнение

$$A\varphi = f, \tag{8.3.1}$$

где f — заданный элемент пространства F, а φ — искомый элемент из $D(A)$.

Пусть уравнение (8.3.1) корректно разрешимо, т. е. оператор A имеет на $R(A)$ ограниченный обратный. Тогда для всякого $f \in R(A)$ существует единственный элемент φ из $D(A)$, являющийся решением уравнения (8.3.1), причем $\|\varphi\| \leqslant k\|f\|$, где $k > 0$ не зависит от φ и f.

При решении тех или иных физических задач обычно нужно получить в результате значение некоторой величины, являющейся функционалом от φ. Мы будем рассматривать линейные ограниченные функционалы от φ. Согласно теореме Рисса, всякий линейный ограниченный функционал в гильбертовом пространстве F записывается в виде скалярного произведения

$$J_p[\varphi] = (\varphi, p), \tag{8.3.2}$$

где p — фиксированный элемент из F.

Введем вместе с оператором A сопряженный к нему оператор A^* с областью определения $D(A^*)$, удовлетворяющий тождеству Лагранжа:

$$(Ah, g) = (h, A^*g) \tag{8.3.3}$$

для любых элементов $h \in D(A)$, $g \in D(A^*)$.

Наряду с уравнением (8.3.1) рассмотрим сопряженное неоднородное уравнение

$$A^*\varphi_p^* = p, \tag{8.3.4}$$

где p — элемент из F, определяющий нужный нам функционал $J_p[\varphi]$ из (8.3.2).

Поскольку уравнение (8.3.1) корректно разрешимо, то сопряженное уравнение (8.3.4) всюду разрешимо, т. е. для любого $p \in F$ уравнение (8.3.4) имеет решение φ_p^*. Подставляя в формулу (8.3.3) вместо элементов h и g решения уравнений (8.3.1) и (8.3.4) φ и φ_p^*, получим

$$(A\varphi, \varphi_p^*) = (\varphi, A^*\varphi_p^*) \tag{8.3.5}$$

или, воспользовавшись уравнениями (8.3.1) и (8.3.4), приходим к равенству

$$(f, \varphi_p^*) = (\varphi, p). \tag{8.3.6}$$

Отсюда и из (8.3.2) следует, что

$$J_p[\varphi] = (f, \varphi_p^*). \tag{8.3.7}$$

Таким образом, если нам нужно найти значение функционала $J_p[\varphi]$, мы можем получить его двумя способами: либо решить уравнение (8.3.1) и определить величину по формуле (8.3.2), либо решить уравнение (8.3.4) и определить ту же величину по формуле (8.3.7).

Следовательно, каждому линейному функционалу $J_p[\varphi] = (\varphi, p)$ может быть поставлен в соответствие элемент φ_p^*, удовлетворяющий уравнению

(8.3.4), причем в качестве свободного члена этого уравнения следует использовать именно элемент $p \in F$, определяющий интересующий нас функционал $J_p[\varphi]$.

Решение φ_p^* (называемое часто *сопряженной функцией* или *функцией ценности*) имеет глубокий физический смысл и играет важную роль при изучении физических процессов.

рассмотренный подход к определению линейных функционалов через сопряженную функцию φ_p^* применяется во многих прикладных задачах. Мы же здесь воспользуемся им для формулировки алгоритмов возмущений.

Пусть свойства оператора A основной задачи (8.3.1) изменились и он переходит в возмущенный оператор $\tilde{A}$:

$$\tilde{A} = A + \delta A, \tag{8.3.8}$$

где δA — некоторый линейный оператор, действующий в X с областью определения $D(\delta A)$. Для простоты будем считать, что $D(\delta A) = D(A)$.

Решение φ и значение функционала $J_p[\varphi]$ также изменяется в этом случае:

$$\varphi \to \tilde{\varphi}, \quad J_p[\varphi] \to \tilde{J}_p = J_p + \delta J_p, \tag{8.3.9}$$

здесь $\tilde{J}_p = J_p[\tilde{\varphi}]$, $\delta J_p = J_p[\tilde{\varphi}] - J_p[\varphi]$. Таким образом, вместо (8.3.1) приходим к возмущенной задаче

$$\tilde{A}\tilde{\varphi} = f. \tag{8.3.10}$$

Установим связь между изменением оператора δA и изменением функционала δL_p.

Предположим, что уравнение (8.3.10) корректно разрешимо, тогда существует единственный элемент $\tilde{\varphi}$ из $D(A)$, являющийся решением уравнения (8.3.10).

Рассмотрим φ_p^* — решение сопряженного невозмущенного уравнения (8.3.4), соответствующее функционалу J_p. Помножив скалярно уравнение (8.3.10) справа на φ_p^*, уравнение (8.3.4) слева на $\tilde{\varphi}$ и вычитая одно из другого, получим

$$(\tilde{A}\tilde{\varphi}, \varphi_p^*) - (\tilde{\varphi}, A^*\varphi_p^*) = (f, \varphi_p^*) - (\tilde{\varphi}, p). \tag{8.3.11}$$

Пользуясь свойством (8.3.3) сопряженного оператора A^*, левую часть последнего равенства перепишем в виде

$$(\tilde{A}\tilde{\varphi}, \varphi_p^*) - (\tilde{\varphi}, A^*\varphi_p^*) = (\delta A\tilde{\varphi}, \varphi_p^*). \tag{8.3.12}$$

В правой части (8.3.11) в соответствии с (8.3.2) и (8.3.6) будем иметь

$$(f, \varphi_p^*) - (\tilde{\varphi}, p) = J_p[\varphi] - J_p[\tilde{\varphi}] = -\delta J_p. \tag{8.3.13}$$

Из (8.3.11)—(8.3.13) получим общее соотношение для вариации функционала

$$\delta J_p = -(\delta A\tilde{\varphi}, \varphi_p^*). \tag{8.3.14}$$

Если вместо уравнений (8.3.10) и (8.3.4) рассмотреть возмущенное сопряженное уравнение

$$(A^* + \delta A^*)\tilde{\varphi}_p^* = p \tag{8.3.15}$$

и невозмущенное основное уравнение (8.3.1), то аналогичным путем можно получить также соотношение

$$\delta J_p = -(\varphi, \delta A^*\tilde{\varphi}_p^*), \tag{8.3.16}$$

которое эквивалентно соотношению (8.3.14).

Отметим одну важную особенность применения формул теории возмущений (8.3.14) и (8.3.16): так как эти формулы пишутся для вариации функционала, погрешность в которой обычно допустима в пределах нескольких процентов, то для вычислений указанных вариаций нет необходимости знать точное решение основной и сопряженной задач, достаточно воспользоваться их приближенными решениями.

Если возмущение операторов A и A^* столь мало, что оно не очень сильно искажает решение φ и φ_p^*, то в формулах (8.3.14) и (8.3.16) можно заменить приближенно $\tilde{\varphi} \approx \varphi$, $\tilde{\varphi}_p^* \approx \varphi_p^*$. При этом мы получим две эквивалентные друг другу формулы теории малых возмущений (т. е. 1-го порядка):

$$\delta J_p \cong -(\delta A\varphi, \varphi_p^*) \equiv \delta J_p^{(1)}, \tag{8.3.17}$$

$$\delta J_p \cong -(\varphi, \delta A^*\varphi_p^*). \tag{8.3.18}$$

Исходя из соотношений (8.3.14), (8.3.16), в случае необходимости можно получать формулы теории возмущений и более высоких порядков, если возмущение оператора A определить с помощью параметра ε, как это делалось в § 8.1. Тогда в предположении, что решение $\tilde{\varphi}$ возмущенной задачи (8.3.10) с оператором $\tilde{A} = A + \varepsilon\delta A$ аналитично по ε в окрестности $|\varepsilon| = 0$, т. е. представимо в виде сходящегося ряда по степеням ε:

$$\tilde{\varphi} = \sum_{i=0}^{\infty} \varepsilon^i \varphi^{(i)}, \quad \varphi^{(0)} = \varphi, \tag{8.3.19}$$

из (8.3.14) можно получить представление в виде ряда по ε и для δJ_p:

$$\delta J_p = \sum_{i=1}^{\infty} \varepsilon^i \delta J_{p,i}, \tag{8.3.20}$$

где

$$\delta J_{p,i} = (\varphi^{(i)}, p).$$

Отыскание элементов $\varphi^{(i)}$ ($i = 1, 2 \dots$) можно осуществить алгоритмом возмущений, сформулированным в § 8.1. Если $\varphi^{(1)}, \dots, \varphi^{(N)}$ найдены, то вычислив $\delta J_{p,i} = (\varphi^{(i)}, p)$ ($i = 1, 2, \dots, N$), найдем также $J_p^{(N)}(\tilde{\varphi})$ — приближение N-го порядка к функционалу $J_p(\tilde{\varphi})$:

$$\delta J_{p,i} = (\varphi^{(i)}, p) = (\varphi, p) + \sum_{i=1}^{N} \varepsilon^i \delta J_{p,i}. \tag{8.3.21}$$

При условии сходимости алгоритма имеем

$$|J_p(\tilde{\varphi}) - J_p^{(N)}(\tilde{\varphi})| \leqslant O(|\varepsilon|^{N+1}).$$

Для нахождения $J_p(\tilde{\varphi})$ можно воспользоваться решением сопряженного уравнения

$$(A^* + \varepsilon \delta A^*)\tilde{\varphi}_p^* = p. \tag{8.3.22}$$

Представляя $\tilde{\varphi}_p^* = \sum\limits_{i=1}^{\infty} \varepsilon^i \varphi_i^*$ и привлекая для вычисления φ_i^* ($i = 1, 2, \dots$) алгоритм возмущений (см. § 8.1), найдем

$$\delta J_{p,i}^* = (f, \varphi_i^*), \quad i = 1, 2, \dots. \tag{8.3.23}$$

Приближение N-го порядка в терминах решений сопряженных уравнений имеет вид

$$J_p^{(N)}(\tilde{\varphi}) = (f, \varphi^*) + \sum_{i=1}^{N} \varepsilon^i \delta J_{p,i}^*. \tag{8.3.24}$$

Если $N = 1$, то получаем формулы теории малых возмущений

$$J_p^{(1)}(\tilde{\varphi}) = (\varphi, p) + \varepsilon \delta J_{p,1}, \tag{8.3.25}$$

$$J_p^{(1)}(\tilde{\varphi}) = (f, \varphi^*) + \varepsilon \delta J_{p,1}^*, \tag{8.3.26}$$

где $(\varphi, p) = (f, \varphi^*)$. Отмечаем также, что $\delta J_{p,1} = \delta J^*_{p,1}$. Действительно, вспоминая вид уравнения для φ^1, φ^*_1, имеем

$$\delta J_{p,1} = (\varphi^{(1)}, p) = (\varphi^{(1)}, A^*\varphi^*) = (A\varphi^{(1)}, \varphi^*) = -(\delta A\varphi, \varphi^*) = \\ = -(\varphi, \delta A^*\varphi^*) = (\varphi, A^*\varphi^*_1) = (A\varphi, \varphi^*_1) = (f, \varphi^*_1) = \delta J^*_{p,1};$$

таким образом,

$$\delta J_{p,1} = \delta J^*_{p,1} = -(\delta A\varphi, \varphi^*),$$

т. е. для отыскания $\delta J_{p,1}$ не нужно вычислять $\varphi^{(1)}$ или φ^*_1. Заметим также, что при $\varepsilon = 1$ формулы (8.3.25), (8.3.26) совпадают с (8.3.17), (8.3.18).

8.4. Алгоритмы возмущений в нестационарных задачах. Применение спектрального метода

Сформулированные в предыдущих параграфах алгоритмы распространяются и на нестационарные задачи. Здесь мы остановимся лишь на одном из них — алгоритме вычисления линейных функционалов.

Пусть A, A^* являются теми же операторами, что и в § 8.1.

Рассмотрим нестационарную задачу в абстрактной форме

$$\frac{\partial \varphi}{\partial t} + A\varphi = f, \quad \varphi = g \text{ при } t = 0 \tag{8.4.1}$$

и поставим ей в соответствие сопряженную задачу

$$-\frac{\partial \varphi^*}{\partial t} + A^*\varphi^* = f^*, \quad \varphi^* = g^* \text{ при } t = T. \tag{8.4.2}$$

Здесь f^* и g^* — пока не определенные элементы, которые будут выбраны позднее. Уравнения (8.4.1), (8.4.2) помножим соответственно на φ^* и φ; полученные результаты вычтем один из другого и проинтегрируем по t на сегменте $0 \leqslant t \leqslant T$. В результате получим

$$\int_0^T \frac{\partial}{\partial t}(\varphi^*, \varphi)\, dt + \int_0^T [((\varphi^*, A\varphi) - (\varphi, A^*\varphi^*)]\, dt = \int_0^T [(f, \varphi^*) - (f^*, \varphi)]\, dt, \tag{8.4.3}$$

Так как операторы A^* и A сопряжены, т. е.

$$(\varphi^*, A\varphi) - (\varphi, A^*\varphi^*) = 0,$$

то выражение (8.4.3) с учетом соответствующих начальных условий представляется в виде

$$(g^*, \varphi_T) - (g, \varphi_0^*) = \int_0^T [(f, \varphi^*) - (f^*, \varphi)]\, dt, \tag{8.4.4}$$

где

$$\varphi_T = \varphi|_{t=T}, \quad \varphi_0^* = \varphi^*|_{t=0}.$$

Теперь предположим, что нам требуется найти линейный функционал от решения (8.4.1), имеющий вид

$$J = (g^*, \varphi_T) + \int_0^T (f^*, \varphi)\, dt. \tag{8.4.5}$$

Предположим, что входные данные в задаче (8.4.1) возмущены, т. е. вместо g и f рассмотрим $\tilde{g} = g + \delta g$ и $\tilde{f} = f + \delta f$. Тогда, учитывая последнее выражение, мы получим формулу для вариации функционала

$$\delta J = (\delta g, \varphi_0^*) + \int_0^T (\delta f, \varphi^*)\, dt. \tag{8.4.6}$$

Таким образом, для вычисления отклонения функционала J, соответствующего различным изменениям во входных данных, не нужно решать большое число задач вида

$$\frac{\partial \tilde{\varphi}}{\partial t} + A\tilde{\varphi} = \tilde{g}, \quad \tilde{\varphi} = \tilde{g} \text{ при } t = 0 \tag{8.4.7}$$

с различными $\tilde{f}$ и $\tilde{g}$. Достаточно решить одно сопряженное уравнение (8.4.2) и воспользоваться формулой (8.4.6). Подчеркнем, что формула (8.4.6) является точной, т. е. она дает нам выражение для δJ без каких-либо погрешностей.

Рассмотрим теперь задачу (8.4.1), в которой возмущены не только f, g, но и оператор A:

$$\frac{\partial \tilde{\varphi}}{\partial t} + \tilde{A}\tilde{\varphi} = \tilde{g}, \quad \tilde{\varphi} = \tilde{g} \text{ при } t = 0, \tag{8.4.8}$$

где $\tilde{f} = f + \delta f$, $\tilde{g} = g + \delta g$, $\tilde{A} = A + \delta A$. Пусть снова требуется вычислить значение функционала:

$$J(\tilde{\varphi}) = (g^*, \tilde{\varphi}_T) + \int_0^T (f^*, \tilde{\varphi})\, dt, \tag{8.4.9}$$

где $\tilde{\varphi}_T = \tilde{\varphi}|_{t=T}$. Проведя рассуждения, аналогичные приведенным выше, несложно получить формулу малых возмущений для J

$$J(\tilde{\varphi}) \cong J(\varphi) + \delta J^{(1)}, \tag{8.4.10}$$

где

$$\delta J^{(1)} = (\delta g, \varphi_0^*) + \int_0^T (\delta f - \delta A\varphi_0, \varphi_0^*)\, dt. \tag{8.4.11}$$

В ряде задач математической физики могут интересовать не только функционалы от решений, но и отклонения от решений невозмущенной задачи. Эти отклонения (в случае, когда оператор не возмущается) удовлетворяют уравнению

$$\frac{\partial}{\partial t}\delta\varphi + A\delta\varphi = \delta f, \quad \delta\varphi = \delta\varphi_0 \text{ при } t = 0. \tag{8.4.12}$$

Если рассматривается стационарная задача, то это уравнение принимает вид

$$A\delta\varphi = \delta f. \tag{8.4.13}$$

Для решения (8.4.12), (8.4.13) можно воспользоваться методами, изложенными в предыдущих главах. Но может оказаться предпочтительным и спектральный метод (метод Фурье), в частности, это может быть в случае, когда необходимо вычислить достаточно большое число отклонений от невозмущенного решения. Сформулируем данный метод в применении к решению уравнения (8.4.13).

Рассмотрим задачу на собственные значения

$$Aw = \lambda w \tag{8.4.14}$$

и соответствующую ей сопряженную задачу

$$A^*w^* = \lambda w^*. \tag{8.4.15}$$

Пусть (8.4.14) определяет полную систему собственных вектор-функций $\{w_n\}$, соответствующих собственным значениям $\{\lambda_n\}$, а задача (8.4.15) определяет $\{w_n^*\}$ и $\{\bar{\lambda}_n\}$. Пусть имеет место ортогональность вектор-функций, а именно:

$$(w_n, w_n^*) = 0 \text{ при } m \neq n. \tag{8.4.16}$$

Эти системы можно пронормировать:

$$(w_n, w_m^*) = \begin{cases} 1, & n = m, \\ 0, & n \neq m. \end{cases} \tag{8.4.17}$$

Решение задачи (8.4.13) представим в виде ряда Фурье по собственным функций (8.4.14):

$$\delta\varphi = \sum_n \delta\varphi_n w_n \tag{8.4.18}$$

и аналогично

$$\delta f = \sum_n \delta f_n w_n, \tag{8.4.19}$$

где

$$\delta\varphi_n = (\delta\varphi, w_n^*), \quad \varphi_n = (\delta f, w_n^*).$$

Подставляя (8.4.18) и (8.4.19) в (8.4.13) и скалярно умножая на w_n^*, получаем соотношение

$$\lambda_m \delta\varphi_m = \delta f_m, \quad m = 1, 2, \ldots. \tag{8.4.20}$$

Таким образом, решение (8.4.13) записывается в виде

$$\delta\varphi = \sum_n \frac{\delta f_n}{\lambda_n} w_n, \tag{8.4.21}$$

или

$$\delta\varphi = \sum_n \frac{(\delta f, w_n^*)}{\lambda_n} w_n. \tag{8.4.22}$$

Итак, отклонение решения от основного состояния мы получили в форме ряда Фурье. Если мы интересуемся линейным функционалом, а не самим решением, то

$$\delta J = (p, \delta\varphi). \tag{8.4.23}$$

В этом случае

$$\delta J = \sum_n \frac{(\delta F, w_n^*)}{\lambda_n} (p, w_n) = (\delta F, \varphi^*), \tag{8.4.24}$$

где

$$\varphi^* = \sum \frac{(p, w_n)}{\bar{\lambda}_n} w_n^*.$$

Здесь φ^* есть решение задачи $A^*\varphi^* = p$.

Мы рассмотрели стационарный случай. Аналогично может быть рассмотрена нестационарная задача и получены соответствующие формулы теории возмущения.

8.5. Формулировка теории возмущений для сложных нелинейных моделей

Обычно довольно сложно конструировать математическую модель, описывающую сложные процессы и явления. Такие модели должны учитывать, как правило, много различных эффектов, не все из которых описаны с требуемой точностью. Это значит, что в тот или иной момент времени мы используем одну или другую упрощенную математическую формулировку, которая позволяет нам описать лишь небольшое число характеристик процесса, абстрагируясь от многих, иногда очень важных деталей. Однако такие рассмотрения позволяют в общем описать физические процессы, как правило, с необходимой точностью. Что касается оценки влияния эффекта, не учтенного математической моделью, то она, как это было показано в предыдущем параграфе, может быть осуществлена с помощью специальным образом определенной теории возмущения.

В данном параграфе мы осуществим переход от нелинейных уравнений к нелинейным. Мы покажем, что в известных приближениях нелинейные задачи также допускают для своего решения и анализа подходы, свойственные задачам нелинейным. Разумеется, такие подходы возможны на базе различных способов линеаризации.

Кроме того, ниже мы осуществим ряд обобщений результатов предыдущих параграфов. Так, мы рассмотрим уравнения, зависящие от набора параметров $\{\alpha_i\}$, $\{\beta_i\}$, а оператор A будем считать матрично-дифференциальным оператором, действующим в некотором гильбертовом пространстве F со скалярным произведением $(\cdot, \cdot)$. Далее, в настоящем параграфе мы допускаем ситуацию, когда оператор A и элемент f в невозмущенном уравнении известны с ограниченной точностью. Уравнения с отмеченными свойствами часто возникают в задачах теории планирования эксперимента в оптимальном управлении, при решении обратных задач.

Рассмотрим стационарный процесс, который описывается уравнением в следующей операторной форме:

$$A\varphi = f, \tag{8.5.1}$$

где A — есть матрично-дифференциальный оператор, зависящий от решения вектор-функции φ и входных данных $\alpha_1, \alpha_2, \ldots, \alpha_n$ — функции координат, f является заданным вектором источников — функций координат и заданных параметров $\beta_1, \beta_2, \ldots, \beta_m$. Следовательно, мы имеем

$$A = A(\varphi, \alpha_1, \alpha_2, \ldots, \alpha_n)$$

и

$$f = f(\beta_1, \beta_2, \ldots, \beta_m).$$

Пусть $\bar{\alpha}_1, \bar{\alpha}_2, \ldots, \bar{\alpha}_n$ и $\bar{\beta}_1, \bar{\beta}_2, \ldots, \bar{\beta}_m$ — данные, соответствующие некоторому стандартному состоянию системы. Это состояние будем называть невозмущенным. Естественно, что невозмущенное состояние системы будет описываться уравнением

$$A(\bar{\varphi}, \bar{\alpha}_1, \bar{\alpha}_2, \ldots, \bar{\alpha}_n)\bar{\varphi} = f(\bar{\beta}_1, \bar{\beta}_2, \ldots, \bar{\beta}_m). \tag{8.5.2}$$

Если ввести обозначения

$$\bar{A} = A(\bar{\varphi}, \bar{\alpha}_1, \bar{\alpha}_2, \ldots, \bar{\alpha}_n), \quad \bar{f} = f(\bar{\beta}_1, \bar{\beta}_2, \ldots, \bar{\beta}_m),$$

то уравнение (8.5.2) перепишется в виде

$$\bar{A}\bar{\varphi} = \bar{f}.$$

Мы предложим, что решение уравнения (8.5.3) нам известно.

Примем, далее, что истинное, или, как мы будем в дальнейшем называть, возмущенное, состояние системы описывается уравнением (8.5.1), входные данные которого мало отличаются от стандартных, т. е.

$$\alpha_i = \bar{\alpha}_i + \delta\alpha_i, \quad \beta_j = \bar{\beta}_j + \delta\beta_j, \tag{8.5.3}$$

где отклонения $\delta\alpha_i$ и $\delta\beta_i$ предполагаются малыми по сравнению с $\bar{\alpha}_i$ и $\bar{\beta}_i$ соответственно.

Тогда вместо (8.5.2) имеем

$$A(\bar{\varphi} + \delta\varphi, \bar{\alpha}_i + \delta\alpha_i)(\bar{\varphi} + \delta\varphi) = f(\bar{\beta}_j + \delta\beta_j). \tag{8.5.4}$$

Здесь используется следующее представление:

$$\varphi = \bar{\varphi} + \delta\varphi.$$

Априори предложим, что $\delta\varphi$ много меньше $\bar{\varphi}$. В предположении достаточной гладкости оператора A, решения φ и входных данных рассмотрим разложения

$$\begin{gathered} A(\bar{\varphi} + \delta\varphi, \bar{\alpha}_i + \delta\alpha_i) = \bar{A} + \overline{\frac{\partial A}{\partial \varphi}}\delta\varphi + \overline{\frac{\partial A}{\partial \alpha_i}}\delta\alpha_i + \ldots, \\ f(\bar{\beta}_j + \delta\beta_j) = f + \overline{\frac{\partial f}{\partial \beta_j}}\delta\beta_j. \end{gathered} \tag{8.5.5}$$

Подставляя (8.5.5) в (8.5.4) и ограничиваясь членами первого порядка, получаем

$$\bar{A}\bar{\varphi} + \left(\bar{A} + \overline{\frac{\partial A}{\partial \varphi}}\bar{\varphi}\right)\delta\varphi + \overline{\frac{\partial A}{\partial \alpha_i}}\bar{\varphi}\delta\alpha_i = f + \overline{\frac{\partial f}{\partial \beta_i}}\delta\beta_j. \tag{8.5.6}$$

Привлекая уравнение (8.5.3), имеем

$$\left(\bar{A} + \overline{\frac{\partial A}{\partial \varphi}}\bar{\varphi}\right)\delta\varphi = \overline{\frac{\partial f}{\partial \beta_i}}\delta\beta_j - \overline{\frac{\partial A}{\partial \alpha_i}}\bar{\varphi}\delta\alpha_i. \tag{8.5.7}$$

Это основное уравнение для определения малых отклонений решения φ от невозмущенного состояния.

Мы предположим теперь, что оператор невозмущенного состояния $\bar{A}$ и источник $\bar{f}$ известны с ограниченной точностью, т. е.

$$\begin{aligned} \bar{A} &= \bar{\Lambda} + \varepsilon, \\ \bar{f} &= \bar{F} + \xi. \end{aligned} \tag{8.5.8}$$

Здесь ε — оператор ошибки модели:

$$\varepsilon = \bar{A} - \Lambda,$$

а ξ — ошибка вектор-функции источников:

$$\xi = \bar{f} - \bar{F}.$$

Пусть

$$\begin{aligned} \|\bar{A}\| &\gg \|\varepsilon\|, \\ \|\bar{f}\| &\gg \|\xi\|. \end{aligned} \tag{8.5.9}$$

Здесь нормы оператора и вектор-функции определены в соответствующих метрических пространствах.

Подставляя (8.5.13) в (8.5.7), имеем

$$\left(\bar{\Lambda} + \overline{\frac{\partial \Lambda}{\partial \varphi}}\bar{\varphi}\right)\delta\varphi = \overline{\frac{\partial F}{\partial \beta_j}}\delta\beta_j - \overline{\frac{\partial \Lambda}{\partial \alpha_j}}\bar{\varphi}\delta\alpha_j + \eta, \tag{8.5.10}$$

где

$$\eta = \overline{\frac{\partial \xi}{\partial \beta_j}}\delta\beta_j - \overline{\frac{\partial \varepsilon}{\partial \alpha_i}}\bar{\varphi}\delta\alpha_i - \left(\varepsilon + \overline{\frac{\partial \varepsilon}{\partial \varphi}}\bar{\varphi}\right)\delta\varphi. \tag{8.5.11}$$

В предположении (8.5.9) уравнение (8.5.10) можно переписать в виде

$$\left(\bar{\Lambda} + \overline{\frac{\partial \Lambda}{\partial \varphi}}\bar{\varphi}\right)\delta\varphi = \overline{\frac{\partial F}{\partial \beta_j}}\delta\beta_j - \overline{\frac{\partial \Lambda}{\partial \alpha_i}}\bar{\varphi}\delta\alpha_i + O(\|\varepsilon\| + \|\xi\|).$$

Следовательно, с точностью до малых величин

$$\left(\bar{\Lambda} + \overline{\frac{\partial \Lambda}{\partial \varphi}}\bar{\varphi}\right)\delta\varphi = \overline{\frac{\partial F}{\partial \beta_j}}\delta\beta_j - \overline{\frac{\partial \Lambda}{\partial \alpha_i}}\bar{\varphi}\delta\alpha_i. \tag{8.5.12}$$

Уравнение, полученные выше, есть модель для вычисления невозмущенного состояния системы при изменении входных данных на величины $\delta\alpha_i$ и $\delta\beta_j$.

Введем обозначения

$$\begin{aligned} L &= \bar{\Lambda} + \overline{\frac{\partial \Lambda}{\partial \varphi}}\bar{\varphi}, \\ \delta F &= \overline{\frac{\partial F}{\partial \beta_j}}\delta\beta_j - \overline{\frac{\partial \Lambda}{\partial \alpha_i}}\bar{\varphi}\delta\alpha_i. \end{aligned} \tag{8.5.13}$$

Тогда окончательно имеем

$$L\delta\varphi = \delta F, \tag{8.5.14}$$

а формальное решение этой задачи записывается в виде

$$\delta\varphi = L^{-1}\delta F. \tag{8.5.15}$$

Формула (8.5.15) в теории возмущений наиболее удобна для случая, когда необходимо определить отклонения решения только для одного набора входных данных. При планировании направленных изменений состояния модели весьма важно бывает сделать серию тестовых вычислений. Мы должны принимать во внимание тот факт, что для оценки чувствительности модели на изменение различных параметров или для получения оптимального соотношения между параметрами нужно иметь большое число различных решений. Следовательно, нам необходимо попытаться построить более универсальные теории возмущений для функционалов, которые позволят нам всесторонне изучить математическую модель при изменении входных данных.

Вместе с основным уравнением

$$L\delta\varphi = \delta F \tag{8.5.16}$$

так же, как и в линейном случае, введем сопряженное уравнение

$$L^*\varphi^* = p, \tag{8.5.17}$$

где L и L^* — операторы, сопряженные в смысле Лагранжа:

$$(Lg, h) = (g, L^*h). \tag{8.5.18}$$

Здесь g и h — элементы гильбертова пространства из области определения операторов L и L^* соответственно. Функцию p пока будем предполагать неопределенной.

Умножим скалярно (8.5.16) и (8.5.17) на φ^* и $\delta\varphi$ соответственно и вычтем результаты один из другого. Тогда

$$(\varphi^*, L\delta\varphi) - (\delta\varphi, L^*\varphi^*) = (\varphi^*, \delta F) - (p, \delta\varphi). \tag{8.5.19}$$

Выражение, стоящее в левой части равенства (8.5.19), с учетом (8.5.18) равно нулю. Следовательно, мы имеем

$$(p, \delta\varphi) = (\varphi^*, \delta F). \tag{8.5.20}$$

Рассмотрим теперь набор линейных функционалов

$$\delta J_n = (p_n, \delta\varphi). \tag{8.5.21}$$

Функцию φ^*, соответствующую p_n, будем обозначать через φ_n^*.

Таким образом, на основе равенства (8.5.20) мы будем иметь набор функционалов

$$\delta J_n = (\varphi_n^*, \delta F). \tag{8.5.22}$$

Предположим, что мы заранее выбрали N функционалов $J_1, J_2, \ldots, J_N$ и соответственно решили N сопряженных задач

$$L^*\varphi_n^* = p_n, \quad n = 1, 2, \ldots, N. \tag{8.5.23}$$

Из (8.5.22) следует, что нет необходимости вычислять вариации δJ_n, соответствующие различным наборам $\delta\alpha_i$ и $\delta\beta_j$, поскольку, задавая φ_n^* ($n = 1, 2, \ldots, N$), можно непосредственно вычислить значение функционалов δJ_n при любых возмущениях во входных данных.

Таким образом, осуществив линеаризацию исходной задачи, мы получили формулы теории малых возмущений, вид которых аналогичен полученным в линейных задачах. Но обратим внимание на тот факт, что значительная часть информации о нелинейности задачи теперь содержится в операторе L. Это обстоятельство и обусловливает эффективность применения сформулированных алгоритмов во многих нелинейных задачах.

8.6. Применения сопряженных уравнений и методов возмущений в прикладных задачах

В данном параграфе мы проиллюстрируем некоторые из возможных приложений сформулированных в данной главе методов теории сопряженных уравнений и алгоритмов возмущений в прикладных задачах.

8.6.1. Задачи теории переноса излучения

Сопряженные уравнения и алгоритмы возмущений широко используются в самых различных задачах теории ядерных реакторов, в частности в задачах для уравнения переноса излучения в веществе. Это уравнение является линейным интегро-дифференциальным уравнением в частных производных первого порядка. В нестационарном случае оно имеет вид

$$\frac{1}{v}\frac{\partial\varphi}{\partial t}+\bar{\Omega}\cdot\bar{\nabla}\varphi+\Sigma(\bar{r},E)\varphi-\int d\bar{\Omega}'\int dE'\Sigma(\Omega'\to\bar{\Omega};E'\to\bar{E})\cdot\varphi(\bar{r},E',\bar{\Omega}',t)=q(\bar{r},E,\bar{\Omega},t), \tag{8.6.1}$$

где неизвестная функция $\varphi(\bar{r},E,\bar{\Omega},t)$ есть поток частиц (например, нейтронов), летящих в направлении $\bar{\Omega}$ с энергией E в точке $\bar{r}$ в момент времени t; $\Sigma(\bar{r},E)$ — полное макроскопическое сечение взаимодействия; $\Sigma(\bar{\Omega}'\to\bar{\Omega},E'\to E)$ — дифференциальное сечение перехода из $(E',\bar{\Omega}')$ в $(E,\bar{\Omega})$ при столкновениях (мы не разделяем здесь это сечение на сечении упругого и неупругого рассеяния, деления и т. д.); $q(r,E,\bar{\Omega},t)$ — распределение источников излучения.

Пусть среда, в которой мы ищем решение уравнения (8.6.1), занимает объем V, ограниченный поверхностью S. Если вне объема V источники отсутствуют и среда вне V не отражает излучения, то естественным граничным условием для φ является

$$\varphi(\bar{r}_s,E,\bar{\Omega},t)=0 \text{ при } \bar{\Omega}\bar{n}<0, \tag{8.6.2}$$

где $\bar{n}$ — внешняя нормаль к поверхности S в точке $\bar{r}_s$.

Считая, что источники действуют не бесконечно долго, получим начальное условие

$$\varphi(\bar{r},E,\bar{\Omega},t)=0 \text{ при } t\to\infty. \tag{8.6.3}$$

Сопряженное к (8.4.1) уравнение имеет вид

$$-\frac{1}{v}\frac{\partial\varphi_p^*}{\partial t}-\bar{\Omega}\cdot\bar{\nabla}\varphi_p^*+\Sigma\varphi_p^*-\int d\Omega'\int dE'\Sigma(\bar{\Omega}\to\bar{\Omega}';E\to E')\cdot\varphi_p^*(\bar{r},E',\bar{\Omega}',t)=p(\bar{r},E,\bar{\Omega},t) \tag{8.6.4}$$

с граничными и начальными условиями

$$\varphi_p^*(\bar{r}_s,E,\bar{\Omega},t)=0 \text{ при } \bar{n}\bar{\Omega}>0, \tag{8.6.5}$$

$$\varphi_p^*(r,E,\bar{\Omega},t)=0 \text{ при } t\to\infty. \tag{8.6.6}$$

Отметим, что уравнение (8.6.4) можно получить исходя из физического смысла сопряженных функций $\varphi_p^*(r,E,\bar{\Omega},t)$ как ценности частиц (например, нейтронов) по отношению к функционалу:

$$J_p[\varphi]=\int\limits_V d\bar{r}\int dE\int d\bar{\Omega}\int\limits_{-\infty}^{+\infty}dt\,\varphi(\bar{r},E,\bar{\Omega},t)p(\bar{r},E,\bar{\Omega},t). \tag{8.6.7}$$

При этом ценность нейтрона $\varphi_p^*(\bar{r},E,\bar{\Omega},t)$ равна тому значению, которое принимает функционал $J_p[\varphi]$ при впускании одного нейтрона в точку $x=(\bar{r},E,\bar{\Omega},t)$.

Предположим теперь, что различные возмущения в рассматриваемой системе приводят к изменению сечений взаимодействия излучения с веществом:

$$\Sigma\to\Sigma+\delta\Sigma,$$
$$\Sigma(\bar{\Omega}'\to\bar{\Omega};E'\to E)\to\Sigma(\bar{\Omega}'\to\bar{\Omega};E'\to E)+\delta\Sigma(\bar{\Omega}'\to\bar{\Omega},E'\to E).$$

При этом изменяется оператор как основного уравнения $A\to A+\delta A$, так и сопряженного уравнения $A^*\to A^*+\delta A^*$, причем

$$\delta A=\delta\Sigma-\int d\bar{\Omega}'\int dE'\delta\Sigma(\bar{\Omega}'\to\bar{\Omega};E'\to E),$$
$$\delta A^*=\delta\Sigma-\int d\bar{\Omega}'\int dE'\delta\Sigma(\bar{\Omega}\to\bar{\Omega}';E\to E'),$$

Используя соотношения (8.3.14) и (8.3.16) из § 8.3, получим формулы теории возмущений в виде

$$\delta J_p=-\int d\bar{r}\int d\bar{\Omega}\int dE\int dt\,\varphi_p^*\left[\delta\Sigma\,\tilde{\varphi}_i-\int d\bar{\Omega}'\int dE'\delta\Sigma(x'\to x)\tilde{\varphi}(x')\right], \tag{8.6.8}$$

$$\delta J_p=-\int d\bar{r}\int d\bar{\Omega}\int dE\int dt\,\varphi\left[\delta\Sigma\,\tilde{\varphi}_p^*-\int d\bar{\Omega}'\int dE'\delta\Sigma(x'\to x)\varphi_p^*(x')\right]. \tag{8.6.9}$$

Как уже отмечалось, соотношения теории возмущений наряду с их прямым использованием (для оценки различных эффектов и для анализа измерений) могут быть полезными для нахождения упрощенной модели сложной системы, такой, в которой некоторая интересующая нас величина J_p имеет то же значение, что и в истинной системе.

В качестве примера могут быть рассмотрены гомогенизация гетерогенной системы или усреднение сечений по энергиям. Формулы (8.6.8) и (8.6.9) позволяют получить метод усреднения сечений по пространству или по энергии.

Действительно, подставляя, например, в соотношение (8.6.8) $\delta\Sigma = \bar{\Sigma} - \Sigma(\bar{r}, E)$ и полагая $\delta J_p = 0$, получим

$$\bar{\Sigma} = \frac{\int d\bar{r} \int dE \Sigma(\bar{r}, E) \int d\bar{\Omega} \tilde{\varphi} \varphi_p^*}{\int d\bar{r} \int dE \int d\bar{\Omega} \tilde{\varphi} \varphi_p^*}, \tag{8.6.10}$$

т. е. сечения следует усреднять с весом $\tilde{\varphi}\varphi_p^*$.

Замечание. К числу таких методов относится и метод эффективных граничных условий, заключающийся в замене истинных условий некоторыми упрощенными условиями, но такими, которые приводят к правильному значению потока излучения вдали от границы.

8.6.2. Задачи охраны окружающей среды

Как следует из предыдущих параграфов, использование сопряженных функций для представления линейных функционалов в форме $J_p = (f, \varphi_p^*)$ позволяет часто просто оценить вариации функционала J_p (или несколько таких функционалов) от вариаций входных параметров исходный задач (например, от вариаций функции f). Это свойство сопряженных функций было положено во многих работах в основу решения важных прикладных задач охраны окружающей среды и позволило предложить конструктивные экономичные подходы к решению этих задач. Изложим идею этих подходов на примере задачи о моделировании экологических ситуаций в акваториях водных бассейнов.

Интенсивное развитие промышленности требует выяснения оптимальных условий размещения новых промышленных предприятий и технологических ограничений на стоки, загрязняющие водные бассейны (моря, озера, заливы и т. д.), с таким расчетом, чтобы загрязнения прибрежных зон было минимальным. Математически эта задача сводится к проблеме минимакса.

Рассмотрим водный бассейн с областью определения D, которую будем считать цилиндром с боковой (береговой) поверхностью S и постоянной

глубиной H. Задача диффузии гидрозоля имеет вид

$$\begin{gathered}\left(\frac{\partial}{\partial t}+u\frac{\partial}{\partial x}+v\frac{\partial}{\partial y}+w\frac{\partial}{\partial z}\right)\varphi+\sigma\varphi-\frac{\partial}{\partial z}v\frac{\partial}{\partial z}\varphi-\mu\Delta\varphi=Qw(r-r_0),\\ \frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}+\frac{\partial w}{\partial z}=0,\\ w=0 \text{ при } z=0,\ z=H,\\ \frac{\partial\varphi}{\partial n}=0 \text{ на } S,\\ \frac{\partial\varphi}{\partial z}=0 \text{ при } z=0,\ z=H,\\ \varphi(r,T)=\varphi(r,0),\end{gathered} \tag{8.6.11}$$

где φ — концентрация загрязняющего гидрозоля, $w(r-r_0)$ — функция источника, $r=(x,y,z)$, $0\leqslant t\leqslant T$, Q — мощность источника, r_0 — точка предполагаемого стока, u,v,w — компоненты вектора $\bar{u}$. Предполагается, что решение задачи (8.6.11) имеет годовой ход, т. е. T — интервал времени, равный году.

Задачу (8.6.11) можно записать в операторной форме

$$A\varphi=f. \tag{8.6.12}$$

Для решения задачи о размещении стоков промышленного предприятия в акватории водного бассейна введем в рассмотрение функционал

$$J_{p_i}=\int\limits_0^T dt\int\limits_D p_i(r)\varphi(r,t)\,dr, \tag{8.6.13}$$

где

$$p_i(r)=\begin{cases}1 \text{ при } r\in D_i,\\ 0 \text{ при } r\notin D_i,\end{cases}$$

D_i — область, подлежащая охране от загрязнений. Введем в рассмотрение ограничения по санитарной норме, т. е. потребуем, чтобы r_0 принадлежало такой области Σ_0, для каждой точки которой выполнялось бы условие

$$J_{p_i}(r_0)\leqslant c_i, \tag{8.6.14}$$

где c_i — константа, связанная с санитарной нормой для района D_i. Итак, задача сводится к решению уравнения (8.6.12) при ограничении (8.6.14).

Для решения уже этой задачи введем в рассмотрение сопряженную задачу

$$A^*\varphi^*_{p_i}=p_i. \tag{8.6.15}$$

Теперь, пользуясь соотношением сопряженности $(A\varphi, \varphi^*_{p_i}) = (\varphi, A^*\varphi^*_{p_i})$ ($(\cdot,\cdot) \equiv (\cdot,\cdot)_{L_2(D\times[0,T])}$), получаем, что функционал J_{p_i} можно представить следующим образом:

$$J_p = Q \int_0^T dt \int_D \varphi^*_{p_i}(r,t) w(r - r_0)\, dr. \tag{8.6.16}$$

Именно на основе (8.6.16) и будем решать задачу. Так, пусть уравнение (8.6.15) решено. Тогда найдем функционал J_{p_i}, параметрически зависящий от r_0. По формуле (8.6.16), меняя значение r_0, мы можем построить линии одинаковых значений функционала $J_{p_i}(r_0)$, а значит, и определить области, где $J_{p_i} \leqslant c_i$. Если окажется, что именно этой области принадлежит интересующая нас D_i, то задача выбора допустимых значений Q и r_0 решена. Если же такой области не найдется, то необходимо, например, уменьшить значение интенсивности и решить задачу выбора Q вновь.

Как легко заметить, сформулированная задача легко обобщается на случай, когда мы имеем n экологически охраняемых зон D_i ($i = 1, 2, \ldots, n$). Здесь рассматривается n функционалов $J_{p_i}(r_0)$ при ограничениях $J_{p_i}(r_0) \leqslant c_i$ ($i = 1, 2, \ldots, n$). Для практической реализации алгоритма решения задачи необходимо будет один раз решить n сопряженных задач, а затем уже решить следующую задачу минимакса:

$$\max_{i=1,2,\ldots,n} B_i(r_0) = \min_{r_0 \in \Sigma_0},$$

где $B_i(r_0) = J_{p_i}(r_0)/c_0$. Заметим, что здесь эта проблема минимакса решается путем явного перебора, который выполняется элементарно. В этом состоит замечательное свойство сопряженных задач, позволяющих проблему минимакса свести к простейшей реализации. Если бы мы не воспользовались сопряженными задачами, то для решения задачи минимакса нам потребовалось бы решать большое число задач с различным положением источников промышленных отходов. Такая задача едва ли была бы разрешима с заданной точностью даже при использовании самых совершенных средств вычислительной техники.

В заключение заметим, что алгоритмы теории возмущений и сопряженные уравнения играют важную роль для постановки и разработки методов решения многих обратных задач математической физики. Данная область приложений теории возмущений и сопряженных уравнений будет рассмотрена в следующей главе.

Глава 9.

Постановка и численные методы решения некоторых обратных задач

В современной литературе термин «обратные задачи» используется для ряда различных типов задач математической физики.

Мы рассмотрим два типа обратных задач. Первый тип — это задачи определения состояния некоторого процесса в предыдущие моменты времени. Примером может служить задача об определении начального распределения температуры в теле, если известно поле тепла к данному моменту времени. Другой тип — это задачи, в которых требуется восстановить оператор с известной структурой, но с неизвестными коэффициентами, подлежащими определению на основе информации о функционалах от решений. Примером может служить обратная задача для уравнения Штурма — Лиувилля, в которой требуется определить коэффициент в дифференциальном уравнении второго порядка по свойствам спектральной функции некоторой краевой задачи.

Обратные задачи математической физики часто оказываются в классическом смысле поставленными некорректно. Малым изменениям в регистрируемых функционалах могут соответствовать большие изменения в решениях задач. Понятие корректности и пример некорректной задачи математической физики — задачи Коши для Лапласа — были приведены в начале нашего века Адамаром.

Долгое время так называемые некорректные задачи математической физики считались неинтересными и исследовались мало. Интенсивное исследование этих задач началось в связи с необходимостью решения задач интерпретации геофизических данных. Большой вклад в развитие теории и

методов решения задач математической физики, не являющихся корректными в классическом смысле (по Адамару), внесли советские математики (см. § 12.3).

Как было показано, решение некорректных в классическом смысле задач становится устойчивым по отношению к изменениям данных, если наложить на множество допустимых решений некоторые дополнительные ограничения. Поэтому задачи такого типа получили название условно корректных.

В связи с необходимостью построения приближенных решений условно корректных задач по приближенным данным было введено понятие регуляризирующего семейства по Тихонову, суть которого состоит в следующем. Условно корректной задаче сопоставляется семейство классических корректных задач (регуляризирующее семейство), зависящее от параметра, причем при стремлении параметра к некоторому пределу последовательность решений классически корректных задач должна стремиться к решению интересующей нас условно корректной задачи. Было показано, что при соответствующем выборе параметра (параметра регуляризации), зависящего от точности данных, решение задачи из регуляризирующего семейства по приближенным данным будет являться приближенным решением нашей корректной задачи.

В работах отмеченных выше авторов исследован широкий круг условно корректных задач. Мы рассмотрим только некоторые из них. § 9.1 вводит в круг основных понятий общей теории условно корректных задач. В § 9.2, 9.3 рассмотрена регуляризация задачи определения входных (начальных) данных эволюционных уравнений. В оставшейся части главы методами теории возмущений исследуются задачи восстановления структуры линейных и нелинейных операторов.

9.1. Основные определения и примеры

Рассмотрим задачу решения уравнения

$$A\varphi = f, \tag{9.1.1}$$

где A — некоторый линейный оператор, действующий в банаховом пространстве F и имеющий неограниченный обратный. В этом случае задача (9.1.1) поставлены некорректно, так как, с одной стороны, для произвольного элемента $f \in F$ решение уравнения (9.1.1) может не существовать, а с другой стороны, малым изменениям правой части f могут соответствовать сколь угодно большие вариации решения φ. Обеспечить уравнению (9.1.1)

устойчивость можно только за счет сужения класса решений $\{\varphi\}$. Пусть M — некоторое множество из пространства F. Задачу (9.1.1) будем называть условно корректной (корректной на M), если оператор A_M — сужение оператора A на M — имеет ограниченный обратный, т. е. для всех $\varphi \in M$ имеет место априорная оценка

$$\|\varphi\| \leqslant \omega(\|A\varphi\|), \quad \varphi \in M, \tag{9.1.2}$$

где $\omega(\varepsilon)$ — непрерывная функция, $\omega(0) = 0$. Множество M называется множеством корректности.

На выбор множества M влияют как физические соображения, связанные с самой постановкой задачи, так и возможности ЭВМ и требования к точности получения результата.

Это определение подсказывает также один из способов устойчивого решения уравнения (9.1.1) — минимизация функционала $\|A\varphi - f\|$ на множестве корректности M. Элемент φ_0, доставляющий минимум этому функционалу, называется квазирешением. Можно показать[1)], что при некоторых дополнительных предположениях (M — выпуклый компакт, пространство F — строго выпукло, в частности, гильбертово) квазирешение существует, единственно и непрерывно зависит от правой части $f \in F$. Тем самым понятие квазирешения как бы возвращает задаче (9.1.1) корректность. Например, на рис. 9.1 изображена ситуация, когда правая часть $f \notin AM$. В этом случае квазирешение φ_0 находится из уравнения $A\varphi_0 = g$, где g — проекция элемента f на множестве AM.

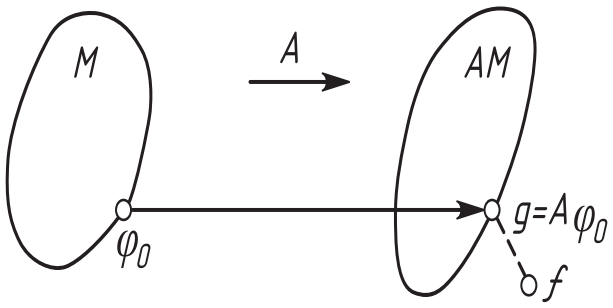


Рис. 9.1.

Часто множество корректности M можно задать с помощью некоторого неотрицательного однородного функционала l:

$$M = \{\varphi \in F | l(\varphi) \leqslant m\}, \quad l(\varphi) \geqslant 0, \quad l(\lambda\varphi) = |\lambda| l(\varphi) \tag{9.1.3}$$

[1)] См. В. К. Иванов [16].

(возможно, что для некоторых элементов $\varphi \in F$ $l(\varphi) = \infty$).

В этом случае естествен также альтернативный к методу квазирешений подход[2)] — минимизация функционала $l(\varphi)$ на множестве $\|A\varphi - f_\varepsilon\| \leqslant \varepsilon$, где ε — число, характеризующее точность входных данных f_ε: $\|f - f_\varepsilon\| \leqslant \varepsilon$ (ошибки в правой части f или в операторе A). Обычно величина функционала l^2 (называемого стабилизирующим) характеризует гладкость решения φ. Типичный пример: $F = L_2$, $l(\varphi) = \|\varphi\|_{W_2^k}$. В этом случае метод выбирает «наиболее гладкое» («до порядка k») решение неравенства $\|A\varphi - f_\varepsilon\| \leqslant \varepsilon$. Можно показать, что он эквивалентен минимизации функционала

$$\psi_\alpha[\varphi] = \|A\varphi - f_\varepsilon\|^2 + \alpha l^2(\varphi)$$

на всем пространстве, причем положительный параметр α определяется по невязке условия

$$\|A\varphi_\alpha - f_\varepsilon\| = \varepsilon,$$

где φ_α — экстремаль функционала $\psi_\alpha[\varphi]$. В этом методе не требуется знания числа m (достаточно только, чтобы множество, определяемое неравенством $l(\varphi) \leqslant 1$, было компактным и на точном решении уравнения (9.1.1) функционал l был конечен).

При выборе множества M в виде (9.1.3) можно показать[3)], что оценка (9.1.2) эквивалентна линейной оценке

$$\|\varphi\| \leqslant \alpha l(\varphi) + c(\alpha)\|A\varphi\|,\ \alpha \in (0, \alpha_0),\ \varphi \in F, \tag{9.1.4}$$

которая справедлива для всех $\varphi \in F$ и всех α из некоторого интервала $(0, \alpha_0)$, причем $c(\alpha)$ — положительная непрерывная функция α, связанная с функцией $\omega(\varepsilon)$ формулой

$$\omega(\varepsilon) = \inf_{\alpha \in (0, \alpha_1)} (\alpha m + c(\alpha)\varepsilon).$$

Обычно в некорректных задачах $\lim_{\alpha \to 0} c(\alpha) = \infty$.

Представляет большой интерес также вырожденный случай множества корректности M вида (9.1.3) с $m = 0$. В этом случае M есть некоторая поверхность Φ в пространстве F, задаваемая уравнением $l(\varphi) = 0$.

Для $\varphi \in \Phi$ оценка (9.1.4) переходит в оценку $\|\varphi\| \leqslant c\|A\varphi\|$.

Чаще всего поверхность Φ есть некоторое (конечномерное) подпространство, т. е. функционал l определен с помощью линейного проектора P : $l(\varphi) = \|P\varphi\|$. Применительно к этой ситуации проанализируем простейший пример организации итерационного алгоритма вычислений, который

[2)] А. Н. Тихонов [16].

[3)] А. Л. Бухгейм [16].

не выводит последовательность приближенных решений из заданного подпространства Φ.

Пусть F — гильбертово пространство, каждый элемент f которого представим в виде ряда Фурье по некоторой полной биортогональной системе функций $\{u_n\}$, $\{u_n^*\}$. Таким образом,

$$f = \sum_{n=1}^{\infty} f_n u_n, \tag{9.1.5}$$

где

$$f_n = (f, u_n^*).$$

Подпространство Φ зададим таким образом, чтобы в него были включены только такие элементы f гильбертова пространства, у которых в разложении (9.1.5) отлично от нуля не больше N гармоник, соответствующих наиболее крупномасштабным возмущениям:

$$f = \sum_{n=1}^{N} f_n u_n.$$

Будем решать уравнение (9.1.1) с помощью итерационного процесса

$$\varphi^{j+1} = \varphi^j - \tau(A\varphi^j - f),\ \varphi^0 = 0, \tag{9.1.6}$$

или, короче,

$$\varphi^{j+1} = T\varphi^j + \tau f, \quad \varphi^0 = 0,$$

где $T = E - \tau A$ — оператор шага.

Предположим, что решение задачи (9.1.1) существует, единственно и принадлежит подпространству Φ. Предположим далее, что на всем гильбертовом пространстве $\|T\|_F$ определена равенством

$$\|T\|_F^2 = \sup_{\varphi \in F} \frac{(T\varphi, T\varphi)}{(\varphi, \varphi)} = \beta(T^*T) > 1,$$

а на подпространстве Φ —

$$\|T\|_\Phi^2 = \sup_{\varphi \in \Phi} \frac{(T\varphi, T\varphi)}{(\varphi, \varphi)} < 1.$$

При этих предположениях итерационный процесс (9.1.6) на функциях φ^j из Φ будет сходится, а на функциях всего пространства F — расходиться. Поэтому, если мы хотим реализовать сходящийся итерационный процесс, нам необходимо позаботиться о том, чтобы на каждом шаге итерационного

процесса приближенное решение φ^j принадлежало Φ. Конструктивно это сделать весьма просто.

Положим на некотором шаге $\varphi^{j-1} \in \Phi$. Тогда с помощью рекуррентного соотношения (9.1.6) получим новый элемент φ^j. Функция $T\varphi^{j-1}$ уже может не принадлежать подпространству Φ. Для того чтобы $\varphi^j \in \Phi$, необходимо разложить эти функции в ряд Фурье:

$$\varphi^j = \sum_{k=1}^{\infty} \varphi_k^j u_k,$$

а затем отбросить в этом ряду все члены с номерами $n > N$. Алгоритмически наиболее просто для этой цели определить только N первых коэффициентов Фурье:

$$\varphi_k^j = (\varphi^j, u_k^*), \quad k = 1, 2, \ldots, N,$$

и построить конечную сумму

$$\varphi^j = \sum_{k=1}^{N} \varphi_k^j u_k.$$

Если описанный процесс продолжить как итерационный, то все функции φ^j — приближенные решения задачи — будут принадлежать подпространству Φ. При некоторых дополнительных предположениях (например, ортогональности базиса $\{u_n\}$) последовательность $\{\varphi^j\}$ будет сходиться к некоторой функции φ^∞, которая принимается за приближенное решение уравнения (9.1.1).

Обратим теперь внимание на другую сторону проблемы решения условно-корректных задач, а именно на точность задания входных данных. Обычно при решении задач математической физики приходится иметь дело по крайней мере с погрешностью за счет аппроксимации задачи (9.1.1) разностной задачей или за счет неточных сведений об операторе A и функции f.

Пусть $\bar{A}$ — точный оператор, $\bar{f}$ — точный вектор, $\bar{\varphi}$ — точное решение задачи

$$\bar{A}\bar{\varphi} = \bar{f},$$

причем

$$A = \bar{A} + \delta A, \ f = \bar{f} + \delta f,$$

и задача $A\varphi = f$ корректна на подпространстве Φ гильбертова пространства F, $A\Phi \subseteq \Phi$, A — симметричный положительный оператор. Относительно $\delta\Phi$

и δf на подпространстве Φ известна их априорная погрешность:

$$\|\delta A\|_\Phi \leqslant \varepsilon_1,\ \|\delta f\|_\Phi \leqslant \varepsilon_2. \tag{9.1.7}$$

Тогда решение уравнения $A\varphi = f$ сводится к организации такого итерационного процесса, который порождал бы новые приближения, принадлежащие подпространству Φ. Если оператором A является положительно определенная матрица, то, как было показано в главе 4, существует целый набор итерационных процессов, сходящихся к решению задачи $A\varphi = f$. При этом подпространство приближенных решений Φ совпадает со всем гильбертовым (евклидовым) пространством F. В этом, кстати сказать, состоит одна из приятных особенностей задач с положительными матрицами. Итерационный процесс (9.1.6) при соответствующем выборе параметра τ будет сходиться, а оптимизация процесса может быть произведена, например, с учетом априорной информации (9.1.7) выбором числа шагов f_0 итерационного процесса (см. § 4.6).

Предположим теперь, что в уравнении $A\varphi = f$ оператор A симметричен и его спектр имеет как положительную, так и отрицательную части.

Анализ показывает, что итерационный процесс (9.1.6) в указанных предположениях будет расходиться. В самом деле, пусть

$$\varphi = \sum_n \varphi_n u_n, \quad f = \sum_n f_n u_n \tag{9.1.8}$$

и $\{u_n\}$ — полная ортонормированная система собственных функций оператора A. Подставляя (9.1.8) в (9.1.6) и скалярно умножая результат на u_n, приходим к рекуррентным соотношениям для коэффициентов Фурье

$$\varphi_n^{j+1} = \varphi_n^j - \tau(\lambda_n \varphi_n^j - f_n), \quad \varphi_n^0 = 0,$$

или для невязки $\xi^j = A\varphi^j - f$,

$$\xi_n^{j+1} = (1 - \tau\lambda_n)\xi_n^j,\ \xi_n^0 = -f_n. \tag{9.1.9}$$

Решая уравнение (9.1.9), получаем

$$\xi_n^j = -(1 - \tau\lambda_n)^j f_n.$$

Следовательно,

$$\xi^j = -\sum_n (1 - \tau\lambda_n)^j f_n u_n.$$

Естественно, что итерационный процесс (9.1.6) будет сходиться только в том случае, если

$$\lim_{j\to\infty} \xi^j = 0.$$

Если оператор A своими собственными числами имеет только положительные числа из промежутка

$$\alpha(A) \leqslant \lambda_n(A) \leqslant \beta(A),$$

то выбором τ

$$0 < \tau < 2/\beta \tag{9.1.10}$$

процесс можно сделать сходящимся.

Однако в рассматриваемом случае симметричная матрица A имеет как положительные, так и отрицательные собственные числа. Пусть τ выбрано из интервала (9.1.10). Тогда все гармоники невязки, соответствующие положительным λ, будут от итерации к итерации подавляться со скоростью T_n^j, где $T_n = (1 - \tau\lambda_n)^j < 1$ и j — показатель степени. Что касается гармоник, соответствующих отрицательным собственным числам, то, так как такие компоненты невязки будут расти, для них

$$T_n^j = (1 - \tau\lambda_n)^j > 1.$$

Это приводит к расходимости итерационного процесса.

Таким образом, итерационный процесс (9.1.6) с последовательностью пробных функций φ^j, принадлежащих всему гильбертовому пространству, расходится.

Примером процесса, который бы не выводил последовательность приближенных решений из Φ, является двухшаговый метод минимальных невязок (см. 4.3.2)

$$\varphi^{j+1} = \varphi^j - \tau_j(A\varphi^j - f) - \gamma_j A^*(A\varphi^j - f).$$

Сделаем несколько замечаний о практическом подходе к численному решению условно корректных задач. Такие задачи обычно сводятся к системам линейных уравнений с плохо обусловленными матрицами общей структуры. Как правило, они решаются с помощью многошагово метода минимальных невязок (см. 4.3.2), который обеспечивает быструю сходимость итерационного процесса. Возможно также применение метода сопряженных градиентов после симметризации уравнений с помощью трансформации Гаусса. Этот второй метод мы рассмотрим ниже в связи с решением обратных эволюционных задач (см. § 9.3). При решении задачи итерационными методами процесс следует оборвать на шаге, где норма невязки при-

ближенно окажется равной априорной погрешности входных данных, т. е. $\|\xi^j\| \approx \varepsilon_1 + \varepsilon_2$. В этом случае, как отмечено в § 4.6, мы приходим к максимально достижимой точности решения при заданных априорных погрешностях.

Приведенные выше итерационные и вариационные способы устойчивого решения (9.1.1) являются частным случаем так называемых регуляризирующих алгоритмов. Дадим общее определение.

Семейство линейных операторов R_α в пространстве F, зависящее от числового параметра α ($0 < \alpha \leqslant \alpha_0$), называется регуляризирующим семейством (алгоритмов) для уравнения (9.1.1) на множестве M_R, если выполняются следующие условия:

$$\|R_\alpha\| < \infty \tag{9.1.11}$$

для любого $\alpha \in (0, \alpha_0]$ и

$$\|R_\alpha A\varphi - \varphi\| \to 0 \tag{9.1.12}$$

при $\alpha \to 0$ и всех $\varphi \in M_R$. При этом множество M_R, на котором имеет место соотношение (9.1.12), может быть шире множества корректности M задачи (9.1.1). В случае же, когда стремление к нулю в соотношении (9.1.12) имеет место равномерно по $\varphi \in M_R$, эти множества обычно совпадают.

Покажем, как с помощью регуляризирующего семейства R_α можно устойчивым образом находить приближенное решение уравнения (9.1.1) при условии, что точное решение $\varphi \in M_R$, а вместо точной правой части f известно ее ε — приближение f_ε: $\|f - f_\varepsilon\| \leqslant \varepsilon$.

Положим $\varphi_{\alpha\varepsilon} = R_\alpha f_\varepsilon$ и оценим $\|\varphi - \varphi_{\alpha\varepsilon}\|$. Из неравенства треугольника имеем

$$\|\varphi - \varphi_{\alpha\varepsilon}\| = \|\varphi - R_\alpha f + R_\alpha(f - f_\epsilon)\| \leqslant$$
$$\leqslant \|\varphi - R_\alpha f\| + \|R_\alpha\|\,\|f - f_\varepsilon\| \leqslant \|\varphi - R_\alpha f\| + \|R_\alpha\|\varepsilon. \tag{9.1.13}$$

Выберем $\alpha(\varepsilon)$ так, чтобы $\|R_\alpha\|\varepsilon \to 0$ и $\alpha(\varepsilon) \to 0$ при $\varepsilon \to 0$.

Тогда первое слагаемое в (9.1.13) стремится к нулю при $\varepsilon \to 0$ в силу условия (9.1.12), а второе — по построению функции $\alpha(\varepsilon)$. В итоге $\varphi_{\alpha(\varepsilon)\varepsilon\to\varphi}$ при $\varepsilon \to 0$.

Таким образом, регуляризирующее семейство R_α дает принципиальную возможность устойчивого решения условно- корректной задачи (9.1.1) с приближенно заданной правой частью. Параметр α носит название параметра регуляризации. В итерационном методе (9.1.6) роль параметра регуляризации играет номер итерации j.

Для одной и той же задачи (9.1.1) может быть обычно найдено бесконечно много регуляризирующих алгоритмов. При практических вычислениях естественно учитывать такие их характеристики, как простота орга-

низации вычислительного процесса, учет особенностей исходной задачи, оптимальность в том или ином смысле и т. д.

Проиллюстрируем введенные выше понятия на простейшем примере интегрального уравнения первого рода — уравнения Вольтерра:

$$A\varphi(t) \equiv \int_0^t A(t,\tau)\varphi(\tau)\,d\tau = f(t), \quad t \in [0,T]. \tag{9.1.14}$$

К решению уравнения (9.1.14) сводятся многие задачи интерполяции показаний физических приборов. Рассмотрим оператор A, определенный формулой (9.1.14), в пространстве $C[0,T]$ непрерывных функций φ с нормой

$$\|\varphi\| \equiv \|\varphi\|_T = \max_{t\in[0,T]} |\varphi(t)|.$$

Ядро $A(t,\tau)$ предполагается непрерывно дифференцируемым по t, непрерывным по τ и отличным от нуля на диагонали $t = \tau$. Для простоты анализа будем считать, что

$$A(t,t) \equiv 1. \tag{9.1.15}$$

При этих условиях уравнение (9.1.14) дифференцированием по t приводится к уравнению Вольтера второго рода

$$\varphi(t) + \int_0^t D_t A(t,\tau)\varphi(\tau)\,d\tau = f'(t), \tag{9.1.16}$$

которое, как известно, может быть решено методом последовательных приближений. Однако если вместо f известно лишь его ε-приближение в норме $C[0,T]$: $\|f - f_\varepsilon\| \leqslant \varepsilon$, то задача дифференцирования становится некорректной. Покажем, что семейство операторов R_α, ставящих в соответствие функции $f(t)$ решение φ_α уравнения

$$\varphi_\alpha(t) + \int_0^t \Delta_\alpha A(t,\tau)\varphi_\alpha(\tau)\,d\tau = \Delta_\alpha f(t), \quad t \in [0,T_0], \tag{9.1.17}$$

$$T_0 = T - \alpha_0, \quad \alpha \in (0,\alpha_0],$$

где

$$\Delta_\alpha f(t) = \frac{f(t+\alpha) - f(t)}{\alpha}, \alpha_0 < T,$$

будет регуляризирующим на интервале $[0,T_0]$ для всех непрерывно дифференцируемых решений φ уравнения (9.1.14). В самом деле, решение φ_α уравнения Вольтерра второго рода (9.1.17) существут, единственно и удовлетво-

ряет оценке

$$\|\varphi_\alpha\|_{T_0} \leqslant e^{KT_0}\|\Delta_\alpha f\|_{T_0},$$

где

$$K = \max_{0\leqslant\tau\leqslant t\leqslant T_0} |\Delta_\alpha A(t,\tau)|,$$

а так как

$$\|\Delta_\alpha f\|_{T_0} \leqslant \frac{2}{\alpha}\|f\|_T,$$

то

$$\|\varphi_\alpha\|_{T_0} \equiv \|R_\alpha f\|_{T_0} \leqslant \frac{2e^{KT_0}}{\alpha}\|f\|_T < \infty \tag{9.1.18}$$

для $\alpha > 0$, а, следовательно, условие (9.1.11) выполнено. Для проверки условия (9.1.12) достаточно показать, что

$$\|\varphi_\alpha - \varphi\|_{T_0} \to 0 \text{ при } \alpha \to 0.$$

Применив к уравнению (9.1.14) оператор Δ_α с учетом условия (9.1.15), легко получить равенство

$$\varphi(t) + \int_0^t \Delta_\alpha A(t,\tau)\varphi(\tau)\,d\tau = \Delta_\alpha f(t) - g_\alpha(t), \tag{9.1.19}$$

где

$$g_\alpha(t) = \frac{1}{\alpha}\int_t^{t+\alpha} (A(t+\alpha,\tau) - A(\tau,\tau))\,\varphi(\tau)\,d\tau + \frac{1}{\alpha}\int_t^{t+\alpha} (\varphi(\tau) - \varphi(t))\,d\tau. \tag{9.1.20}$$

Вычитая равенство (9.1.19) из (9.1.17), приходим к уравнению для невязки $u_\alpha = \varphi_\alpha - \varphi$:

$$u_\alpha + \int_0^t \Delta_\alpha A(t,\tau)u_\alpha(\tau)\,d\tau = g_\alpha(t), \quad t \in [0,T_0]. \tag{9.1.21}$$

Из (9.1.21), (9.1.20) следует, что

$$\|\varphi - \varphi_\alpha\|_{T_0} \equiv \|u_\alpha\|_{T_0} \leqslant e^{KT_0}\|g_\alpha\|_{T_0}, \tag{9.1.22}$$

$$\|g_\alpha\|_{T_0} \leqslant \frac{1}{\alpha}K\alpha\|\varphi\|_T\alpha + \frac{1}{\alpha}\|\varphi'\|_T\alpha^2 = \alpha\left(K\|\varphi\|_T + \|\varphi'\|_T\right) \to 0 \tag{9.1.23}$$

при $\alpha \to 0$. Таким образом, мы показали, что операторы $\{R_\alpha\}$ образуют регуляризирующее семейство для уравнения (9.1.14) на множестве непрерывно дифференцируемых функций φ. Важной особенностью этого регуляризиру-

ющего алгоритма является то, что он в отличие от общих вариационных алгоритмов сохраняет свойство вольтерровости исходного уравнения.

Определим функционал l формулой

$$l(\varphi) \equiv (K\|\varphi\|_T + \|\varphi'\|_T)e^{KT_0}.$$

Так как $\|\varphi\|_{T_0} \leqslant \|\varphi_\alpha\|_{T_0} + \|\varphi_\alpha - \varphi\|_{T_0}$, то из оценок (9.1.18), (9.1.22), (9.1.23) следует, что

$$\|\varphi\|_{T_0} \leqslant \alpha l(\varphi) + \frac{2e^{KT}}{\alpha}\|f\|_T, \quad 0 < \alpha \leqslant \alpha_0.$$

Таким образом, согласно (9.1.4) задача является условно- корректной на множестве $M = \{\varphi : l(\varphi) \leqslant m\}$.

9.2. Решение обратных эволюционных задач с постоянными коэффициентами

В этом параграфе мы рассмотрим два устойчивых метода решения обратной эволюционной задачи

$$\frac{d\varphi}{dt} - A\varphi = 0,\ \varphi(0) = g,\ A \geqslant 0.$$

Первый из них основан на методе Фурье и сводится фактически к решению некоторой спектральной задачи, а во втором исходная некорректная задача редуцируется к решению последовательности корректных (прямых) эволюционных задач

$$\frac{d\varphi}{dt} + A\varphi = 0,\ \varphi(0) = g,\ A \geqslant 0.$$

9.2.1. Метод Фурье

Пусть A — положительно определенная матрица, не зависящая от времени, имеющая вещественный спектр в промежутке $\alpha(A) \leqslant \lambda \leqslant \beta(A)$, а вектор-функция φ — решение следующей задачи Коши:

$$\frac{d\varphi}{dt} - A\varphi = 0, \quad 0 \leqslant t \leqslant t_0, \tag{9.2.1}$$

$$\varphi = g \text{ при } = 0,$$

где g — заданное значение вектора в начальный момент времени.

Рассмотрим две спектральные задачи:

$$Au = \lambda u, \quad A^* u^* = \lambda u^*. \tag{9.2.2}$$

Предположим, что они определяют два биортогональных базиса собственных функций $\{u_n\}$ и $\{u_n^*\}$. Функции φ и g представимы в виде сумм Фурье:

$$\varphi = \sum_n \varphi_n u_n, \quad g = \sum_n g_n u_n. \tag{9.2.3}$$

Подставим эти суммы в (10.2.1) и результат скалярно умножим на u_n. Получим систему обыкновенных дифференциальных уравнений для коэффициентов Фурье:

$$\begin{gathered} \frac{d\varphi_n}{dt} - \lambda_n \varphi_n = 0, \\ \varphi_n = g_n \text{ при } t = 0, \\ n = 1, 2, \ldots, N. \end{gathered} \tag{9.2.4}$$

Решение каждого уравнения (9.2.4) имеет вид

$$\varphi_n = g_n e^{\lambda_n t}, \quad n = 1, 2, \ldots, N, \tag{9.2.5}$$

и, следовательно, решение задачи (9.2.1) представимо суммой

$$\varphi(t) = \sum_{n=1}^{N} g_n e^{\lambda_n t} u_n. \tag{9.2.6}$$

Итак, мы установили, что решение задачи (9.2.1) представлено в виде суммы Фурье, каждый член которой по времени экспоненциально растет в зависимости от величины n-го собственного числа λ_n.

Предположим, что нас интересует физически определенное решение этой задачи в интервале времени $0 \leqslant t \leqslant t_0$. Рассмотрим задачу, аналогичную (9.2.1), но уже корректную:

$$\begin{gathered} \frac{d\varphi}{dt} - A\varphi = 0, \quad 0 \leqslant t \leqslant t_0, \\ \varphi = h \text{ при } t = t_0. \end{gathered} \tag{9.2.7}$$

Поступая аналогично предыдущему, получим

$$\varphi = \sum_{n=1}^{N} h_n e^{-\lambda_n (t - t_0)} u_n. \tag{9.2.8}$$

Потребуем, чтобы решение (9.2.8) при $t = 0$ совпадало с вектором g из задачи (9.2.1). Отсюда получаем связь между коэффициентами Фурье функции h и функции g:

$$g_n = h_n e^{-\lambda_n t_0}. \tag{9.2.9}$$

Таким образом, функция g восстанавливается с помощью функции h вполне просто:

$$g = \sum_{n=1}^{N} h_n e^{-\lambda_n t_0} u_n. \tag{9.2.10}$$

Более того, малым ошибкам в h (или h_n) будут соответствовать малые ошибки в функции g. Однако наша задача обратна к рассмотренной. Мы располагаем информацией о функции g, а нам требуется восстановить функцию h по формуле

$$h = \sum_{n=1}^{N} g_n e^{\lambda_n t_0} u_n. \tag{9.2.11}$$

Если бы мы располагали точной информацией о функции g и имели возможность вести расчет с бесконечным числом значащих цифр, то восстановление функции h по формуле (9.2.11) не представляло бы труда. В данной ситуации, однако, функцию g мы знаем с определенной погрешностью, которая априори считается известной, и расчет проводится на ЭВМ с ограниченным числом знаков (словом), поэтому в процессе вычислений появляются ошибки округления. Эти два обстоятельства делают задачу вычисления h по формуле (9.2.11) уже не такой простой.

Прежде всего предположим, что исследователю, пытающемуся произвести обработку экспериментальных данных на основе решения обратной эволюционной задачи (9.2.1), заранее известна система собственных функций u_n и он имеет возможность в результате разложения исходных данных (функций g) по этой системе выделить полезную информацию и с достаточной точностью оценить погрешность в каждой компоненте Фурье — g_n.

Если задача носит статистический характер и допускает многократное повторение, то на основе хорошо разработанных методов корреляционного анализа в этом случае удается существенно повысить точность данных в g_n, даже если в единичном измерении погрешность значительно превышает полезную информацию. Во всяком случае, предварительная обработка материалов наблюдения позволяет сделать заключения о величине систематической (или случайной, если речь идет о единичном измерении) погрешности в g_n. Поэтому для любого n будем иметь

$$g_n = \bar{g}_n(1 + \delta_n),$$

где $\bar{g}_n$ — точное значение (априори нам неизвестное!), а δ_n — относительная погрешность, которую будем считать известной.

Обычно погрешность δ_n оказывается минимальной для наиболее длинных волн возмущений и быстро растет в направлении высоких гармоник, как правило, описывающих мелкомасштабные особенности решения. Поэтому, начиная с некоторого номера, коэффициенты g_n в основном описывают погрешность во входных данных. Из формулы (9.2.11) следует, что именно самые высокочастотные компоненты имеют наибольший экспоненциальный вес. Следовательно, если мы не позаботимся заранее о том, чтобы исключить из рассмотрения эти паразитические гармоники, то в итоге можем получить заведомо неверный результат, так как для таких гармоник g_n практически не содержит полезной информации, но, будучи умноженными на большие коэффициенты $e^{\lambda_n t_0}$, они могут внести крупный вклад в h и тем самым исказить, иногда непоправимо, решение задачи. Таким образом, первая и основная задача состоит в определении информативности коэффициентов g_n.

Предположим, что на основе априорной информации установлено, что n_0 первых коэффициентов g_n имеют относительную погрешность меньше η, т. е. $\delta_n < \eta$, где η — максимально допустимая погрешность. Тогда алгоритм восстановления функции (9.2.11) оказывается аналогичным уже рассмотренному при построении элементов подпространства Φ в задаче (9.2.1). Нам просто нужно исключить из ряда (9.2.11) те гармоники, которые являются паразитическими. В результате будем иметь

$$h = \sum_{n=1}^{n_0} g_n e^{\lambda_n t_0} u_n. \tag{9.2.12}$$

За основу алгоритма решения частной спектральной задачи примем итерационный метод, сформулированный в § 1.1. Если необходимо построить набор первых (наиболее крупномасштабных) собственных функций u_n или u_n^* и соответствующих им собственных чисел, то для этой цели можно воспользоваться алгоритмом ортогонализации, описанным в § 1.1.

9.2.2. Редукция к решению прямой задачи[4)]

Рассмотрим опять задачу Коши

$$\frac{d\varphi}{dt} - A\varphi = 0,\ \varphi(0) = g,\ 0 \leqslant t \leqslant t_0, \tag{9.2.13}$$

[4)] С. П. Шишатский [16].

где A — самосопряженный неограниченный положительный оператор, действующий в гильбертовом пространстве F. Другими словами, в отличие от предыдущего пункта мы не делаем предварительно конечно-разностную аппроксимацию по пространственным переменным. В этом случае задача (9.2.13) классически некорректна. Мы считаем, что решение φ задачи (9.2.13) существует и принадлежит множеству $M = \{\varphi(t) : \|\varphi(t)\| \leqslant m, t \in [0, t_0]\}$, однако вместо точного начального условия g нам задано его приближение g_ε:

$$\|g - g_\varepsilon\| \leqslant \varepsilon. \tag{9.2.14}$$

Известно, что[5)]

$$\|\varphi(t)\| \leqslant \|\varphi(0)\|^{1-\frac{t}{t_0}} \|\varphi(t_0)\|^{\frac{t}{t_0}}, \tag{9.2.15}$$

поэтому для любого фиксированного $t \in (0, t_0)$ задача определения $\varphi(t)$ условно корректна на множестве M, так как в силу (9.2.15)

$$\|\varphi(t)\| \leqslant m^{\frac{t}{t_0}} \|g\|^{1-\frac{t}{t_0}}, \quad \varphi \in M. \tag{9.2.16}$$

Покажем, что операторы R_α, определенные формулой

$$R_\alpha = (e^{-At_0} + \alpha E)^{-t/t_0}, \quad \alpha > 0, \tag{9.2.17}$$

образуют регуляризирующее семейство на множестве корректности M. В самом деле, так как оператор R_α согласно (9.2.17) есть функция сопряженного положительного оператора A, то используя спектральное разложение, получаем

$$\|R_\alpha\| \leqslant \max_{\lambda \geqslant 0} (e^{-\lambda t_0} + \alpha)^{t/t_0} = \alpha^{t/t_0} < \infty. \tag{9.2.18}$$

Для проверки условия (9.1.12) (см. § 7.1; роль оператора A в нем играет оператор e^{-AT} и $f \equiv g$, так как $e^{-At}\varphi(t) = g$) достаточно показать, что

$$|R_\alpha e^{-At}\varphi(t) - \varphi(t)| \to 0 \text{ при } \alpha \to 0,\ \varphi \in M. \tag{9.2.19}$$

Снова, используя спектральное разложение оператора A, имеем

$$\|\varphi(t) - R_\alpha e^{-At}\varphi(t)\| = \|e^{At}g - R_\alpha g\| \leqslant \max_{\lambda \geqslant 0} e^{\lambda(t-t_0)}[1 - (1 + \alpha e^{\lambda t_0})^{-t/t_0}]\|e^{At_0}g\|. \tag{9.2.20}$$

Для оценки этой величины воспользуемся неравенством

$$(1+x)^\tau - 1 \leqslant \tau x (1+x)^\tau (1+\tau x)^{-1},$$

[5)] С. Г. Крейн [16].

которое справедливо для $x \geqslant 0$ и $\tau \in [0,1]$. Учитывая также, что $\varphi \in M$, т. е. $\|\varphi(t_0)\| = \|e^{At_0}g\| \leqslant m$, получаем

$$\max_{\lambda \geqslant 0} e^{\lambda(t-t_0)}[1-(1+\alpha e^{\lambda t_0})^{-\frac{t}{t_0}}]\|e^{At_0}g\| \leqslant \alpha^{1-\frac{t}{t_0}} \max_{\alpha \leqslant x < \infty} x^{\frac{t}{t_0}-1}[1-(1+x)^{-\frac{t}{t_0}}]m \leqslant$$
$$\leqslant \alpha^{1-\frac{t}{t_0}} \max_{\alpha \leqslant x < \infty} \frac{t}{t_0} x^{\frac{t}{t_0}} \left(1+\frac{t}{t_0}x\right)^{-1} m = \frac{t}{t_0}\left(1-\frac{t}{t_0}\right)^{1-\frac{t}{t_0}} \alpha^{1-\frac{t}{t_0}} m.$$

Итак,

$$\|\varphi(t) - R_\alpha g\| \leqslant \frac{t}{t_0}\left(1-\frac{t}{t_0}\right)^{1-\frac{t}{t_0}} \alpha^{1-\frac{t}{t_0}} m \to 0 \tag{9.2.21}$$

при $\alpha \to 0$, $t < t_0$. Тем самым доказано, что $\{R_\alpha\}$ — регуляризирующее семейство. Используя оценки (9.2.14), (9.2.18), (9.2.21) и неравенство треугольника, имеем

$$\|\varphi(t) - R_\alpha g_\varepsilon\| \leqslant \|\varphi(t) - R_\alpha g\| + \|R_\alpha\| \cdot \|g - g_\varepsilon\| \leqslant$$
$$\leqslant \frac{t}{t_0}\left(1-\frac{t}{t_0}\right)^{1-\frac{t}{t_0}} \alpha^{1-\frac{t}{t_0}} m + \alpha^{-\frac{t}{t_0}}\varepsilon. \tag{9.2.22}$$

Элементарные выкладки показывают, что наименьшее значение правой части этого неравенства достигается при

$$\alpha_0 = \left(1-\frac{t}{t_0}\right)^{-2+\frac{t}{t_0}} \frac{\varepsilon}{m}.$$

Так как для $t \in (0, t_0)$

$$\left(1-\frac{t}{t_0}\right)^{-\left(1-\frac{t}{t_0}\right)^2} \leqslant e^{1/2e},$$

то

$$\|\varphi(t) - R_{\alpha_0} g\| \leqslant e^{1/2e} m^{\frac{t}{t_0}} \varepsilon^{1-\frac{t}{t_0}}. \tag{9.2.23}$$

Оценка (9.2.23) уклонения приближенного решения задачи (9.2.13), построенного с помощью оператора R_{α_0}, от точного решения $\varphi(t)$ отличается от априорной оценки (9.2.16) устойчивости на M лишь множителем $e^{1/2e} \approx 1,21$. В этом смысле предлагаемый метод решения задачи (9.2.13) оптимален.

Для практического вычисления элемента

$$R_{\alpha_0} g_\varepsilon \equiv (e^{-At_0} + \alpha_0)^{-t/t_0} g_\varepsilon$$

уместно положить

$$R_\alpha g_\varepsilon \approx Q_n(e^{-At_0}) g_\varepsilon,$$

где $Q_n(x)$ — многочлен наилучшего приближения степени n к функции $(x+\alpha_0)^{-t/t_0}$ на отрезке $0 \leqslant x \leqslant 1$. Оператор $Q_n(e^{-At_0})$ есть многочлен по e^{-At_0}, т. е. для вычисления элемента

$$Q_n(e^{-At_0})g_\varepsilon$$

достаточно уметь вычислять элементы $e^{-kAt_0}g_\varepsilon$ при $k = 1, 2, \ldots, n$. Но $e^{kAt_0}g_\varepsilon$ есть решение корректной задачи Коши

$$\frac{d\psi}{dt} + A\psi = 0, \quad \psi(0) = g_\varepsilon, \quad t \geqslant 0, \tag{9.2.24}$$

при $t = kt_0$. Таким образом, мы свели некорректную задачу (9.2.13) к последовательности корректных задач (9.2.24), эффективные методы решения которых были рассмотрены в главе 5. Можно показать, что окончательная погрешность $\|\varphi(t) - Q_n(e^{-At_0})g_\varepsilon\|$ метода оценивается сверху величиной

$$\left(1 - \frac{t}{t_0}\right)^{-\left(1-\frac{t}{t_0}\right)^2} \varepsilon^{1-\frac{t}{t_0}}\, m^{\frac{t}{t_0}} + \frac{2^{\frac{t}{t_0}+1}}{\Gamma\left(\frac{t}{t_0}\right)} (n+1)^{\frac{t}{t_0}-1} \frac{\beta^{n+1+\frac{t}{t_0}}}{(1-\beta^2)^{1+\frac{t}{t_0}}} \|g_\varepsilon\|, \tag{9.2.25}$$

где $\beta = 1 + 2\alpha_0 - 2\sqrt{\alpha_0 + \alpha_0^2}$, Γ — гамма-функция. Важно подчеркнуть, что в практически интересных случаях степень многочленов Q_n невелика. Так из (9.2.25) следует, что, например, при

$$\frac{\varepsilon}{\|g_\varepsilon\|} = 0,1; \quad \frac{\|g_\varepsilon\|}{m} = 0,1$$

для получения точности $2\varepsilon^{1-\frac{t}{t_0}}\, m^{\frac{t}{t_0}}$ достаточно воспользоваться многочленом Q_2, а при

$$\frac{\varepsilon}{\|g_\varepsilon\|} = 0,05; \quad \frac{\|g_\varepsilon\|}{m} = 0,1$$

— многочленом Q_4. Для вычисления коэффициентов многочлена $Q_n(x)$ имеются явные формулы, которые зависят только от отношений ε/m, t/t_0 и не зависят от оператора A и входных данных g_ε.

9.3. Обратная эволюционная задача с оператором, зависящим от времени

Рассмотрим эволюционную задачу

$$\frac{d\varphi}{dt} - A(t)\varphi = 0, \quad 0 \leqslant t \leqslant t_0, \tag{9.3.1}$$

$$\varphi = g \text{ при } t = 0$$

с оператором $A > 0$, зависящим от времени. Как и раньше, предполагается, что задача (9.3.1) является результатом редукции задачи математической физики по пространственным переменным к системе обыкновенных дифференциальных уравнений. В этом случае метод Фурье уже неприменим, и для решения задачи (9.3.1) необходимо использовать численные методы.

Переходим к обсуждению одного из возможных алгоритмов численного решения. Задаче (9.3.1) поставим в соответствие модельную задачу, в известном смысле близкую:

$$\frac{d\bar{\varphi}}{dt} - \bar{A}\bar{\varphi} = 0, \quad 0 \leqslant t \leqslant t_0, \tag{9.3.2}$$

$$\bar{\varphi} = g \text{ при } t = 0,$$

где $\bar{A} > 0$ — оператор, не зависящий от времени, имеющий положительный спектр

$$\alpha(\bar{A}) \leqslant \lambda(\bar{A}) \leqslant \beta(\bar{A}),$$

в некотором смысле близкий к оператору $A(t)$. Ради определенности будем полагать, что

$$A(t) = \bar{A} + \delta A(t), \tag{9.3.3}$$

где

$$\|\delta A(t)\| \ll \|\bar{A}\| \tag{9.3.4}$$

для любых t из интервала $0 \leqslant t \leqslant t_0$.

Задача (9.3.26) в дальнейшем позволит нам получить необходимую априорную информацию для организации вычислительного процесса решения основной задачи (9.3.1).

Методами, изложенными в § 1.1 и § 9.2, определим m информативных (с точки зрения ошибок во входных данных) собственных элементов u_n, u_n^* и собственных чисел λ_n ($n = 1, 2, \dots m$). Остальные гармоники ряда Фурье для g_n ($n = m+1, m+2, \dots N$) должны быть отброшены, так как ошибки при определении этих коэффициентов превышают (иногда весьма значительно) полезную информацию. Тогда получим

$$\bar{g} = \sum_{n=1}^{m} g_n u_n, \tag{9.3.5}$$

где

$$g_n = (g, u_n^*).$$

В результате решение модельной задачи $\bar{\varphi}$ на промежутке $0 \leqslant t \leqslant t_0$ может быть представлено в виде

$$\bar{\varphi}(t) = \sum_{n=1}^{m} g_n e^{\lambda_n t} u_n. \tag{9.3.6}$$

Попытаемся решить модельную задачу (9.3.26) численно. С этой целью рассмотрим, например, разностную схему второго порядка точности относительно $\Delta t = \tau$:

$$\frac{\bar{\varphi}^{j+1} - \bar{\varphi}^j}{\tau} - \bar{A}\,\frac{\bar{\varphi}^{j+1} + \bar{\varphi}^j}{2} = 0, \quad j = 1, 2, \ldots, j_0, \tag{9.3.7}$$

$$\bar{\varphi}^0 = g.$$

Решение задачи (9.3.7) будем искать с помощью метода Фурье, предположив, что мы располагаем всем набором собственных элементов u_n и u_n^*. Такое предположение делается только с целью теореитического анализа и получения некоторой априорной информации о поведении решения. Тогда будем иметь

$$\bar{\varphi}^j = \sum_{n=1}^{N} \bar{\varphi}_n^j u_n. \tag{9.3.8}$$

В результате для коэффициентов Фурье с помощью (9.3.7) приходим к рекуррентным соотношениям

$$\bar{\varphi}_n^{j+1} = \frac{1 + \dfrac{\tau\lambda_n}{2}}{1 - \dfrac{\tau\lambda_n}{2}} \bar{\varphi}_n^j, \quad j = 1, 2, \ldots, j_0, \tag{9.3.9}$$

$$\bar{\varphi}_n^0 = g_n.$$

Следовательно,

$$\bar{\varphi}_n^j = \left[\frac{1 + \dfrac{\tau\lambda_n}{2}}{1 - \dfrac{\tau\lambda_n}{2}}\right]^j g_n. \tag{9.3.10}$$

Таким образом, имеем

$$\bar{\varphi}^j = \sum_{n=1}^{N} T_n^j g_n u_n, \tag{9.3.11}$$

где

$$T_n = \frac{1 + \dfrac{\tau\lambda_n}{2}}{1 - \dfrac{\tau\lambda_n}{2}}.$$

Предположим, что шаг τ выбран из условия, чтобы знаменатель в выражении для T_n не обращался в нуль ни для одного значения n. Например,

$$\tau < \frac{2}{\beta(\bar{A})}. \tag{9.3.12}$$

Заметим, что это условие согласовано с условием аппроксимации крупномасштабных возмущений.

Формальный анализ решения модельной задачи в виде (9.3.11) показывает, что все $T_n > 1$ высокочастотные гармоники, соответствующие большим номерам n, имеют быстро растущие с номером амплитуды. Следовательно, для них $T_n \gg 1$ и тем более $T_n^j \gg 1$. Поскольку при обработке входных данных g мы отбросили все гармоники ряда Фурье (9.3.5), начиная с $n = m + 1$, то на первый взгляд кажется, что этого достаточно для того, чтобы сумма Фурье

$$\bar{g} = \sum_{n=1}^{m} g_n u_n$$

порождала решение с таким же числом членов

$$\bar{\varphi}^j = \sum_{n=1}^{m} T_n^j g_n u_n. \tag{9.3.13}$$

Такое положение было бы в действительности, если бы наши ЭВМ позволяли вести расчет с бесконечным числом значащих цифр. Однако из-за ограниченности машинного слова в процессе вычисления вследствие ошибок округления сразу же появятся компоненты g_n для $n > m$. И хотя они малы, но имеют большой «вес» в решении, пропорциональный $T_n^j \gg 1$. Эти ошибки в конце концов могут существенно исказить основное решение задачи. Чтобы избежать катастрофического роста ошибок высокочастотных компонент ряда Фурье, необходимо найти такую конструкцию, которая автоматически переводила бы любой элемент векторного пространства F в элемент некоторого подпространства Φ.

Определим Φ следующим образом: будем считать, что элемент принадлежит подпространству Φ, если амплитуды последних $N - m$ гармоник суммы Фурье этого элемента по системе функции u_n в процессе численного решения задачи возрастают от шага к шагу не быстрее, чем несколько амплитуд последней информативной гармоники с номером m. При конструкции такого подпространства ошибки округления на его элементах будут возрастать не быстрее амплитуды m-й гармоники. Это обеспечит корректность вычислительной схемы. Ряд авторов предложили вместо оператора $\bar{A}$ в модельной задаче (9.3.26) рассматривать оператор $\bar{A}_\varepsilon = A - \varepsilon \bar{A}^2$. В этом

случае вместо задачи (9.3.26) будем иметь

$$\frac{d\bar{\varphi}_\varepsilon}{dt} - \bar{A}\bar{\varphi}_\varepsilon = -\varepsilon\bar{A}^2\bar{\varphi}_\varepsilon, \quad 0 \leqslant t \leqslant t_0, \tag{9.3.14}$$

$$\bar{\varphi}_\varepsilon = g \text{ при } t = 0,$$

где ε — пока произвольный параметр. Этот параметр выберем из условия, чтобы решение задачи не выходило из множества Φ. Ради простоты анализа предположим, что $\bar{A} = \bar{A}^*$. Рассмотрим разностную схему

$$\frac{\bar{\varphi}_\varepsilon^{j+1} - \bar{\varphi}_\varepsilon^{j}}{\tau} - (\bar{A} - \varepsilon\bar{A}^2)\,\frac{\bar{\varphi}_\varepsilon^{j+1} + \bar{\varphi}_\varepsilon^{j}}{2} = 0, \quad \bar{\varphi}_\varepsilon^0 = g. \tag{9.3.15}$$

Решение задачи (9.3.15) будем искать с помощью метода Фурье по собственным функциям оператора A. Тогда получим

$$\bar{\varphi}_\varepsilon^j = \sum_{n=1}^{N} \left[\frac{1 + \dfrac{\tau\lambda_n}{2} - \varepsilon\dfrac{\tau\lambda_n^2}{2}}{1 - \dfrac{\tau\lambda_n}{2} + \varepsilon\dfrac{\tau\lambda_n^2}{2}}\right]^j g_n u_n. \tag{9.3.16}$$

Параметр ε выберем из условия, чтобы относительная ошибка в гармонике с номером m за счет введения оператора $\varepsilon\bar{A}^2$ не превышала η (обычно в качестве η можно брать $\eta < 1$ в зависимости от того, каково соотношение в гармонике $n = m$ между полезной информацией и неучитываемыми погрешностями («шумом»)). Из этого условия получаем соотношение

$$\eta\frac{\tau\lambda_m}{2} = \varepsilon\frac{\tau\lambda_m^2}{2}, \tag{9.3.17}$$

откуда

$$\varepsilon = \frac{\eta}{\lambda_m}. \tag{9.3.18}$$

Таким образом, мы приходим к определению одной из важнейших априорных величин, необходимых для дальнейшего численного расчета. Легко видеть, что при заданном параметре ε из (9.3.18) амплитуды всех гармоник $n > m$ будут возрастать со временем не быстрее, чем T_m^j.

Наконец, нам понадобится еще одна априорная величина. Для ее нахождения рассмотрим

$$\bar{\varphi}^j = \sum_{n=1}^{m} g_n e^{\lambda_n t} u_n, \tag{9.3.19}$$

$$\bar{\varphi}_\varepsilon^j = \sum_{n=1}^{N} g_n T_n^j(\varepsilon) u_n, \tag{9.3.20}$$

где

$$T_n(\varepsilon) = \frac{1 + \dfrac{\tau\lambda_n}{2} - \varepsilon\dfrac{\tau\lambda_n^2}{2}}{1 - \dfrac{\tau\lambda_n}{2} + \varepsilon\dfrac{\tau\lambda_n^2}{2}}.$$

Поскольку решение $\bar{\varphi}_\varepsilon^j$ принадлежит Φ, то без большой погрешности его можно заменить на

$$\bar{\varphi}_\varepsilon^j = \sum_{n=1}^{m} g_n T_n^j(\varepsilon) u_n, \tag{9.3.21}$$

где мы ограничились только первыми m членами. Решение в виде (9.3.21) находится конструктивно с помощью уже полученной системы функций u_n и u_n^* ($n = 1, 2, \ldots, m$). Из выражений (9.3.19) и (9.3.21) найдем величины φ^j и $\bar{\varphi}_\varepsilon^j$ при $j = 1, 2 \ldots, j_0$. После этого введем в рассмотрение векторы

$$\bar{\varphi} = \begin{Vmatrix} \bar{\varphi}^1 \\ \bar{\varphi}^2 \\ \ldots \\ \bar{\varphi}^{j_0} \end{Vmatrix}, \quad \bar{\varphi}_\varepsilon = \begin{Vmatrix} \bar{\varphi}_\varepsilon^1 \\ \bar{\varphi}_\varepsilon^2 \\ \ldots \\ \bar{\varphi}_\varepsilon^{j_0} \end{Vmatrix}$$

и подсчитаем норму

$$\|\bar{\varphi} - \bar{\varphi}_\varepsilon\| = \delta. \tag{9.3.22}$$

Это и будет последняя из искомых априорных величин. Две другие — τ и ε — определены формулами (9.3.12) и (9.3.18).

Сформулируем численный алгоритм решения исходной задачи (9.3.1). С учетом изложенного выше анализа построим следующую аппроксимацию задачи:

$$\frac{\varphi^{j+1} - \varphi^j}{\tau} - (A_j - \varepsilon A_j^2)\,\frac{\varphi^{j+1} + \varphi^j}{2} = 0, \quad \varphi^0 = g, \tag{9.3.23}$$

где τ и ε выбираются на основе анализа априори изученной простой модели:

$$\tau < \frac{2}{\beta(\bar{A})},\ \varepsilon = \frac{\eta}{\lambda_m(\bar{A})}. \tag{9.3.24}$$

Введем в рассмотрение векторы

$$\varphi = \begin{Vmatrix} \varphi^1 \\ \varphi^2 \\ \ldots \\ \varphi^{j_0} \end{Vmatrix}, \quad f = \begin{Vmatrix} -R_0 g \\ 0 \\ \ldots \\ 0 \end{Vmatrix}$$

и матрицу

$$\Lambda = \left\| \begin{array}{ccccccc} -S_0 & 0 & 0 & 0 & \dots & 0 & 0 \\ R_1 & -S_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & R_2 & -S_2 & 0 & \dots & 0 & 0 \\ 0 & 0 & R_3 & -S_3 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & R_{j_0-1} & -S_{j_0-1} \end{array} \right\|,$$

где

$$S_j = E - \frac{\tau}{2}(A_j - \varepsilon A_j^2), \quad R_j = E + \frac{\tau}{2}(A_j - \varepsilon A_j^2),$$

$$A_j = A(t_{j+1/2}).$$

Тогда приходим к задаче

$$\Lambda\varphi = f. \tag{9.3.25}$$

Задачу (9.3.25) симметризуем, умножив на Λ^*:

$$\Lambda^*\Lambda\varphi = \Lambda^* f, \tag{9.3.26}$$

а затем сформулируем некоторый итерационный процесс. В частности, для этой цели используется метод сопряженных градиентов, не требующий априорного знания границ спектра $\Lambda^*\Lambda$.

Формулировкой метода последовательных приближений описание алгоритма не исчерпывается. Необходимо еще определить оптимальное число итераций k_0, которые приводят к максимально достижимой точности при заданных априорных условиях. Поскольку такое число может быть найдено не с очень большой точностью, будем полагать, что априорная оценка аппроксимации (9.3.22), полученная для модельной задачи, оказывается применимой и для задачи (9.3.1). Предположим поэтому, что

$$\|\varphi - \varphi_\varepsilon\| = \delta, \tag{9.3.27}$$

где φ — точное решение задачи (9.3.1) в узлах сетки, а φ_ε — решение разностной задачи с регуляризирующим оператором. Тогда используемый итерационный процесс естественно продолжать до тех пор, пока ошибка итерационного процесса оказывается большей, чем ошибка аппроксимации (9.3.27), и процесс следует закончить при равенстве этих ошибок. Алгоритмически это сделать наиболее просто следующим образом. Введем в рассмотрение вектор невязки ξ^k по формуле

$$\xi^k = \Lambda^*(\Lambda\varphi^k - f) = \Lambda^*\Lambda(\varphi^k - \varphi). \tag{9.3.28}$$

Тогда имеет место оценка

$$\|\xi^k\| \leqslant \|\Lambda^*\Lambda\| \, \|\varphi^k - \varphi\|. \tag{9.3.29}$$

Очевидно, величина $\|\varphi - \varphi_\varepsilon\|$ должна быть эквивалентна величине δ, что приводит к требованию

$$\|\xi^k\| \leqslant \delta \|\Lambda^*\Lambda\|. \tag{9.3.30}$$

Это означает, что вычислительный процесс следует продолжать до тех пор, пока норма невязки $\|\xi^k\|$ не будет сравнима с величиной в правой части (9.3.30). Таким образом, приходим к параметрической оценке для k_0:

$$\|\xi^{k_0}\| \leqslant \beta(\Lambda^*\Lambda)\delta. \tag{9.3.31}$$

Как видно, решение обратных эволюционных задач требует большой подготовительной работы по изучению различных простых моделей, которые позволяют получать необходимую априорную информацию для конструирования качественного вычислительного алгоритма. В отдельных случаях возникают и более сложные ситуации. Однако проведенное рассмотрение дает представление о некоторых принципах формирования численных методов на основе изучения возникающих погрешностей и анализа алгоритма с помощью простых моделей. Нами обсуждена только одна точка зрения на процесс регуляризации, но и она уже дает представление о возможных подходах к численному решению обратных задач.

В заключение следует отметить, что изложенные методы и идеи могут быть также применены к численному решению задачи Коши для уравнения эллиптического типа. Эти задачи в классическом смысле поставлены некорректно и для своего решения требуют привлечения методов, разработанных в теории условно корректных задач.

9.4. Постановка обратных задач на основе методов теории возмущений

Постановки некоторых обратных задач на основе теории сопряженных функций и методов теории возмущений начинают играть все большую роль в формировании вычислительных алгоритмов, особенно при решении сложных задач математической физики, в которых априори трудно оценить влияние тех или иных факторов на решение задачи. Особое значение эти

проблемы приобретают в планировании экспериментов с целью получения наиболее информативного набора функционалов.

9.4.1. Некоторые вопросы линейной теории измерений

В настоящее время теория измерений приобретает большее значение в организации информационной системы. Измерительная техника позволяет получить набор сведений (функционалов) о процессе, анализировать процесс и направлять его. С помощью таких функционалов интерпретируется физический процесс.

Мы не будем говорить об отдельных элементарных измерениях, таких как измерение напряжения, силы тока в отдельных участках электрической цепи и т. д. Нас будут интересовать только сложные физические явления и процессы, которые должны быть поняты и количественно оценены с требуемой точностью. Аналогичные задачи возникают постоянно, особенно в новых областях техники. К примеру, нельзя разработать методы измерений коэффициента размножения нейтронов в реакторе, если в деталях не ясен физический процесс цепной реакции и диффузии нейтронов, неизвестны уравнения, описывающие поведение ядерного реактора при изменении различных условий.

Несомненно, методы измерений и сами приборы существенно совершенствуются вместе с развитием теории физического процесса. Разработка теории и эксперимента, как правило, сопровождается созданием новых или усовершенствованием прежних методов измерений.

Возникает вопрос, нельзя ли в настоящее время сформулировать более или менее общий подход к методам измерения применительно к различным процессам с возможностью формального математического описания алгоритма. Оказывается, такой подход можно сформулировать по крайней мере для задач с линейными операторами. В дальнейшем речь будет идти именно об этом классе задач.

Представляется, что основой теории измерений вариаций физических величин может служить теория возмущений. Суть дела состоит в следующем. Предположим, что мы изучаем сложный физический процесс с помощью прибора, имеющего определенные физические характеристики. Показания такого прибора связаны с исследуемым полем физической величины и являются функционалами поля. В большинстве случаев, однако, экспериментатора интересует не само поле физической величины, а отклонения от него под влиянием обычно малых возмущений. Это значит, что измерения должны быть проведены с достаточной точностью, чтобы зарегистри-

ровать указанные отклонения поля от некоторого «стандартного» состояния. Предположим, что это первое необходимое требование к прибору выполнено и мы располагаем измерениями отклонений показания прибора от нормы с требуемой точностью. Спрашивается, достаточно ли этой информации для удовлетворительной интерпретации эксперимента и можем ли мы с достаточной точностью восстановить информацию о возмущенном состоянии системы. К сожалению, на этот вопрос обычно дать ответ очень трудно. Объясняется это тем, что задачи восстановления информации о поле физической величины с помощью измерительных приборов являются, как правило, некорректно поставленными задачами математической физики.

Для того чтобы обойти эту принципиальную трудность обработки экспериментальных данных, необходимо с самого начала связать отклонения показания прибора непосредственно с отклонениями изучаемых физических параметров процесса. В этом случае ошибка в исследуемой характеристике будет прямо пропорциональна ошибке в отклонении показания прибора — вариации функционала — и, следовательно, при интерпретации мы используем максимальную информацию измерительного прибора. Именно с этих позиций мы приходим к излагаемой ниже теории.

9.4.2. Сопряженные функции и понятие ценности

Рассмотрим функцию $\varphi(x)$, удовлетворяющую уравнению

$$L\varphi(x) = q(x), \tag{9.4.1}$$

где L — некоторый линейный оператор, а $q(x)$ — распределение источника в среде. При этом под x будем понимать совокупность всех переменных задачи (временная и пространственные координаты, энергия, направление скорости), считая, что функции φ и q являются действительными.

Ради определенности будем полагать, например, что исследуемый процесс связан с диффузией или переносом субстанции, хотя выводы теории выходят далеко за рамки такого рода задач.

Введем гильбертово пространство функций со скалярным произведением

$$(g, h) = \int g(x)h(x)\,dx, \tag{9.4.2}$$

где интегрирование ведется по всей области D определения функций g и h.

При решении тех или иных физических задач обычно нужно получить в результате значение некоторой величины, являющейся функционалом от $\varphi(x)$. Любая величина, связанная с $\varphi(x)$, может быть выражена в виде такого

скалярного произведения. Например, если нас интересует результат измерения некоторого процесса в среде с характеристикой прибора $\Sigma(x)$, то это значение есть

$$J_\Sigma[\varphi] = \int \varphi(x)\Sigma(x)\,dx = (\varphi, \Sigma). \tag{9.4.3}$$

Таким образом, будем рассматривать физические величины, которые могут быть выражены в виде линейного функционала от $\varphi(x)$:

$$J_p[\varphi] = (\varphi, p),$$

где величина p характеризует интересующий нас физический процесс. Введем вместе с оператором L сопряженный к нему оператор L^*, определяющийся из условия

$$(g, Lh) = (h, L^*g), \tag{9.4.4}$$

для любых функций g и h. Наряду с уравнением (9.4.1), которое будем называть основным, введем сначала формально неоднородное сопряженное уравнение

$$L^*\varphi_p^* = p(x), \tag{9.4.5}$$

где $p(x)$ — некоторая произвольная пока функция, а $\varphi_p^* \in \Phi^*$. Подставляя в формулу (9.4.4) вместо функций h и g решения уравнений (9.4.1) и (9.4.5) φ и φ_p^*, получим

$$(\varphi_p^*, L\varphi) = (\varphi, L^*\varphi_p^*) \tag{9.4.6}$$

или, воспользовавшись уравнениями (9.4.1) и (9.4.5),

$$(\varphi_p^*, q) = (\varphi, p); \tag{9.4.7}$$

иначе говоря,

$$J_q[\varphi_p^*] = J_p[\varphi].$$

Поэтому, если нам нужно найти значение функционала $J_p[\varphi]$, мы можем получить его двояко: либо решить уравнение (9.4.1) и определить эту величину по формуле

$$J_p[\varphi] = (\varphi, p), \tag{9.4.8}$$

либо решить уравнение (9.4.5) и определить ту же величину по формуле

$$J_p[\varphi] = J_q[\varphi_p^*] = (\varphi_p^*, q). \tag{9.4.9}$$

Следовательно, каждому линейному функционалу $J_p[\varphi] = (\varphi, p)$ может быть поставлена в соответствие функция $\varphi_p^*(x)$, удовлетворяющая уравнению (9.4.5), причем в качестве свободного члена этого уравнения следует

использовать именно функцию $p(x)$, характеризующую интересующий нас процесс.

Пусть в среде имеется «источник единичной мощности», помещенный в точку x_0, т. е.

$$q(x) = \delta(x - x_0). \tag{9.4.10}$$

Так как

$$(\varphi(x), \delta(x - x_0)) = \varphi(x_0), \tag{9.4.11}$$

то в этом случае

$$J_p[\varphi] = J_{q-\delta(x-x_0)}[\varphi_p^*] = \varphi_p^*(x_0). \tag{9.4.12}$$

Следовательно, сопряженная функция $\varphi_p^*(x)$ описывает зависимость функционала $J_p[\varphi] = (\varphi, p)$ от точки помещения «источника единичной мощности».

Представим себе физическую систему (или прибор), в которой измеряется некоторая величина $J_p[\varphi]$, являющаяся линейным функционалом от решения, связанного, например, с плотностью частиц субстанции φ. Если в некоторую точку системы впустить определенное количество частиц (или, наоборот, извлечь эти частицы), то измеряемое значение величины $J_p[\varphi]$ будет соответственно увеличиваться или уменьшаться, причем это изменение будет зависеть от той точки, в которой мы производим изменение числа частиц. Как видно из предыдущего, эта зависимость описывается сопряженной функцией φ_p^*, удовлетворяющей уравнению (9.4.5). Следовательно, сопряженная функция $\varphi_p^*(x)$ дает вклад частиц, находящихся в той или иной точке системы, в интересующий нас функционал J_p. Поэтому функцию $\varphi_p^*(x)$ можно назвать ценностью субстанции в точке x по отношению к функционалу $J_p[\varphi] = (\varphi, p)$[6)].

Толкование сопряженной функции $\varphi_p^*(x)$ как ценности субстанции позволяет дать ясную трактовку и теории возмущений для любого функционала $J_p[\varphi]$. Действительно, если в элементе объема Δx около точки x мы изменим число частиц на величину δN, то соответствующее изменение величины J_p будет выражено следующим уравнением:

$$\delta J_p = \delta N \varphi_p^*(x). \tag{9.4.13}$$

Если в рассматриваемой системе произведены некоторые малые изменения параметров, так что оператор L переходит в оператор $L + \delta L$, то это соответствует изменению числа частиц в каждом элементе Δx на величину $\delta N = -\Delta x \delta L_\varphi$. Общее изменение функционала L_p при таком изменении

[6)] Термин «ценность» весьма удачен в задачах теории переноса излучения. Возможно, что в других задачах будет найден более подходящий термин.

запишем в виде

$$\delta J_p = -\int \varphi_p^*(x)\delta L\varphi(x)\,dx. \tag{9.4.14}$$

Строгий вывод этого соотношения будет дан ниже.

Соотношение (9.4.13) позволяет измерять распределение функции ценности в системе, изменяя известным образом число частиц в разных точках x системы и измеряя при этом соответствующее изменение величины J_p. Введенное понятие ценности может быть полезным в теории различных измерительных приборов. Действительно, прибор обычно предназначен для измерения какой-либо одной величины J_p. Поэтому для каждого прибора может быть введена вполне определенная функция ценности $\varphi_p^*(x)$, которая может быть однажды измерена или сосчитана. Если распределение субстанции и ее ценности известны, то соотношение (9.4.14) может быть использовано для измерения двояким образом. Во-первых, измеряя величины δJ_p при различных изменениях параметров среды δL, мы можем при помощи соотношения (9.4.14) определять величины δL, т. е. различные характеристики взаимодействия частиц с веществом. Например, таким образом можно измерить (по существу, так и делается) сечения взаимодействия нейтронов с веществом для различных образцов, помещая эти образцы в прибор и определяя $\delta\Sigma = \delta L$ по изменению величины J_p. Во-вторых, соотношение (9.4.14) позволяет вводить поправки в измеряемую величину J_p за счет различных возмущающих факторов в приборе. Наконец, определение понятия ценности позволяет получать уравнения для функции φ_p^* за счет различных возмущающих факторов в приборе. Наконец, определение понятия ценности позволяет получать уравнения для функции $\varphi_p^*(x)$, исходя непосредственно из физического смысла этой величины, точно так же, как уравнение для потока нейтронов получается из закона сохранения числа нейтронов.

Приведенные выше формулы позволяют также получать теорему взаимности для функции Грина основного и сопряженного уравнений $G(x, x_0)$ и $G^*(x, x_1)$. Функция $G(x, x_0)$ удовлетворяет уравнению (9.4.1) при $q(x) = \delta(x - x_0)$, а функция $G^*(x, x_1)$ — уравнению (9.4.5) при $p(x) = \delta(x - x_1)$.

Подставляя в формулу (9.4.7)

$$\varphi(x) = G(x, x_0), \quad \varphi_p^* = G^*(x, x_1)$$

и приведенные выражения для q и p, получим

$$G(x_1, x_0) = G^*(x_0, x_1),$$

что и является формулировкой теоремы взаимности.

9.4.3. Теория возмущений для линейных функционалов

Если свойства среды, с которой взаимодействует поле, изменяются, т. е. если оператор уравнения (9.4.1) переходит в

$$L' = L + \delta L,$$

то изменяются и поле $\varphi(x)$, и значение функционала $J_p[\varphi]$:

$$\varphi(x) \to \varphi'(x), \quad J_p[\varphi] \to J'_p + \delta J_p.$$

Установим связь между изменением оператора δL и изменением функционала δJ_p. Возмущенная система описывается уравнением

$$L'\varphi' = (L + \delta L)\varphi' = q. \tag{9.4.15}$$

Сопряженная функция невозмущенной системы, соответствующая функционалу J_p, описывается уравнением

$$L^*\varphi^*_p = p. \tag{9.4.16}$$

Помножив скалярно уравнение (9.4.15) на φ^*, уравнение (9.4.16) — на φ', вычитая одно из другого и пользуясь определением сопряженного оператора уравнения (9.4.4), получим слева

$$(\varphi^*_p, L'\varphi') - (\varphi', L^*\varphi^*_p) = (\varphi^*_p, \delta L\varphi'), \tag{9.4.17}$$

а справа в соответствии с уравнением (9.4.7) будем иметь

$$(\varphi^*_p, q) - (\varphi', p) = J_p[\varphi] - J_p[\varphi'] = -\delta J_p. \tag{9.4.18}$$

Приравнивая выражения (9.4.17) и (9.4.18), получим общее соотношение для приращения функционала

$$\delta J_p = -(\varphi^*_p, \delta L\varphi'). \tag{9.4.19}$$

Если вместо уравнений (9.4.15) и (9.4.16) рассмотреть возмущенное сопряженное уравнение

$$(L^* + \delta L^*)\varphi^{*'}_p = p \tag{9.4.20}$$

и невозмущенное основное уравнение (9.4.1), то аналогичным путем можно получить также соотношение

$$\delta J_p = -(\varphi, \delta L^* \varphi_p^{*\prime}), \tag{9.4.21}$$

которое, конечно, эквивалентно соотношению (9.4.19).

Отметим важную особенность применения формул теории возмущений: так как эти формулы пишутся для вариации функционала, погрешность в которой обычно допустима в пределах нескольких процентов, то для вычисления указанных вариаций нет необходимости знать точное решение основной и сопряженной задач, достаточно воспользоваться их приближенными решениями.

Если возмущение оператора L (а следовательно, и L^*) столь мало, что оно не очень сильно искажает функции φ и φ_p^*, то в формулах (9.4.19) и (9.4.21) можно заменить приближенно $\varphi' = \varphi$, $\varphi^{*\prime} = \varphi^*$. При этом мы получим две эквивалентные друг другу формулы теории малых возмущений[7)]:

$$\delta J_p = -(\varphi_p^*, \delta L \varphi), \tag{9.4.22}$$

$$\delta J_p = -(\varphi, \delta L^* \varphi_p^*). \tag{9.4.23}$$

Получение формулы теории возмущений, кроме их прямого использования для оценки различных эффектов и для анализа измерений, могут иметь и еще одно весьма важное применение.

При теоретическом рассмотрении и в практических расчетах часто пользуются методом замены исследуемой сложной системы упрощенной моделью. Необходимым условием такой замены является, очевидно, требование, чтобы она не приводила к изменению некоторых основных для рассматриваемого вопроса характеристик системы. Примером может служить замена в дифференциальных уравнениях переменных коэффициентов коэффициентами постоянными. К числу таких методов относится и метод эффективных граничных условий, заключающийся в замене истинных условий некоторыми упрощенными, но такими, которые приводят к правильному значению некоторого избранного функционала.

Полученные выше формулы теории возмущений позволяют сформулировать весьма общий подход к различным задачам. Пусть рассматриваемая система характеризуется оператором L, причем наиболее существенной величиной является функционал $J_p[\varphi]$. Если искомая простая модель характеризуется оператором $L' = L + \delta L$, то для того, чтобы величина J_p

[7)] Ради простоты в дальнейшем формулы для приращения функционалов (9.4.19) в (9.4.21) будем также называть формулами теории возмущений.

не изменялась при переходе от истинной системы к модели, необходимо выполнение условия

$$\delta J_p = -(\varphi_p^*, [L' - L]\varphi') = 0, \text{ т. е. } (\varphi_p^*, L'\varphi') = (\varphi_p^*, L\varphi'). \tag{9.4.24}$$

Если мы интересуемся несколькими величинами J_{p_1}, J_{p_2} и т. д., то получим несколько условий типа (9.4.24) с решениями $\varphi_{p_1}^*$, $\varphi_{p_1}^*$ и т. д.

Требование (9.4.24) не определяет однозначно искомой эквивалентной модели, но является ее необходимым условием и вместе с другими может помочь ее нахождению. В частности, если оператор модели L', вид которого может быть найден из физических соображений, содержит один или несколько параметров, то условия (9.4.24) могут быть использованы для определения значений этих параметров.

9.4.4. Численные методы решения обратных задач и планирование эксперимента

Предположим, что мы располагаем набором функционалов (измерений) J_{p_i} $(i = 1, 2, \ldots, n)$. Будем считать, что измерения по своему характеру разнообразны, например, измерение производится одним и тем же прибором в разных «точках» области определения решения или приборы регистрируют разные характеристики исследуемого явления. Ради простоты полагаем, что статистические погрешности в измерениях устранены и мы имеем дело уже с предварительно обработанной системой данных.

Каждому функционалу J_{p_i} поставим в соответствие функцию ценности для невозмущенной задачи, т. е. модели, в которой оператор L и область его определения считаются известными. Решим n различных задач

$$L^*\varphi_{p_i}^* = p_i, \quad i = 1, 2, \ldots, n. \tag{9.4.25}$$

Найдем заранее n функций ценности $\varphi_{p_i}^*$ и решим одну основную задачу с модельным «невозмущенным» оператором L, сопряженным L^*:

$$L\varphi = q. \tag{9.4.26}$$

Будем считать, что $\varphi \in \Phi$ и $\varphi^* \in \Phi^*$, где Φ и Φ^* — области определения операторов L и L^* соответственно. Далее построим n формул теории малых возмущений

$$(\varphi_{p_i}^*, \delta L\varphi) = -\delta J_{p_i}, \quad i = 1, 2, \ldots, n, \tag{9.4.27}$$

где δL — разность между изучаемым оператором L' и модельным L. Предположим, что оператор L известен:

$$L = \sum_{k=1}^{m} [\alpha_k A_k + B_k(\beta_k C_k)], \tag{9.4.28}$$

где A_k, B_k и C_k — элементарные линейные операторы, например, дифференцирования или интегрирования либо комбинации тех и других; $\alpha_k(x)$ и $\beta_k(x)$ — искомые коэффициенты, обычно в грубом приближении известны для невозмущенной (модельной) задачи.

Теперь нашей целью является восстановление коэффициентов α'_k и β'_k в выражении

$$L' = \sum_{k=1}^{m} [\alpha'_k A_k + B_k(\beta'_k C_k)]. \tag{9.4.29}$$

С помощью выражений (9.4.28) и (9.4.29) получим

$$\delta L = \sum_{k=1}^{m} [\delta\alpha_k A_k + B_k(\delta\beta_k C_k)], \tag{9.4.30}$$

где

$$\delta\alpha_k = \alpha'_k - \alpha_k, \quad \delta\beta_k = \beta'_k - \beta_k.$$

Подставим (9.4.30) в (9.4.27). Тогда при соответствующих условиях приходим к системе уравнений

$$\sum_{k=1}^{m} [(\varphi^*_{p_i}, \delta\alpha_k A_k \varphi) + (B^*_k \varphi^*_{p_i}, \delta\beta_k C_k \varphi)] = -\delta J_{p_i}, \tag{9.4.31}$$

$$i = 1, 2, \ldots, n.$$

Дальнейшая задача состоит в параметризации вариации $\delta\alpha_k$ и $\delta\beta_k$. Сначала рассмотрим простейший случай, когда $\delta\beta_k = 0$, а $\delta\alpha_k$ — постоянные. В этих условиях (9.4.31) переходит в задачу линейной алгебры

$$\sum_{k=1}^{m} \delta\alpha_k (\varphi^*_{p_i}, A_k \varphi) = -\delta J_{p_i}, \quad i = 1, 2, \ldots, n. \tag{9.4.32}$$

Здесь $(\varphi^*_{p_i}, A_k\varphi)$ — элементы матрицы, которые при заданных φ, $\varphi^*_{p_i}$ и A_k могут быть рассчитаны.

Пусть y — вектор с компонентами $\delta\alpha_k$, F — вектор с компонентами — δJ_{p_i} и $a_{ik} = (\varphi^*_{p_i}, A_k\varphi)$ — элементы матрицы Λ. Тогда приходим к уравнению

$$\Lambda y = F. \tag{9.4.33}$$

Если число функционалов n равно числу определяемых вариаций коэффициентов α_k, то система (9.4.33) в принципе позволяет найти $\delta\alpha_k$. Если число n больше m, то система (9.4.33) оказывается переопределенной, а ее решение (если оно существует) обычно находится с помощью метода наименьших квадратов в предположении, что y доставляет минимум квадратичному функционалу

$$\|\Lambda y - F\|^2 = \min. \tag{9.4.34}$$

Вектор y, минимизирующий этот функционал, иногда называют квазирешением уравнения (9.4.33). Если $n = m$, то для решения системы (9.4.33) можно использовать методы, которые были нами обсуждены в главе 4 в связи с анализом итерационных процессов при неточных входных данных.

Если $\delta\alpha_k(x)$ и $\delta\beta_k(x)$ — функции, то решение обратной задачи можно находить с помощью тех или иных методов параметризации, сущность которых состоит в следующем. Предположим, что на основании априорного анализа поведения физических параметров (обычно в результате статистического и корреляционного анализа) находятся некоторые полные ортогональные системы функций $u_{k,l}(x)$ и $v_{k,l}(x)$, такие, что с их помощью возможно достаточно хорошее приближение к функциям α_k и β_k при малом числе $n(k)$, так что

$$\delta\alpha_k(x) = \sum_{l=1}^{n(k)} a_{k,l} u_{k,l}(x),$$

$$\delta\beta_k(x) = \sum_{l=1}^{n(k)} b_{k,l} v_{k,l}(x), \tag{9.4.35}$$

где $a_{k,l}$ и $b_{k,l}$ — коэффициенты, подлежащие определению. Подставим выражения (9.4.35) в (9.4.31) и получим

$$\sum_{k=1}^{m} \sum_{l=1}^{n(k)} [a_{k,l}(\varphi^*_{p_i}, u_{k,l} A_k \varphi) + b_{k,l}(B^*_k \varphi^*_{p_i}, v_{k,l} C_k \varphi)] = -\delta J_{p_i}. \tag{9.4.36}$$

Упорядочим теперь величины $a_{k,l}$, $b_{k,l}$ и переобозначим их через y_j ($j = 1, 2, \dots$). Введем в рассмотрение такую матрицу Λ, что уравнение

$$\Lambda y = F$$

эквивалентно системе (9.4.36). Тогда мы снова приходим к задаче линейной алгебры (9.4.36), решая которую находим $a_{k,l}$ и $\delta_{k,l}$, а следовательно, $\delta\alpha_k$ и $\delta\beta_k$.

Мы рассмотрели только тот случай, когда решение модельной задачи близко к реальной, т. е. можно заменить φ' на φ и, таким образом, восполь-

зоваться теорией малых возмущений. Если невозмущенное (модельное) состояние процессов существенно отличается от истинного, то рассмотренный выше алгоритм можно считать только первым приближением к решению обратной задачи. После того как вариации $\delta\alpha_k$ и $\delta\beta_k$ найдены, можно исправить коэффициенты α_k, β_k и найти

$$\alpha'_k = \alpha_k + \delta\alpha_k,$$

$$\beta'_k = \beta_k + \delta\beta_k.$$

После этого необходимо решить «невозмущенную» задачу

$$L'\varphi' = f \tag{9.4.37}$$

с оператором

$$L' = \sum_{k=1}^{\infty}[\alpha'_k A_k + B_k(\beta'_k C_k)],$$

перейти к новому приближению в решении обратной задачи, приняв вместо (9.4.31) более общую формулу возмущений

$$\sum_{k=1}^{m}[(\varphi^*_{p_i}, \delta\alpha_k A_k\varphi') + (B^*_k\varphi^*_{p_i}, \delta\beta_k C_k\varphi')] = -\delta J_{p_i},\ i = 1, 2, \ldots, n, \tag{9.4.38}$$

и повторить цикл вычислений для уточнения вариации $\delta\alpha_k$, $\delta\beta_k$. Это мы будем называть вторым приближением в решении обратной задачи. Разумеется, указанный процесс может быть продолжен. Сходимость методов последовательных приближений может быть доказана с учетом конкретной информации об элементарных операторах задачи A_k и области определения операторов L, L^*.

Проиллюстрируем наш алгоритм на простом примере. Предположим, что рассматривается задача

$$-\frac{d}{dx}\beta(x)\frac{d\varphi'}{dx} + \alpha(x)\varphi' = f(x), \quad \varphi'(0) = \varphi'(1) = 0 \tag{9.4.39}$$

с неизвестными коэффициентами $\alpha(x)$ и $\beta(x)$, относительно которых делается априорное допущение. Предполагается, например, что они являются непрерывными функциями в области определения решения $0 \leqslant x \leqslant 1$ и известны их приближенные значения $\bar\alpha$ и $\bar\beta$, т. е.

$$\alpha(x) = \bar\alpha + \delta\alpha(x),\ \beta(x) = \bar\beta + \delta\beta(x). \tag{9.4.40}$$

Если на основе априорной информации можно выбрать значения $\alpha(x)$ и $\beta(x)$ в модели более точно, то не обязательно предполагать, что они равны постоянным $\bar{\alpha}$ и $\bar{\beta}$.

Кроме того, на основе предварительного изучения делается вывод о возможности представления $\delta\alpha(x)$ и $\delta\beta(x)$ в виде конечной суммы:

$$\delta\alpha(x) = \sum_{l=1}^{n(1)} a_l u_l(x), \quad \delta\beta(x) = \sum_{l=1}^{n(1)} b_l v_l(x), \tag{9.4.41}$$

где $\{u_l(x)\}$ и $\{v_l(x)\}$ — некоторые полные ортонормированные системы функций (например, тригонометрические полиномы, полиномы Лежандра и т. д.).

Пусть $p_1(x), p_2(x), \ldots, p_n(x)$ — характеристики измерений, так что в каждом из измерений регистрируется функционал

$$J'_{p_i}[\varphi'] = \int_0^1 p_i(x)\varphi'(x)\,dx, \quad i = 1, 2, \ldots, n. \tag{9.4.42}$$

Функции $p_i(x)$ можно назвать характеристиками прибора в данном измерении.

Введем в рассмотрение невозмущенную (модельную) задачу, соответствующую задаче (9.4.39):

$$-\frac{d}{dx}\bar{\beta}\frac{d\varphi}{dx} + \bar{\alpha}\varphi = f, \quad \varphi(0) = \varphi(1) = 0. \tag{9.4.43}$$

Наряду с задачей (9.4.43) сформулируем n сопряженных задач, соответствующих избранной модели:

$$-\frac{d}{dx}\bar{\beta}\frac{d\varphi^*_{p_i}}{dx} + \bar{\alpha}\varphi^*_{p_i} = p_i(x), \tag{9.4.44}$$

$$\varphi^*_{p_i}(0) = \varphi^*_{p_i}(1) = 0, \quad i = 1, 2, \ldots, n.$$

Согласно общей теории будем иметь

$$J_{p_i}[\varphi] = \int_0^1 p_i(x)\varphi(x)\,dx = \int_0^1 f(x)\varphi^*_{p_i}(x)\,dx. \tag{9.4.45}$$

Предположим теперь, что модельные задачи (9.4.43) и (9.4.44) решены. Далее находим вариации функционала δJ_{p_i} по формуле

$$\delta J_{p_i} = J'_{p_i} - J_{p_i}, \quad i = 1, 2, \ldots, n, \tag{9.4.46}$$

где J'_{p_i} — измерение прибора с характеристикой p_i, соответствующее (9.4.42) (в котором φ' нам неизвестно); J_{p_i} — функционал, теоретически рассчитываемый на основе любого соотношения в (9.4.45). Здесь требуется такая точность в измерении, которая бы гарантировала расчет вариаций δJ_{p_i}.

Рассмотрим теперь формулы теории малых возмущений (9.4.32):

$$A = E, \quad B = -\frac{d}{dx}, \quad C = \frac{d}{dx}.$$

С учетом граничных условий для $\varphi^*_{p_i}$ и φ получим

$$\int\limits_0^1 \left(\delta\alpha\varphi\varphi^*_{p_i} + \delta\beta \frac{d\varphi}{dx}\,\frac{d\varphi^*_{p_i}}{dx} \right) dx = -\delta J_{p_i}. \tag{9.4.47}$$

Подставив выражение для $\delta\alpha(x)$ и $\delta\beta(x)$ из (9.4.41), (9.4.47) и получим, что

$$\sum_{l=1}^{n(1)} \left(a_l \int\limits_0^1 u_l\varphi\varphi^*_{p_i}\,dx + b_l \int\limits_0^1 v_l \frac{d\varphi}{dx}\,\frac{d\varphi^*_{p_i}}{dx}\,dx \right) = -\delta J_{p_i},\ i = 1, 2, \ldots, n. \tag{9.4.48}$$

Если $n = 2n(1)$, то система уравнений (9.4.48) полностью определена.

Решая эту систему, находим коэффициенты a_l, b_l и на основе представления (9.4.41) получаем первое приближение для величин α' и β'. Эти величины можно уточнять, используя метод последовательных приближений, рассмотренный выше. Точно так же могут быть поставлены и решены более сложные обратные задачи, в том числе задача по определению возмущений δf в источниках.

Обсудим теперь проблему планирования сложного эксперимента. Ее можно сформулировать следующим образом. Среди всевозможного (практически реализуемого) набора измерений необходимо выбрать тот, который оказывается наиболее информативным с точки зрения решения конкретной обратной задачи по восстановлению требуемых характеристик среды (коэффициентов уравнений). В общем плане оптимизации эта задача оказывается очень сложной. Однако можно рассмотреть некоторые частные подходы к ее решению.

Допустим, что перед осуществлением эксперимента строится модель невозмущенной задачи, с ее помощью описываются линейные функционалы от решения и с учетом априорной информации о точности измерений делается вывод о необходимой точности измерения функционалов. Предположим, что необходимые требования к точности измерения функционалов δJ_{p_i} обеспечены. Далее рассматриваются различные совокупности измере-

ний и выбираются те из них, которые приводят к наилучшей обусловленности матрицы Λ. Полученная система линейных уравнений в этом случае хорошо решается, и такой план эксперимента является в известном смысле оптимальным из данной совокупности (конечно, здесь не рассматриваются экономические вопросы, оказывающиеся иногда решающими при планировании эксперимента). Если же при заданном наборе информативных функционалов, обеспечивающих наилучшую обусловленность матрицы Λ, не удается реализовать высоких требований к точности измерений функционалов J_{p_i}, то возникает более сложная задача о планировании эксперимента при заданных ограничениях на точность измерений, допускаемых разрешением приборной техники. Это уже своеобразная задача на оптимизацию с ограничениями.

Мы не останавливаемся здесь на вопросах статистической обработки экспериментальных материалов. Эти вопросы, освещенные достаточно подробно в литературе, не вносят в теорию постановки и решение обратных задач каких-либо принципиальных трудностей.

Глава 10.

Методы оптимизации

Бурное развитие методов математического моделирования задач науки и техники стимулировало широкий поиск эффективных алгоритмов оптимизации параметров математических моделей по отношению к тем или иным функционалам задач. Некоторые общие подходы к проблеме оптимизации возникли в теории вариационного исчисления; они частично уже описаны в § 2.2. Однако набор оптимизационных алгоритмов в настоящее время так значителен, что можно говорить о специальном направлении в вычислительной математике. Это, прежде всего, методы линейного и квадратичного программирования, это выпуклое и динамическое программирование, это и принцип максимума Понтрягина. Все эти методы имеют своеобразие как в постановках задач оптимизации, так и в методах реализации. Указанные методы и будут изложены в настоящей главе.

10.1. Выпуклое программирование

Напомним некоторые факты. Выпуклой называется функция f, определенная на выпуклом множестве X, которая при любых $\alpha \in (0,1)$ и $x', x'' \in X$ удовлетворяет неравенству

$$f(\alpha x' + (1-\alpha)x'') \leqslant \alpha f'(x') + (1-\alpha) f(x'').$$

Если выпуклая функция f непрерывна и дифференцируема, то для любых $x, x_* \in X$ выполняется дифференциальное неравенство

$$f(x) \geqslant f(x_*) + \left(f'(x_*), x - x_*\right). \tag{10.1.1}$$

Здесь в правой части стоит скалярное произведение векторов $f'(x_*)$ и $x - x_*$. И наоборот, если дифференцируемая функция f при любых $x, x_* \in X$ удовлетворяет неравенству (10.1.1), то это выпуклая функция.

Геометрически неравенство (10.1.1) означает следующее. Построим график функции f и в точке графика $(x_*, f(x_*))$ проведем к нему касательную гиперплоскость. Тогда для каждого $x \in X$ правая часть неравенства (10.1.1) совпадает с соответствующей ординатой касательной гиперплоскости, то есть график выпуклой функции расположен выше (точнее, не ниже) касательной гиперплоскости. В одномерном случае две типичные ситуации изображены на рис. 10.1. В случае 10.1a точка x_* является точкой абсолютного минимума функции f, производная $f'(x_*)$ равна нулю, а касательная горизонтальна. Ситуация общего положения изображена на рис. 10.1b.

Для рассматриваемого одномерного случая неравенство (10.1.1) принимает вид

$$f(x) \geqslant f(x_*) + f'(x_*)(x - x_*). \tag{10.1.2}$$

В правой части неравнества (10.1.1) и соответственно (10.1.2) присутствуют два члена разложения функции f по формуле Тейлора в окрестности точки $x = x_*$. Заметим, что для дважды дифференцируемой функции можно написать представление

$$f(x) = f(x_*) + f'(x_*)(x - x_*) + \frac{1}{2} f''(\xi)(x - x_*)^2, \tag{10.1.3}$$

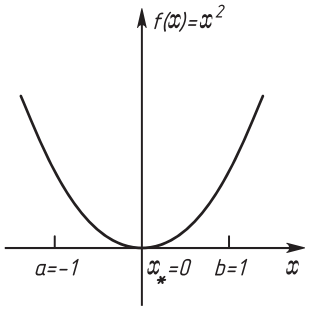
где ξ — некоторая точка отрезка с концами x_* и x. Поэтому в предположении двухкратной дифференцируемости функции f на X условием ее выпуклости является неравенство

$$f''(x) \geqslant 0. \tag{10.1.4}$$

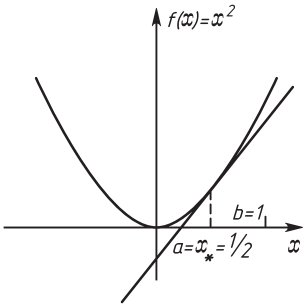
Аналогичные рассуждения можно провести и для многомерного случая. При этом условие (10.1.4) заменяется требованием положительной полуопределенности матрицы вторых производных $f''(x)$.

При описании допустимого множества, на котором ищется минимум функции f, помимо ее области задания X обычно вводятся те или иные ограничения. Например, в рассматриваемых на рис. 10.1 случаях такими ограничениями на всей вещественной оси $-\infty < x < \infty$ могут быть неравенства $x \geqslant a$ и $x \leqslant b$, которые можно записать в единообразном виде:

$$a - x \leqslant 0, \quad x - b \leqslant 0. \tag{10.1.5}$$



(a)



(b)

Рис. 10.1.

В дальнейшем для ограничений, которых может быть несколько (например, m), будем пользоваться универсальным обозначением

$$g_i(x) \leqslant 0, \quad i = 1, 2, \ldots, m. \tag{10.1.6}$$

В случае (10.1.5) $g_1(x) = a - x$ и $g_2(x) = x - b$. Для широкого класса задач при заданном множестве X с ограничениями (10.1.6) удобно пользоваться формулой

$$D = \{x \in X : \ g_i(x) \leqslant 0, \ i = 1, 2, \ldots, m\}.$$

Здесь через D обозначено допустимое множество. Заметим, что выпуклость множества D можно гарантировать, если все функции $g_i(x)$ выпуклые. Задача о поиске

$$\inf\{f(x) : x \in D\} \tag{10.1.7}$$

называется задачей выпуклого программирования. При этом в качестве X фигурирует либо все пространство, либо его положительный ортант (т. е. множество векторов с неотрицательными компонентами).

Перейдем к рассмотрению важного понятия — функции Лагранжа, определим ее следующим образом:

$$L(x,\lambda)=f(x)+\sum_{i=1}^{m}\lambda_i g_i(x),\quad x\in X,\quad \lambda_i\geqslant 0,\quad i=1,2,\ldots,m. \tag{10.1.8}$$

Здесь $\lambda=(\lambda_1,\lambda_2,\ldots,\lambda_m)$. Условие неотрицательности всех компонент такого набора в дальнейшем будем выражать векторным неравенством $\lambda\geqslant 0$. Чтобы избежать рассмотрения различных вырожденных случаев, которые могут возникнуть при нелинейных левых частях системы (10.1.6), мы предположим в дальнейшем выполненным следующее

Условие регулярности. Существует точка $\xi\in D$, в которой все ограничения (10.1.6) обращаются в строгие неравенства:

$$g_i(\xi)<0,\quad i=1,2,\ldots,m.$$

Справедливо следующее важное утверждение: для того чтобы $x_*\in X$ доставлял минимум функции f при ограничениях (10.1.6), достаточно, а при условии регулярности и необходимо, чтобы существовал такой набор $\lambda^*=(\lambda_1^*,\lambda_2^*,\ldots,\lambda_m^*)$, что при всех $x\in X$ и $\lambda\geqslant 0$ выполнялись бы неравенства

$$L(x_*,\lambda)\leqslant L(x_*,\lambda^*)\leqslant L(x,\lambda^*). \tag{10.1.9}$$

Это утверждение известно как теорема Куна — Таккера[1)]. Пару (x_*,λ^*) принято при этом называть седловой точкой функции Лагранжа (на множестве $X\times R_+^m$). Важно отметить, что множество, на котором ищется седловая точка функции Лагранжа, имеет простой вид: ограничения (10.1.6) в его описании не участвуют и выполняются для x — компоненты седловой точки автоматически. Заметим также, что в случае, когда ограничения (10.1.6) линейные, теорема Куна — Таккера верна без условия регулярности.

Если пара (x_*,λ^*) является седловой точкой функции Лагранжа (и, следовательно, x_* решает задачу выпуклого программирования), то выполнены условия дополнительности:

$$\lambda_i^* g_i(x_*)=0,\quad i=1,2,\ldots,m. \tag{10.1.10}$$

Действительно, если бы для некоторого k оказалось $\lambda_k^*>0$ и $g_k(x_*)<0$, то, положив $\lambda_i=\lambda_i^*$ при $i\neq k$ и $\lambda_k=\lambda_k^*/2$, мы нарушили бы левое неравенство в (10.1.9). Допустимости x_* и условий дополнительности, очевидно, достаточно для выполнения левого неравенства в (10.1.9). Условия (10.1.10)

[1)] Доказательство этой теоремы можно найти, например, в работе В. Г. Карманова «Математическое программирование» [21].

показывают, что

$$L(x_*, \lambda^*) = f(x_*). \tag{10.1.11}$$

Введем теперь важное понятие двойственности в задачах выпуклого программирования. При этом будем предполагать, что у функции Лагранжа имеется седловая точка (x_*, λ^*). Положим

$$\varphi(\lambda) = \inf\{L(x, \lambda) : x \in X\},\ \lambda \geqslant 0. \tag{10.1.12}$$

При некоторых λ функция φ может принимать и несобственное значение — ∞. Однако в силу (10.1.9) и (10.1.11) $\varphi(\lambda^*) = f(x_*)$. Теперь если положить

$$\tilde{f}(x) = \begin{cases} +\infty, & x \notin D, \\ f(x), & x \in D, \end{cases}$$

то окажется, что

$$\tilde{f}(x) = \sup\{L(x, \lambda) : \lambda \geqslant 0\},\ x \in X.$$

При этом задача $\inf\{\tilde{f}(x) : x \in X\}$ и задача (10.1.7) эквивалентны. Соотношения (10.1.9) и (10.1.11) показывают, что в рассмотренном случае

$$f(x_*) = \min\{\tilde{f}(x) : x \in X\} = \max\{\varphi(\lambda) : \lambda \geqslant 0\}. \tag{10.1.13}$$

Таким образом, если нашей задачей является отыскание $f(x_*) = \min\{f(x) : x \in D\}$, то можно решать вместо этого двойственную задачу, т. е. искать $\max\{\varphi(\lambda);\ \lambda \geqslant 0\}$.

В заключение проиллюстрируем изложенное выше простейшим примером, взяв $X = R$ и $f(x) = x^2$. В случае *а* на рис. 10.1 $g_1(x) = -1 - x$, $g_2(x) = x - 1$, а функция Лагранжа имеет вид

$$L(x, \lambda) = x^2 - \lambda_1(1 + x) + \lambda_2(x - 1). \tag{10.1.14}$$

Поскольку в точке минимума $x_* = 0$ оба ограничения выполняются как строгие неравенства, то согласно (10.1.10) $\lambda_1^* = 0$ и $\lambda_2^* = 0$. Проверим необходимые и достаточные условия (10.1.9). В нашем случае

$$L(x_*, \lambda) = L(0, \lambda) = -\lambda_1 - \lambda_2,\ L(x_*, \lambda^*) = L(0, 0) = 0,$$
$$L(x, \lambda^*) = L(x, 0) = x^2,$$

и условия (10.1.9) превращаются в очевидные (при $\lambda_1 \geqslant 0,\ \lambda_2 \geqslant 0$) неравенства: $-(\lambda_1 + \lambda_2) \leqslant 0 \leqslant x^2$. В случае *б* $g_1(x) = 1/2 - x$, $g_2(x) = x - 1$, а функция

Лагранжа принимает вид

$$L(x,\lambda) = x^2 + \lambda_1\left(\frac{1}{2} - x\right) + \lambda_2(x-1). \tag{10.1.15}$$

Теперь $x_* = 1/2$, и в строгое неравенство обращается только второе ограничение. Поэтому следует положить $\lambda_2^* = 0$, а λ_1^* определить из условия, что $L(x,\lambda^*)$ достигает максимума при $x = x_* = 1/2$. Ввиду дифференцируемости всех функций это приводит нас к системе

$$\lambda_2^* = 0,$$
$$\frac{\partial}{\partial x}\left[x^2 + \lambda_1^*\left(\frac{1}{2} - x\right) + \lambda_2^*(x-1)\right]\Bigg|_{x=1/2} = 0,$$

откуда легко находим, что $\lambda_1^* = 1$. Таким образом, условие (10.1.9) выглядит следующим образом: $1/4 - 1/2 \cdot \lambda_2 \leqslant 1/4 \leqslant x^2 + 1/2 - x$. Левое неравенство выполняется в силу неотрицательности при всех x выражения $x^2 - x + 1/4 = (x - 1/2)^2$.

Применим теперь для решения этих задач понятие двойственности. Функции (10.1.14) и (10.1.15) достигают минимума по x при любых λ_1 и λ_2.

Для вычисления функции φ в первой задаче нужно при фиксированных λ_1 и λ_2 найти минимум по x (так как $X = R$) выражения (10.1.14). Приравнивая нулю производную по x, найдем, что следует положить $x = (\lambda_1 - \lambda_2)/2$, поэтому

$$\varphi(\lambda) = -\left[\frac{(\lambda_2 - \lambda_1)^2}{4} + \lambda_1 + \lambda_2\right]. \tag{10.1.16}$$

Среди неотрицательных λ максимум функции (10.1.16), как нетрудно видеть, доставляют значения $\lambda_1^* = \lambda_2^* = 0$. Это и есть решение двойственной задачи, причем $\varphi(0) = 0 = f(0)$, т. е. мы пришли к соотношению (10.1.13). Совершенно так же для функции (10.1.15) найдем, что

$$\varphi(\lambda) = -\left[\frac{(\lambda_2 - \lambda_1)^2}{4} + \lambda_2 - \frac{1}{2}\lambda_1\right]. \tag{10.1.17}$$

Двойственная задача теперь состоит в максимизации величины (10.1.17) при ограничениях $\lambda_1 \geqslant 0$, $\lambda_2 \geqslant 0$. Можно проверить, что этот максимум достигается при $\lambda_1 = 1$ и $\lambda_2 = 0$, причем снова выполнено соотношение (10.1.13).

10.2. Линейное программирование

Задача линейного программирования состоит в нахождении максимума или минимума линейной функции при конечном числе линейных ограничений. Эта задача возникает во многих приложениях. Она же обычно является составной частью методов оптимизации в нелинейном случае при поэтапной линеаризации задачи. Для задачи линейного программирования принято несколько канонических форм записи. Для многих практических задач условие $x_i \geqslant 0$ ($i = 1, 2, \ldots, n$) является естественным, поэтому широко распространена следующая форма:

$$\begin{aligned} &\sum_{j=1}^{n} c_j x_j \to \min, \\ &\sum_{j=1}^{n} a_{ij} x_j \geqslant b_i, \quad i = 1, 2, \ldots, l, \\ &x_i \geqslant 0, \quad i = 1, 2, \ldots, n. \end{aligned}$$

Однако при дальнейшем изложении нам удобнее пользоваться формой

$$\begin{aligned} &\sum_{j=1}^{n} c_j x_j \to \min, \\ &\sum_{j=1}^{n} a_{ij} x_j \geqslant b_i, \quad i = 1, 2, \ldots, m, \quad m \geqslant n, \end{aligned} \tag{10.2.1}$$

включив условия неотрицательности (если они присутствуют) в общий список неравенств. Введя в рассмотрение векторы $b = (b_1, \ldots, b_m)^T$, $c = (c_1, \ldots, c_n)^T$ и $x = (x_1, x_2, \ldots, x_n)^T$ и матрицу $A = (a_{ij})$ ($i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$), задачу (10.2.1) можно переписать в виде

$$\min\{(c, x) : Ax \geqslant b\}. \tag{10.2.2}$$

Применим результаты предыдущего параграфа для построения двойственной задачи. Для этого заметим, что в нашем случае

$$f(x) = (c, x), \quad g_i(x) = b_i - (Ax)_i, \quad i = 1, 2, \ldots, m.$$

Поэтому функция Лагранжа для поставленной задачи линейного программирования принимает вид

$$L(x,\lambda)=\sum_{j=1}^{n}c_jx_j+\sum_{i=1}^{m}\lambda_i\left(b_i-\sum_{j=1}^{n}a_{ij}x_j\right) \tag{10.2.3}$$

или, если ввести вектор множителей Лагранжа $\lambda=(\lambda_1,\lambda_2,\ldots,\lambda_m)^T$, то

$$L(x,\lambda)=(c,x)+(\lambda,b-Ax)=(\lambda,b)+(x,c-A^T\lambda). \tag{10.2.4}$$

Функция $\varphi(\lambda)$ из (10.1.12) в наших условиях имеет вид

$$\varphi(\lambda)=\inf\{(\lambda,b)+(c-A^T\lambda,x):x\in\mathbf{R}^n\},\quad \lambda\geqslant 0.$$

Поэтому, в силу произвольности x, конечное значение функция φ может принимать лишь при λ, удовлетворяющих системе линейных уравнений

$$c-A^T\lambda=0$$

или

$$\sum_{i=1}^{m}a_{ij}\lambda_i=c_j,\quad j=1,2,\ldots,n. \tag{10.2.5}$$

Таким образом, двойственной для задачи (10.2.1) оказывается следующая задача линейного программирования:

$$\begin{aligned}&\sum_{i=1}^{m}b_i\lambda_i\to\min,\\&\sum_{i=1}^{m}a_{ij}\lambda_i\geqslant c_j,\quad j=1,2,\ldots,n,\\&\lambda_i\geqslant 0,\quad i=1,2,\ldots,m,\end{aligned} \tag{10.2.6}$$

или

$$\max\{(\lambda,b):A^T\lambda=c,\lambda\geqslant 0\}. \tag{10.2.7}$$

Рассмотрим теперь вопрос о численном решении задачи линейного программирования, которую в дальнейшем будем считать невырожденной. Прежде всего, это означает, что система уравнений (10.2.5) линейно независима ($\operatorname{rang}A=n$), а система ограничений задачи (10.2.1) имеет угловые точки. Каждая такая точка удовлетворяет всем неравенствам задачи (10.2.1), и некоторые n из этих неравенств с линейно независимыми левыми частями обращаются в равенства, т. е. для угловой точки $\bar{x}$ найдется

такое подмножество индексов $I \equiv \{i_1, i_2, \ldots, i_n\}$, что

$$\begin{aligned} \sum_{j=1}^{n} a_{ij}\bar{x}_j &= b_i, \quad i \in I, \\ \sum_{j=1}^{n} a_{ij}\bar{x}_j &\leqslant b_i, \quad i \notin I, \end{aligned} \tag{10.2.8}$$

причем квадратная подматрица (a_{ij}) $(i \in I, j = 1, 2, \ldots, n)$ неособенная. Если при некотором $i \notin I$ в (10.2.8) имеет место равенство, то такая угловая точка называется вырожденной. В невырожденной задаче линейного программирования все ее угловые точки предполагаются невырожденными (т. е. при $i \notin I$ в (10.2.8) имеют место строгие неравенства). Формально описываемый ниже симплекс-метод может применяться и к вырожденным задачам, но гарантировать его конечность можно лишь при отсутствии ситуации вырождения.

Основой симплекс-метода является то обстоятельство, что линейная функция достигает максимума или минимума в угловой точке многогранного тождества. Поэтому можно организовать перебор этих угловых точек (их конечное число) и выбрать ту, что нас интересует. Поскольку в реальных задачах число угловых точек очень велико, то полный их перебор оказывается нереализуемым. Симплекс-метод и состоит в направленном переборе угловых точек, при котором практически приходится перебирать лишь их незначительную часть.

Итак, пусть выбрана угловая точка $\bar{x}$ и соответствующее множество $I \equiv \{i_1, i_2, \ldots, i_n\}$. Так как в седловой точке должны выполняться условия дополнительности (10.1.10), а это при сделанном выше предположении о невырожденности всех угловых точек равносильно равенству $\lambda_i = 0$, $\notin I$, будем искать именно такие векторы $\bar{\lambda}$, которые удовлетворяют соотношениям

$$\bar{\lambda}_i = 0, \ i \notin I.$$

Согласно (10.2.5), остальные компоненты $\bar{\lambda}_i$, $i \in I$, должны удовлетворять системе

$$\sum_{i \in I} a_{ij}\lambda_i = c_j, \quad j = 1, 2, \ldots, n, \tag{10.2.9}$$

матрица которой транспонирована по отношению к матрице системы (10.2.8) и, следовательно, вместе с последней — квадратная и неособенная. Это приводит к следующему правилу проверки $\bar{x}$ на оптимальность: нужно решить систему (10.2.9) и проверить неотрицательность ее единственного решения. Если $\bar{\lambda}_i \geqslant 0$ при всех $i \in I$, то согласно предыдущему пункту пара $(\bar{x}, \lambda)$ является седловой точкой функции Лагранжа, и задача решена: $x_* = \bar{x}$

— решение задачи (10.2.6). Если же для некоторого $i_k \in I$ оказалось $\bar{\lambda}_{i_k} < 0$, то следует перейти к другой угловой точке, а следовательно, и к другому множеству индексов для системы (10.2.8). Опишем этот переход. Исключим из множества I номер i_k, для которого оказалось $\bar{\lambda}_{i_k} < 0$. Тогда вместо системы (10.2.8) получим систему

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, \quad i \in \{i_1, \ldots, i_{k-1}, i_{k+1}, \ldots, i_n\},$$

множеством решений которой является прямая, проходящая через точку $\bar{x}$. Направляющий вектор z удовлетворяет соответствующей однородной системе и после добавления условия нормировки может быть определен из квадратной неособенной системы

$$\sum_{j=1}^{n} a_{ij} z_j = 0, \quad i \in \{i_1, \ldots, i_{k-1}, i_{k+1}, \ldots, i_n\}, \tag{10.2.10}$$

$$\sum_{j=1}^{n} a_{i_k j} z_j = 1.$$

Выбором условия нормировки мы распорядились так, что система (10.2.10) лишь правой частью отличается от системы (10.2.8). Поскольку

$$(c, z) = \sum_{j=1}^{n} c_j z_j = \sum_{j=1}^{n} \left(\sum_{i=1}^{m} a_{ij} \bar{\lambda}_i \right) z_j = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} a_{ij} z_j \right) \bar{\lambda}_i = \bar{\lambda}_{i_k} < 0,$$

то при возрастании параметра ε от нуля до $+\infty$ значение минимизируемой функции на точках $\bar{x} + \varepsilon z$ строго убывает:

$$(c, \bar{x} + \varepsilon z) = (c, \bar{x}) + \varepsilon \bar{\lambda}_{i_k}.$$

Возможны две ситуации. Если

$$\sum_{j=1}^{n} a_{ij} z_j \geqslant 0, \quad i = 1, 2, \ldots, m,$$

то вместе с $\bar{x}$ системе ограничений задачи (10.2.1) удовлетворяют и точки луча

$$\{\bar{x} + \varepsilon z : \varepsilon \geqslant 0\}. \tag{10.2.11}$$

Поскольку вдоль этого луча минимизируемая функция убывает до $-\infty$, то задача линейного программирования не имеет решения ввиду неограниченности снизу минимизируемой функции (двойственная задача при этом

имеет несовместную систему ограничений). Во втором случае, когда множество

$$S = \{i : \sum_{j=1}^{n} a_{ij} z_j < 0\}$$

непусто, вдоль луча (10.2.11), не покидая допустимого множества, можно сдвинуться лишь до значения

$$\bar{\varepsilon} = \min\{\Delta_i / g_i : i \in S\}, \tag{10.2.12}$$

где для краткости введены обозначения

$$\Delta_i = \sum_{j=1}^{n} a_{ij} x_j - b_i, \quad g_i = -\sum_{j=1}^{n} a_{ij} z_j.$$

Заметим, что $S \cap I = \emptyset$ и при $i \in S$ всегда $g_i > 0$, $\Delta_i > 0$. Если минимум в (10.2.12) достигается при $i' \in S$, то новая угловая точка

$$\bar{\bar{x}} = \bar{x} + \bar{\varepsilon} z$$

и соответствующее ей новое множество

$$I = \{i_1, \ldots, i_{k-1}, i', i_{k+1}, \ldots, i_n\},$$

получающееся из I заменой элемента i_k на элемент i', могут быть приняты в качестве исходных для следующего шага метода. Значение минимизируемой функции при таком переходе уменьшилось на величину $(-\bar{\varepsilon}\bar{\lambda})$, которая строго положительна при $\bar{\varepsilon} > 0$ (что гарантировано в невырожденном случае). Таким образом, в невырожденном случае значение минимизируемой функции от шага к шагу строго убывает, угловые точки не могут повторяться, и, ввиду конечности их числа, метод оказывается также конечным.

Проиллюстрируем описанный метод на примере задачи

$$\begin{aligned} x_1 + x_2 \to \min,\ -x_1 &\geqslant -1, \\ x_1 &\geqslant 0, \\ -x_2 &\geqslant -1, \\ x_2 &\geqslant 0. \end{aligned}$$

Здесь допустимое множество D — квадрат с узлами $(0,0)$, $(0,1)$, $(1,0)$ и $(1,1)$, а решением задачи является начало координат. Пусть нами выбрана начальная вершина $(1,1)$. Этой вершине соответствует множество $I = \{1,3\}$, так как именно первое и третье ограничения обращены в равенства. Систе-

ма (10.2.9) для определения $\bar{\lambda}_1$ и $\bar{\lambda}_3$ имеет вид

$$-\lambda_1 = 1,$$
$$-\lambda_3 = 1,$$

так что $\bar{\lambda}_1 = \bar{\lambda}_3 = -1$, и в качестве i_k можно взять как индекс 1, так и индекс 3. Положив $i_k = 1$, придем к системе

$$-z_1 = 1,$$
$$-z_2 = 0,$$

в которую превращается система (10.2.10). Таким образом, $z = (-1, 0)$, $g_2 = 1$, $g_4 = 0$, $S = \{2\}$ и согласно (10.2.12)

$$\bar{\varepsilon} = \Delta_2 / g_2 = 1/1 = 1.$$

Поскольку множество S оказалось одноэлементным, то минимум (10.2.12) фактически выбирать не пришлось и нужно принять единственную возможность $i' = 2$. В результате получаем

$$\bar{\bar{x}} = \bar{x} + \bar{\varepsilon} z = (1, 1) + 1 \cdot (-1, 0) = (0, 1), \ \bar{I} = \{2, 3\}.$$

Для нового множества $\bar{I}$ получаем систему (10.2.9):

$$\lambda_2 = 1, \quad -\lambda_3 = 1,$$

так что $i_k = 3$. Вектор z теперь определится из системы

$$z_1 = 0, \quad -z_2 = 1,$$

получим $z = (0, -1)$, $g_1 = 0$, $g_4 = 1$. Поскольку $\Delta_4 = 1$, то $\bar{\varepsilon} = 1$, $i' = 4$ и

$$x_* = \bar{\bar{x}} + \bar{\varepsilon} z = (0, 1) + 1 \cdot (0, -1) = (0, 0), \quad I^* = \{2, 4\}.$$

Наконец, определив по множеству I^* из системы

$$\lambda_2 = 1, \quad \lambda_4 = 1,$$

что $\lambda_2^* = \lambda_4^* = 1 \geqslant 0$, заключаем, что $x_* = (0, 0)$ является решением поставленной задачи. Набор $\lambda = (0, 1, 0, 1)$ при этом решает соответствующую двой-

ственную задачу:

$$-\lambda_1 - \lambda_3 \to \max,$$
$$-\lambda_1 + \lambda_2 = 1,$$
$$-\lambda_3 + \lambda_4 = 1,$$
$$\lambda_1, \lambda_2, \lambda_3, \lambda_4 \geqslant 0.$$

В заключение пункта отметим, что для нахождения исходной угловой точки (если она неизвестна из каких-либо дополнительных соображений) разработаны специальные приемы. Кроме того, симплекс-метод может применяться к задачам линейного программирования в другой форме, а для его реализации могут использоваться различные вычислительные схемы[2)].

10.3. Квадратичное программирование

Задача выпуклого программирования состоит в минимизации многочлена второго порядка при наличии конечного числа линейных ограничений на неизвестные. При этом обычно предполагается, что квадратичная форма, входящая в минимизируемую функцию, положительно полуопределена, что обеспечивает выпуклость этой функции. Для упрощения дальнейшего изложения мы потребуем, чтобы эта квадратичная форма была положительно полуопределена. Хотя задача квадратичного программирования является несколько более сложной, чем задача линейного программирования, однако и здесь, как мы увидим, можно построить конечный метод решения, применив подходы, сходные с использованными в предыдущем пункте.

Нам будет удобно рассматривать задачу квадратичного программирования в следующей форме:

$$f(x) = \frac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{n} c_{jk}x_j x_k + \sum_{j=1}^{n} d_j x_j \to \min,$$

$$\sum_{j=1}^{n} a_{ij}x_j \geqslant b_i, \quad i = 1, 2, \ldots m,$$

или в матричном виде:

$$f(x) = \frac{1}{2}(x, Cx) + (d, x) \to \min, \quad Ax \geqslant b. \tag{10.3.1}$$

[2)]Более подробно с вопросом можно ознакомиться по монографиям В. А. Булавского, Р. А. Звягиной, М. А. Яковлевой [21]; Д. Б. Юдина, В. Г. Гольштейна [21].

Здесь $C = (c_{jk})$ — симметричная положительно определенная $n \times n$ матрица, $A = (a_{ij})$ — $m \times n$ — матрица, d — вектор размерности n и b — вектор размерности m.

Функция Лагранжа в нашем случае имеет вид

$$L(x, \lambda) = \frac{1}{2}(x, Cx) + (d, x) + (\lambda, b - Ax), \tag{10.3.2}$$

при этом седловую точку нужно искать при $\lambda \geqslant 0$ и свободном x (как и в предыдущем пункте, $X = \mathbf{R}^n$). Поэтому признак оптимальности (10.1.9) (правое неравенство) приводит к условию

$$\frac{\partial}{\partial x} L(x, \lambda^*)\big|_{x=x^*} = 0,$$

или

$$Cx_* + d - A^T\lambda^* = 0. \tag{10.3.3}$$

Левое же неравенство признака оптимальности (10.1.9), как и раньше, сводится к допустимости x_* и выполнению условий дополнительности (10.1.10). В нашем случае это означает, что

$$Ax_* \geqslant b, \tag{10.3.4}$$

$$\lambda_i^*(b_i - (Ax_*)_i) = 0, \quad i = 1, 2, \ldots, m. \tag{10.3.5}$$

Кроме того, множители Лагранжа должны быть неотрицательны:

$$\lambda_i^* \geqslant 0, \quad i = 1, 2, \ldots, m. \tag{10.3.6}$$

Таким образом, если мы найдем пару (x_*, λ^*), удовлетворяющую условиям (10.3.3)—(10.3.6), то x^* будет решением задачи (10.3.1), а λ^* — решением соответствующей двойственной задачи.

Опишем теперь метод решения, аналогичный методу, рассмотренному в предыдущем пункте. В общих чертах метод основан на следующих соображениях. Многогранное множество, описываемое системой ограничений задачи (10.3.1), разбивается на грани различной размерности (включая внутренность — грань максимальной размерности).

Каждая грань состоит из допустимых точек, удовлетворяющих некоторой системе уравнений вида

$$A_I x = b_I. \tag{10.3.7}$$

Здесь через A_I обозначена матрица, составленная из строк матрицы A с номерами $i \in I \equiv \{i_1, i_2, \ldots, i_I\}$, то есть

$$A_I = \{(A)_i, i \in I\}.$$

Аналогично составлен столбец b_I. Уменьшив, если нужно, множество I, можно считать строки A_I линейно независимыми. Наличие граней, которые описываются линейной системой (10.3.7) с зависимыми уравнениями, мы снова будем считать вырождением, и конечность метода в этом случае обосновываться не будет (так же, как это было для линейного программирования).

Поскольку искомая точка x_* принадлежит (относительной) внутренности некоторой грани (в частности, это может быть нульмерная грань, то есть угловая точка), то она должна доставлять минимум нашей квадратичной функции на всем множестве решений системы (10.3.7). Поэтому можно поступить следующим образом. Будем перебирать различные подсистемы (10.3.7), находить на множестве их решений минимум квадратичной функции и проверять, удовлетворяет ли этот минимум ограничениями и признаку оптимальности. Как и в случае линейного программирования перебор множеств I должен быть направленным, чтобы избежать полного перебора всех граней (это сделало бы метод нереализуемым). Кроме того, намеченная схема корректна лишь для случая положительно определенной матрицы C, так как иначе для некоторых множеств I минимум может и не достигаться.

Перейдем теперь к формальному описанию метода. Предположим, что имеется допустимая точка $\bar{x}$ и выделено множество $I = \{i_1, i_2, \ldots, i_l\}$, для которого строки матрицы A_I линейно независимы, а $\bar{x}$ удовлетворяют системе (10.3.7).

Возможны следующие случаи:

а) точка $\bar{x}$ доставляет минимум функции f при ограничениях (10.3.7). Согласно классическому признаку Лагранжа, в этом случае существуют множители $\bar{\lambda}_I = \{\bar{\lambda}_i, i \in I\}$, для которых

$$\frac{\partial}{\partial x}\left[f(x) + \bar{\lambda}_I(b_I - A_I x)\right]_{x=\bar{x}} = 0,$$

или, в нашем случае,

$$C\bar{x} + d - A_I^T \bar{\lambda}_I = 0. \tag{10.3.8}$$

Заметим, что в силу линейной независимости строк матрицы A_I равенством (10.3.8) вектор $\bar{\lambda}_I$ определяется однозначно. Если найденные $\bar{\lambda}_I$ неотрицательны, то, определив $\bar{\lambda}_i = 0$ при $i \notin I$, найдем, что при $x_* = \bar{x}$ и

$\lambda^* = \bar{\lambda}$ выполнены все соотношения (10.3.3)—(10.3.6), так что решение найдено. Если же при некотором $i_k \in I$ оказалось, что $\bar{\lambda}_{i_k} < 0$, то этот индекс из множества I удаляется, и для той же точки $\bar{x}$ мы получаем новое множество индексов $\bar{I} = \{i_1, \ldots, i_{k-1}, i_{k+1}, \ldots, i_l\}$, т. е. переходим к новой грани, размерность которой на единицу больше исходной;

б) функция f достигает минимума на множестве решений системы (10.3.7) в точке $x_0 \neq \bar{x}$. В этом случае мы попытаемся сдвинуться в точку x_0 или, если помешают ограничения задачи (10.3.1), как можно ближе к x_0. Алгоритмически это осуществляется так. Положим $z = x_0 - \bar{x}$,

$$g_i = -(Az)_i = -\sum_{j=1}^{n} a_{ij} z_j, \qquad i \notin I, \tag{10.3.9}$$

$$\Delta_i = (A\bar{x})_i - b_i = \sum_{j=1}^{n} a_{ij} \bar{x}_j - b_i, \quad i \notin I, \tag{10.3.10}$$

и, определив ε_0 по формуле

$$\varepsilon_0 = \min\{\Delta_i / g_i : g_i > 0\}, \tag{10.3.11}$$

положим $\bar{\varepsilon} = \min\{\varepsilon_0, 1\}$.

Если $\bar{\varepsilon} = 1$, то мы сместились в точку x_0 и множество I в дальнейшем сохраним. Если же $\bar{\varepsilon} < 1$, то номер i', на котором реализовался (10.3.11), мы включим в множество I, т. е. перейдем к новому множеству индексов $I = \{i_1, i_2, \ldots, i_l, i'\}$. В обоих случаях точка $\bar{x}$ заменяется на точку $\bar{x} + \bar{\varepsilon} z$.

Отметим, что для нахождения точки x_0 нужно решить систему

$$\begin{aligned} Cx + d - A_I^T \lambda_I &= 0, \\ A_I x &= b_I, \end{aligned} \tag{10.3.12}$$

которая в силу положительной определенности матрицы C и линейной независимости строк матрицы A_I имеет неособенную матрицу.

Метод решения в целом состоит из последовательности шагов перечисленных двух типов. Поскольку на шаге б) число элементов в множестве I возрастает, то между двумя соседними шагами типа а) может располагаться лишь конечное число шагов типа б). С другой стороны, для каждого шага типа а) значение функции f в точке $\bar{x}$ однозначно определяется множеством I, и — если окажется, что от одного шага к другому функция f строго убывает, то одно и то же множество I на шагах типа а) не может встретиться дважды, и метод оказывается конечным. Потребуем, чтобы выполнялось следующее условие невырожденности: через точку минимума функции f на множестве решений каждой системы (10.3.7) не проходит

граничных гиперплоскостей — ограничений с номерами $i \notin I$. В этом случае вслед за каждым шагом типа а) должен следовать шаг типа б) с $\bar{\varepsilon} > 0$ (поскольку все $\Delta_i > 0$ при $i \notin I$), и значение функции f на этом шаге строго убывает (мы фактически смещаемся из точки $\bar{x}$). Поскольку на остальных шагах значение функции f не возрастает, то конечность метода обеспечена.

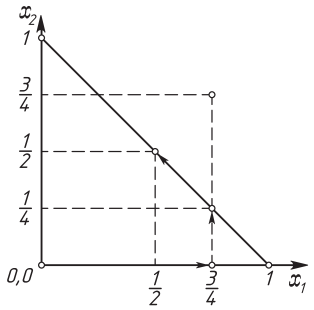


Рис. 10.2.

Проиллюстрируем метод на следующем примере:

$$\min\left\{\frac{1}{2}\left(x_1-\frac{3}{4}\right)^2+\frac{1}{2}\left(x_2-\frac{3}{4}\right)^2 : x_1 \geqslant 0, x_2 \geqslant 0, -x_1-x_2 \geqslant -1\right\}. \qquad (10.3.13)$$

Начав с точки $(0,0)$, что соответствует $I=\{1,2\}$ и системе $x_1=0$, $x_2=0$, мы совершаем шаг типа а), так как наша грань состоит из единственной точки и в точке $(0,0)$ тривиально достигается минимум функции f. Поскольку в нашем случае

$$C=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}, \quad d=\begin{pmatrix}-3/4\\ -3/4\end{pmatrix},$$

$$A_I=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}, \quad b_I=\begin{pmatrix}0\\ 0\end{pmatrix},$$

то из системы (10.3.8) найдем, что $\lambda_1=\lambda_2=-3/4$, так что оба индекса в I можно взять в качестве i_k. Положим $i_k=1$ и перейдем к следующему шагу с $\bar{x}=(0,0)$ и $I=\{2\}$. Теперь система (10.3.7) будет совпадать с осью абсцисс. Минимум функции f на этом множестве достигается в точке $(3/4,0)$, удовлетворяющей системе ограничений. Поэтому в результате шага типа б) мы сместимся в эту точку, сохранив $I=\{2\}$. В результате следующим снова оказывается шаг типа а), причем система (10.3.8) принимает вид

$$\frac{3}{4}-\frac{3}{4}=0, \quad 0-\frac{3}{4}-\bar{\lambda}_2=0,$$

откуда получаем, что $\bar{\lambda}_2 = -3/4$, поэтому следует положить $i_k = 2$. К следующему шагу мы переходим с $\bar{x} = (3/4, 0)$ и $I = \emptyset$. В соответствии с этим минимум функции f нужно искать на всей плоскости, и поэтому $x_0 = (3/4, 3/4)$. Однако точка x_0 не удовлетворяет третьему ограничению. Поэтому, двигаясь по направлению к ней, мы определим $z = (0, 3/4)$ и при $\bar{\varepsilon} = 1/3$ перейдем к точке $\bar{x} = (3/4, 1/4)$ и множеству $I = \{3\}$. Мы снова должны совершить шаг типа б), поскольку точка минимума $x_0 = (1/2, 1/2)$ на множестве решений уравнения $x_1 + x_2 = 1$ не совпадает с $\bar{x}$. Поскольку, однако, эта точка является допустимой, то окажется, что $\varepsilon = 1$, и мы получим решение задачи $(1/2, 1/2)$. Получаемая последовательность точек изображена на рис. 10.2. Стрелками указаны переходы от одной точки к другой.

10.4. Численные методы для задачи выпуклого программирования

Здесь мы рассмотрим бесконечные итерационные методы, рассчитанные в основном на более общий случай, чем рассмотренные в двух предыдущих пунктах. При этом для каждого из приводимых методов мы изложим лишь основную схему и проиллюстрируем их на примере (10.3.13) предыдущего пункта. Всюду функция считается дифференцируемой.

1. *Метод условного градиента.* Метод состоит в следующем. Для данной точки $\bar{x}$ из допустимого множества D линеаризируем минимизируемую функцию и рассмотрим задачу

$$\min\{f(\bar{x}) + (f'(\bar{x}), x - \bar{x}) : x \in D\}. \tag{10.4.1}$$

Если x^0 — решение этой задачи, то на отрезке, соединяющем точки $\bar{x}$ и x^0, ищется точка минимума функции f, то есть решается следующая задача одномерной минимизации:

$$\min\{f(\bar{x} + \alpha(x^0 - \bar{x})) : \alpha \in [0, 1]\}. \tag{10.4.2}$$

Если этот минимум реализуется при $\alpha = \bar{\alpha}$, то в качестве следующего приближения берется точка

$$\bar{x} + \bar{\alpha}(x^0 - \bar{x}). \tag{10.4.3}$$

В силу выпуклости множества D точка (10.4.3) — допустимая вместе с $\bar{x}$ и x^0. Заметим, что если множество D имеет сложное задание (функции g_i — нелинейные), то задача (10.4.1) немногим легче, чем исходная. Поэтому

практически этот метод применяется при нелинейных ограничениях, когда вся нелинейность сосредоточена в функции f. В этом случае задача (10.4.1) является задачей линейного программирования и может быть эффективно решена (в обозначениях 10.2: $c = f'(\bar{x})$; постоянное слагаемое $f(\bar{x}) + (f'(\bar{x}), \bar{x})$, конечно, не влияет на оптимизацию).

Характер работы этого метода рассмотрим на примере (10.3.13). Если в точке $\bar{x} = (\bar{x}_1, \bar{x}_2)$ оказывается, что $\bar{x}_1 > \bar{x}_2$ (т. е. эта точка расположена ниже диагонали первого квадрата), то в качестве решения задачи (10.4.1), имеющей вид (после отбрасывания постоянного слагаемого)

$$\min\left\{\left(\bar{x}_1 - \frac{3}{4}\right)x_1 + \left(\bar{x}_2 - \frac{3}{4}\right)x_2 : x_1 \geqslant 0, x_2 \geqslant 0, x_1 + x_2 \leqslant 1\right\},$$

мы получим точку $x_1^0 = 0$, $x_2^0 = 1$. Если же $\bar{x}_1 < \bar{x}_2$, то наоборот, $x_1^0 = 1$ и $x_2^0 = 0$. В соответствии с этим итерационный процесс будет идти так, как показано на рис. 10.3, где отмечены четыре последовательных точки и пути перехода от одной к другой.

2. *Метод проекции градиента.* Известно, что направление градиента $f(x)$ является направлением возрастания (если $f'(x) \neq 0$) функции f. Поэтому для нахождения минимума часто используют метод, заключающийся в смещении из очередной точки $\bar{x}$ в направлении «антиградиента» $z = -f'(\bar{x})$. При наличии ограничений такое движение может вывести нас из допустимой области. Поэтому следует предусмотреть процедуру возвращения, пример, проектирование на множестве D. Таким образом, описанная вычислительная процедура состоит в следующем. Выбирая длину шага α (каждый раз разную), вычисляем точку $x^0 = \bar{x} - \alpha f'(\bar{x})$ и затем решаем задачу

$$\min\left\{\frac{1}{2}(x - x^0, x - x^0) : x \in D\right\}. \tag{10.4.4}$$

Задача (10.4.4) при произвольном D может быть достаточно сложной. Однако если D — многогранное множество, то эта задача квадратичного программирования, которая может быть решена в конечное число действий. Заметим, что у задачи (10.4.4) есть дополнительная специфика: матрица C квадратичной формы — единичная. Это может быть учтено при решении задачи квадратичного программирования. Характер работы метода для задачи (10.3.13) изображен на рис. 10.4. Смещаясь по направлению к точке $(3/4, 3/4)$ (для нашей задачи это и есть направление антиградиента) и проектируя на допустимое множество, мы последовательно приближаемся к решению $(1/2, 1/2)$. Следует отметить, что здесь, как и вообще во многих итерационных методах, тактика выбора длины шага играет очень

существенную роль как для достижения сходимости, так и для обеспечения приемлемой скорости.

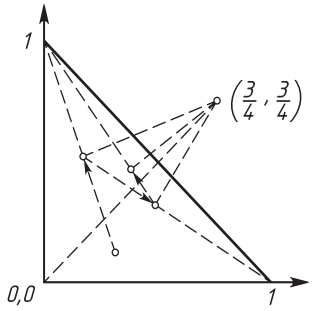


Рис. 10.3.

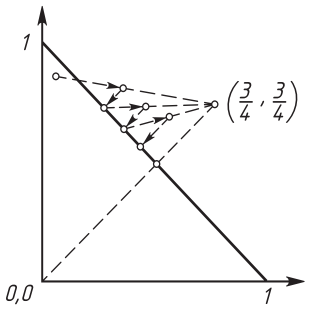


Рис. 10.4.

3. *Метод возможных направлений.* Оба предыдущих метода были в основном ориентированы на задачи с линейными ограничениями. Теперь мы познакомимся с методом, применимым и для общего нелинейного случая. Пусть имеется допустимая точка $\bar{x}$. Положим

$$S(\bar{x}) = \{i : g_i(\bar{x}) = 0\}. \tag{10.4.5}$$

Ограничения с номерами из множества (10.4.5) принято называть активными в точке $\bar{x}$. Будем искать направление z из точки $\bar{x}$, которое составляет тупой угол как с градиентом функции f, так и с внешними нормалями активных ограничений (т. е. с векторами $g_i^{'}(\bar{x}), i \in S$). Такой выбор направления z обеспечивает, с одной стороны, убывание минимизируемой функции вдоль z, а с другой, — возможность сдвинуться вдоль этого направления, не выходя за (криволинейную) границу области D. Технически эта идея реализуется путем наложения линейных ограничений

$$\begin{gathered}(f'(\bar{x}), z) + \sigma|f'(\bar{x})| \leqslant 0, \\ (g_i'(\bar{x}), z) + \sigma|g_i'(\bar{x})| \leqslant 0, \quad i \in S.\end{gathered} \tag{10.4.6}$$

Здесь $|f'(\bar{x})|$ и $|g_i'(\bar{x})|$ — евклидовы длины соответствующих векторов. Величину σ следует максимизировать. Однако, ввиду однородности ограничений (10.4), нужно еще добавить какое-нибудь условие нормировки. Чаще всего добавляют ограничение либо на евклидову норму вектора z

$$(z, z) \leqslant 1,$$

либо на кубическую норму

$$-1 \leqslant z_j \leqslant 1,\ j = 1, 2, \ldots, n.$$

В последнем случае для определения z и σ получается задача линейного программирования. Определив направление z, решают задачу одномерной минимизации:

$$\min_{\alpha}\{f(\bar{x}+\alpha z) : g_i(\bar{x}+\alpha z) \leqslant 0, i = 1, 2, \ldots, m\}.$$

Заметим, что в таком чистом виде метод возможных направлений может оказаться несходящимся. Для обеспечения сходимости нужно вместо множества в системе использовать множество

$$S = \{i : g_i(\bar{x}) \geqslant -\varepsilon\},$$

где ε — некоторое положительное (малое) число. По мере продвижения к решению число ε можно уменьшать. Качественный характер работы метода на задаче изображен на рисунке 10.5.

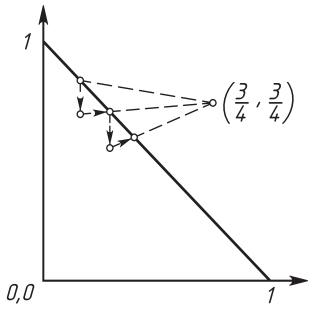


Рис. 10.5.

4. *Метод штрафных функций.* Это также метод, предназначенный для использования, вообще говоря, при нелинейных ограничениях. Основная идея состоит в том, что вместо явного учета ограничений к минимизируемой функции добавляется член, штрафующий за их нарушение. Если величина штрафа достаточно велика, то можно ожидать, что точка свободного минимума такой модифицированной функции будет не слишком нарушать ограничения. Часто эта схема реализуется следующим образом. Положим

$$\Phi(x) = \sum_{i=1}^{m} \left(\max\{g_i(x), 0\}\right)^2$$

и будем искать минимум функции

$$f(x) + r\Phi(x), \tag{10.4.7}$$

где r — некоторое достаточно большое положительное число.

Для задачи (10.4) функция Φ принимает вид

$$\Phi(x) = (\max\{-x_1, 0\})^2 + (\max\{-x_2, 0\})^2 + (\max\{x_1 + x_2 - 1, 0\})^2. \tag{10.4.8}$$

Поскольку, однако, условия неотрицательности переменных не входят в противоречие с минимизацией функции f, то в точке минимума величины (10.4.7) эти ограничения будут выполнены. Поэтому фактически в (10.4.8) первые два слагаемых в (10.4.8) будут равны нулю. Третье же слагаемое, наоборот, будет отлично от нуля, поскольку именно третье ограничение препятствует продвижению в точке абсолютного минимума функции f. Эти качественные (впрочем, не вполне строгие) рассуждения показывают, что минимум величины (10.4.7) фактически будет совпадать с минимумом функции

$$\frac{1}{2}\left(x_1 - \frac{3}{4}\right)^2 + \frac{1}{2}\left(x_2 - \frac{3}{4}\right)^2 + r(x_1 + x_2 - 1)^2.$$

Приравняв нулю частные производные по x_1 и по x_2, получим систему

$$x_1 - \frac{3}{4} + 2r(x_1 + x_2 - 1) = 0,$$

$$x_2 - \frac{3}{4} + 2r(x_1 + x_2 - 1) = 0,$$

откуда найдем, что

$$x_1 = x_2 = \frac{3/4 + 2r}{1 + 4r}.$$

При больших r имеем

$$x_1 = x_2 \approx \frac{1}{2} + \frac{1}{16r},$$

что приближенно дает решение задачи.

10.5. Динамическое программирование

В последующих двух параграфах рассматриваются задачи оптимального управления. Любая задача оптимального управления в соответствии с принятой математической моделью задается уравнением состояния и предельными условиями, описывающими поведение объекта. При этом всегда в уравнениях задачи можно выделить группу зависимых переменных, описывающих состояние объекта и группу управляющих функций, которые доступны непосредственно измерению извне и имеют значения, принадлежащие заданному множеству допустимых управлений. Задача оптимального управления состоит в том, что требуется из множества допустимых управ-

лений выбрать такие, которые придают заданному функционалу (зависящему в общем случае от решения уравнений и управлений) наименьшее возможное значение.

Будем рассматривать задачи оптимального управления на примере управляемого объекта, поведение которого описывается системой обыкновенных дифференциальных уравнений:

$$\begin{gathered}\frac{d\varphi}{dt} = f(\varphi, u),\ 0 \leqslant t \leqslant T, \\ \varphi(0) = \varphi_0,\end{gathered} \tag{10.5.1}$$

где

$$\varphi = \{\varphi_1, \ldots, \varphi_n\},\ f = \{f_1, \ldots, f_n\},\ u = \{u_1, \ldots, u_m\}.$$

Допустимыми управлениями будем считать произвольные кусочно-непрерывные измеримые функции $u = u(t)$, принимающие значения из замкнутой области $U \subset E^m$.

В классе допустимых управлений требуется найти такое управление $u(t)$ и соответствующее ему решение $\varphi(t)$ задачи (10.5.1), чтобы функционал

$$J[u] = \int\limits_0^T f_0(\varphi, u)\, dt = \min_{u \in U}. \tag{10.5.2}$$

При этом предполагается, что каждое допустимое управление определяет единственное решение задачи (10.5.1).

В основе метода динамического программирования лежит принцип оптимальности Р. Беллмана, который может быть сформулирован следующим образом.

Оптимальное управление в любой момент времени не зависит от предыстории системы и определяется только целью управления и состоянием системы в этот момент.

Если ввести обозначение

$$Q(\varphi, t) = \min_{u \in U} \int\limits_t^T f_0(\varphi, u)\, d\tau, \tag{10.5.3}$$

то из принципа оптимальности имеем

$$Q\left(\varphi(t), t\right) = \min_u \left\{ \int\limits_t^{t+\Delta t} f_0(\varphi, u)\, d\tau + \min_u \int\limits_{t+\Delta t}^T f_0(\varphi, u)\, d\tau \right\}. \tag{10.5.4}$$

Второе слагаемое в скобках по определению есть $Q(\xi+\Delta\xi, t+\Delta t)$, где $\Delta\xi = \int\limits_{t}^{t+\Delta t} f(\varphi, u)\, d\tau$. Предполагая возможным разложение обоих членов в скобках по формуле Тейлора и устремляя затем $\Delta t \to 0$, получим из (10.5.4) уравнение в частных производных (уравнение Беллмана):

$$-\frac{\partial Q}{\partial t} = \min_{u\in U}\left[f_0(\varphi, u) + \left(f(\varphi, u), \frac{\partial Q}{\partial \varphi}\right)\right], \tag{10.5.5}$$

$$Q(\varphi, T) = 0. \tag{10.5.6}$$

Пусть минимум в правой части (10.5.5) достигается лишь в единственной точке $u^* \in U$, тогда u^* есть функция от φ и $\partial Q/\partial\varphi$:

$$u^* = u^*\left(\varphi, \frac{\partial Q}{\partial \varphi}\right). \tag{10.5.7}$$

Подставляя эту функцию в уравнение (10.5.5), будем иметь нелинейную систему уравнений

$$-\frac{\partial Q}{\partial t} = f_0\left(\varphi, u^*\left(\varphi, \frac{\partial Q}{\partial \varphi}\right)\right) + \left(f\left(\varphi, u^*\left(\varphi, \frac{\partial Q}{\partial \varphi}\right)\right), \frac{\partial Q}{\partial \varphi}\right). \tag{10.5.8}$$

Если считать u^* некоторой функцией φ, t, то система (10.5.8) будет гиперболической системой уравнений, характеристики которой направлены от $t = 0$ к $t = T$. Строгое обоснование метода динамического программирования (применительно к непрерывным задачам оптимального управления) было дано В. Г. Болтянским [21], получившим необходимое и достаточное условие оптимальности в терминах $Q(\varphi, t)$.

В основе метода динамического программирования лежит идея погружения данной конкретной задачи в семейство более простых задач. Нагляднее всего это можно проиллюстрировать при выводе уравнений динамического программирования для процессов, описываемых системой разностных уравнений

$$\varphi_{i+1} = g(\varphi_i, u_i), \quad i = 0, 1, \ldots, N-1. \tag{10.5.9}$$

Здесь $\varphi_i \in E_n$ — n — мерный вектор состояния; $u_i \in E_m$ — m — мерный вектор управления.

Разностные уравнения (10.5.9) могут возникать как из физического описания процесса, так и при дискретизации системы (10.5.1). На решениях системы (10.5.9) требуется минимизировать функционал вида

$$J(u) = \sum_{i=0}^{N-1} f_0(\varphi_i, u_i) \to \min_{\{u_0, \ldots, u_{N-1}\}}. \tag{10.5.10}$$

Из постановки задачи видно, что оптимальное значение функционала, если решение задачи (10.5.9), (10.5.10) существует, зависит от начального состояния φ_0 и числа шагов N. Обозначив это оптимальное значение через $Q_N(\varphi_0)$, запишем задачу минимизации следующим образом:

$$Q_N(\varphi_0) = \min_{u_0} \min_{\{u_1,\dots,u_{N-1}\}} \left[f_0(\varphi_0, u_0) + \sum_{i=1}^{N-1} f_0(\varphi_i, u_i) \right]. \tag{10.5.11}$$

Поскольку в силу структуры системы (10.5.9) изменения $(u_1, \dots, u_{N-1})$ не влияют на φ_0 и выбор u_0, то (10.5.12) можно переписать следующим образом:

$$Q_N(\varphi_0) = \min_{u_0} \left[f_0(\varphi_0, u_0) + \min_{\{u_1,\dots,u_{N-1}\}} \sum_{i=1}^{N-1} f_0(\varphi_i, u_i) \right]. \tag{10.5.12}$$

По определению второй член в фигурных скобках есть $Q_{N-1}(\varphi_1)$, и мы получаем

$$Q_N(\varphi_0) = \min_{u_0} [f_0(\varphi_0, u_0) + Q_{N-1}(\varphi_1)]. \tag{10.5.13}$$

Рассуждая аналогичным образом, получаем следующие рекуррентные соотношения:

$$\varphi_0 \text{ задано},$$

$$Q_{N-j}(\varphi_j) = \min_{u_j \in U} [f_0(\varphi_j, u_j) + Q_{N-j-1}(\varphi_{j+1})], \tag{10.5.14}$$

$$j = 0, 1, \dots, N-2; \quad \varphi_{j+1} = g(\varphi_j, u_j),$$

$$Q_1(\varphi_{N-1}) = \min_{u_{N-1} \in U} [f_0(\varphi_{N-1}, u_{N-1})], \tag{10.5.15}$$

$$\varphi_{N-1} = g(\varphi_{N-2}, u_{N-2}).$$

Из системы (10.5.14), (10.5.15) следует, что, считая известной φ_{N-1} и решая относительно простую задачу минимизации функции m переменных, мы можем из (10.5.14) последовательно найти $u_{N-2}, \dots, u_0$ и $Q_N(\varphi_0)$. Но поскольку система (10.5.9) определяет последовательно $\varphi_1, \varphi_2, \dots, \varphi_{N-1}$, то на самом деле получается типичная для оптимального управления двухточечная краевая задача. Уравнения (10.5.14), (10.5.15), дающие необходимые и достаточные условия оптимальности управления $(u_1, \dots, u_{N-1})$, являются следствиями структуры системы (10.5.9), которая не зависит от $\varphi_{j-1}, u_{j-1}, \dots$ (так называемые марковские системы) и аддитивности функционала (10.5.10).

Рассмотрим на простом примере использование динамического программирования. Процесс описывается одним уравнением

$$\begin{gathered}\dot{\varphi} = u, \\ \varphi(0) = \varphi_0, \quad |u| < 1,\end{gathered} \tag{10.5.16}$$

и требуется минимизировать функционал

$$J[u] = \int_0^T \varphi^2\, dt \to \min_{|u|\leqslant 1}.$$

Уравнение Беллмана (10.5.6) запишется следующим образом:

$$-\frac{\partial Q}{\partial t} = \min_{|u|\leqslant 1}\left(\varphi^2 + u\frac{\partial Q}{\partial \varphi}\right). \tag{10.5.17}$$

Поскольку линейная функция достигает минимума на границе отрезка, то

$$u^* = -sign\left(\frac{\partial Q}{\partial \varphi}\right).$$

Дискретизируем задачу (10.5.16) следующим образом ($T = 5, N = 5, \tau = T/N = 1$):

$$\begin{gathered}\varphi_{i+1} = \varphi_i + \tau u_i, \\ i = 0, 1, \ldots, N-1, \quad |u_i| \leqslant 1, \\ J = \sum_{i=0}^{N-1} \tau\varphi_i^2 \to \min_{\{u_0,\ldots,u_{N-1}\}}.\end{gathered} \tag{10.5.18}$$

Выражение для $Q_1(\varphi_{N-1})$ имеет вид

$$Q_1(\varphi_4) = \min_{|u_4|\leqslant 1} \tau\varphi_4^2 = \tau\varphi_4^2, \tag{10.5.19}$$

u_4^* — любое, $|u_4^*| \leqslant 1$.

Запишем уравнение для $Q_2(\varphi_3)$:

$$Q_2(\varphi_3) = \min_{|u_3|\leqslant 1}[\varphi_3^2 + Q_1(\varphi_3 + \tau u_3)]. \tag{10.5.20}$$

Будем менять φ_i $(i = 3, 2, 1, 0)$ от $\varphi = +5$ до $\varphi = -5$ с шагом -1. При каждом φ_3 найдем из (10.5.20) u_3^* и соответствующее значение $Q_2(\varphi_3)$. Аналогичным образом найдем таблицу значений u_i^* и $Q_{5-i}(\varphi_i)$ для $i = 2, 1, 0$. Полученные значения приведены в табл. 10.1. Найдем решение задачи (10.5.18) при $\varphi = 3$. Из таблицы получаем $Q_5(3) = 14$, $u_0^*(\varphi_0 = 3) = -1$. Далее из уравнения получа-

ем $\varphi_1 = \varphi_0 + u_0^* = 3 - 1 = 2$ и из таблицы $u_1^*(2) = -1$. Действуя последовательно, получаем $\varphi_2 = 1$, $u_2^*(1) = -1$, $\varphi_3 = 0$, $u_3^*(0) = 0$, $\varphi_4 = 0$, $u_4^*(0) = 0$.

Таблица 10.1.

φ	$Q_5(\varphi_0)$	u_0^*	$Q_4(\varphi_1)$	u_1^*	$Q_3(\varphi_2)$	u_2^*	$Q_2(\varphi_{N-2})$	u_{N-2}^*	$Q_1(\varphi_{N-1})$	u_{N-1}^*
5	55	-1	54	-1	50	-1	41	-1	25	
4	30	-1	30	-1	29	-1	25	-1	16	
3	14	-1	14	-1	14	-1	13	-1	9	
2	5	-1	5	-1	5	-1	5	-1	4	
1	1	-1	1	-1	1	-1	1	-1	1	
0	0	0	0	0	0	0	0	0	0	
-1	1	1	1	$+1$	1	$+1$	1	1	1	
-2	5	1	5	$+1$	5	$+1$	3	1	4	
-3	14	1	14	$+1$	14	$+1$	13	1	0	
-4	30	1	30	$+1$	29	$+1$	25	1	16	
-5	55	1	54	$+1$	50	$+1$	41	1	25	

Приведенная таблица дает одновременно решение задачи (10.5.18) для $\varphi_0 = +5 \ldots, -5$. Как и в других областях численного анализа, при решении задач динамического программирования возникают вопросы аппроксимации, устойчивости и сходимости алгоритмов. Сложность возникающих в методе динамического программирования уравнений затрудняет их практическое использование. К тому же задачи типа (10.5.1) и (10.5.2) могут быть решены использованием более простого принципа максимума.

10.6. Принцип максимума Понтрягина

Пусть движение некоторого объекта описывается системой дифференциальных уравнений

$$\frac{d\varphi}{dt} = f(\varphi, u), \quad 0 \leqslant t \leqslant T, \tag{10.6.1}$$

с краевыми условиями

$$\varphi(0) \in S_0, \ \varphi(T) \in S_1, \tag{10.6.2}$$

где $\varphi = (\varphi_1, \ldots, \varphi_n)$, $f = (f_1, \ldots, f_n)$, $u = (u_1, \ldots, u_n)$, S_0 и S_1 — заданные многообразия, которые, в частности, могут независимо друг от друга вырождаться в точку или совпадать со всем n-мерным евклидовым пространством E_n. Пусть $U \subset E_m$ — заданное замкнутое множество и требуется определить момент времени T и такое кусочно-непрерывное управление $u = u(t) \in U$, что-

бы соответствующая траектория $\varphi = \varphi(t, u)$ удовлетворяла условиям (10.6.1), (10.6.2) и функционалу

$$J[u] = \int_0^T f_0(\varphi, u)\, dt = \min_{u \in U}. \tag{10.6.3}$$

Будем предполагать, что функции $f_i(\varphi, u)$ определены и непрерывны по совокупности переменных (φ, u) вместе со своими частными производными $\partial f_i / \partial \varphi_j$, $i = 0, 1, \ldots, n$, $j = 1, 2, \ldots, n$, а многообразия S_0 и S_1 заданы соотношениями

$$S_0 = \{\varphi : \varphi_i(0) = \varphi_i^0,\ i = 1, 2, \ldots, n\}, \tag{10.6.4}$$

$$S_1 = \{\varphi : h_k(\varphi(T)) = 0,\ k = 1, 2, \ldots, l,\ l \leqslant n\}, \tag{10.6.5}$$

где $h_k(x)$ — функции, имеющие непрерывные частные производные, причем система векторов

$$\frac{\partial h_k}{\partial x} \equiv \text{grad } h_k(x), \quad k = 1, 2, \ldots, l,$$

линейно независима для любого $x \in S_1$. В частности, если $l = n$, то из системы (10.6.5), вообще говоря, можно определить отдельные изолированные точки $\varphi = (\varphi_1, \ldots, \varphi_n)$, которые могут быть координатами правого конца траектории. Поэтому естественно считать, что случай $l = n$ соответствует задаче оптимального управления (10.6.1)—(10.6.3) с закрепленным правым концом. Наконец, при $0 < l < n$ говорят об оптимальной задаче (10.6.1)—(10.6.3) с подвижным правым концом. При сделанных предположениях размерность многообразия S_1 равна $n - l$ независимо от того, рассматриваем ли мы задачу (10.6.1)—(10.6.3) с закрепленными, подвижными или свободными концами.

Теорема 1 (принцип максимума). Пусть для управляемого объекта

$$\begin{gathered} \frac{d\varphi}{dt} = f(\varphi, u), \quad u \in U, \quad S_0 = \{\varphi(0) = \varphi^0\}, \\ S_1 = \{h_k(\varphi(T)) = 0, \quad k = 1, 2, \ldots, l\} \end{gathered} \tag{10.6.6}$$

выполнены все предположения, сформулированные выше. Пусть $\{\varphi(t), u(t)\}$, $0 \leqslant t \leqslant T$, — оптимальный процесс, переводящий объект из заданного состояния φ^0 в состояние $\varphi^1 \in S_1$, и пусть введена вспомогательная функция — функция Гамильтона

$$H(\varphi, \psi, u) = \sum_{i=0}^{n} \psi_i f_i(\varphi, u). \tag{10.6.7}$$

Тогда существует нетривиальная вектор функция

$$\psi(t) = \{\psi_0, \psi_1(t), \ldots, \psi_n(t)\}, \quad \psi_0 = \text{const} \leqslant 0,$$

удовлетворяющая системе уравнений

$$\frac{\partial \psi_i}{\partial t} = -\frac{\partial H(\varphi(t), \psi, u(t))}{\partial \varphi_i}, \quad i = 1, 2, \ldots, n, \tag{10.6.8}$$

с граничными условиями

$$\psi_i(T) = \sum_{k=1}^{l} \gamma_k \frac{\partial h_k(\varphi(T))}{\partial \varphi_i}, \quad i = 1, 2, \ldots, n, \tag{10.6.9}$$

где числа $\gamma_1, \gamma_2, \ldots, \gamma_l$ такие, что для любого момента времени t, $0 \leqslant t \leqslant T$, выполнено условие максимума

$$H\left(\varphi(t), \psi(t), u(t)\right) = \max_{u \in U} H\left(\varphi(t), \psi(t), u\right), \tag{10.6.10}$$

и, если конечный момент времени T не фиксирован, имеет место дополнительное соотношение[3)]

$$H_T = H\left(\varphi(T), \psi(T), u(T)\right) = 0. \tag{10.6.11}$$

Эта теорема занимает центральное место в теории оптимального управления. Из нее могут быть получены различные варианты принципа максимума при различных способах задания граничных условий и функционалов. Отметим еще, что если бы левый конец был подвижен, то для него имели бы место соотношения, аналогичные условиям (10.6.9) на правом конце.

Часто встречающиеся и важные в технических приложениях задачи на оптимальное быстродействие являются частным случаем задачи (10.6.1)—(10.6.3), когда $f_0(\varphi, u) \equiv 1$. Переформулируем теорему 1 для этого случая. Заметим, что функция Гамильтона принимает вид

$$\mathscr{H} = \sum_{i=1}^{n} \psi_i f_i(\varphi, u) + \psi_0 \cdot 1 = H + \psi_0. \tag{10.6.12}$$

Сопряженная система в силу $\partial \mathscr{H} / \partial \varphi_i = \partial H / \partial \varphi_i$ будет выглядеть так:

$$\frac{d\psi_i}{dt} = -\frac{\partial H}{\partial \varphi_i}, \quad i = 1, 2, \ldots, n,$$

[3)] Доказательство этой теоремы можно найти в книге: *Понтрягин Л. С.* и др. Математическая теория оптимальных процессов. — М.: Наука, 1976.

$$H = \sum_{i=1}^{n} \psi_i f_i(\varphi, u). \tag{10.6.13}$$

Условия трансверсальности (10.6.9) и (10.6.11) не меняют своего вида:

$$\psi_i(T) = \sum_{k=1}^{l} \gamma_k \frac{\partial h_k(\varphi(T))}{\partial \varphi_i}, \quad i = 1, 2, \ldots, n, \tag{10.6.14}$$

$$\mathscr{H}_T = H_T + \psi_0 = 0.$$

Так как $\psi_0 \leqslant 0$, то последнее условие можно записать в форме неравенства

$$H(\varphi(T), \psi(T), u(T)) \geqslant 0. \tag{10.6.15}$$

И, наконец, условие максимума (10.6.10) для функции $\mathscr{H} = H + \psi_0$ можно переписать следующим образом:

$$H(\varphi(t), \psi(t), u(t)) = \max_{u \in U} H(\varphi(t), \psi(t), u). \tag{10.6.16}$$

В результате приходим к следующей теореме.

Теорема 2. Если $\{u(t), \varphi(t)\}$, $0 \leqslant t \leqslant T$, — оптимальное по быстродействию решение задачи (10.6.1)—(10.6.3) (здесь $f_0(\varphi, u) \equiv 1$), то существует такая нетривиальная вектор-функция $\psi = (\psi_1, \ldots, \psi_n)$, удовлетворяющая системе уравнений (10.6.13) и условиям (10.6.14), (10.6.15), что в любой момент времени t имеет место условие максимума (10.6.16).

Нетривиальность вектора $(\psi_1, \ldots, \psi_n)$ доказывается путем несложных рассуждений.

Принцип максимума дает необходимые условия оптимальности и в сочетании с различными численными методами, позволяющими найти траектории управления, приближенно удовлетворяющие принцип максимума, является одним из важнейших средств практического решения разных задач оптимального управления[4)]. Здесь мы ограничимся разбором одного примера.

Рассмотрим управляемый объект

$$\frac{d^2\varphi}{dt^2} = u \tag{10.6.17}$$

с ограничением на управление

$$-1 \leqslant u \leqslant 1. \tag{10.6.18}$$

[4)] *Будак Б. М., Васильев Ф. П.* Приближенные методы решения задач оптимального управления. Вып. 1 (тексты лекций). — М.: МГУ, 1967.

Если ввести новые фазовые переменные

$$\varphi_1 = \varphi, \quad \varphi_2 = \frac{d\varphi}{dt},$$

то получим систему уравнений, эквивалентную (10.6.17):

$$\frac{d\varphi_1}{dt} = \varphi_2, \quad \frac{d\varphi_2}{dt} = u. \tag{10.6.19}$$

Для этого объекта поставим задачу быстрейшего попадания из произвольной точки фазового пространства $(\varphi_1^0, \varphi_2^0)$ в начало отсчета $\varphi_1 = 0$, $\varphi_2 = 0$, т. е. наш объект (10.6.17) должен на кратчайшее время прийти в начало отсчета и там остановиться.

Воспользуемся теоремой 2. Для этого выпишем функцию Гамильтона

$$H = \psi_1\varphi_2 + \psi_2 u \tag{10.6.20}$$

и сопряженную систему уравнений

$$\frac{d\psi_1}{dt} = -\frac{\partial H}{\partial \varphi_1} = 0, \quad \frac{d\psi_2}{dt} = -\frac{\partial H}{\partial \varphi_2} = -\psi_1. \tag{10.6.21}$$

Отсюда

$$\psi_1 = d_1, \quad \psi_2 = -d_1 t + d_2,$$

где d_1, d_2 — константы интегрирования. Из условия максимума (10.6.10) сразу же следует, что

$$u = \begin{cases} +1, & \text{если} \quad \psi_2 > 0, \\ -1, & \text{если} \quad \psi_2 < 0, \end{cases} \tag{10.6.22}$$

а так как ψ_2 — линейная функция, т. е. только один раз может менять свой знак, то искомое управление $u(t)$ может иметь только одну точку переключения.

Рассмотрим случай, когда $u = +1$. Интегрирование уравнений (10.6.19) дает

$$\varphi_2 = c_2 + t, \quad \varphi_1 = \frac{1}{2}(c_2 + t)^2 + c_1,$$

или

$$\varphi_1 = \frac{1}{2}\varphi_2^2 + c_1, \tag{10.6.23}$$

т. е. под действием управления $u = +1$ движение происходит по параболам (10.6.23), изображенным на рис. 10.6, причем движение происходит снизу вверх, так как $d\varphi_2/dt = +1$, т. е. φ_2 возрастает.

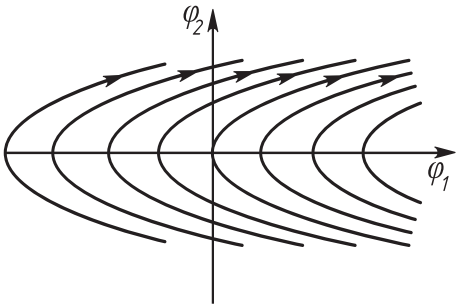


Рис. 10.6.

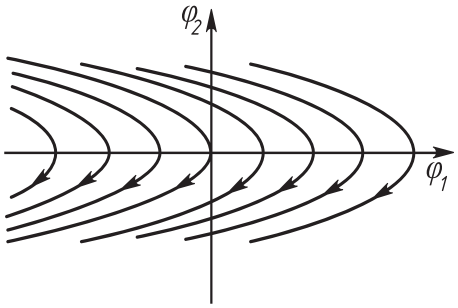


Рис. 10.7.

Аналогичные рассуждения для случая $u = -1$ показывают, что движение должно происходить по параболам вида

$$\varphi_1 = -\frac{1}{2}\varphi_2^2 + c_1, \tag{10.6.24}$$

как это изображено на рис. 10.7.

Далее, так как наш объект должен попасть в конечный момент времени в начало координат, т. е. двигаться по параболе

$$\varphi_1 = \frac{1}{2}\varphi_2^2, \text{ или } \varphi_1 = -\frac{1}{2}\varphi_2,$$

то окончательная картина возможных движений примет вид, указанный на рис. 10.8. Таким образом, если начальная точка (φ_1, φ_2) лежит выше линии AOB, то объект должен двигаться под воздействием управления $u = -1$ до тех пор, пока не попадет на параболу $\varphi_1 = 1/2\varphi_2^2$, после чего управление должно переключиться на $u = +1$, и движение будет происходить по параболе $\varphi_1 = 1/2\varphi_2^2$, пока мы не попадем в начало координат. Легко разобрать случай, когда начальная точка находится ниже кривой AOB.

Таким образом, оптимальное управление может иметь лишь следующий вид:

$$u = \begin{cases} +1, & \text{ниже} \smile AOB \text{ и на параболе } \varphi_1 = \frac{1}{2}\varphi_2^2, \\ -1, & \text{выше} \smile AOB \text{ и на параболе } \varphi_1 = -\frac{1}{2}\varphi_2^2. \end{cases}$$

Из приведенных рассуждений, зная координаты начальной точки, легко получить фазовые траектории, время движения и момент переключения управления с одного крайнего значения на другое.

Итак, мы показали, что если оптимальное решение существует, то оно должно иметь вид, изображенный на рис. 10.8, поскольку принцип максимума — необходимое условие оптимальности. Можно показать, что эти траектории действительно являются оптимальными.

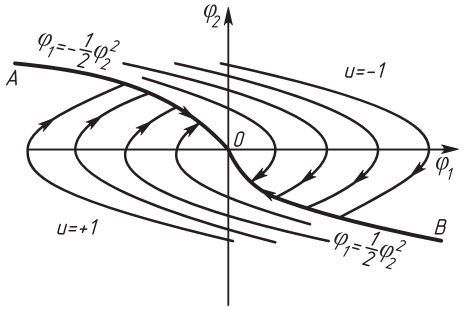


Рис. 10.8.

В заключение отметим, что классическая задача вариационного исчисления, состоящая в минимизации функционала

$$J = \int_0^T f_0\left(\varphi, \frac{d\varphi}{dt}, t\right)\, dt \tag{10.6.25}$$

на классе кусочно-гладких функций, удовлетворяющих некоторым граничным условиям

$$\varphi(t_0) \in S_0, \quad \varphi(T) \in S_1,$$

является простым частным случаем задачи (10.6.1)—(10.6.3), а именно: требуется найти минимум функционала

$$J = \int_{t_0}^T f_0\left(\varphi, u, t\right)\, dt \tag{10.6.26}$$

при условиях

$$\frac{d\varphi}{dt} = u, \quad u \in U \equiv E_n.$$

С помощью принципа максимума могут быть получены все необходимые условия для этой задачи, известные из классического вариационного исчисления: уравнения Эйлера, условия Вейерштрасса — Эрдмана, имеющие место в точках излома экстремали, условие Лежандра, условие Вейерштрасса.

Наряду со сформулированными выше задачами в теории оптимального управления рассматриваются также задачи с запаздыванием, с интегральными ограничениями, с параметрами, с дискретным временем, а также аналогичные задачи оптимального управления для уравнений с частными производными.

10.7. Экстремальные задачи с ограничениями и вариационные неравенства

В главе 2 нами были изучены различные вариационные постановки задач математической физики и приближенные методы решения соответствующих экстремальных задач. Отличительной особенностью при этом являлось рассмотрение экстремальной задачи во всем гильбертовом пространстве без каких-либо дополнительных ограничений на искомое решение. В то же время существует широкий круг проблем, возникающих в физике, механике, геофизике, оптимальном управлении и экономике, приводящих к экстремальным задачам, решение которых ищется на более узком классе функций, чем это допускает область определения соответствующих функционалов.

Ниже будут даны некоторые сведения о существовании и единственности решений экстремальных задач с ограничениями и дана характеристика этих решений с помощью вариационных неравенств, приведен ряд примеров использования вариационных неравенств для анализа свойств решений экстремальных задач с ограничениями и обсуждены возможные подходы к приближенному решению подобных задач.

10.7.1. Элементы общей теории

Пусть U — гильбертово пространство над полем действительных чисел со скалярным произведением $(\ ,\ )_U$ и нормой $\|\cdot\|_U$. Мы предположим, что на U заданы непрерывная симметричная билинейная форма $\pi(u,v)$, непрерывная линейная форма $L(v)$ и выпуклое замкнутое множество $U_d \subseteq U$. Введем в рассмотрение квадратичный функционал

$$J(v) = \pi(v,v) - 2L(v), \tag{10.7.1}$$

предполагая при этом, что функционал $\pi(v,v)$ положительно определен в U, т. е. существует такая константа $\alpha > 0$, что

$$\pi(v,v) \geqslant \alpha \|v\|_U^2 \tag{10.7.2}$$

для любого $v \in U$.

Справедливо следующее утверждение, подробное доказательство которого приведено в монографии Лионса [2].

Если выполнены сделанные выше предположения, то существует и единственен элемент $u \in U_d$, являющийся решением задачи

$$J(u) = \inf_{v \in U_d} J(v). \tag{10.7.3}$$

Одним из наиболее универсальных и развитых подходов к наглядному описанию экстремальной задачи (10.7.3) с целью изучения свойств ее решения является использование вариационных неравенств. Сформулируем и докажем утверждение, которое лежит в основе этого подхода.

Пусть выполнены сделанные выше предположения. Тогда элемент $u \in U_d$ в том и только том случае является решением задачи (10.7.3), если для любого $v \in U_d$ выполняется неравенство

$$\pi(u, v-u) \geqslant L(v-u). \tag{10.7.4}$$

Докажем необходимость. Предположим, что элемент u является решением задачи (10.7.3). Тогда для любых $v \in U_d$ и $\Theta \in (0,1)$ выполняется неравенство (здесь мы используем выпуклость U_d, т. е. принадлежность элемента $\varphi = (1-\Theta)w + \Theta v$ множеству U_d для любых $w, v \in U_d$ и $\Theta \in (0,1)$)

$$J(u) \leqslant J((1-\Theta)u + \Theta v). \tag{10.7.5}$$

Отсюда имеем

$$\frac{J(u + \Theta(v-u)) - J(u)}{\Theta} \geqslant 0. \tag{10.7.6}$$

Устремляя Θ к нулю и используя конкретный вид функционала $J(v)$, получим

$$\begin{aligned} \lim_{\Theta \to 0} \frac{J(u + \Theta(v-u)) - J(u)}{\Theta} &= \\ &= \lim_{\Theta \to 0} \{2[\pi(u, v-u) - L(v-u)] + \Theta \pi(v-u, v-u)\} = \\ &= 2[\pi(u, v-u) - L(v-u)] \geqslant 0 \end{aligned} \tag{10.7.7}$$

для любого $v \in U_d$. Тем самым неравенство (10.7.4) доказано.

Предположим теперь, что элемент u удовлетворяет неравенству (10.7.4). Квадратичный функционал $J(v)$ является выпуклым, т. е. для любых v, $w \in U$ и $\Theta \in (0,1)$ выполняется неравенство

$$J((1-\Theta)w + \Theta v) \leqslant (1-\Theta)J(w) + \Theta J(v)$$

или эквивалентное ему неравенство

$$J(v) - J(w) \geqslant \frac{J((1-\Theta)w + \Theta v) - J(w)}{\Theta}. \tag{10.7.8}$$

Подставляя в (10.7.8) элемент u вместо w, получим

$$J(v) - J(u) \geqslant \frac{J((1-\Theta)u + \Theta v) - J(u)}{\Theta}.$$

Устремляя в последнем неравенстве Θ к нулю и учитывая соотношение (10.7.7) и неравенство (10.7.4), имеем

$$J(v) - J(u) \geqslant 2[\pi(u, v-u) - L(v-u)] \geqslant 0.$$

Следовательно,

$$J(u) \leqslant J(v)$$

для любого $v \in U_d$, т. е. элемент u является решением задачи (10.7.3).

Рассмотрим два конкретных случая использования доказанного утверждения. Пусть U_d является подпространством пространства U, замкнутым в смысле нормы пространства U. Тогда, полагая в неравенстве (10.7.4) $v = u \pm \varphi$, где φ — произвольный элемент из U_d, приходим к двум неравенствам:

$$\pi(u, \varphi) \geqslant L(\varphi),$$

$$-\pi(u, \varphi) \geqslant -L(\varphi),$$

Отсюда следует, что при слдеанном предположении элемент u в том и только том случае является решением задачи (10.7.3), если для любого $\varphi \in U_d$ выполняется равенство

$$\pi(u, \varphi) = L(\varphi). \tag{10.7.9}$$

Полученное соотношение является не чем иным, как уравнением Эйлера для вариационной задачи

$$J(u) = \inf_{v \in U_d} J(v). \tag{10.7.10}$$

Предположим теперь, что множество U_d является конусом. (Напомним, что конусом называется замкнутое выпуклое множество $K \subset U$, содержащее вместе с каждым $\varphi \in K$ элемент $\gamma\varphi$ для любого $\gamma \geqslant 0$ и такое, что если φ, $-\varphi \in K$, то φ — нулевой элемент U.) Подставляя в (10.7.7) вместо

элемента v элемент $v+u$, приходим к неравенству

$$\pi(u,v) \geqslant L(v), \tag{10.7.11}$$

которое выполняется для любого $v \in U_d$. Здесь мы использовали тот факт, что вместе с любыми векторами $v, w \in U_d$ принадлежит также их сумма. Далее, полагая в (10.7.7) $v=0$, а в (10.7.11) — $v=u$, соответственно получаем

$$-\pi(u,u) \geqslant -L(u),$$

$$\pi(u,u) \geqslant L(u),$$

откуда

$$\pi(u,u) = L(u). \tag{10.7.12}$$

Покажем, что (10.7.11), (10.7.12) эквивалентны (10.7.4). Для этого достаточно показать, что (10.7.4) следует из (10.7.11), (10.7.12). Вычитая из (10.7.11) равенство (10.7.12), мы получаем

$$\pi(u,v) - \pi(u,u) \geqslant L(v) - L(u),$$

откуда, в силу билинейности функционала π и линейности функционала L, приходим к неравенству

$$\pi(u, v-u) \geqslant L(v-u),$$

которое выполняется для любого $v \in U_d$. Таким образом, если U_d — конус, то (10.7.4) и (10.7.11), (10.7.12) эквивалентны.

10.7.2. Примеры экстремальных задач

При иллюстрации результатов предыдущего пункта для наглядности ограничимся одномерным случаем, когда элементы $u \in U$ являются функциями одной переменной x, а U является пространством Соболева $W_2^1(0,1)$. Скалярное произведение и норма в U задаются соответственно соотношениями

$$(u,v)_U = \int_0^1 \left[\frac{du}{dx}\frac{dv}{dx} + uv\right] dx, \tag{10.7.13}$$

$$\|u\|_U = (u,u)_U^{1/2}.$$

Рассмотрим билинейный функционал

$$\pi(u,v) = (u,v)_U \tag{10.7.14}$$

и линейный функционал

$$L(v) = \int_0^1 fv\,dx, \tag{10.7.15}$$

где $f = f(x)$ — некоторый фиксированный элемент пространства $L_2(0,1)$. Очевидно, что $\pi(u,v)$ будет симметричным непрерывным положительно определенным функционалом, а $L(v)$ — непрерывным функционалом. Это следует из соотношений

$$\begin{gathered}\pi(u,v) \leqslant \|u\|_U \cdot \|v\|_U, \\ \pi(v,v) = \|v\|_U^2,\end{gathered} \tag{10.7.16}$$

$$L(v) \leqslant \|f\|_{L_2(0,1)} \cdot \|v\|_{L_2(0,1)} \leqslant \|f\|_{L_2(0,1)} \cdot \|v\|_U. \tag{10.7.17}$$

Таким образом, для квадратичного функционала

$$J(v) = (v,v)_U - 2\int_0^1 fv\,dx \tag{10.7.18}$$

будут выполнены все условия утверждений предыдущего пункта, и, следовательно, если $U_d \subseteq U$ — замкнутое выпуклое множество, то элемент $u \in U_d$, являющийся решением экстремальной задачи

$$J(u) = \inf_{v \in U_d} J(v), \tag{10.7.19}$$

существует и единственен. При этом он однозначно определяется неравенством

$$\pi(u, v-u) \geqslant L(v-u), \tag{10.7.20}$$

которое выполняется для любого $v \in U_d$. Необходимо отметить, что поскольку любая функция $v(x) \in W_2^1(0,1)$ в то же время принадлежит банахову пространству непрерывных функций $C(0,1)$, то решение $u(x)$ задачи (10.7.19) всегда будет непрерывной функцией. Кроме того, существует такая константа α, что для любой функции $v(x) \in U = W_2^1(0,1)$ выполняется неравенство

$$\|v\|_{C(0,1)} \leqslant \alpha \|v\|_U, \tag{10.7.21}$$

т. е. из сходимости последовательности функций $\{v_k(x)\}$ к функции $v^\infty(x)$ в норме пространства $W_2^1(0,1)$ следует сходимость последовательности $\{v_k(x)\}$

к $v^{\infty}(x)$ в норме пространства $C(0,1)$. Перейдем к рассмотрению конкретных случаев.

Пусть $U_d = U$. Тогда U_d является несобственным (совпадающим с U) подпространством U и, согласно первому примеру из 10.7.1, решение $u(x)$ задачи (10.7.19) для любого $v(x) \in U$ удовлетворяет равенству

$$(u, v)_U = L(v).$$

Отсюда с помощью стандартных рассуждений приходим к выводу, что $u(x)$ является обобщенным решением краевой задачи Неймана

$$-\frac{d^2u}{dx^2} + u = f(x), \quad 0 < x < 1,$$

$$\begin{aligned} &\frac{du}{dx} = 0 \text{ при } x = 0, \\ &\frac{du}{dx} = 0 \text{ при } x = 1. \end{aligned} \tag{10.7.22}$$

Если, например, $f(x)$ — непрерывная функция, то $u(x)$ является классическим решением этой задачи.

Предположим теперь, что $U_d = \overset{\circ}{W}{}_2^1(0,1)$. Известно, что $\overset{\circ}{W}{}_2^1(0,1)$ является замкнутым подпространством в пространстве $W_2^1(0,1)$ (так как $\overset{\circ}{W}{}_2^1(0,1)$ — полное пространство в метрике пространства $W_2^1(0,1)$). Поэтому, согласно первому примеру из 10.7.1, решение $u(x) \in \overset{\circ}{W}{}_2^1$ задачи (10.7.19) удовлетворяет для любого $v(x) \in \overset{\circ}{W}{}_2^1$ равенству

$$(u, v)_U = L(v).$$

Отсюда аналогично предыдущему приходим к выводу, что $u(x)$ является обобщенным решением краевой задачи Дирихле

$$-\frac{d^2u}{dx^2} + u = f(x), \quad 0 < x < 1, \tag{10.7.23}$$

$$u(0) = u(1) = 0.$$

Рассмотрим случай, когда множество U_d не является подпространством пространства U, т. е. имеет более сложную структуру. Определим множество U_d соотношением

$$U_d = \{u(x) : u(x) \in \overset{\circ}{W}{}_2^1, u(x) \geqslant 0 \text{ для } x \in (0,1)\}, \tag{10.7.24}$$

т. е. U_d является множеством функций из $W_2^1(0,1)$, неотрицательных на интервале $(0,1)$ и обращающихся в нуль на границах этого интервала. Пока-

жем, что U_d является конусом в $W_2^1(0,1)$. Для этого нам нужно доказать следующее: 1) U_d выпукло; 2) U_d замкнуто; 3) если $v(x) \in U_d$, то $\gamma v(x) \in U_d$ для любого $\gamma \geqslant 0$; 4) если $v(x)$ и $-v(x)$ принадлежат U_d, то $v(x) \equiv 0$.

Требование выпуклости означает, что вместе с $u(x)$ и $v(x)$ из U_d элемент $w(x) = (1-\Theta)u(x) + \Theta v(x)$ принадлежит U_d для любого $\Theta \in (0,1)$. Выполнение этого требования вытекает из неотрицательности функции $w(x)$, так как она является линейной комбинацией неотрицательных функций с положительными коэффициентами.

Докажем выполнение второго требования. Пусть $v^k(x)$, $k = 0, 1, \ldots$, — последовательность функций из U_d, сходящаяся в норме пространства $W_2^1(0,1)$ к некоторой функции $v^\infty(x) \in W_2^1(0,1)$. Так как $v^\infty(x) \in C(0,1)$, то, согласно (10.7.21), последовательность $\{v^k(x)\}$ сходится в норме пространства $C(0,1)$ к функции v^∞. Таким образом, функция $v^\infty(x)$ будет неотрицательна, обращаясь в нуль на границе интервала $(0,1)$, и, следовательно, будет принадлежать множеству U_d.

Выполнение третьего и четвертого требований устанавливается тривиально.

Таким образом, если множество U_d определено соотношением (10.7.24), то решение $u(x)$ задачи (10.7.19) существует, единственно и характеризуется тем, что для любого $v(x) \in U_d$ выполняется неравенство

$$\pi(u, v) \geqslant L(v) \tag{10.7.25}$$

и, кроме того,

$$\pi(u, u) = L(u). \tag{10.7.26}$$

Исследуем свойства решения $u(x)$ сформулированной задачи. Пусть в некоторой точке $\xi \in (0,1)$ функция $u(x)$ положительна. Тогда, поскольку $u(x)$ непрерывна, то существует такой интервал $(\xi^-, \xi^+) \subseteq (0,1)$, что функция $u(x)$ положительна на этом интервале и $u(\xi^-) = u(\xi^+) = 0$ (в частности, может оказаться, что $\xi^- = 0$ и $\xi^+ = 1$). Обозначим через $D(\xi^-, \xi^+)$ множество бесконечно дифференцируемых на $(0,1)$ функций, финитных в интервале (ξ^-, ξ^+) (т. е. тождественно равных нулю в точках $x \notin (\xi^-, \xi^+)$). Как известно, множество функций $D(\xi^-, \xi^+)$ плотно в $\overset{\circ}{W}{}_2^1(\xi^-, \xi^+)$, обладает тем свойством, что для всех достаточно малых $\varepsilon > 0$ функции

$$v_\varepsilon(x) = u(x) \pm \varepsilon\varphi(x) \tag{10.7.27}$$

принадлежит множеству U_d. Подставляя функции $v(x) = v_\varepsilon(x)$ в неравенство (10.7.26), имеем

$$\pi(u, u) \pm \varepsilon\pi(u, \varphi) \geqslant L(u) \pm \varepsilon L(\varphi). \tag{10.7.28}$$

Отсюда, используя равенство (10.7.26), получаем

$$\pi(u,\varphi) \geqslant L(\varphi), \tag{10.7.29}$$

$$-\pi(u,\varphi) \geqslant -L(\varphi), \tag{10.7.30}$$

а также, следовательно,

$$\pi(u,\varphi) = L(\varphi) \tag{10.7.31}$$

для любого $\varphi \in D(\xi^-,\xi^+)$.

Перепишем (10.7.31), учитывая, что $\varphi(x) \equiv 0$ для $x \notin (\xi^-,\xi^+)$:

$$\int\limits_{\xi^-}^{\xi^+} \left[\frac{du}{dx}\frac{d\varphi}{dx} + u\varphi\right] dx = \int\limits_{\xi^-}^{\xi^+} f\varphi\, dx, \tag{10.7.32}$$

где $\varphi(x)$ — произвольная функция из $D(\xi^-,\xi^+)$, а $u(x) \in \overset{\circ}{W}{}_2^1(\xi^-,\xi^+)$. Так как множество $D(\xi^-,\xi^+)$ плотно в $\overset{\circ}{W}{}_2^1(\xi^-,\xi^+)$, то равенство (10.7.32) будет также выполняться для любой $\varphi \in \overset{\circ}{W}{}_2^1(\xi^-,\xi^+)$. Отсюда, согласно второму примеру из настоящего пункта, вытекает, что на отрезке $[\xi^-,\xi^+]$ функция $u(x)$ будет являться обобщенным решением задачи Дирихле

$$\begin{gathered} -\frac{d^2u}{dx^2} + u = f(x),\ \xi^- < x < \xi^+, \\ u(\xi^-) = u(\xi^+) = 0. \end{gathered} \tag{10.7.33}$$

Таким образом, любой точке $\xi \in (0,1)$, где решение $u(x)$ задачи (10.7.19) оказывается положительным, ставится в соответствие некоторый интервал (ξ^-,ξ^+), на котором функция $u(x)$ будет являться обобщенным решением задачи (10.7.33).

Резюмируя изложенное выше, можно сделать вывод, что интервал $\Omega = (0,1)$ представим в виде

$$\Omega = \Omega^0 \cup \Omega^+, \tag{10.7.34}$$

где Ω^+ — открытое множество, являющееся объединением некоторого числа интервалов (возможно, счетного числа), на которых функция $u(x)$ положительна и является обобщенным решением задач вида (10.7.33), и Ω^0 — дополнение Ω^+ до Ω (возможно, пустое), на котором функция $u(x)$ тождественно равна нулю.

Рассмотренная нами задача относится к задачам со свободными границами, которые возникают, например, в механике сплошной среды. Неизвестными здесь являются как границы интервалов, составляющих множество Ω^+, так и функция $u(x)$ на этих интервалах.

В заключение рассмотрим один интересный пример, когда U_d не является ни подпространством пространства U, ни конусом в этом подпространстве. Для этого предположим, что на интервале $(0,1)$ заданы две функции: $F_1(x)$ и $F_2(x)$, принадлежащие пространству $W_2^1(0,1)$ и удовлетворяющие для всех $x \in (0,1)$ неравенству

$$F_1(x) < F_2(x) \tag{10.7.35}$$

и неравенствам

$$\begin{aligned} F_1(0) \leqslant 0 \leqslant F_2(0), \\ F_1(1) \leqslant 0 \leqslant F_2(1). \end{aligned} \tag{10.7.36}$$

Определим множество U_d соотношением

$$U_d = \{u(x) : u(x) \in \overset{\circ}{W}{}_2^1(0,1),\ F_1(x) \leqslant u(x) \leqslant F_2(x) \text{ для всех } x \in (0,1)\}. \tag{10.7.37}$$

Множество U_d состоит из функций пространства $W_2^1(0,1)$, обращающихся в нуль на границах интервала $(0,1)$ и удовлетворяющих на этом интервале неравенству

$$F_1(x) \leqslant u(x) \leqslant F_2(x). \tag{10.7.38}$$

Анализ, аналогичный использованному в предыдущем пункте, показывает, что U_d является замкнутым выпуклым множеством пространства $W_2^1(0,1)$. Отсюда и их (10.7.1) следует, что при таком выборе U_d задача (10.7.19) имеет единственное решение.

Предположим, что в некоторой точке $\xi \in (0,1)$

$$F_1(\xi) < u(\xi) < F_2(\xi). \tag{10.7.39}$$

Тогда в силу непрерывности $u(x)$ существуют такие точки $\xi^-,\ \xi^+ \in [0,1]$, что $\xi \in (\xi^-, \xi^+)$ и функция $u(x)$ на интервале (ξ^-, ξ^+) удовлетворяет неравенствам

$$F_1(x) < u(x) < F_2(x), \tag{10.7.40}$$

а на границах интервала (ξ^-, ξ^+) — краевым условиям

$$u(\xi^-) = \begin{cases} \text{либо } F_1(\xi^-), \\ \text{либо } F_2(\xi^-), \end{cases} \qquad u(\xi^+) = \begin{cases} \text{либо } F_1(\xi^+), \\ \text{либо } F_2(\xi^+). \end{cases} \tag{10.7.41}$$

Таким образом, интервал $\Omega = (0,1)$ представим в виде

$$\Omega = \Omega^0 \cup \Omega^+, \tag{10.7.42}$$

где Ω^+ является объединением интервалов, на которых функция $u(x)$ удовлетворяет неравенству (10.7.40), а Ω^0 дополняет Ω^+ до Ω, причем на Ω^0 функция $u(x)$ равна либо $F_1(x)$, либо $F_2(x)$. Заметим, что если функции $F_1(x)$ и $F_2(x)$ не обращаются в нуль на границах интервала (0,1), то Ω^0 является замкнутым множеством, т. е. является объединением конечного числа отрезков и изолированных точек, принадлежащих интервалу $(0,1)$. При этом на каждом из отрезков, входящих в Ω^0, функция $u(x)$ равна либо только функции $F_1(x)$, либо только функции $F_2(x)$.

Исследуем теперь свойства функции $u(x)$ на некотором интервале $(\xi^-,\xi^+)\subset\Omega^+$. Так как для всех достаточно малых $\varepsilon>0$ и для любой функции $\varphi(x)\in D(\xi^-,\xi^+)$ вместе с функцией $u(x)$ множеству U_d будут принадлежать функции $v_\varepsilon(x)=u(x)\pm\varepsilon\varphi(x)$, то, подставляя в неравенство (10.7.20) вместо функции $v(x)$ функцию $v_\varepsilon(x)$, приходим к неравенствам

$$\pi(u,\varphi)\geqslant L(\varphi),$$

$$-\pi(u,\varphi)\geqslant -L(\varphi).$$

Отсюда делаем вывод, что для любого $\varphi(x)\in D(\xi^-,\xi^+)$ выполняется равенство

$$\pi(u,\varphi)=L(\varphi) \tag{10.7.43}$$

или эквивалентное ему

$$\int\limits_{\xi^-}^{\xi^+}\left[\frac{du}{dx}\frac{d\varphi}{dx}+u\varphi\right]dx=\int\limits_{\xi^-}^{\xi^+}f\varphi\,dx, \tag{10.7.44}$$

где $u(x)$ удовлетворяет граничным условиям

$$\begin{aligned}u(\xi^-)&=g^-,\\ u(\xi^+)&=g^+,\end{aligned} \tag{10.7.45}$$

а g^- и g^+ определяются соотношениями

$$g^-=\begin{cases}\text{либо } F_1(\xi^-),\\ \text{либо } F_2(\xi^-),\end{cases}\qquad g^+=\begin{cases}\text{либо } F_1(\xi^+),\\ \text{либо } F_2(\xi^+).\end{cases} \tag{10.7.46}$$

Теперь на основании рассмотренных ранее примеров, заключаем, что на каждом из интервалов (ξ^-,ξ^+) функция $u(x)$ является обобщенным решением задачи Дирихле

$$-\frac{d^2u}{dx^2}+u=f(x),\quad \xi^-<x<\xi^+, \tag{10.7.47}$$

$$u(\xi^-) = g^-,$$

$$u(\xi^+) = g^+.$$

Таким образом, решение сформулированной задачи, в отличие от предыдущей, заключается не только в нахождении границ интервалов множества Ω^+ и вычислении функции $u(x)$ на этих интервалах, но и в определении значений функции $u(x)$ на границах этих интервалов. Рассмотренная задача также относится к задачам со свободными границами.

10.7.3. Численные методы для экстремальных задач

Использование того или иного численного метода решения экстремальной задачи тесно связано с конкретными свойствами квадратичного функционала $J(v)$ и множества $U_d \subseteq U$, на котором ищется решение этой задачи. Так, в случае первого и второго примеров, рассмотренных в предыдущем пункте, когда отыскание приближенного решения экстремальной задачи сводится к нахождению обобщенного или классического решения линейной краевой задачи для дифференциального уравнения, целесообразнее решать непосредственно полученную краевую задачу. Такой подход оказался весьма затруднительным с практической точки зрения для третьего и четвертого примеров, поскольку границы областей, где нужно искать решение краевой задачи, и сами краевые условия являются неизвестными. Тем не менее существует много подходов для решения указанных задач, использующих, например, регуляризацию, метод штрафов и методы нелинейного программирования. В настоящем пункте мы остановимся на другом подходе, который заключается в предварительном сведении задачи (10.7.3) к некоторой конечномерной задаче и последующем решении экстремальной задачи для квадратичной функции n переменных на выпуклом ограниченном множестве пространства n измерений.

Пусть U^h — некоторая последовательность конечномерных подпространств размерности $n = n(h)$ пространства U ($n(h) \to \infty$ при $h \to 0$), аппроксимирующих U в следующем смысле: для любого $v(x) \in U$ и любого $\Delta > 0$ существует такое h_Δ, что для всех $h < h_\Delta$ в U^h найдется элемент $v^h(x)$, аппроксимирующий $v(x)$ с точностью Δ, т. е. для всех $h < h_\Delta$ выполняется неравенство

$$\inf_{\varphi \in U^h} \|u - \varphi\| \leqslant \Lambda. \tag{10.7.48}$$

Задачу (10.7.3) заменим приближенной задачей

$$J(u^h) = \inf_{v \in U_d^h} J(v), \tag{10.7.49}$$

где $U_d^h = U_d \cap U^h$ — замкнутое выпуклое множество. Очевидно, что задача (10.7.49) также имеет решение (см. 10.7.1), которое, в силу предположения полноты последовательности U^h в U, сходится при $h \to 0$ к решению исходной задачи (10.7.3).

Предположим, что функция $\psi_k^h(x)$, $k = 1, 2, \ldots, n$, образует базис пространства U^h и для некоторого $v(x)$ справедливо соотношение

$$v(x) = \sum_{k=1}^{n} \alpha_k^h \psi_k^h(x). \tag{10.7.50}$$

При этих предположениях

$$J(v) \equiv \Phi(\alpha) = (A\alpha, \alpha) - 2(g, \alpha), \tag{10.7.51}$$

где A — симметричная положительно определенная матрица порядка n с элементами

$$a_{ij} = \pi(\psi_i^h, \psi_j^h), \tag{10.7.52}$$

g и α — векторы размерности n с компонентами

$$\begin{aligned} g_i &= L(\psi_i^h), \\ \alpha_i &= \alpha_i^h \end{aligned} \tag{10.7.53}$$

соответственно, а через $(\ ,\ )$ обозначено обычное скалярное произведение в пространстве n-мерных вещественных векторов. Теперь задача (10.7.49) принимает вид

$$\Phi(\beta) = \inf_{\alpha \in V_d} \Phi(\alpha), \tag{10.7.54}$$

где V_d определяется следующим образом: $\alpha \in V_d$ в том и только том случае, если функция

$$v(x) = \sum_{k=1}^{n} \alpha_k \psi_k^h(x) \tag{10.7.55}$$

принадлежит множеству U_d^h. Очевидно, что V_d является замкнутым выпуклым множеством пространства n-мерных вещественных векторов и задача (10.7.55) имеет единственное решение β, причем функция

$$u^h(x) = \sum_{k=1}^{n} \beta_k \psi_k^h(x) \tag{10.7.56}$$

является решением задачи (10.7.49).

В настоящее время для решения задачи (10.7.54) разработано и исследовано большое количество алгоритмов. К наиболее распространенным относятся различные варианты метода спуска, градиентные методы, методы возможных направлений, метод локальных вариаций и другие методы.

Опишем сначала один из вариантов метода релаксации.

Пусть $\alpha^{(s)} = (\alpha_1^{(s)}, \ldots, \alpha_n^{(s)}) \in V_d$ является s-м приближением к искомому решению β задачи (10.7.54). Тогда вычисление $(s+1)$-го приближения состоит из n промежуточных шагов, которые реализуются по формулам

$$\alpha^{s+\frac{i}{n}} = \alpha^{s+\frac{i-1}{n}} - q_i^s r_i^s e_i, \quad i = 1, 2, \ldots, n, \tag{10.7.57}$$

где e_i — вектор с компонентами $e_{ij} = \delta_{ij}$ $(i = 1, 2, \ldots, n)$ (δ_{ij} — символ Кронекера),

$$r_i^s = \sum_{j=1}^{n} a_{ij} \alpha_j^{s+\frac{i-1}{n}} - g_i = \left.\frac{\partial \Phi(\alpha)}{\partial \alpha_i}\right|_{\alpha^{s+\frac{i-1}{n}}}, \tag{10.7.58}$$

а q_i^s является решением следующей одномерной экстремальной задачи:

$$\begin{gathered} \Phi\left(\alpha^{s+\frac{i}{n}}\right) = \inf_q \Phi\left(\alpha^{s+\frac{i-1}{n}} - q r_i^s e_i\right), \\ \alpha^{s+\frac{i}{n}} \in V_d. \end{gathered} \tag{10.7.59}$$

Этот релаксационный метод относится к методам покоординатного спуска и в случае, когда V_d совпадает со всем пространством, переходит в известный метод Гаусса — Зейделя решения систем линейных уравнений $A\alpha = g$ с симметричной и положительно определенной матрицей A.

Поясним изложенное выше на примере решения задачи (10.7.19), когда множество U_d определяется соотношением (10.7.37) при дополнительном предположении, что функция $F_1(x)$ вогнута, а функция $F_2(x)$ выпукла, т. е. для любых $x', x'' \in [0,1]$ и $\Theta \in [0,1]$ выполняются неравенства

$$\begin{aligned} F_1((1-\Theta)x' + \Theta x'') &\geqslant (1-\Theta)F_1(x') + \Theta F_1(x''), \\ F_2((1-\Theta)x' + \Theta x'') &\leqslant (1-\Theta)F_2(x') + \Theta F_2(x''). \end{aligned} \tag{10.7.60}$$

Рассмотрим на отрезке $[0,1]$ для заданного $n \geqslant 1$ равномерную сетку

$$\Omega_h = \left\{ x_k : \; x_k = k \times h, \; k = 0, 1, \ldots, n+1, \; h = \frac{1}{n+1} \right\} \tag{10.7.61}$$

и введем по формуле (2.3.14) из главы 2 систему базисных функций $\psi_k(x)$ $(k = 1, \ldots, n)$. Пространство U^h определяется как линейная оболочка системы функций $\{\psi_k(x)\}_{k=1}^n$. Иначе говоря, пространство U^h является простран-

ством кусочно-линейных восполнений сеточных функций, определенных в узлах сетки Ω_h и обращающихся в нуль в точках $x_0 = 0$ и $x_{n+1} = 1$.

При сделанных предположениях относительно функций $F_1(x)$ и $F_2(x)$ и пространства U^h множество $U_d^h = U_d \cap U^h$ является множеством функций $v(x) \in U^h$, удовлетворяющих неравенствам

$$F_1(x_i) \leqslant v(x_i) \leqslant F_2(x_i), \quad i = 1, 2, \ldots, n.$$

Соответственно, множеством V_d будет n-мерный прямоугольный параллелепипед, координаты $\alpha = (\alpha_1, \ldots, \alpha_n)$ которого удовлетворяют неравенствам

$$F_1(x_i) \leqslant \alpha_i \leqslant F_2(x_i), \quad i = 1, 2, \ldots, n.$$

Если же относительно функций $F_1(x)$ и $F_2(x)$ не предполагать соответственно их вогнутости и выпуклости, то требования принадлежности функции $v(x)$ множеству U_d^h (векторов α множеству V_d) могут оказаться значительно сложнее.

Для иллюстрации метода локальных вариаций рассмотрим простейшую одномерную вариационную задачу. Пусть имеется вариационный функционал

$$J(\varphi) = \int_0^1 \pi(x, \varphi, \varphi_x)\, dx. \tag{10.7.62}$$

Требуется найти функцию $\varphi(x)$, доставляющую минимум функционалу (10.7.62) и удовлетворяющую следующим ограничениям:

$$F_1(x) \leqslant \varphi(x) \leqslant F_2(x). \tag{10.7.63}$$

Заметим, что в некоторых точках отрезка $[0, 1]$ ограничения могут отсутствовать. В этом случае мы формально считаем, что имеется естественное ограничение:
либо

$$-\infty < \varphi,$$

либо

$$\varphi < \infty.$$

В такой трактовке граничные условия для функции $\varphi(x)$ при $x = 0$ и $x = 1$ также удобно считать ограничениями. Это мы и будем делать в дальнейшем.

Простейшим вариантом реализации метода является следующий. Разобьем интервал $[0, 1]$ на n равных частей точками $x_i = i\Delta x$, $\Delta x = 1/n$

($i = 0, 1, \ldots, n$). Вариационный функционал (10.7.62) представим в виде суммы частных функционалов:

$$J = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} \pi(x, \varphi, \varphi_x)\, dx. \tag{10.7.64}$$

Каждый из интегралов в (10.7.64) аппроксимируем по формуле прямоугольников

$$\int_{x_i}^{x_{i+1}} \pi(x, \varphi, \varphi_x)\, dx \approx \Delta x \pi(x_{i+1/2}, \varphi_{i+1/2}, \varphi'_{i+1/2}),$$

где

$$\varphi_{i+1/2} = \frac{\varphi_i + \varphi_{i+1}}{2}, \quad \varphi'_{i+1/2} = \frac{\varphi_{i+1} - \varphi_i}{\Delta x}, \quad x_{i+1/2} = x_i + \frac{\Delta x}{2}.$$

Таким образом, в выражение для частных функционалов войдут только значения искомой функции в точках φ_i и φ_{i+1}. Введем следующее обозначение:

$$V_i(\varphi_i, \varphi_{i+1}) = \Delta x \pi(x_{i+1/2}, \varphi_{i+1/2}, \varphi'_{i+1/2}).$$

Приближенно будем иметь

$$J(\varphi) \approx V = \sum_{i=0}^{n-1} V_i(\varphi_i, \varphi_{i+1}). \tag{10.7.65}$$

Мы пришли к задаче отыскания такого набора чисел $\{\varphi_i\}$, который при заданных ограничениях

$$F_1(x_i) \leqslant \varphi_i \leqslant F_2(x_i), \tag{10.7.66}$$

доставляет минимум функционалу V.

Для решения задачи воспользуемся методом последовательных приближений. Пусть нам известны приближения до номера k: φ_i^k. Новые приближения φ_i^{k+1} будем искать следующим образом. Введем в рассмотрение некоторое число h и для каждого x_i рассмотрим совокупность трех чисел: $\varphi_i^k - h$, φ_i^k, $\varphi_i^k + h$. Величина φ_i^k по предложению удовлетворяет ограничениям. Необходимо теперь проверить выполнение ограничений для двух остальных величин. Если какое-нибудь ограничение не выполняется, то соответствующее число из рассмотрения исключается, а мы имеем дело только с парой чисел: либо $\varphi_i^k - h$, φ_i^k, либо φ_i^k, $\varphi_i^k + h$.

Аналогично методу Гаусса — Зейделя предположим, что нам уже известны значения решения на $(k+1)$-м приближении для всех индексов, меньших i, т. е. $\varphi_0^{k+1}, \varphi_1^{k+1}, \ldots, \varphi_{i-1}^{k+1}$. Естественно эту информацию использовать для ускорения метода. Поэтому наряду с частными функционалами V_i

удобно ввести в рассмотрение другие:

$$\begin{aligned} G_i =& V_{i-1}(\varphi_{i-1}^{k+1}, \varphi_i^k) + V_i(\varphi_i^k, \varphi_{i+1}^k), \\ U_i^+ =& V_{i-1}(\varphi_{i-1}^{k+1}, \varphi_i^k + h) + V_i(\varphi_i^k + h, \varphi_{i+1}^k), \\ U_i^- =& V_{i-1}(\varphi_{i-1}^{k+1}, \varphi_i^k - h) + V_i(\varphi_i^k - h, \varphi_{i+1}^k). \end{aligned} \tag{10.7.67}$$

Эти выражения учитывают в сумме (10.7.65) те два члена, которые зависят от величины φ_i и соответствуют трем возможным ее вариациям: $\varphi_i^k - h$, φ_i^k, $\varphi_i^k + h$.

Дальнейшее сводится к определению того из этих трех чисел, которое удовлетворяет заданным ограничениям и доставляет минимум соответствующему функционалу в (10.7.67). Именно это число и объявляется новым приближенным значением φ_i^{k+1}. После того, как будут найдены значения φ_i^{k+1} для всех $i = 0, 1, \ldots, n$, можно переходить к очередному приближению.

В том случае, когда при данных параметрах Δx и h приближенное решение получено с заданной точностью, этот процесс нужно проводить с новым значением h, например, уменьшенным вдвое и т. д. Естественно, что при организации этого нового процесса следует использовать полученную информацию о приближенном решении, взяв его в качестве начального. Затем необходимо уменьшить Δx и добиваться лучшей аппроксимации в решении задачи.

Мы рассмотрели наиболее простой случай ограничений, когда известные требования накладываются на само решение φ. В более общем случае, когда ограничение задается соотношением

$$B(x, \varphi, \varphi_x) \leqslant 0,$$

необходимо воспользоваться для любого x_i той или иной аппроксимацией и каждый раз проверять выполнение этого приближенного условия.

Метод локальных вариаций распространяется и на многомерные вариационные задачи с ограничениями, причем принципиально сохраняются все этапы изложенного алгоритма. Например, в случае двумерной области вместо совокупности чисел, подлежащих сравнению на основе функционалов и ограничений $\varphi_i - h$, φ_i, $\varphi_i + h$, теперь будем иметь $\varphi_{ij} - h$, φ_{ij}, $\varphi_{ij} + h$, а частный функционал U_{ij} будет состоять из четырех слагаемых:

$$U_{ij} = V_{ij} + V_{i-1,j} + V_{i,j-1} + V_{i-1,j-1},$$

которые содержат величину φ_{ij}^k. Аналогичным образом определим U_{ij}^+ и U_{ij}^-, где вместо φ_{ij} следует подставить величины $\varphi_{ij} + h$ и $\varphi_{ij} - h$ соответственно.

Алгоритм для задач с большим числом переменных формируется аналогичным образом.

В заключение следует подчеркнуть, что метод локальных вариаций не связан с выбором базиса и поэтому применим для областей произвольной формы. Однако следует иметь в виду, что решение вариационной задачи методом локальных вариаций может привести не к абсолютному минимуму, а к локальному. Что касается сходимости метода к локальному минимуму, то она гарантируется существом алгоритма.

Глава 11.

Вычислительные тензорные методы

В ряде задач квантовой физики и химии число координатных осей определяется числом атомов или электронов в изучаемой системе. Такие задачи являются *существенно многомерными*, в них естественным образом возникают массивы с большим числом измерений (мод). В настоящем разделе описываются современные вычислительные методы работы с многомерными массивами (тензорами). Их развитие связано с тем, что стандартные алгоритмы решения задач математической физики плохо масштабируются при увеличении количества координатных осей. Кроме того, d-мерный массив с числом отсчетов n по каждому измерению содержит n^d элементов и, как правило, не может задаваться перечислением всех элементов. По существу, способ задания, или, другими словами, *формат представления* многомерных массивов, является центральной проблемой при организации вычислений. Необходимы такие представления и аппроксимации многомерных массивов, которые определяются относительно малым числом параметров и позволяют проводить вычисления без явного получения всех элементов массива.

11.1. Основные понятия и обозначения

В алгебре под тензором понимается полилинейная форма на векторных пространствах. После введения базисов в этих пространствах тензор получает представление в виде d-мерного массива – в полной аналогии с представлением линейных операторов или билинейных форм с помощью матриц. В дальнейшем под d-мерным тензором понимается d-мерный массив. Элементы тензора $\mathbf{A}$ обозначаются $A(i_1, \ldots, i_d)$. Каждый индекс i_k меня-

ется от 1 до n_k. Числа n_k называются размерами тензора. Тензоры одних и тех же размеров можно естественным образом складывать и умножать на числа. Таким образом, тензоры образуют векторное пространство. Фробениусова норма тензора определяется как

$$||\mathbf{A}||_F = \sqrt{\sum_{i_1,\ldots,i_d} |A(i_1,\ldots,i_d)|^2}. \tag{11.1.1}$$

С заменой базиса в пространстве k-го измерения связана операция умножения тензора на матрицу. Если U – матрица размеров $n_k \times m$, то результатом будет тензор $B(i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_d)$ с элементами

$$B(i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_d) = \sum_{i_k=1}^{n_k} A(i_1,\ldots,i_{k-1},i_k,i_{k+1},\ldots,i_d)U(i_k,j).$$

Под тензорным форматом понимается малопараметрическое представление тензора, часто в виде так называемых *тензорных разложений*.

11.2. Тензорные разложения

Мы рассмотрим следующие форматы для записи тензоров: *каноническое разложение*, *разложение Таккера*, *TT-разложение* (тензорный поезд – от английского *tensor train*) и *HT-разложение* (иерархическое разложение Таккера).

11.2.1. Каноническое разложение и его свойства

Каноническим разложением называется представление тензора в виде

$$A(i_1,\ldots,i_d) = \sum_{\alpha=1}^{r} U_1(i_1,\alpha)\ldots U_d(i_d,\alpha). \tag{11.2.1}$$

Минимальное число слагаемых r в разложениях вида (11.2.1) называется *каноническим рангом*, или просто *рангом* тензора. Будем писать $r = \operatorname{rank}\mathbf{A}$. Канонический формат является дискретным аналогом разделения переменных в функции от d переменных. Он задается матрицами U_k размеров $n_k \times r$, составленными из элементов $U_k(i_k,\alpha)$. Часто используется обозначение

$$\mathbf{A} = (U_1, \ldots, U_d).$$

Если $n_1 = \ldots = n_d = n$, то для хранения матриц U_k достаточно dnr ячеек памяти. В то же время общее число элементов тензора равно n^d.

При определенных условиях на матрицы U_k каноническое разложение является единственным с точностью до одновременной перестановки столбцов в U_k и их масштабирования. Довольно общие достаточные условия единственности даются *теоремой Крускала*. Приведем ее в трехмерном случае. Для формулировки требуется определить понятие *ранга Крускала*.

Определение 11.2.1. *Рангом Крускала $k(U)$ матрицы U размера $n \times r$ называется минимальное k, для которого в матрице существует $k+1$ линейно зависимых столбцов.*

Теорема 1. Если

$$k(U_1) + k(U_2) + k(U_3) \geq 2r + 2,$$

то каноническое разложение является единственным.

Обратим внимание на то, что при $d = 2$, т. е. в случае матриц, каноническое разложение является существенно неединственным. Каноническое разложение при $d = 3$ называется также *трилинейным разложением*. Его единственность объясняет ту роль, которое оно играет в многочисленных приложениях как модель данных. Условия Крускала не являются необходимыми для единственности. Известны и другие, более слабые условия, гарантирующие единственность.

При $d = 2$ тензор является матрицей, а его каноническое разложение совпадает со скелетным разложением матрицы и может быть вычислено, например, с помощью метода исключения Гаусса. Приближение заданного ранга может быть найдено с помощью алгоритма сингулярного разложения (SVD). При $d \geq 3$ задача построения точного или приближенного канонического разложения становится существенно более тяжелой. Задача вычисления канонического ранга тензора при $d \geq 3$ является NP-полной. Задача аппроксимации заданного тензора $\mathbf{A}$ тензором $\mathbf{C} = (U_1, \ldots, U_d)$ ранга не выше r формулируется как задача минимизации следующего функционала ошибки:

$$\mathbf{C} = \arg \min_{\text{rank}(\mathbf{B}) \leq r} ||\mathbf{A} - \mathbf{B}||_F. \tag{11.2.2}$$

Трудность ее решения связана с тем, что множество тензоров канонического ранга не выше r может не быть замкнутым. Например, рассмотрим последовательность тензоров

$$T(h) = \frac{P(h) - P(0)}{h},$$

где

$$P(h) = (a + hb) \otimes \ldots \otimes (a + hb).$$

Тензор $P(h)$ имеет канонический ранг 1, а ранг тензора $T(h)$ не выше 2. При этом

$$T(h) = \frac{dP}{dh}(0) + \mathcal{O}(h),$$

поэтому при $h \to 0$ тензор $T(h)$ имеет предел, равный

$$T = b \otimes a \otimes \ldots \otimes a + \ldots + a \otimes \ldots \otimes a \otimes b, \tag{11.2.3}$$

который представлен в виде суммы d слагаемых ранга 1. В случае линейно независимых векторов a и b можно доказать, что ранг T в точности равен d. Это означает, что для тензора T существует приближение ранга 2 с любой наперед заданной точностью, т. е. минимум не достигается ни в одной точке. Поэтому в практических вычислениях необходимо использовать некоторые регуляризации.

Алгоритмы вычисления канонического разложения

Существуют различные алгоритмы вычисления канонического разложения и канонической аппроксимации. Классическим подходом является метод переменных направлений, или *переменных наименьших квадратов* (в англоязычной литературе — ALS: Aternating Least Squares). Метод основан на полилинейности разложения и является итерационным. Зафиксируем все матрицы $U_1, \ldots, U_d$, кроме k-й. Тогда задача минимизации по U_k становится задачей минимизации квадратичного функционала, т. е. стандартной задачей линейной алгебры, сводящейся к решению системы линейных алгебраических уравнений. Ее решение дает новое значение U_k', на котором значение функционала, очевидно, не увеличивается. Расчетные формулы метода ALS имеют очень простой вид:

$$\begin{aligned} M_k U_k' &= F_k, \\ M_k = \circ_{j \neq k} \Gamma_j, \quad \Gamma_j &= U_j^* U_j, \\ F_k(i_k, \alpha) = \sum_{\substack{1 \leq i_s \leq n_s \\ s \neq k}} A(i_1, \ldots, i_d) &\prod_{\substack{1 \leq l \leq d \\ l \neq k}} U_l(i_l, \alpha). \end{aligned} \tag{11.2.4}$$

Символом $\circ$ обозначена операция адамарова (поэлементного) умножения матриц. Стоимость одной итерации для каждого k составляет $O(n^d r)$ для вычисления правой части, $O(nr^3)$ — для окончательного получения решения. Можно доказать, что метод переменных направлений обладает ло-

кальной сходимостью, однако в целом свойства его сходимости изучены плохо.

11.2.2. Разложение Таккера и его свойства

Разложением Таккера называется представление тензора в виде

$$A(i_1, \ldots, i_d) = \sum_{\alpha_1, \ldots, \alpha_d} G(\alpha_1, \ldots, \alpha_d) U_1(i_1, \alpha_1) \ldots U_d(i_d, \alpha), \tag{11.2.5}$$

индекс α_k меняется от 1 до r_k. Тензор $\mathbf{G}$ называется *ядром Таккера*, матрицы U_k размеров $n_k \times r_k$ называются *факторами Таккера*. Минимально возможные значения r_k в представлениях вида (11.2.5) называются *рангами Таккера*. Вектор из целых чисел $(r_1, \ldots, r_d)$ иногда называют *мультилинейным рангом*. В разложении Таккера исходный тензор получается путем умножения ядра Таккера на матрицы $U_1, \ldots, U_d$.

Разложение Таккера активно используется в многофакторном анализе. Недостатком разложения Таккера является наличие вспомогательного d-мерного массива $\mathbf{G}$, что приводит к экспоненциальной сложности при больших d. Тем не менее при небольших d разложение Таккера может быть достаточно эффективным. В общем случае ранги Таккера и канонический ранг связаны неравенствами $r_k \leq r$, т. е. ранги Таккера дают оценку снизу для канонического ранга, однако канонический ранг может быть существенно выше.

Предположим теперь, что существует разложение Таккера с рангами r_k. Чтобы получить представление вида (11.2.5), введем матрицы A_k, элементы которых определяются следующим образом:

$$A_k(i_k, i_1 \ldots i_{k-1} i_{k+1} \ldots i_d) = A(i_1, \ldots, i_d).$$

Матрица A_k имеет размеры $n_k \times \frac{N}{n_k}$, где $N = n_1 \ldots n_d$, и составлена из элементов исходного тензора, индекс i_k нумерует ее строки, а значения остальных индексов определяют столбцы. Матрица A_k называется *k-й матрицей развертки* тензора $\mathbf{A}$. Аналогично определяются матрицы развертки G_k для ядра Таккера $\mathbf{G}$. Матрицы A_k и G_k связаны соотношением

$$A_k = U_k G_k V_k^\top, \quad V_k = \otimes_{l \neq k} U_l, \tag{11.2.6}$$

откуда следует, что ранг матрицы A_k не выше r_k. Разложение (11.2.6) не является единственным. Одно из таких разложений можно вычислить с по-

мощью редуцированного сингулярного разложения матрицы A_k:

$$A_k = U_k \Lambda_k V_k^\top.$$

Таким образом можно определить матрицы U_k. После этого ядро разложения Таккера получается с помощью умножения тензора $\mathbf{A}$ на матрицы $U_1^\top, \ldots, U_d^\top$. Сложность описанного вычисления экспоненциальна по d. Если для тензора ищется приближение в формате Таккера с заданными рангами, то данный алгоритм (называемый также HOSVD: Higher Order Singular Value Decomposition) дает квазиоптимальное приближение, что подтверждается следующей теоремой.

Теорема 2. Пусть матрицы развертки A_k тензора $\mathbf{A}$ могут быть представлены в виде $A_k = R_k + E_k$, $||E_k||_F = \varepsilon_k$, $\operatorname{rank} R_k = r_k$. Тогда алгоритм HOSVD построит приближение $\mathbf{B}$ с рангами Таккера не выше r_k и при этом

$$||\mathbf{A} - \mathbf{B}|| \leq \sqrt{\sum_{k=1}^{d} \varepsilon_k^2}. \tag{11.2.7}$$

11.2.3. TT-формат и его свойства

Редукция к матрицам

В отличие от канонического разложения, вычисление точного или приближенного разложения Таккера сводится к получению известных матричных разложений. Однако при большом числе измерений формат Таккера, очевидно, применять нельзя. Поиск тензорного формата, который сводится к известным матричным разложениям и может использоваться при большом числе измерений, привел почти одновременно к TT-формату и HT-формату. Оба формата основаны на одной и той же схеме редукции размерности, которая используется в них тем не менее по-разному. TT-формат имеет очень простой вид:

$$A(i_1, \ldots, i_d) = G_1(i_1) \ldots G_d(i_d), \tag{11.2.8}$$

где $G_k(i_k)$ — матрица размера $r_{k-1} \times r_k$ при фиксированном i_k, а также считается, что $r_0 = r_d = 1$. Элементы этой матрицы удобно обозначать через $G_k(\alpha_{k-1}, i_k, \alpha_k)$. Тогда

$$A(i_1, \ldots, i_d) = \sum_{\alpha_0, \ldots, \alpha_d} G_1(\alpha_0, i_1, \alpha_1) \ldots G_d(\alpha_{d-1}, i_d, \alpha_d), \tag{11.2.9}$$

где α_k меняется от 1 до r_k. Числа r_k называются *TT-рангами*, а $G_k(i_k)$ — *ядрами*, или *вагонами*, TT-разложения.

Представление (11.2.8) можно рассматривать как блочное обобщение канонического разложения ранга 1. Аналогичные представления были известны ранее в различных областях науки как *линейные тензорные сети* и MPS-разложения (MPS: Matrix Product States). Для хранения матриц $G_k(i_k)$, т. е. для записи TT-разложения, требуется

$$\sum_{k=1}^{d} r_{k-1} n_k r_k = O(dnr^2)$$

ячеек памяти (параметров). В отличие от канонического ранга, вычисление которого является NP-полной задачей, минимально возможные TT-ранги определяются рангами некоторых матриц, ассоциированным с данным тензором. Эти матрицы также называются матрицами развертки, хотя они и отличаются от матриц развертки, связанных с разложением Таккера. Мы сохраним для них обозначение A_k, но теперь матрица A_k имеет размер $N_k \times N/N_k$, где $N_k = n_1 \dots n_k$, а ее элементы определяются формулой

$$A_k = A_k(i_1 \dots i_k; i_{k+1} \dots i_d) = A(i_1, \dots, i_d). \tag{11.2.10}$$

Первые k индексов тензора $\mathbf{A}$ нумеруют строки матрицы A_k, а последние $d-k$ индексов задают столбцы.

Теорема 3. Для любого TT-разложения заданного тензора выполняются неравенства

$$r_k \geq \operatorname{rank} A_k, \quad 1 \leq k \leq d-1, \tag{11.2.11}$$

и существует такое TT-разложение, для которого

$$r_k = \operatorname{rank} A_k, \quad 1 \leq k \leq d-1. \tag{11.2.12}$$

Доказательство. Неравенство (11.2.11) для матрицы A_k вытекает из скелетного разложения

$$A_k(i_1 \dots i_k\,;\, i_{k+1} \dots i_d) = U_k(i_1 \dots i_k\,;\, \alpha_k)\, V_k(\alpha_k\,;\, i_{k+1} \dots i_d),$$

где

$$\begin{aligned} U_k(i_1 \dots i_k\,;\, \alpha_k) &= \sum_{\alpha_0, \dots, \alpha_{k-1}} G_1(\alpha_0, i_1, \alpha_1) \dots G_k(\alpha_{k-1}, i_k, \alpha_k), \\ V_k(\alpha_k\,;\, i_{k+1} \dots i_d) &= \sum_{\alpha_{k+1}, \dots, \alpha_d} G_{k+1}(\alpha_k, i_{k+1}, \alpha_{k+1}) \dots G_d(\alpha_{d-1}, i_d, \alpha_d). \end{aligned}$$

Покажем, как можно получить ТТ-разложение с рангами $r_k = \operatorname{rank} A_k$. Для этого проведем индукцию по d. Матрица A_1 допускает скелетное разложение $A_1 = UV$, в котором U имеет $r_1 = \operatorname{rank} A_1$ линейно независимых столбцов, а V — столько же линейно независимых строк, причем составленных из некоторых строк матрицы A_1. Это разложение можно записать в виде

$$A_1(i_1; i_2, \ldots, i_d) = \sum_{\alpha_1=1}^{r_1} G_1(i_1, \alpha_1) V(\alpha_1, i_2, \ldots, i_d), \tag{11.2.13}$$

где d-мерный тензор $\mathbf{V}$ является подтензором исходного тензора $\mathbf{A}$. Объединив индексы α_1 и i_2 в мультииндекс, рассмотрим $\mathbf{V}$ как тензор размерности $d-1$ и запишем для него ТТ-разложение с рангами, равными $\operatorname{rank} V_k$:

$$V(\alpha_1 i_2, i_3, \ldots, i_d) = \sum_{\alpha_2, \ldots, \alpha_d} G_2(\alpha_1 i_2, \alpha_2) \ldots G_d(\alpha_{d-1}, i_d, \alpha_d). \tag{11.2.14}$$

Искомый тензорный поезд получается объединением формул (11.2.13) и (11.2.14). Остается только заметить, что V_k является подматрицей в A_{k+1}. Поскольку ранг подматрицы не может быть больше ранга матрицы, получаем

$$\operatorname{rank} V_k = \operatorname{rank} A_{k+1}, \quad 1 \le k \le d-1.$$

Базовые операции в ТТ-формате

Доказательство теоремы 3 является конструктивным: тензорный поезд для $\mathbf{A}$ строится путем последовательного вычисления скелетных разложений матриц, являющихся подматрицами матриц A_k. Для вычисления скелетных разложений можно применить алгоритм сингулярного разложения (SVD). На его основе легко получается также приближенное ТТ-разложение с меньшими ТТ-рангами и с полным контролем погрешности.

Теорема 4. Если приближенное ТТ-разложение строится с помощью алгоритма сингулярного разложения и на каждом шаге фробениусова норма погрешности не превышает $\varepsilon_k,\ k = 1, \ldots, d-1$, то суммарная ошибка оценивается сверху как

$$||\mathbf{A} - \mathbf{B}|| \le \sqrt{\sum_{k=1}^{d-1} \varepsilon_k^2}.$$

Доказательство этого утверждения можно найти в работах И.В. Оселедца [25], И.В. Оселедца и Е.Е. Тыртышникова [25].

При сложении двух d-мерных тензоров, заданных в ТТ-формате, и при их поэлементном умножении можно получить результат в ТТ-формате без

использования d-мерных массивов. Например, если

$$A(i_1,\ldots,i_d) = A_1(i_1)\ldots A_d(i_d), \quad B(i_1,\ldots,i_d) = B_1(i_1)\ldots B_d(i_d),$$

то для тензора $\mathbf{C} = \mathbf{A}\circ\mathbf{B}$ (поэлементное произведение) ТТ-разложение имеет вид

$$C(i_1,\ldots,i_d) = C_1(i_1)\ldots C_d(i_d), \quad C_k(i_k) = A_k(i_k)\otimes B_k(i_k),$$

т. е. результат представим в ТТ-формате с ТТ-рангами, равными произведению ТТ-рангов $\mathbf{A}$ и $\mathbf{B}$. Если такие операции повторять многократно, то рост рангов приведет к существенному росту числа параметров. Поэтому необходимо иметь *надежную и быструю* процедуру аппроксимации ТТ-разложения другим ТТ-разложением с меньшими ТТ-рангами. Надежность означает гарантированную оценку погрешности, а быстрота определяется тем, что полные d-мерные массивы в ходе вычислений возникать не должны. Такая процедура существует и называется *ТТ-округлением*. Она содержит всего $O(dnR^3)$ операций, где R — верхняя оценка ТТ-рангов исходного ТТ-разложения.

В самых разных задачах векторы и матрицы могут рассматриваться как d-мерные массивы, а ТТ-формат может применяться для их точного или приближенного представления. Линейное отображение на пространстве $\mathbb{R}^{n\times\ldots\times n}$ задается *d-уровневой матрицей* с элементами $A(i_1,\ldots,i_d;j_1,\ldots,j_d)$, строки и столбцы естественным образом нумеруются с помощью d-мерных индексов. ТТ-формат для матрицы определяется как представление вида

$$A(i_1,\ldots,i_d;j_1,\ldots,j_d) = A_1(i_1,j_1)\ldots A_d(i_d,j_d), \tag{11.2.15}$$

где матрица $A_k(i_k,j_k)$ имеет размер $r_{k-1}\times r_k$. Элементы матриц $A_k(i_k,j_k)$ удобно обозначать через $A_k(\alpha_{k-1},i_k,j_k,\alpha_k)$.

Введем также $n_k\times n_k$-матрицы $A^{(k)}_{\alpha_{k-1},\alpha_k}$. Определим их элементы правилом

$$[A^{(k)}_{\alpha_{k-1},\alpha_k}]_{i_k,j_k} := A_k(\alpha_{k-1},i_k,j_k,\alpha_k).$$

Тогда возникает формула

$$A = \sum_{\alpha_0,\ldots,\alpha_d} A^{(1)}_{\alpha_0,\alpha_1}\otimes\ldots\otimes A^{(d)}_{\alpha_{d-1},\alpha_d}.$$

Чтобы придать ей еще более компактный вид, определим матрицы $A^{(k)}$ как блочные матрицы, составленные из блоков $A^{(k)}_{\alpha_{k-1},\alpha_k}$, и введем операцию *квазикронекерова произведения* блочных матриц, которая отличается от обычного умножения блочных матриц тем, что при умножении блоков ис-

пользуется кронекерово произведение:

$$[A^{(k)} \bowtie A^{(k+1)}]_{\alpha_{k-1},\alpha_{k+1}} := \sum_{\alpha_k} A^{(k-1)}_{\alpha_{k-1},\alpha_k} \otimes A^{(k)}_{\alpha_k,\alpha_{k+1}}.$$

Тогда для исходной матрицы получается компактная формула

$$A = A^{(1)} \bowtie \ldots \bowtie A^{(d)}.$$

Если $r_k = 1$ при всех k, то A задается одним кронекеровым произведением

$$A = A^{(1)} \otimes \ldots \otimes A^{(d)}.$$

Операция квазикронекерова произведения удобна для явного представления классических матриц специального вида. Например, матрица d-мерного оператора Лапласа

$$\Delta_d = \Delta_1 \otimes I \otimes \ldots \otimes I + \ldots + I \otimes \ldots \otimes I \otimes \Delta_1,$$

где Δ_1 — матрица одномерного оператора Лапласа, имеет TT-ранги, равные 2, и записывается в виде

$$\Delta_d = \begin{bmatrix} I \otimes \ldots \otimes I & \Delta_{d-1} \end{bmatrix} \bowtie \begin{bmatrix} \Delta_1 \\ I \end{bmatrix} =$$
$$= \begin{bmatrix} I \otimes \ldots \otimes I & \Delta_{d-2} \end{bmatrix} \bowtie \begin{bmatrix} I & \Delta_1 \\ 0 & I \end{bmatrix} \bowtie \begin{bmatrix} \Delta_1 \\ I \end{bmatrix} = \ldots =$$
$$= \begin{bmatrix} I & \Delta_1 \end{bmatrix} \bowtie \begin{bmatrix} I & \Delta_1 \\ 0 & I \end{bmatrix} \bowtie \ldots \bowtie \begin{bmatrix} I & \Delta_1 \\ 0 & I \end{bmatrix} \bowtie \begin{bmatrix} \Delta_1 \\ I \end{bmatrix}.$$

11.2.4. HT-формат и его свойства

Тензорные разложения могут естественным образом строиться на основе разложения Таккера. Например, в TT-формате можно записывать не весь тензор, а лишь его ядро Таккера. HT-формат использует разложение Таккера непосредственно при редуцировании размерности.

Пусть уже получено разложение Таккера с рангами r_k. Ядро $\mathbf{G}$ будет представлять собой d-мерный массив $G(\alpha_1, \ldots, \alpha_d)$. Разобьем все индексы α_k на пары (какие-то индексы могут остаться без пары). Для примера рассмотрим случай $d = 6$, объединим индексы $\beta_1 = (\alpha_1, \alpha_2)$, $\beta_2 = (\alpha_3, \alpha_4)$, $\beta_3 = (\alpha_5, \alpha_6)$ и будем рассматривать $\mathbf{G}$ как трехмерный массив. Для него также можно построить разложение Таккера. Факторы Таккера будут представлять собой матрицы размера $r^2 \times r$, а ядро будет размера $r \times r \times r$ (в предположении, что

все ранги ограничены r). Фактически HT-формат представляет собой последовательное применение разложения Таккера к ядрам. Его можно интерпретировать с точки зрения линейных пространств. Если разложение Таккера выбирает некоторый базис вида $U_1 \otimes \ldots \otimes U_d$, где U_k — некоторые подпространства размерности r_k, то объединение индексов и вторичное применение разложения Таккера соответствуют выбору подпространств в тензорных произведениях $U_i \otimes U_j$. Схема может быть обобщена на объединение произвольного количества индексов, наиболее просто она выглядит в том случае, когда индексы объединяются попарно.

Пример бинарного дерева, порождающего HT-разложение, приведен на рис. 11.1.

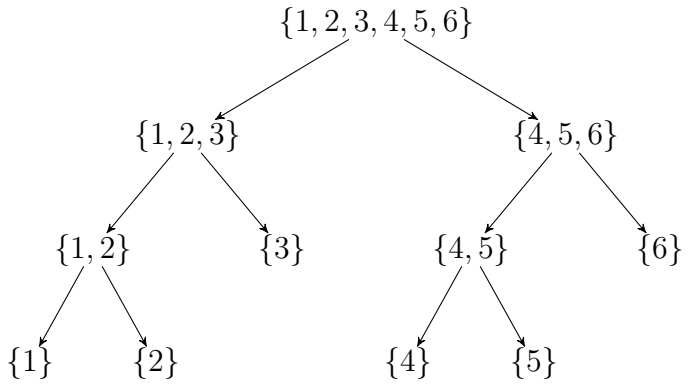


Рис. 11.1. Редукция размерности в HT-формате

Аналогично TT-формату HT-разложение может быть посчитано с помощью последовательных сингулярных разложений — это следует из его интерпретации как последовательного разложения Таккера. Число параметров HT-разложения определяется HT-рангами. Каждому узлу дерева соответствует ранг, и суммарная память равна $\mathcal{O}(dnr + r^3)$, если считать, что $r_t = r$ и $n_k = n$. TT-разложение можно получить как специальный случай HT-формата с полностью несбалансированным деревом. Для HT-формата можно построить аналоги основных алгоритмов TT-формата, включая вычисление основных операций линейной алгебры, округление. На практике обычно TT-формат оказывается более эффективным за счет более простой структуры данных, однако существуют примеры, когда HT-формат может иметь меньшее количество параметров.

11.3. Вычислительные методы построения TT-аппроксимации

11.3.1. Общие принципы

Предположим, что решение интересующей нас задачи можно представить в виде тензора, и известно, что соответствующий тензор можно приблизить в ТТ-формате. В качестве модельного примера рассмотрим уравнение Пуассона в кубе $[0,1]^d$ с граничными условиями Дирихле, дискретизированное с помощью центральных разностей на тензорной сетке. В качестве правой части возьмем вектор из одних единиц:

$$\Delta_d u = f. \tag{11.3.1}$$

В данном случае оператор представляется в ТТ-формате с ТТ-рангами 2, а правая часть имеет ТТ-ранги, равные 1. Вектору неизвестных u естественным образом ставится в соответствие d-мерный тензор $\mathbf{U}$. Задача состоит в построении аппроксимации U с помощью ТТ-формата. Такая задача является частным случаем класса линейных систем, когда и матрица, и правая часть приближены в ТТ-формате:

$$Au = f,$$

где решению u и правой части f по некоторому правилу ставятся в соответствие тензоры U и F, причем

$$A = A_1 \bowtie A_2 \bowtie \ldots \bowtie A_d,$$

$$U(i_1, \ldots, i_d) \approx U_1(i_1) \ldots U_d(i_d), \quad F(i_1, \ldots, i_d) \approx F_1(i_1) \ldots F_d(i_d).$$

Существуют два принципиально различных подхода к нахождению малоранговых приближений к решению многомерных задач. Первых подход основан на использовании классических итерационных методов (например, методов Крылова) для решения линейных систем. На каждом шаге требуется вычисление матрично-векторных произведений, сложений и вычисление скалярных произведений. Все эти операции можно реализовать в рамках ТТ-формата. Проблема состоит в том, что при выполнении таких операции (в частности, при вычислении матрично-векторного произведения) ТТ-ранги будут расти, однако эта проблема легко решается: все операции необходимо проводить приближенно, используя алгоритмы округления. При этом нужно использовать методы и подходы, созданные в теории

неточных итерационных методов. В случае, если число итераций невелико, такой подход часто оказывается эффективным и простым в реализации. Однако он имеет существенный недостаток: число итераций может быть большим, а большинство известных методов построения предобуславливателей (например, методы неполной факторизации) не имеют тензорных аналогов. Поэтому активно развивается другой подход, на котором мы и сосредоточимся.

11.3.2. Методы оптимизации в малоранговых форматах: линейные системы

У нас есть существенная априорная информация: решение может быть приближено в ТТ-формате, что существенно сокращает число свободных параметров. Покажем, как это можно использовать, на примере решения линейных систем с симметричной положительно определенной матрицей. Система

$$Au = f,$$

где $A = A^* > 0$, эквивалентна минимизации функционала

$$J(u) = (Au, u) - 2(f, u). \tag{11.3.2}$$

Рассмотрим множество TT_r с ТТ-рангами, не превосходящими r. Тогда приближенное решение можно найти, минимизируя функционал (11.3.2) по TT_r. Это оптимизационная задача с $\mathcal{O}(dnr^2)$ параметрами, которую можно решать различными методами, однако наиболее эффективными оказываются те, которые напрямую используют структуру ТТ-формата. Первый подход состоит в использовании метода ALS. Мы фиксируем все ядра, кроме $F_k(i_k)$. Получившаяся минимизационная задача оказывается квадратичной и сводится к решению линейной системы с $\mathcal{O}(nr^2)$ неизвестными. Весь оптимизационный процесс сводится к последовательной минимизации по $U_1(i_1) \dots U_d(i_d)$. Гарантировано, что значение функционала не увеличивается и на практике оказывается, что сходимость достаточно быстрая. Эффективная реализация приближенной минимизации функционала (11.3.2) требует многих технических деталей: как выглядят «локальные» линейные системы, как пересчитывать матрицы, как поддерживать ортогональность ядер — однако все они решаемы. Более принципиальная трудность состоит в следующем. В методе переменных направлений необходимо заранее знать ранги решений, что часто неудобно, так как необходимо «угадать» $(d-1)$ параметр. Здесь на помощь приходит специальная структура ТТ-формата,

которая позволяет создавать эффективные алгоритмы с автоматической адаптацией ранга. В физике твердого тела при исследовании спиновых систем был предложен алгоритм ренормгруппы матрицы плотности (density matrix renormalization group, DMRG), в котором содержится базовая идея такой адаптации. Рассмотрим наш d-мерный тензор как $(d-1)$-одномерный тензор, в котором моды k и $k+1$ считаются за одну большую моду длины n^2. Для этого достаточно вместо пары ядер $U_k(i_k)$ и $U_{k+1}(i_{k+1})$ рассмотреть суперблок

$$U_{k,k+1}(i_k, i_{k+1}) = U_k(i_k)U_{k+1}(i_{k+1}).$$

Теперь сделаем один шаг ALS с оптимизацией функционала по суперблоку $U_{k,k+1}$; реализация алгоритма при этом не меняется. После того как новый суперблок посчитан, k-й TT-ранг можно вычислить из приближенной факторизации

$$U'_{k,k+1}(i_k, i_{k+1}) \approx U'_k(i_k)U'_{k+1}(i_{k+1}), \tag{11.3.3}$$

которая вычисляется с помощью сингулярного разложения. В алгоритме DMRG ранги определяются адаптивно, и необходимо задать лишь параметр точности ε. Отметим, что метод DMRG не является классическим методом оптимизации, так как пространство параметров меняется на каждом шаге итерационного процесса, что делает его особенно интересным с математической точки зрения.

Основным недостатком метода DMRG является как раз его основное отличительное свойство — использование суперблока. Если для спиновых систем размер мод небольшой ($n = 2$ или $n = 4$), то для задач математической физики (как, например, для уравнения (11.3.1)) n может быть порядка $100 - 1000$, и сложность $\mathcal{O}(n^2)$ становится слишком высокой. Альтернативой DMRG и наиболее эффективными на практике на данный момент являются алгоритмы, которые сочетают оба подхода, классический итерационный подход и минимизацию напрямую в TT-формате. Такие методы получили название *alternating minimal energy* (AMEN). Кратко опишем основную идею. При использовании метода ALS базисы не расширяются. На самом деле, достаточно построить некоторое расширение $\widehat{U}_k(i_k) \in \mathbb{C}^{r_{k-1} \times p}$ для текущего базиса, где p – некий параметр. Откуда брать такое расширение? В работах Долгова и Савостьянова было предложено использовать для расширения приближенную проекцию невязки $Au - f$ на текущее ядро. Таким образом удается получить адаптацию ранга при условии сохранения сложность метода переменных направлений. Для такого метода можно получить оценки сходимости (в условиях отсутствия аппроксимации, т. е. ранги могут расти достаточно быстро): AMEN-метод сходится не медленнее алгоритма наискорейшего спуска для исходной системы. Конечно, методы наискорейше-

го спуска могут иметь очень плохую сходимость, однако на практике сходимость AMEN-метода очень хорошая: обычно требуется всего несколько проходов по ядрам для получения необходимой точности.

На примере уравнения (11.3.1) продемонстрируем работу AMEN-подхода. Отметим, что отображения вектора неизвестных на тензоры может быть разным. Очень эффективным подходом является идея QTT-представления (Quantized Tensor Train). В одномерном случае она состоит в преобразовании вектора длины 2^d в тензор размера $2 \times 2 \times \ldots 2$ путем естественного упорядочивания. Оператор Лапласа в таком представлении становится TT-матрицей с TT-рангами не выше 3. В многомерном случае ситуация аналогичная. В нашем случае одномерный размер сетки — 2^D, где $D = 8$. Вычислялись приближенные решения с помощью метода AMEN для различных размерностей и для точности $\varepsilon = 10^{-6}$. Интерес представляет не только время счета, но и память, необходимая для хранения приближения. Результаты представлены на рис. 11.2.

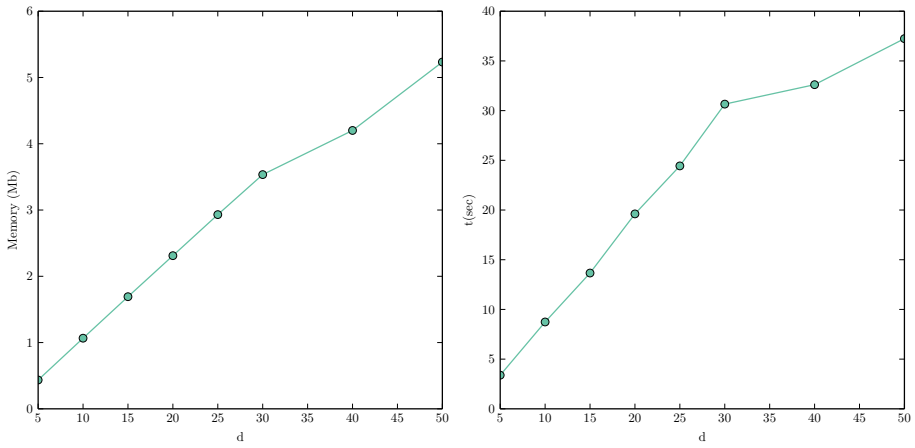


Рис. 11.2. Слева: память для хранения приближенного решения в зависимости от размерности задачи. Справа: время решения (в секундах) оптимизационной задачи методом AMEN в зависимости от размерности задачи

11.3.3. Задачи на собственные значения

Описанные выше подходы (ALS, DMRG, AMEN) могут быть применены к решению других задач, которые сводятся к задачам оптимизации, в частности к решению задач на собственные значения с симметричной матрицей. Задача нахождения минимального собственного значения сводится

к минимизации отношения Рэлея

$$\frac{(Au, u)}{(u, u)} \to \min .$$

При этом алгоритм остается без изменений, за одним исключением: вместо линейной системы на каждом локальном шаге решается задача на собственные значения с $\mathcal{O}(nr)$ неизвестными. Аналогичным образом можно использовать AMEN-идею. Минимизацию отношения Рэлея можно обобщить и на блочный случай. Тогда решение представляет собой набор из B тензоров, соответствующих минимальным собственным значениям. Эти тензоры можно записать в виде $(d+1)$-мерного тензора, который можно приблизить в *блочном TT-формате*:

$$U_(i_1, \ldots, i_d, \beta) \approx U_1(i_1) \ldots U_d^{>}(i_d, \beta), \quad \beta = 1, \ldots B,$$

и ортогональность такого вектора легко поддержать, если ортогонализовать все ядра слева направо. Для нахождения ядер можно применить минимизацию блочного отношения Рэлея:

$$\mathrm{Tr}(U^\top AU).$$

Адаптация метода переменных направлений в этом случае носит, однако, нетривиальный характер из-за наличия дополнительного индекса и условий ортогональности векторов. Но этого можно избежать, если после каждого шага оптимизации делать «перенос» индекса β с $k+1$ на k ядро по формуле

$$\widehat{U}(i_{k-1}, \beta)\widehat{U}_k(i_k) \approx U_{k-1}(i_{k-1})U_k(i_k, \beta),$$

что можно сделать с помощью сингулярного разложения. В таком представлении у нас всего получается представление для текущей аппроксимации вида

$$u = QU_k,$$

где Q – унитарная матрица размера $n^d \times (r_{k-1}n_kr_k)$, а матрица U_k имеет размер $(r_{k-1}n_kr_k) \times B$. Локальная задача после этого становится стандартной задачей на собственные значения. При этом «перенос индекса» имеет побочный эффект: при $B > 1$ он может быть использован для адаптации ранга даже без использования подходов DMRG и AMEN.

11.3.4. Динамические задачи и принцип Дирака — Френкеля

Третий класс задач, которые необходимо решать – это динамические задачи вида

$$\frac{dy}{dt} = Ay, \quad y(0) = y_0, \tag{11.3.4}$$

где A и y_0 представлены в ТТ-формате, а решение необходимо аппроксимировать в ТТ-формате. Здесь даже минимизационная постановка является нетривиальной задачей. Очень эффективный подход берет свое начало в работах Дирака и Френкеля по квантовой механике (*принцип Дирака — Френкеля*). Он имеет следующий вид. Для простоты пока рассмотрим не задачу (11.3.4), а задачу *динамической малоранговой аппроксимации*. Пусть у нас задана известная функция $Y(t)$ и ставится задача приближения $Y(t)$ тензором $\widehat{Y}(t)$ в ТТ-формате с рангами не выше r. Такую аппроксимацию можно строить поточечно, однако при этом теряется вся информация об «истории». Принцип Дирака — Френкеля состоит в следующем. Искомая траектория $\widehat{U}(t) \in \mathrm{TT}_r$ должна удовлетворять условию

$$\langle \frac{dY}{dt} - \frac{d\widehat{Y}}{dt}, v \rangle = 0 \tag{11.3.5}$$

для всех векторов v из касательного пространства к многообразию TT_r в данной точке Y. Смысл такого принципа простой: если скорость $\frac{d\widehat{Y}(t)}{dt}$ лежит в касательном пространстве, то решение всегда будет находиться на многообразии. Для уравнения вида (11.3.4) достаточно заменить $\frac{dY}{dt}$ на $A(Y)$.

Принцип Дирака — Френкеля приводит к системе дифференциальных уравнений для параметров ТТ-разложения $\widehat{Y}(t)$. Для простоты приведем эти уравнения для случая $d = 2$. В этом случае мы имеем многообразие матриц ранга не выше r. Рассмотрим следующую их параметризацию:

$$\widehat{Y}(t) = U(t)S(t)V^*(t),$$

где матрицы $U(t)$ и $V(t)$ являются унитарными матрицами с r столбцами, а матрица $S(t)$ (необязательно диагональная!) имеет размер $r \times r$. Тогда уравнения на $U(t)$, $S(t)$ и $V(t)$ имеют вид

$$\frac{dU}{dt}S = \frac{dY}{dt}V, \frac{dV}{dt}S^* = \frac{dY^*}{dt}U, \frac{dS}{dt} = U^*\frac{dY}{dt}V. \tag{11.3.6}$$

Отметим, что уравнения (11.3.6) достаточно сложно проинтегрировать, так как это система нелинейных обыкновенных дифференциальных

уравнений, при этом матрица S, входящая перед первой производной, может быть плохо обусловленной или даже вырожденной.

Однако для системы (11.3.6) (и ее тензорного аналога) можно предложить очень эффективную численную схему интегрирования, основанную на схеме дробных шагов. Для этого необходимо записать уравнение в виде

$$\frac{d\widehat{Y}}{dt} = P_Y(\frac{dY}{dt}),$$

где P_Y – проектор на касательное пространство в заданной точке, который имеет простой вид:

$$P_Y(Z) = UU^*Z + ZVV^* - UU^*ZVV^* = P_1(Z) + P_2(Z) - P_3(Z).$$

Дальше применяем схему дробных шагов. Оказывается, что уравнения с каждым из проекторов P_1, P_2, P_3 можно явно проинтегрировать. При этом порядок, в котором берется расщепление, играет ключевую роль. Шаг по $P_1(Z) = UU^*Z$ сводится к системе

$$\frac{dU}{dt} = 0, \quad \frac{d(VS^*)}{dt} = \frac{dY^*}{dt}U,$$

решением которой является $VS^* = L = (Y(t+\tau) - Y(t))U$. Для нахождения новых значений V, S достаточно использовать любую ортогональную факторизацию. Аналогичным образом выполняется и интегрирование по $P_2(Z) = ZVV^*$. Интересно выглядит интегрирование по $P_3(Z) = -UU^*ZVV^*$: матрицы U и V не меняются, а матрица S пересчитывается по формуле $S := S - U^*(Y(t+\tau) - Y(t))V$.

Отметим, что наиболее вычислительно эффективной является схема, когда дробные шаги выполняются в порядке P_1, P_3, P_2: тогда можно получить ряд интересных оценок. Получившаяся схема имеет первый порядок по шагу по времени τ. Легко получить схему второго порядка, если использовать схему Марчука: сначала делается шаг по P_1, потом по P_3, потом по P_2, потом по P_3 и, наконец, опять по P_1.

Предложенный подход можно обобщить и на d-мерный случай. При этом проход слева направо в одну сторону даст схему первого порядка, тогда как проход слева направо и справа налево даст схему расщепления второго порядка без дополнительной вычислительной сложности! При этом структура алгоритма остается точно такой же, как и для линейных систем и задач на собственные значения. Для линейных задач вида (11.3.4) каждая локальная задача будет просто линейной динамической задачей с $\mathcal{O}(nr^2)$

переменными, для решения которой можно использовать, например, методы, основанные на пространствах Крылова.

Глава 12.

Обзор методов вычислительной математики

Высокопроизводительные электронные вычислительные машины создали основу для алгоритмических построений и широких математических экспериментов во многих областях науки и техники. Это способствовало привлечению новых научных кадров к проблемам машинной математики. Ценный опыт, накопленный при решении прикладных задач, в дальнейшем был использован для построения эффективных методов и алгоритмов вычислительной математики. В данной главе мы кратко перечислим основные направления в вычислительной математике, которые сложились к настоящему времени, и отметим основные тенденции их развития.

12.1. Теория аппроксимации, устойчивости и сходимости разностных схем

Широкое использование метода конечных разностей для решения дифференциальных уравнений математической физики вызвало необходимость детального изучения тех свойств разностных уравнений, которые непосредственно влияют на качество разностных схем. Такими свойствами прежде всего являются устойчивость и сходимость.

Развитие теории устойчивости и сходимости началось после того, как расчеты на быстродействующих машинах показали, что разностная схема, аппроксимирующая корректную дифференциальную задачу, может оказаться неустойчивой (некорректной). Неустойчивая схема чувствительна к ошибкам округления, допускаемым в процессе счета, и поэтому может привести к решению, значительно отличающемуся от решения дифференци-

альной задачи. Эта отличительная и своеобразная черта разностных уравнений вызвала усиленные теоретические исследования по установлению связи между сходимостью и устойчивостью.

В середине 50-х гг. А. Ф. Филиппов [6], Лакс [6, 7], Рихтмайер [3, 6, 7], В. С. Рябенький [6], Дж. Нейман [7] почти одновременно и с разных позиций сформулировали следующий основной результат, получивший название теоремы эквивалентности: если линейная однородная дифференциальная задача корректна и разностная схема аппроксимирует эту задачу, то устойчивость разностной схемы является необходимым и достаточным условием сходимости решения разностной задачи к решению исходной задачи. В окончательном виде для абстрактного уравнения в банаховом пространстве эта теорема сформулирована и доказана В. С. Рябеньким и А. Ф. Филипповым [6]. Теорема эквивалентности сформулирована в терминах одной и той же нормы. Сходимость в других нормах, как правило, может быть установлена на основе теорем вложения С. Л. Соболева [1]. Повышая требования гладкости к начальным данным, можно ослабить требования на устойчивость схемы, на что первыми обратили внимание В. С. Рябенький и А. Ф. Филиппов [6]. Эта идея последовательно проведена в теореме эквивалентности Стрэнга [7] с использованием понятия слабой устойчивости.

Из исследований, связанных с поиском эффективных признаков устойчивости, в первую очередь следует отметить работу Неймана и Рихтмайера [7], в которой сформулирован локальный критерий устойчивости. Однако этот критерий верен лишь для уравнений с постоянными коэффициентами в случае самосопряженных задач, в связи с чем начались усиленные поиски границ применимости локального критерия.

Лакс и Ниренберг [6, 7] разработали теорию устойчивости гиперболических разностных схем в терминах так называемого символа разностной схемы. В случае явных разностных уравнений символ совпадает с обычной матрицей перехода, получаемой методом Фурье, при этом локальный критерий устойчивости оказывается справедливым, если коэффициенты имеют ограниченные вторые производные по x.

Стрэнг [7] сформулировал теорему сходимости для систем квазилинейных гиперболических уравнений при условии локальной устойчивости разностных уравнений, соответствующих первой вариации дифференциальной системы, и достаточной гладкости решения.

Изучение разностных схем с переменными коэффициентами связано с использованием понятия диссипативности. Прежде всего здесь следует отметить работу Крайса [6], который сформулировал теоремы о связи порядка диссипативности разностных уравнений, аппроксимирующих системы гиперболических уравнений, с порядком их точности. При этом матрич-

ные коэффициенты разностных уравнений предполагаются эрмитовыми и Липшиц-непрерывными как функции от x.

Интересный подход к исследованию устойчивости предложен в работах Н. Н. Яненко и Ю. И. Шокина [7]: вместо разностного уравнения рассматривается некое сопутствующее ему дифференциальное уравнение, называемое первым дифференциальным приближением, из некорректности которого следует неустойчивость разностной схемы.

Весьма важный класс разностных схем составляют схемы с положительными операторами, рассмотренные Фридрихсом [2], который ввел общее понятие положительных схем и установил для них достаточный критерий устойчивости в L_2.

С. К. Годунов и В. С. Рябенький [6, 7] ввели понятие спектра семейства разностных операторов, что позволило им сформулировать необходимое условие устойчивости разностных уравнений, хорошо вскрывающее сущность неустойчивости. Введено новое понятие ядра спектра семейства разностных операторов. В терминах радиусов ядер спектра даны оценки норм степеней операторов семейства, причем эти оценки оказываются равномерными для всего семейства и могут быть использованы для исследования устойчивости.

Все рассмотренные выше признаки устойчивости можно назвать спектральными, так как они основаны на изучении спектра разностных операторов. Эти признаки могут установить сходимость в норме L_2. Доказательства устойчивости разностных схем в C проводились С. И. Сердюковой [6], Томэ [6], А. А. Самарским [7] и др.

Из неспектральных подходов к изучению устойчивости разностных аналогов уравнений параболического и гиперболического типов необходимо указать на весьма общую теорию, построенную А. А. Самарским [3, 6, 7] на основе энергетических неравенств и априорных оценок. В этой теории для широкого класса двухслойных и трехслойных схем содержатся достаточные условия устойчивости, сформулированные в виде неравенств между операторными коэффициентами разностных схем. Эти условия весьма конструктивны и позволяют не только исследовать схемы на устойчивость, но и строить новые устойчивые схемы. Большое развитие в настоящее время получил энергетический метод. Идея метода состоит в выборе такой нормы для вектора решения, значения которой возрастают от шага к шагу не быстрее, чем $1 + O(\Delta t)$, что означает устойчивость в этой норме.

Энергетический метод исследования устойчивости появился еще в работе Куранта, Фридрихса и Леви [7] и был с успехом развит О. А. Ладыженской [7], Лизом [7], Лаксом [7], Крайсом [6], А. А. Самарским [3], А. Н. Коноваловым [15] и др.

Для гиперболических краевых задач весьма полное исследование проблемы устойчивости проведено Крайсом [6]. Им установлены достаточные условия устойчивости разностных аналогов при весьма общих предположениях относительно входных данных задач.

Теория аппроксимации и сходимости с общих позиций функционального анализа развита Л. В. Канторовичем и Г. П. Акиловым [1], которые рассмотрели широкий класс операторных уравнений, уделив особое внимание проблеме численного решения интегральных уравнений.

Важное значение для теории сходимости имеет разработанная С. Л. Соболевым ЦТ теория замыкания вычислительного алгоритма, широко используемая для теоретического обоснования приближенных методов решения задач математической физики.

12.2. Методы численного решения задач математической физики

Понятия аппроксимации, устойчивости и сходимости создали необходимую базу широкого поиска эффективных разностных схем для решения задач математической физики. Алгоритмы решения задач с помощью конечно-разностных методов, как правило, представляют собой сочетание методов построения разностных аналогов задач и методов их решения. Поэтому прогресс в конструктивной теории конечно-разностных методов обязан взаимосогласованному развитию указанных двух направлений исследований. Если попытаться просуммировать богатейший опыт в развитии конечно-разностных методов последних лет, то условно можно выделить следующие важные направления.

Построение разностных схем. Одно из таких направлений связано с разработкой методов построения консервативных разностных схем, основанных на законах сохранения, свойственных большинству физических процессов. Для конструирования консервативных разностных схем исходят из уравнений балансов, записанных для отдельной ячейки сеточной области, с последующим использованием квадратурных и интерполяционных формул. Построенные разностные уравнения после необходимых преобразований и суммирования по всем точкам сеточной области удовлетворяют дискретным аналогам интегральных законов сохранения.

Такие подходы рассмотрены О. А. Ладыженской [7], в работах которой построены разностные операторы для уравнений с разрывными коэффициентами, имеющие единый вид для любой внутренней точки области. Для обоснования алгоритма использовано понятие обобщенного решения и до-

казано, что решение разностной задачи образует некоторый функционал, переходящий при $h \to 0$ в функционал дифференциальной задачи.

Консервативные разностные схемы сквозного счета в гидродинамике разработаны С. К. Годуновым [4], Лаксом и Вендрофом [6] на основе явных разностных аппроксимаций. Большое значение для решения задач гидродинамики имеет метод интегральных соотношений, предложенный А. А. Дородницыным [3] и развитый О. М. Белоцерковским, П. И. Чушкиным [4] и др., в котором использована частичная разностная аппроксимация уравнений, записанных в дивергентной форме, на основе метода прямых. Эти методы сыграли существенную роль в формировании общего взгляда на конструкцию разностных схем для квазилинейных уравнений. Интересные общие подходы к интегрированию уравнений гидродинамики также предложены в работах К. II. Бабенко, В. В. Русанова [12], Фромма [4], Кроули [4], В. Ф. Куропатенко [4].

В последнее время большое внимание уделяется построению решений задач математической физики высокого порядка точности. Здесь в основном определились два направления. Первое связано с повышенным порядком аппроксимации разностных уравнений. Такие идеи рассмотрены в исследованиях А. А. Самарского [3], Н. Н. Яненко и А. Н. Валпуллина [3], Митчелла [3] и др. Второе направление связано с построением решений на основе разностных уравнений сравнительно невысокого порядка на последовательности сеток с изменяющимися шагами. Эти методы получили название экстраполяции по Ричардсону и нашли отражение в исследованиях Е. А. Волкова [4], Фокса [4], В. В. Шайдурова [4] и др.

Методы построения разностных схем для уравнений эллиптического и параболического типов в классе разрывных коэффициентов разработаны на основе интегро-интерполяционного метода А. Н. Тихоновым, А. А. Самарским [4] и др.

Вариационные и проекционные методы. В последние годы заметно повысился интерес к вариационным и проекционным методам решения задач математической физики. Эти методы давно заняли в вычислительной математике важное место. Особенно эффективны они в тех задачах, где искомыми являются функционалы от решения. Оказалось, что уже при сравнительно невысоких приближениях функционалы получаются с большой точностью. Наиболее полное теоретическое обоснование методов дано в исследованиях С. Г. Михлина [11], который установил необходимые и достаточные условия устойчивости вариационных методов в пространствах с энергетической нормой. Активное развитие вариационных методов обнаружило и некоторые их недостатки, связанные с трудностью построения базисных функций.

Новое направление в методе построения разностных уравнений для задач математической физики было развито на основе использования вариационных и проекционных методов в сочетании со специальной конструкцией базисных функций, отличных от нуля в некоторых сравнительно небольших областях, принадлежащих всей области определения решения. Первые работы Куранта [5], Л. А. Оганесяна [5], Лионса, Темама [5], Сеа [5], Обэна [5], Биркгофа, Шульца, Варги [5], Брембела положили начало развитию этого направления. Развитие этих методов обязано прежде всего работам Бабушки [5], Стрэнга и Фикса [5], Зламала [5], Дугласа и Дюпона [5], В. Я. Ривкинда [4], В. И. Агошкова [5], Г. И. Марчука, В. И. Агошкова [5], А. М. Мацокина [4], В. В. Шайдурова [3] и др.

Решение многомерных стационарных задач. Интенсивное развитие методов решения линейных алгебраических уравнений с якобиевыми и блочно-трехдиагональными матрицами приводит к созданию ряда первоклассных численных алгоритмов решения стационарных задач, основанных на факторизации разностного оператора задачи. Среди методов факторизации особое место занимают различные варианты безытерационных методов матричной факторизации, разработанные В. С. Владимировым [12], М. В. Келдышем [5], И. М. Гельфандом, О. В. Локуциевским [12], К. И. Бабенко, В. В. Русановым [12], Н. Н. Ченцовым, С. К. Годуновым [12], А. А. Абрамовым, В. Б. Андреевым [12] и др.

На основе работ М. И. Вишика, С. Л. Соболева и Л. А. Люстерника для решения краевых задач в областях сложной геометрии В. К. Саульевым [4] был предложен метод фиктивных областей, исследованный впоследствии Мино [2], В. Д. Копченовым [4], Л. А. Руховцом [4], А. Н. Коноваловым [4] и др. Исследованию разностных уравнений, аппроксимирующих краевую задачу, возмущенную по методу фиктивных областей, посвящены, например, работы В. И. Лебедева [4] и В. Я. Ривкинда [4].

В последнее время интенсивно развиваются безытерационные методы решения разностных уравнений, соответствующих дифференциальным задачам, допускающим разделение переменных. Разработанная специальная техника быстрого преобразования Фурье и циклической редукции позволяет существенно сократить объем вычислительной работы. Это направление представлено работами Кули и Таки [13], Бужби, Голуба и Нилсона [12], Хокни [13] и др.

Вместе с развитием методов точной факторизации активно развиваются методы приближенной факторизации, в которых факторизация оператора комбинируется с методом последовательных приближений. Необходимость в таких алгоритмах обнаружилась сразу же, как только задачи математической физики начали редуцироваться к большим алгебраическим

системам. Первые работы Н. И. Булеева [15], Бейкера и Олифанта [15] дали толчок к развитию новых методов решения многомерных задач на основе быстросходящихся процессов.

Начало 60-х гг. ознаменовалось крупным вкладом в вычислительную математику, связанным с именами Дугласа, Писсмана и Рэчфорда [15], предложившими метод переменных направлений. Успех метода был обеспечен использованием простой редукции многомерной задачи к последовательности одномерных с матрицами якобиевого типа, легко обращаемыми на ЭВМ. В конечном итоге метод продольно-поперечных направлений сводится к итерационному методу, в котором оптимизация вычислений осуществляется специальным подбором оператора сжатия, состоящего из произведения более простых операторов и ряда свободных параметров релаксации. При этом последовательное обращение простых операторов, как правило, осуществляется на основе одномерной факторизации. Такие итерационные схемы весьма экономичны и эффективны при незначительном, по сравнению с явным методом Ричардсона, увеличении объема вычислительной работы в расчете на одну итерацию. Метод переменных направлений оказал существенное влияние на построение алгоритмов в различных областях прикладной математики и развитие исследований по нелокальным и блочно-итерационным процессам. Теоретическим исследованиям этого и родственных ему методов посвящены работы Дугласа [15], Е. Г. Дьяконова [15], А. А. Самарского [3], Биркгофа, Варги, Янга [15], Вакспресса [15], Келлога [15], Ганна [15], Ю. В. Воробьева [15] и др.

Развиваются методы, основанные на однородных и неоднородных аппроксимациях. В случае неоднородной аппроксимации каждая из вспомогательных задач может и не аппроксимировать исходную задачу, но в совокупности и в специальных нормах такая аппроксимация имеет место. Эти методы были названы методами расщепления, они развиты в работах советских математиков Н. Н. Яненко [3, 15], Е. Г. Дьяконова [3, 15], А. А. Самарского [3, 15], В. К. Саульева [3], Г. И. Марчука [15] и др.

Большой цикл исследований посвящен выбору оптимизационных параметров схем расщепления на основе спектральных и вариационных методов. Это работы А. А. Самарского [3, 15], Е. Г. Дьяконова [15], В. П. Ильина [3, 15] и др.

Различные аспекты теории метода попеременных направлений и метода расщеплений рассмотрены в работах В. Б. Андреева [15], Видлунда [15], Фейрвезера, Митчелла [15] и др.

Решение многомерных нестационарных задач. Опыт решения одномерных задач подготовил основу для формирования алгоритмов решения более сложных задач математической физики. Важным этапом в развитии

методов решения нестационарных двумерных задач явился метод попеременных направлений, основанный на однородной аппроксимации; первоначально он был применен для решения многомерных уравнений параболического типа и затем получил широкое распространение во многих задачах математической физики.

Развитие методов решения многомерных нестационарных задач связано с методами расщепления, основанными, как правило, на неоднородных разностных аппроксимациях исходной задачи. Сущность метода расщепления состоит в редукции сложного оператора к простейшим. При таком подходе интегрирование данного уравнения сводится к последовательному интегрированию уравнений более простой структуры. При этом разностные схемы обязаны удовлетворять условиям аппроксимации и устойчивости только в конечном итоге. Это дает возможность гибкого построения схем по существу для всех основных задач математической физики. Для явных аппроксимаций метод расщепления был предложен К. А. Багриновским и С. К. Годуновым [15]. Схемы расщепления для неявных аппроксимаций предложены Н. Н. Яненко [15], Е. Г. Дьяконовым [15], А. А. Самарским [15], Г. И. Марчуком [3] и др.

Эти методы нашли широкое применение для разнообразных по своему характеру задач и стимулировали формирование более общего подхода к решению задач математической физики на основе метода слабой аппроксимации, разработанного Н. Н. Яненко [3, 15], А. А. Самарским [3, 15]. Оказалось, что метод расщепления можно толковать как метод слабой аппроксимации исходного уравнения некоторым другим, более простым. Условия, при которых имеет место сходимость решения для метода слабой аппроксимации, сформулированы в теореме Яненко — Демидова [15] и в работах В. И. Лебедева [15] и Е. Г. Дьяконова [3, 15]. Метод слабой аппроксимации нашел естественное применение в задачах гидродинамики, метеорологии, океанологии, теории переноса излучения и т. д. (Г. И. Марчук [3, 17], Н. Н. Яненко [3]).

Широкое применение в задачах гидродинамики, метеорологии, океанологии получила оригинальная схема типа предиктор-корректор Лакса — Вендрофа, в которой предиктор предложен в виде явной разностной схемы. Эта схема является условно устойчивой, она проста в реализации и имеет второй порядок аппроксимации по всем переменным. Подробное исследование схемы приведено в книге Рихтмайера и Мортона [3].

Различные варианты метода предиктор-корректор на основе неявных разностных аппроксимаций предложены Брайеном [4], Дугласом [15], И. Д. Софроновым [12], Г. И. Марчуком, Н. Н. Яненко [15] и др. Оказалось, что все эти схемы в известном смысле эквивалентны и различаются только ме-

тодами реализации. В последней из перечисленных работ в качестве предиктора применена неявная схема расщепления первого порядка точности с факторизованным оператором. Для задач гидродинамики в качестве предиктора используются неявные мажорантные схемы.

Особый интерес представляет сформулированный Лионсом и Темамом [5], а также Бенсусаном, Лионсом и Темамом [15] метод декомпозиции и децентрализации, который примыкает к методам расщепления и слабой аппроксимации.

Метод частиц в ячейке. В последние годы интенсивно развивается новый метод решения многомерных задач математической физики, связанный с именем Харлоу [19]. Этот метод получил название метода больших частиц. Он широко применяется для расчета многомерных гидродинамических течений с сильными деформациями жидкости, большими относительными перемещениями и соударяющимися поверхностями раздела. Сущность метода состоит в следующем. Уравнения гидродинамики на основе слабой аппроксимации на каждом малом временном интервале сводятся к двум более простым системам, первая из которых описывает адаптацию гидродинамических полей между собой без учета адвективных членов и интегрируется обычными способами в неподвижной эйлеровой сетке, а вторая описывает перенос субстанций в лагранжевой системе координат. Именно при решении второй системы используется феноменологическое упрощение модели сплошной среды на основе замены ее системой частиц в каждой ячейке эйлеровой системы, так что суммарный баланс массы, импульса и энергии частиц в ячейке отождествляется с соответствующими характеристиками для сплошной среды. Как только некоторая частица, «несущая» определенную массу в соответствии со своей траекторией, рассчитываемой индивидуально, пересекает границу ячейки, масса, импульс и энергия этой частицы вычитаются из покинутой ячейки и добавляются в новую ячейку, где теперь находится частица. Схема Харлоу основана на явных методах решения уравнений первого и второго этапов, в целом она условно устойчива. Особенно плодотворным является использование в расчетах первого шага неявных схем. В этом случае критерий устойчивости всей схемы совпадает с известным условием Куранта. До сих пор еще не получены абсолютно устойчивые схемы метода частиц, однако в ближайшие годы можно рассчитывать на существенный прогресс в этом направлении.

В последнее время в работах В. Ф. Дьяченко [19], О. М. Белоцерковского и Ю. М. Давыдова [19], Н. Н. Яненко, Н. Н. Анучиной, В. Е. Петренко, Ю. И. Шокина [19] даны различные модификации метода, которые существенно уменьшили свойственные ему флюктуации плотности и давления,

увеличили «запас устойчивости», и рассмотрены различные схемы реализации.

Можно надеяться, что применение абсолютно устойчивых методов и устранение флюктуации позволят распространить метод частиц на слабосжимаемые течения жидкости. В ближайшие годы можно ожидать существенного расширения сферы влияния этого метода на решение многомерных задач.

Метод Монте-Карло, предложенный Нейманом и Уламом, активно развивается уже более двух десятилетий. Первоначальный оптимизм в применении метода через некоторое время уступил место столь же необоснованному пессимизму. Дело в том, что уже на первых этапах развития оказалось, что метод Монте-Карло эффективен только при реализации на быстродействующих ЭВМ с миллионами операций в секунду, поскольку он требует выполнения большого числа статистических проб, понижающих среднюю квадратичную ошибку в получаемом результате.

Однако несмотря на трудности в осуществлении метода на ЭВМ среднего класса, а может быть, и благодаря им, в теорию метода были внесены усовершенствования, которые существенно повысили эффективность метода в решении большого круга задач науки и техники. Наиболее значительные усовершенствования связаны с привлечением для расчетов условных вероятностей процессов и статистических весов, определяемых на основе информации о решениях сопряженных уравнений по отношению к существенным функционалам задач. Такие методы на два порядка, а в некоторых случаях и на четыре уменьшили дисперсию ошибки и, следовательно, на два и четыре порядка сократили время счета по сравнению с методами прямого статистического моделирования.

В настоящее время ЭВМ третьего поколения создали необходимую базу для активного применения этого метода к различным сложным задачам математической физики. Метод Монте-Карло уже имеет солидные позиции в теории переноса излучения, задачах массового обслуживания, кубатурных и интерполяционных процессах, решениях интегральных уравнений и систем алгебраических уравнений. В последнее время он начинает использоваться для решения нелинейных уравнений Больцмана, в задачах линейного программирования и т. д.

Большой вклад в теорию и алгоритмы решения задач математической физики методом Монте-Карло внесли работы В. С. Владимирова [18], Н. С. Бахвалова [18], И. М. Соболя [18], Н. Н. Ченцова [18], Фано, Спансера, Бергера [18], С. М. Ермакова, В. Г. Золотухина [18], Г. А. Михайлова [18], Н. П. Бусленко, Д. И. Голенко [18] и др. Простой и универсальный метод Монте-

Карло, несомненно, станет активным средством вычислительной математики.

12.3. Условно корректные задачи

При решении задач математической физики численными методами важную роль играет корректность постановки исследуемой задачи. Понятие корректности было введено Адамаром. Известно большое число классических задач математической физики, поставленных корректно по Адамару. В связи с более глубоким изучением различных задач естествознания и техники возникла проблема решения так называемых условно корректных задач. А. Н. Тихонов [16] сформулировал требования, которые оказываются естественными в постановке задач, некорректных по Адамару. Сущность этих требований состоит в том, что в условия постановки задачи добавляется априорное предположение о существовании решения и принадлежности его заданному компакту. Для установления условной корректности необходимо доказать теорему единственности.

Широкий цикл исследований по условно корректным задачам проведен М. М. Лаврентьевым [16] и В. К. Ивановым [16]. Различные аспекты теории условно корректных задач математической физики рассмотрены в трудах Джона [16], С. Н. Мергеляна [16], Дугласа [16], С. Г. Крейна [16] и др.

А. Н. Тихонов [16] ввел понятие регуляризации. Сущность его состоит в том, что вместо неограниченного оператора, дающего точную формулу решения некорректно поставленной задачи, рассматривается такая последовательность (регуляризующее семейство) непрерывных операторов, что на каждом элементе, принадлежащем области существования решения, соответствующая последовательность сходится к решению.

Одним из интересных подходов к постановке задач, некорректных по Адамару, является применение понятий и методов теории вероятности. В наиболее полной форме такие исследования были развиты М. М. Лаврентьевым и В. Г. Васильевым [16]. В работах этого направления устанавливается понятие устойчивости, конструируются оптимальные в определенном смысле алгоритмы решения различных классов задач при некоторых предположениях о вероятностных свойствах погрешностей во входных данных и о вероятностных свойствах множества искомых решений.

Лионс и Латтес [16] сформулировали численный метод решения обратных эволюционных уравнений на основе так называемого квазиобращения. К эволюционному уравнению добавляется регуляризующий оператор

с малым параметром, являющийся произведением исходного оператора на его сопряженный. Малый параметр выбирается на основе специальным образом разработанных оптимальных оценок в решении. Метод квазиобращения весьма прост в реализации для решения эволюционных задач математической физики.

Автором и С. А. Атанбаевым [16] разработан метод решения условно корректных задач эволюционного типа на основе применения метода минимальных невязок для всей пространственно-временной области определения решения. Регуляризация в этом методе производится за счет выбора оптимального числа шагов итерационного процесса на основе априорной оценки погрешностей во входных данных. Весьма полное исследование по теории некорректно поставленных задач и методов регуляризации дано в работах В. А. Морозова [14, 16].

Тенденция развития методов решения условно корректных задач свидетельствует о том, что используемые методы тесно примыкают к методам оптимизации вычислительного процесса.

12.4. Вычислительные методы в линейной алгебре

Необходимо отметить все возрастающий интерес к решению больших систем линейных алгебраических уравнений как с разреженными, так и с плотными матрицами, решению плохо обусловленных систем и спектральных задач для матриц произвольной структуры. Большое внимание при этом уделяется использованию априорной и апостериорной информации о задаче в ходе ее решения. Существенное влияние на пересмотр старых вычислительных методов линейной алгебры оказали ЭВМ, которые стимулировали интерес к новым алгоритмам, приспособленным для автоматического счета.

Прямые методы линейной алгебры. Под прямым методом линейной алгебры обычно понимают метод, которым можно решить задачу за конечное число арифметических действий. В вычислительной линейной алгебре прямые методы играют важную роль при решении систем линейных уравнении, вычислении обратных матриц и определителей. Прямые методы позволяют с помощью ряда элементарных преобразований получить разложение исходной матрицы в произведение двух, каждая из которых легко обращается.

Классическими примерами прямых методов служат метод исключения Гаусса, методы вращения и отражения. Вторую группу составляют так

называемые методы сопряженных направлений: метод сопряженных градиентов Хестенса и Штифеля [11] и метод минимальных итераций Ланцоша [3]. Работы этих авторов положили начало развитию методов, основанных на ортогонализации.

В последние годы прямые методы получили значительное развитие в первую очередь благодаря исследованиям Д. К. Фаддеева, В. Н. Фаддеевой, В. Н. Кублановской [18], Бауэра [8], Хаусхолдера [3], Уилкинсона [8], Хенричи [3], Форсайта, Молера [8], Голуба [12], В. В. Воеводина [8] и др.

Большой проблемой по-прежнему остается решение систем уравнений с плохо обусловленными матрицами, которая тесно связана с решением условно корректных задач математической физики. Сложность проблемы связана с сильной чувствительностью решения к точности задания элементов матрицы и компонент вектора правой части системы. Хотя уже получен ряд важных результатов, тем не менее это только начало большого научного поиска, который должен завершиться созданием общей теории.

Итерационные методы. Важнейшим средством решения задач линейной алгебры являются итерационные методы, активное развитие которых привело к созданию ряда хороших алгоритмов, эффективно реализуемых на ЭВМ. Этот прогресс в первую очередь был вызван необходимостью решать задачи математической физики, экономики и управления, приводящие к системам большого порядка с матрицами специального вида. Прямые методы в большинстве случаев оказываются малоэффективными при решении таких задач, хотя каждый новый этап в развитии вычислительной техники и расширяет их возможности.

К настоящему времени определились некоторые направления в построении итерационных методов; мы ограничимся рассмотрением только двух из них. Первое основано на использовании спектральных характеристик операторов, участвующих в процессе. Методы этого типа можно описать следующим образом: строится итерационный процесс с матрицей перехода, зависящей от совокупности параметров, и эти параметры выбираются либо одинаковыми для всех шагов из условия минимизации спектрального радиуса матрицы перехода, либо строится последовательность значений параметров, зависящих от номера итерации так, чтобы вектор ошибки стремился как можно быстрее к нулю равномерно по всем начальным приближениям. Оба способа используют априорную информацию о спектрах участвующих матриц. Выбор таких параметров является составной частью проблемы оптимизации вычислительного алгоритма. Наибольшая трудность этапа состоит в определении границ спектров участвующих матриц.

Активный прогресс в области спектральной оптимизации итерационных методов стимулирует постановку ряда проблем. Следует иметь в виду, что спектральные методы оптимизации особенно эффективны в случае решения серии задач с одним и тем же оператором, но разными входными данными.

Второе направление связано с применением вариационных принципов. Методы этого класса осуществляют последовательную минимизацию некоторого функционала (как правило, квадратичного), который достигает минимального значения на искомом решении системы. Основы вариационного подхода к построению итерационных методов заложены Л. В. Канторовичем [11], Ланцошем [3], Хестенсом, Штифелем [11], М. А. Красносельским, С. Г. Крейном [11] и др. Из последних исследований нужно отметить работы Петришина [9, 10], Форсайта [11], Даниеля [11], Г. И. Марчука, Ю. А. Кузнецова [3, 11], С. К. Годунова, Г. П. Прокопова [11], В. И. Лебедева [9], Н. И. Горбенко, В. П. Ильина [11] и др.

Достоинство вариационных методов типа наискорейшего спуска и итерационного процесса с минимальными невязками состоит в том, что параметры релаксации выбираются за счет использования апостериорной информации, получаемой на каждом шаге. Скорость сходимости таких многошаговых методов не ниже, чем для методов, использующих полиномы Чебышева. Существенным является также то, что такие методы сходятся как для симметричных, так и для несимметричных матриц при условии их положительной определенности. В последнее время удалось построить ряд эффективных методов типа минимальных невязок для положительно полуопределенных матриц.

Важным обстоятельством, сдерживающим до настоящего времени развитие нестационарных вариационных методов, является необходимость хранить большее, по сравнению с чебышевскими методами оптимизации, количество промежуточной информации.

В последнее время развиваются итерационные методы, в которых сочетается подход спектральных и вариационных оптимизаций. В. И. Лебедев сформулировал условия на операторы задач, для которых итерационный процесс имеет неулучшаемую оценку числа арифметических операций. Развивается еще один метод выбора оптимальных параметров итерации, основанный на вероятностном подходе. Ряд интересных результатов в этой области получен Ю. В. Воробьевым [9]. До сих пор не утратил своего большого значения ставший уже классическим метод верхней релаксации Янга — Франкела [10]. Исследования этого метода обобщены в монографиях Вазова, Форсайта [3], Варги [3], Изаксона, Келлера [3], Янга [10] и др.

Обзор и систематизация итерационных методов даны в книге Г. И. Марчука и В. И. Лебедева [17].

Большой круг исследований был выполнен по итерационным методам решения линейных систем с особенными матрицами. Для случая совместных систем автором и Ю. А. Кузнецовым [8, 11] был предложен общий подход к исследованию сходимости стационарных и нестационарных итерационных методов. Этот подход позволил не только расширить область применимости известных итерационных методов, но и дал возможность разработать новый класс методов, получивших название матричных аналогов метода фиктивных областей (см. Ю. А. Кузнецов, А. М. Мацокин [4]). Итерационные методы решения несовместных систем были предложены в работах Ю. А. Кузнецова [8] и др.

Остановимся на итерационных методах решения полной проблемы собственных значений для общих матриц. Рассмотрим только степенные методы, поскольку именно здесь в последнее время получены существенные результаты, чем мы обязаны исследованиям Уилкинсона [8], Бауэра [8], Коллатца [3], В. В. Воеводина [8], Френсиса [8], В. Н. Кублановской [8], Эберлейна [8] и многих других.

Степенные методы основаны на последовательном приведении исходной матрицы с помощью унитарных преобразований подобия (метод Якоби, QR-алгоритм) или преобразований подобия с треугольными матрицами (LR-алгоритм) к матрице, собственные значения которой легко вычисляются. Такими матрицами являются диагональная, треугольная или блочно-треугольная, порядки диагональных блоков которой не выше двух.

До последнего времени существовали эффективные алгоритмы решения проблемы собственных значений лишь для симметричных матриц, такие как метод Якоби (Д. К. Фаддеев, В. Н. Фаддеева [8], Рутисхаузер [8]). Разработка QR-алгоритма (В. Н. Кублановская [8], Френсис [8], Уилкинсон [8]) и обобщенного метода вращений (В. В. Воеводин [8]) позволяет говорить о решении проблемы для произвольных матриц. Наиболее интенсивно в настоящее время разрабатываются различные модификации QR-алгоритма.

Прогресс в развитии проблемы собственных чисел имеется также в связи с работами в области расчета ядерных реакторов, стимулировавшими изучение итерационных методов решения частной проблемы собственных чисел для неотрицательных матриц. Основы теории заложены в трудах Перрона, Фробениуса и значительно развиты в исследованиях Варги [3], Трауба [8], Марека [8] и др.

Анализ ошибок округления. Если до последнего времени вычислительные методы сравнивались между собой по количеству арифметических

действий и объему памяти, которые требовались для их реализации, то теперь к этим характеристикам добавилась точность. Это означает, что анализ ошибок округления при реализации метода на ЭВМ стал одной из составных частей алгоритма.

Начало исследованиям в этой области положено работами Неймана. Систематическое изучение ошибок впервые было проведено Уилкинсоном [8]. Основу математического аппарата Уилкинсона составил метод эквивалентных возмущений, с помощью которых получены оценки норм возмущений для всех преобразований линейной алгебры и построены оценки норм эквивалентных возмущений для большого числа методов.

Параллельно с методом эквивалентных возмущений интенсивно развивалась статистическая теория анализа ошибок. Результаты, полученные Н. С. Бахваловым [8], В. В. Воеводиным [8], Г. Д. Ким [8] и др., положили начало исследованию действительного распределения ошибок округления.

Комплексы стандартных программ. Следствием успехов, достигнутых в вычислительной линейной алгебре, явилась разработка высококачественных стандартных программ для решения систем линейных уравнений и нахождения собственных значений. Так, например, в журнале Numerische Mathematik уже опубликовано большое число различных процедур, которые широко используются как для решения общих задач линейной алгебры, так и для ряда специальных задач математической физики, экономики и т. д., связанных с матрицами специального вида.

Указанная проблема, несомненно, привлечет внимание исследователей, и результатом их усилий должно явиться создание универсальной вычислительной системы решения задач линейной алгебры. Можно указать по крайней мере на две тенденции, которые уже наметились в развитии этого направления: одна связана с тщательной отработкой комплексов алгоритмов и программ решения общих задач, другая состоит в создании универсальных методов, адаптирующихся к конкретным особенностям классов задач. Обе тенденции крайне интересны, поскольку прокладывают пути к системе универсального проблемно-ориентированного математического обеспечения для ЭВМ четвертого и последующих поколений.

12.5. Вопросы оптимизации численных методов

Важной целью вычислительной математики является отыскание наиболее быстрых и экономически выгодных методов решения задач, т. е. оптимизация вычислительных алгоритмов. Проблему оптимизации решения

при заданных ограничениях необходимо изучать с помощью общих математических теорем и оценивать минимально возможные затраты на решение конкретной задачи из заданного класса или суммы задач. Рассмотрение одной изолированной математической задачи оптимизации большей частью не решает практического вопроса. Однако, умея находить условный экстремум, т. е. наилучший способ решения при заданных возможностях и средствах вычислений каждой локальной задачи, мы тем самым подходим к решению общей проблемы. Эта концепция теории оптимизации вычислительных методов, сформулированная Н. С. Бахваловым [20], С. М. Никольским [3], И. Бабушкой и С. Л. Соболевым [20], достаточно хорошо отражает существо поставленной проблемы.

Во многих случаях, однако, построить оптимальный алгоритм не удается, хотя и оказывается возможным построить алгоритм, близкий к оптимальному. Такая ситуация типична, например, при построении асимптотически оптимальных алгоритмов. Можно отметить, что в настоящее время именно теория асимптотических оценок является эффективным средством решения проблем оптимизации алгоритмов для различных классов задач.

К настоящему времени наиболее развитой с точки зрения теории оптимизации является теория кубатурных формул, разработанная С. Л. Соболевым [1, 20], И. Бабушкой [20]. В их работах задача оценки кубатурных формул приводится к решению задачи отыскания минимума линейного функционала ошибок. Получены оценки ошибки кубатурных формул с правильной решеткой на классах периодических финитных и бесконечно дифференцируемых функций. Методы исследования существенно используют асимптотические оценки приближений. Теории кубатурных формул посвящены исследования Н. С. Бахвалова [20] и И. М. Соболя [18], связанные с оптимальными оценками сходимости кубатурных процессов, с методами интегрирования типа метода Монте-Карло, а также с отысканием наилучших способов численного интегрирования.

Несколько иной подход к построению кубатурных формул на основе теоретико-числового анализа, заложенный трудами И. М. Виноградова [20], развивается в работах Н. М. Коробова [20], К. К. Фролова [20], где строятся формулы, точные для конечных тригонометрических полиномов, и даются оценки погрешностей на классе периодических функций.

А. Н. Колмогоров [20] ввел в рассмотрение ряд понятий теории множеств общего характера, позволяющих найти оценку границы для необходимого числа действий при решении вычислительных задач. Особое значение такие оценки имеют для направленного поиска алгоритмов в тех случаях, когда асимптотики сверху и снизу расходятся. Для линейных дифференциальных операторов, имеющих вполне непрерывный обратный, им

дана оценка числа необходимых для решения действий. Эта оценка позволяет находить алгоритмы, асимптотически близкие к оптимальным по числу арифметических действий.

Н. С. Бахваловым [20] исследован комплекс алгоритмов решения задач математической физики конечно-разностными методами. В частности, им даны оценки снизу количества действий при решении задачи Дирихле для уравнения Лапласа.

Интересное направление в оптимизации решения задач математической физики развивается в работах В. И. Лебедева [9]. В качестве основного минимизируемого функционала рассматривается цена алгоритма и энтропия. С помощью этого метода рассмотрены некоторые задачи теории переноса.

При решении проблемы оптимизации зачастую приходится абстрагироваться от многих факторов, таких как методы округления чисел в процессе реализации алгоритма, особенности осуществления арифметических операций в регистрах конкретных ЭВМ и т. д. Между тем именно эти факторы в ряде случаев определяют эффективность алгоритма. Следовательно, здесь необходимо говорить уже об оптимизации вычислительного процесса.

Изучению теории вычислительных процессов и их оптимизации посвящено значительное число исследований Бабушки [20], Дальквиста [20], Хенричи [3] и др. Бабушка, Витасек и Прагер [3] ввели понятие α_k-последовательностей вычислительных процессов, которое отражает тот факт, что при увеличении длины последовательности вычислений точность вычислений должна увеличиваться по степенному закону. На основе теории α_k-последовательностей было введено понятие локальной и глобальной устойчивости численных процессов, которое позволило провести анализ большого круга реальных алгоритмов вычислительной математики.

В последние годы возникло направление в теории оценки точности реального алгоритма на ЭВМ, получившее название интервальной арифметики, разработанное в трудах Мура [20], Никела [20] и др. Основная цель интервальной арифметики состоит в получении апостериорных оценок погрешности, получаемых двукратным счетом на одной и той же ЭВМ.

12.6. Методы оптимизации

С задачами на связанный экстремум человечество знакомо с глубокой древности. История развития этой области знаний в рамках математики может быть разделена, конечно, весьма условно, на ряд этапов. Сначала рас-

сматривались отдельные задачи, главным образом геометрического происхождения, причем среди них имеются весьма знаменитые, не потерявшие своего значения до наших дней. Например, решение изопериметрической задачи было известно еще древним грекам (Зенодор). Общий метод решения аналитических задач на связанный максимум или минимум принадлежит Лагранжу. Наличие техники множителей Лагранжа позволяет многие задачи решать стандартным путем, не подвергая их порой весьма сложному индивидуальному исследованию. Однако при этом решаются лишь задачи на гладких многообразиях. Границы многообразии, то есть ограничения типа неравенств, должны исследоваться отдельно.

Следующий этап можно связать с именем Минковского. Укажем, в частности, на его теоремы о возможности получить всякое следствие из системы неравенств путем их комбинирования с неотрицательными множителями и о конечной порождаемости множества решений системы линейных неравенств. В этих теоремах, по сути дела, заложен фундамент для всей теории линейного программирования. Но фактическое создание линейного программирования началось значительно позже и связано с именами Канторовича и Данцига. После этого началось активное развитие методов оптимизации. В разных странах появились работы по линейному, выпуклому, общему нелинейному, динамическому программированию. Большое число работ посвящено численным методам решения таких задач и их всевозможным приложениям. В настоящее время эта работа продолжается, причем на первое место выходит проблема корректной и экономной алгоритмизации методов, их реализация на ЭВМ в виде пакетов прикладных программ, удобных для решения больших классов практических задач.

Что касается создания современной теории оптимального уравнения, то, по-видимому, впервые вопросы получения оптимальных законов управления серьезно были поставлены в теории движения ракет. Однако с точки зрения общей теории полученные результаты носили частный характер. Дело в том, что большинство задач приводит к неклассическим вариационным постановкам — задачам с ограничениями.

Впервые наиболее важные результаты были получены в теории систем, обеспечивающих минимальное время регулирования.

В 1952—1955 гг. главным образом в работах Фельдбаума были заложены основы теории оптимальных по быстродействию процессов для линейных систем.

В 1956 г. Л. С. Понтрягиным был сформулирован принцип, ведущий к решению общей задачи о нахождении оптимального по быстродействию процесса регулирования. Этот принцип, получивший наименование принципа максимума, был проведен сначала для отдельных типов систем и,

в частности, доказан Р. В. Гамкрелидзе для случая линейных систем. В. Г. Болтянский полностью доказал принцип максимума Понтрягина в качестве необходимого условия оптимальности по быстродействию. Затем принцип максимума был распространен на общий случай минимизации произвольного функционала типа интеграла функции от переменных системы.

При рассмотрении вопросов, касающихся теории оптимальных процессов, необходимо отметить многочисленные работы Р. Беллмана. Метод динамического программирования, развитый им, дает новый аппарат для решения вариационных задач и тесно связан с принципом максимума Понтрягина.

Как уже отмечалось в главе 10, в теории оптимального управления рассматриваются такие задачи с запаздыванием, с дискретным временем, с параметрами, с изопериметрическими условиями и аналогичные задачи оптимального управления для уравнений с частными производными. Работы, в которых содержатся другие концепции и дальнейшее развитие теории оптимального управления, приведены в списке литературы [22].

Многие задачи математической физики допускают естественную вариационную постановку, когда задача сводится к отысканию экстремума некоторого функционала. Вариационный подход позволяет снять ограничения гладкости на искомое решение, не вызванные физической природой явления, и, кроме того, дает возможность строить заведомо устойчивые разностные схемы (см. гл. 2).

В приложениях часто приходится иметь дело с задачами, которые приводятся к экстремальным, но на более узком множестве, чем традиционные, причем соответствующие функционалы могут не обладать гладкостью, необходимой для применения классических методов вариационного исчисления. Для исследования такого рода задач с ограничениями были привлечены вариационные неравенства, что позволило решить довольно сложные задачи механики и физики, до того не поддававшиеся решению.

Вариационные неравенства возникают в ряде разделов механики сплошных сред, в задачах со свободной границей, во многих задачах оптимального управления и т. д. Вопросы теории данного подхода развивались в основном в работах французских математиков [21]. Систематическое изложение численных методов исследования вариационных неравенств, возникающих в различных приложениях, можно найти в книге Гловинского, Лионса, Трепольера «Численное исследование вариационных неравенств»,которая вышла в русском переводе в 1979 г.

12.7. Методы Шварца и разделения области

Альтернирующий метод Шварца известен в вычислительной математике уже многие годы (В. И. Смирнов [2], С. Л. Соболев [23], С. Г. Михлин [23], С. К. Годунов [2] и др.). Он всегда рассматривался прежде всего как метод, позволяющий сводить решение исходной задачи в области со сложной формой границы к последовательности задач в подобластях, форма которых достаточно простая. В последние годы интерес к этому методу значительно возрос (Е. А. Волков [23], С. Е. Романова [23], А. М. Мацокин [23], А. М. Мацокин, С. В. Непомнящих [23], П.-Л. Лионс [23] и др.). Одной из причин этого является то обстоятельство, что в настоящее время разработаны достаточно эффективные алгоритмы численного решения ряда задач математической физики именно для случая простых областей. С учетом этого обстоятельства в последние годы осуществляется поиск модификаций метода Шварца, обладающих более высокой скоростью сходимости, по сравнению с его классическим вариантом. Другим методом, позволяющим часто сводить решения задач к решениям последовательности задач в подобластях, имеющих простую форму границы, является метод разделения области (Э. И. Матеева, Б. В. Пальцев [23], В. Э. Кацнельсон. В. В. Меньшиков [23], Л. Б. Цвик [23], В. В. Смелов [23], А. М. Мацокин [23], В. И. Лебедев, В. И. Агошков [23], Г. И. Марчук, Ю. А. Кузнецов [23], Р. Гловинский [23], М. Дрыя [23], Е. Г. Дьяконов [23] и др.). В первых же исследованиях по данному методу была отмечена необходимость введения в итерационный процесс, лежащий в основе метода, некоторых параметров. Эти параметры необходимо было выбирать так, чтобы процесс сходился. Некоторые подходы по выбору оптимальных значений этих параметров осуществлены В. Э. Кацнельсоном, В. В. Меньшиковым [23]. Л. Б. Цвиком [23] параметры итерационного процесса выбирались на основе одношаговой минимизации квадратичного функционала. В работах М. Е. Дмитриенко, Л. А. Оганесяна [23], В. Г. Осмоловского, В. Я. Ривкинда [23], А. М. Мацокина [23] рассматривались стационарные итерационные процессы метода разделения области в применении к дифференциальным уравнениям и их разностным аналогам. В работах В. И. Лебедева, В. И. Агошкова [23] исследованы эффективные нестационарные итерационные алгоритмы метода разделения области с переменными параметрами, найдены оптимальные наборы этих параметров. В. И. Агошковым, В. И. Лебедевым введен специальный класс операторов — операторов Пуанкаре — Стеклова, изучен ряд свойств этих операторов и на их основе разработана общая методология конструирова-

ния и исследования алгоритмов разделения области. В. В. Смеловым [23] и В. И. Агошковым [23] методы разделения области изучались в применении к задачам теории переноса излучения. Исследованием данных методов в нестационарных задачах посвящены работы Ю. А. Кузнецова [23], Г. И. Марчука, Ю. А. Кузнецова [23], С. Н. Булеева, В. И. Агошкова [23] и др.

В настоящее время число научных работ по методу разделения области постоянно растет как в нашей стране, так и за рубежом. Одной из причин этого роста является то, что данный метод часто позволяет осуществить распараллеливание процесса решения задачи при использовании многопроцессорных ЭВМ, которые интенсивно внедряются в практику вычислений в последние годы.

12.8. Сопряженные уравнения и алгоритмы возмущений

Алгоритмы возмущений возникли в XIX в. в связи с решением задач небесной механики. Математическая теория возмущений берет свое начало из работ А. Пуанкаре [24], А. М. Ляпунова [24], Реллиха [24]. Дальнейшее свое развитие она получила в работах К. Фридрихса, Т. Като, Н. Н. Боголюбова, Ю. А. Митропольского, А. Б. Васильевой, В. Ф. Бутузова, М. И. Вишика, Л. А. Люстерника, Секефальви-Надя, С. А. Ломова, Н. Н. Моисеева, В. П. Маслова, В. А. Треногина, Р. Беллмана, Ван Дейка и многих других ученых как в нашей стране, так и за рубежом. В настоящее время методы возмущений широко применяются для исследования и численного решения различных прикладных задач. Одновременно с их развитием была выявлена значительная роль сопряженных уравнений в теории возмущений.

Определенные Лагранжем сопряженные операторы нашли широкое применение при решении задач математической физики. Однако истинное значение теории сопряженных уравнений, по-видимому, было впервые оценено при развитии квантовой механики. Уравнение Шредингера потребовало развития аппарата сопряженных уравнений по крайней мере для задач на собственные значения. Здесь впервые сопряженные уравнения становятся необходимым математическим аппаратом для формулирования теории малых возмущений в спектральных проблемах.

Новый подход к сопряженным задачам был сформулирован в работах Б. Б. Кадомцева [16], Л. Н. Усачева [24], а также в работе Г. И. Марчука, В. В. Орлова [24], в которой была дана общая формулировка сопряженных задач по отношению к избранным линейным функционалам задач.

В дальнейшем в работах автора [24] было дано развитие теорем сопряженных задач по отношению к заданным функционалам для широкого класса задач математической физики. Оно оказалось плодотворным и для многих других направлений науки. В результате появились более или менее общие подходы к исследованию сложных систем и математических моделей теории диффузии охраны окружающей среды, теории климата и его изменений, иммунологии и др. Вместе с развитием методов сопряженных уравнений формировался рациональный подход к решению обратных задач и к планированию математического эксперимента (Г. И. Марчук [16], [24], Г. И. Марчук, Ю. П. Дробышев [16]).

В последние годы сопряженные уравнения и алгоритмы возмущений интенсивно исследуются в применении к нелинейным задачам (Н. Н. Боголюбов, Ю. А. Митропольский [24], В. С. Владимиров, И. В. Волович [24], М. М. Вайнберг, В. А. Треногий [24], В. П. Маслов [24], Г. И. Марчук [24], Г. И. Марчук, В. И. Агошков [24]). В этих задачах возникают те или иные обобщения теории сопряженных уравнений. Изложение этих обобщений можно найти в книге Г. И. Марчука, В. И. Агошкова, В. П. Шутяева [24]. В ней также рассмотрены вопросы обоснования различных аспектов теории сопряженных уравнений и алгоритмов возмущений применительно как к линейным, так и к нелинейным задачам. В данной книге также приведен обзор многих приложений сопряженных уравнений и алгоритмов возмущений в математической физике, вычислительной математике, геофизике и др.

12.9. Вычислительные тензорные методы

Вычислительные тензорные методы в значительной степени используют тот или иной формат представления многомерных матриц (канонические), разложение Таккера, TT-разложения и др.

Канонический формат был впервые рассмотрен в работах Ф. Л. Хичкока [25]. Вычислительные алгоритмы для работы с каноническим форматом стали появляться значительно позже. Первым серьезным приложением стал факторный анализ, где канонический формат рассматривался как формат представления данных (Р. Б. Каттелл [25], Р. А. Харшман [25], Дж.Д. Кэролл, Дж.Дж. Чанг [25]). Формат Таккера впервые предложен также в факторном анализе (статистические методы в психологии, химии) [25]. В дальнейшем развитие происходило в основном в этих областях (Р. Бро [25], С. Люгранс, Р. Т. Росс [25]). Значительный интерес тензорные разложения привлекли в теории сложности при построении оптимальных алгоритмов

умножения матриц (В. Штрассен [25], Д. Бини [25]), где вычисление канонического ранга небольших тензоров напрямую связано с оценкой асимптотической сложности. Одними из первых работ по изучению математических свойств разложения Таккера стали работы де Л. Латаувера , Б. де Мура и Дж. Вандевалле [25], где были предложены алгоритм HOSVD и интерпретация разложения Таккера как многомерного обобщения сингулярного разложения матриц. Идея использовать каноническое представление в качестве формата для представления данных, в котором можно выполнять все базовые операции, была предложена в работах Г. Бейлкина и М. Дж. Моленкампа [25]. В дальнейшем исследовались возможности приближения функций многих переменных и решений различных уравнений в сепарабельном виде (Е. Е. Тыртышников [25], И. П. Гаврилюк, В. Хакбуш, Б. Н. Хоромский [25], Л. Граседик [25], В. Хакбуш, Б. Н. Хоромский, Е. Е. Тыртышников [25], В. Хакбуш, Б. Н. Хоромский [25], И. В. Оселедец, Д. В. Савостьянов, Е.Е. Тыртышников [25]). Интересно отметить, что в работе Е.Е. Тыртышникова [25] фактически получены оценки TT — и QTT-рангов для асимптотически гладких функций, но эти оценки использовались как вспомогательные. Разложение Таккера и его применения исследовались в работах И. В. Оселедца, Д. В. Савостьянова, Е. Е. Тыртышникова [25], а также Б. Н. Хоромского и В. Хоромской [25].Для вычисления канонического разложения и разложения Таккера были предложены различные матричные и оптимизационные алгоритмы (Л. де Латаувер, Б. де Мур, Дж. Вандевалле [25], И. В. Оселедец, Д. В. Савостьянов, Е. Е. Тыртышников [25], Л. Элден, Б. Савас [25], М. Иштева, Л. да Латаувер, П. А. Абсил, S. van Huffel [25]). Своеобразным итогом исследования этих двух форматов в различных областях науки стал обзор Т. Г. Кольды и Б. В. Бадера [25]. Однако их недостатки были достаточно существенными, что стимулировало исследование других представлений. Отметим, что вопросы сходимости основных алгоритмов вычисления канонического разложения мало изучены: известна лишь оценка на локальную сходимость метода ALS (А. Ушмаев [25]).

TT-формат появился в работах И. В. Оселедца и Е. Е. Тыртышникова [25]. HT-формат был предложен практически одновременно также в работах И. В. Оселедца и Е. Е. Тыртышникова. В работах И. В. Оселедца была предложена идея квантизации для матриц, а в работе Хоромского Б. Н. была предложена аналогичная идея для функций и получены оценки на QTT-ранги для простых функций. Дальнейшее развитие и исследование QTT-формата происходило в работах В. А. Казеева, Б. Н. Хоромского [25], Б. Н. Хоромского, И. В. Оселедца [25], И. В. Оселедца [25], И. В. Оселедца, Е. Е. Тыртышникова [25] и Л. Граседика [25]. В работе И. В. Оселедца и Е. Е. Тыртышникова [25] было построено многомерное обобщение скелетного

разложения на тензоры и предложен алгоритм, для которого в дальнейшем Д. В. Савостьяновым [25], Дж. Баллани, Л. Граседиком, М. Клюге [25] были построены более эффективные модификации и обобщение на HT-формат. Оценка квазиоптимальности многомерного крестового метода получена в работе Д.В. Савостьянова. Методы оптимизации в TT-формате как методы оптимизации на многообразиях был изучены в работах С. Хольца, Т. Роведдера, Р. Шнайдера [25], С. В. Долгова, И. В. Оселедца [25] и И. В. Оселедца [25]. AMEN-подход предложен в работах С. В. Долгова и Д. В. Савостьянова [25]. Минимизация блочного отношения Рэлея исследована в работе С. В. Долгова, И. В. Хоромского, И. В. Оселедца, Д.В. Савостьянова и др. Уравнения динамической малоранговой аппроксимации для матриц и формата Таккера получены в работах О. Коха и Х. Любиха. Для HT-и TT-форматов уравнения получены в работе Х. Любиха, Т. Роведдера, Р. Шнайдера и Б. Вандерейкена [25], а эффективная схема расщепления для интегрирования полученных уравнений исследована в работах Х. Любиха, И. В. Оселедца [25], Д. Хаегемана, Х. Любиха, И. В. Оселедца [25] и др.

В различных областях физики и химии аналогичные представления были известны достаточно давно. Представление MPS (Matrix Product States) (У. Шоллвек [25], М. Фаннес, Б. Нахтергаеле, Р. Ф. Вернер [25]), которое используется в физике твердого тела, квантовой теории информации, статистической физике, имеет ту же алгебраическую структуру, что и TT-формат. Представления, аналогичные HT-формату, встречались в работах М. Фаннеса, Б. Нахтергаеле, Р.Ф. Вернера [25], У. Манте [25], однако до появления TT- и HT-форматов эти результаты были полностью неизвестными в вычислительной математике и численном анализе.

Отметим, что MPS-представления стали возникать в физике твердого тела, начиная с модели AKLT (Т. Кеннеди и др. [25]). В AKLT-модели собственная функция многомерного оператора для спиновой системы точно представима в MPS-формате. Одним из наиболее эффективных алгоритмов вычисления основных состояний является алгоритм DMRG (С. Райт [25]), но его связь с MPS-представлением была обнаружена в работе С. Остлунда и С. Роммера [25]; в дальнейшем результаты были распространены на другие классы задач, включая решение линейных систем (Е. Джекельман [25]), нестационарные задачи (Д. Хаегеман, Дж. И. Кирак, Т. Дж. Осборн [25]), задачи интерполяции (Д. В. Савостьянов, И. В. Оселедец [25]). Эти подходы представляют большой интерес для вычислительных тензорных методов, в частности, для приближенного нахождения собственных функций (Б. Н. Хоромский, И. В. Оселедец [25], Д. Кресснер, К. Тоблер [25]), решения линейных систем (С. Хольц, Т. Роведдер, Р. Шнайдер [25], С. В. Долгов, И. В. Оселедец [25]), решения задачи интерполяции (Д. В. Савостьянов, И. В.

Оселедец [25]). Исследованию структуры TT/HT-многообразий и свойствам разложений в общих пространствах посвящена работа С. Хольца, Т. Роведдера и Р. Шнайдера [25]. Более подробную информацию о новых тензорных разложениях и методах работы с ними можно найти в книге В. Хакбуша [25] и в подробном литературном обзоре Л. Граседика, Д. Клесснера и К. Тоблера [25], который содержит практически все известные работы по данной теме.

Литература

[1] **Функциональный анализ и вычислительная математика**

Анселон (Anselone P. H. M.). Collectively Compact Operator Approximation Theory and Applications to Integral Equations.— Englewood Cliffs: Prentice-Hall Inc., 1967.

Балакришнан (Balakrishnan A. V.). Applied Functional Analysis.— N. Y.: Springer-Verlag, 1976.

Варга Р. Функциональный анализ и теория аппроксимации в численном анализе.— М.: Мир, 1974.

Вейнштейн, Стенгер (Weinstein A., Stenger W.). Methods of Intermediate Problems for Eigenvalues, Theory and Ramifications.— L., N. Y.: Acad. Press, 1972.

Канторович Л. В. Функциональный анализ и прикладная математика // УМН. —1948. — Т. 3, № 6.

Канторович Л. В., Акилов Г. П. Функциональный анализ в нормированных пространствах. — М.: Физматгиз, 1959.

Келдыш М. В., Лидский В. Б. Вопросы спектральной теории несамосопряженных операторов // Труды IV Всесоюзного математического съезда, Ленинград, 3—12 июля, 1961 г. Т. 1. Пленарные доклады. — Л.: Изд-во АН СССР, 1963.

Коллатц Л. Функциональный анализ и вычислительная математика. — М.: Мир, 1969.

Колмогоров А. Н., Фомин С. В. Элементы теории функций и функционального анализа. — М.: Наука, 1976.

Красносельский М. А., Вайникко Г. М., Забрейко П. П. и др. Приближенное решение операторных уравнений. — М.: Наука, 1969.

Крейн С. Г. Линейные дифференциальные уравнения в банаховом пространстве. — М.: Наука, 1967.

Лаврентьев М. А., Шабат Б. В. Методы теории функций комплексного переменного. — М.: Физматгиз, 1965.

Лионс (Lions J.). Equations dilferentiellcs operationnelles. — Berlin — Gottingen — Heidelberg: Springer-Verlag, 1961.

Люстерник Л. А., Соболев В. И. Элементы функционального анализа. — М.: Наука, 1965.

Михлин С. Г. Вариационные методы в математической физике. — М.: Наука, 1970.

Михлин С. Г. Проблема минимума квадратичного функционала. — М.: Гостехиздат, 1952.

Михлин С. Г., Смолицкий Х. Л. Приближенные методы решения дифференциальных и интегральных уравнений. — М.: Наука, 1965.

Натансон И. П. Теория функций вещественной переменной. — М.: Наука, 1974.

Никольский С. М. Приближение функций многих переменных и теоремы вложения. — М.: Наука, 1969.

Ректорис К. Вариационные методы в математической физике и технике. — М.: Мир., 1985.

Соболев С. Л. Некоторые применения функционального анализа в математической физике. — Л.: Изд-во ЛГУ, 1950.

Соболев С. Л. Введение в теорию кубатурных формул. — М.: Наука, 1974.

Треногин В. А. Функциональный анализ. — М.: Наука, 1980.

Флетчер К. Численные методы на основе метода Галеркина. — М.: Мир, 1988.

[2] **Дифференциальные уравнения в частных производных и математическая физика**

Бицадзе А. В. Краевые задачи для эллиптических уравнений второго порядка. — М.: Наука, 1966.

Боссави (Bossavit A.). Regularisation d’equations variationnelles et applications. — Centre national de la Recherche Scientifique, Institute Blaise Pascal, Juin 1970.

Векуа И. Н. Новые методы решения уравнений эллиптического типа. — М.: Гостехиздат, 1948.

Вишик М. И., Люстерник Л. А. Асимптотическое поведение решений линейных дифференциальных уравнений с большими или быстроменяющимися коэффициентами и граничными условиями // УМН. — 1964. — Т. 15, № 4.

Владимиров В. С. Уравнения математической физики. — М.: Наука, 1967.

Владимиров В. С. Обобщенные функции в математической физике. — М.: Наука, 1976.

Годунов С. К. Уравнения математической физики. — М.: Наука, 1971.

Канторович Л. В., Крылов В. И. Приближенные методы высшего анализа.— М.: Физматгиз, 1962.

Кондратьев В. А. Краевые задачи для эллиптических уравнении в областях с коническими или угловыми точками // Труды Московского математического общества. — 1967. — Т. 16.

Курант Р. Уравнения с частными производными. — М.: Мир, 1964.

Курант Р., Гильберт Д. Методы математической физики. Т. 1. — М.: Гостехиздат, 1953.

Лаврентьев М. А. Вариационный метод в краевых задачах для систем уравнений эллиптического типа. — М.: Изд-во АН СССР, 1952.

Лаврентьев М. М. О некоторых некорректных задачах математической физики. — Новосибирск: Изд-во СО АН СССР, 1962.

Ладыженская О. А. Смешанная задача для гиперболического уравнения. — М.: Гостехиздат, 1953.

Леонс Ж.-Л. Некоторые методы решения нелинейных краевых задач. — М.: Мир, 1972.

Лионс Ж. -Л., Маженес Э. Неоднородные граничные задачи и их приложения. — М.: Мир, 1971.

Миллер (Miller J. H.), ed. Topics in Numerical Analysis. Procs of the Royal Irish Acad. Conf. on Numer. Analysis 1972. — L.; N. Y.: Acad. Press, 1973.

Миллер (Miller J. H.), ed Topics in Numerical Analysis. II, Procs of the Royal Irish Acad. Conf. on Numer. Analysis, 1974. — L.; N. Y.: Acad. Press, 1975.

Мино (Mignot A.). Methodes d’approximation des solutions de problems aux limites // Rend. del som. Mat. della Univ. di Padova. — 1968. — V. 11.

Петровский И. Г. Лекции об уравнениях с частными производными. — М.: Физматгиз, 1961.

Рождественский Б. Л., Яненко Н. Н. Системы квазилинейных уравнений. — М.: Наука, 1968.

Смирнов В. И. Курс высшей математики. Т. 1—5. — М.: Гостехиздат, 1948.

Соболев С. Л. Уравнения математической физики. — М.: Наука, 1966.

Тихонов А. Н., Самарский А. А. Уравнения математической физики. — М.: Наука, 1966.

Фихтенгольц Г. М. Курс дифференциального и интегрального исчисления. Т. 3. — М.: Физматгиз, 1963.

Фридрихс (Friedrichs K.). Non-linear hyperbolic differential equations for functions of two independent variables // Amer. J. Math. —1948. — V. 70.

[3] **Численные методы (монографии и учебные пособия)**

Бабенко К. И. Основы численного анализа. — М.; Наука, 1986.

Бабушка И., Витасек Э., Прагер М. Численные процессы решения дифференциальных уравнений. — М.: Мир, 1969.

Балакришнан, Нойштадт (Balakrishnan A. V., Newstadt L. W.). Computing Methods in Optimisation Problems. — Academic Press, 1964.

Бахвалов Н. С. Численные методы. Т. 1. — М.: Наука, 1973.

Бахвалов Н. С., Жидков Н. П., Кобельков Г. М. Численные методы. — М.: Наука, 1987.

Беллман, Калаба, Локет (Bellman R., Kalaba R., Lockett J.). Numerical Inversion of the Laplace Transform. — N. Y.: Am. Elsevier Publ. Co. Inc., 1966.

Березин И. С., Жидков Н. П. Методы вычислений. Т. 1, 2. — М.: Физматгиз, 1966.

Блюм (Blum E. K.). Numerical Analysis and Computation: Theory and Practice. — L.: Addison-Wesley Publ. Cотр. Inc., 1972.

Брэмбл (Bramble J. H.), ed. Numerical Solution of Partial Differential Equations // Proc. of a Symp. Held at the Univ. of Maryland. — L., N. Y.: Acad. Press, 1966.

Вазов В., Форсайт Дж. Разностные методы решения дифференциальных уравнений в частных производных. — М.: ИЛ, 1963.

Валиуллин А. Н. Схемы повышенной точности для задач математической физики. Лекции для студентов НГУ. — Новосибирск: Изд-во НГУ, 1973.

Варга (Varga R. S.). Matrix Iterative Analysis. — N. Y., 1963.

Воеводин В. В. Математические модели и методы в параллельных процессах. — М.: Наука, 1986.

Волков Е. А. Численные методы. — М.: Наука, 1982.

Воробьев Ю. В. Метод моментов в прикладной математике. — М.: Физматгиз, 1958.

Гельфонд А. О. Исчисление конечных разностей. — М.: Наука, 1967.

Годунов С. К. Разностные методы решения уравнений газовой динамики. — Новосибирск: Изд-во НГУ, 1962.

Годунов С. К., Забродин А. В., Иванов М. Я. и др. Численное решение многомерных задач газовой динамики. — М.: Наука, 1976.

Годунов С. К., Рябенький В. С. Введение в теорию разностных схем. — М.: Физматгиз, 1962.

Годунов С. К., Рябенький В. С. Разностные схемы. — М.: Наука, 1973.

Дальквист, Бьйорк (Dahlquist G., Bjorck A.). Numerical Methods. — Englewood Cliffs: Prentice Hall Inc., 1974.

Дородницын А. А. Об одном методе численного решения некоторых нелинейных задач аэродинамики // Труды III Всесоюзного математического съезда. Т. II. —М.: Изд-во АН СССР, 1956.

Дородницын А. А. Лекции по численным методам решения уравнений вязкой жидкости. — М.: ВЦ АН СССР, 1969.

Дробышевич В. И., Дымников В. П., Ривин Г. С. Задачи по вычислительной математике. — М.: Наука, 1980.

Дьяконов Е. Г. Итерационные методы решения разностных аналогов краевых задач для уравнений эллиптического типа. — Киев: Из-во ИК АН УССР, 1970.

Дьяконов Е. Г. Разностные методы решения краевых задач. — М.: МГУ, 1971. — Вып. 1 (стационарные задачи); 1972. — Вып. 2 (нестационарные задачи).

Изаксон, Келлер (Isaacson E., Keller H. B.). Analysis of Numerical Methods. — N. Y.: Wiley, 1966.

Ильин В. П. Разностные методы решения эллиптических уравнений. — Новосибирск: Изд-во НГУ, 1970.

Ильин В. П. Численные методы решения задач электрооптики. — Новосибирск: Наука, 1974.

Канторович Л. В., Крылов В. И. Приближенные методы анализа. — М.;Л.: Физматгиз, 1962.

Келлер (Keller H. B.). Numerical Methods for Two-Point Boundary Value Problems. — N. Y.: Blaisdell Publ. Covnp., 1968.

Коллатц Л. (Kollatz L.). Численные методы решения дифференциальных уравнений. — М.: ИЛ, 1953.

Коновалов А. Н. Численное решение задач теории упругости. — Новосибирск: Изд-во НГУ, 1968.

Ланцош К. Практические методы прикладного анализа. — М.: Физматгиз, 1961.

Лебедев В. И., Бахвалов Н. С., Агошков В. И., Бабурин О. В., Князев А. В., Шутяев В. П. Параллельные алгоритмы решения некоторых стационарных задач математической физики. — М.: ОВМ АН СССР, 1984.

Лионс, Марчук (Lions J. L., Marchook G. I.). Sur les methodes numeriques en sciences physiques et economiques. — P.: Dunod, 1974.

Марчук Г. И. Методы расчета ядерных реакторов. — М.: Атомиздат, 1961.

Марчук Г. И. Численные методы в прогнозе погоды. — Л.: Гидрометиздат, 1967.

Марчук Г. И. (Marchuk G. I.). Methods and problems of computational mathematics // Article from the Proceedings of the International congress of mathematicians. Nice, September. 1970.

Марчук Г. И. Методы и проблемы вычислительной математики // Международный конгресс математиков в Ницце. 1970. Доклады советских математиков. — М.: Наука, 1972.

Марчук Г. И. Методы вычислительной математики. — Новосибирск: Наука, 1973.

Миллер, Стрэнг (Miller J., Strong G.). Matrix theorems for partial differential and difference equations // Math. Scandinavica. — 1966. — V. 18, № 2.

Митчелл (Mitchell A. B.). Computational Methods in Partial Differential Equations. — L.: Wiley, 1970.

Мысовских И. П. Лекции по методам вычислений. — М.: Физматгпз, 1962. — Некоторые проблемы вычислительной и прикладной математики. // Под ред. М. М. Лаврентьева. — Новосибирск: Наука, 1975.

Никольский С. М. Квадратурные формулы. — М.: Наука, 1974.

Обэн Ж.-П. Приближенное решение эллиптических краевых задач. — М.: Мир, 1977.

Положий Г. Н. Численное решение двумерных и трехмерных краевых задач математической физики и функций дискретного аргумента. — Киев: Изд-во КГУ, 1962.

Понтрягин Л. С., Болтянский В. Г., Гамкрелидзе Р. В., Мищенко Е. Ф. Математическая теория оптимальных процессов. — М.: Физматгиз, 1961.

Рихтмайер Р. Д. (Richtmyer R.). Разностные методы решения краевых задач. — М.: ИЛ, 1960.

Рихтмайер, Мортон (Richtmyer R. D., Morton K. W.). Difference Methods for Initial-Value Problems. — N. Y.: Wiley, 1967.

Рябенький В. С. Метод разностных потенциалов для некоторых задач механики сплошных сред. — М.: Наука, 1987.

Самарский А. А. Введение в численные методы. — М.: Наука, 1982.

Самарский А. А. Теория разностных схем. — М.: Наука, 1982.

Самарский А. А., Андреев В. Б. Разностные методы для эллиптических уравнений. — М.: Наука, 1976.

Самарский А. А., Гулин А. В. Устойчивость разностных схем. — М.: Наука, 1973.

Самарский А. А., Попов Ю. П. Разностные схемы газовой динамики. — М.: Наука, 1975.

Саульев В. К. Интегрирование уравнений параболического типа методом сеток. — М.: Физматгиз, 1960.

Сеа (Cea J.). Optimisation theorie et algorithmes. — P.: Dunod, 1971.

Стрэнг Г., Фикс Дж. Теория метода конечных элементов. — М.: Мир, 1977.

Трауб (Traub J. P.). Iterative Methods for the Solution of Equations. — Englewood Cliffs: Prentice-Hall Inc., 1964.

Фокс Л., Майерс Д. Ф. (Fox L., Mayers D. F.). Computing Methods for Scientists and Engineers. — Oxford, 1968.

Хаусхолдер А. С. Основы численного анализа. — М.: ИЛ, 1956.

Хенричи (Henrici P.). Error Propagation for Difference Methods. — N. Y.: John Willey and Sons, 1963.

Шайдуров В. В. Многосеточные методы конечных элементов. — М.: Наука, 1989.

Яненко Н. Н. Метод дробных шагов решения многомерных задач математической физики. — Новосибирск: Наука, 1967.

[4] **Метод сеток**

Белоцерковский О. М., Чушкин П. И. Численный метод интегральных соотношений // ЖВМ и МФ. — 1962. — Т. 2, № 5.

Брайен (Brayn K.). A scheme for numerical integration of the equations of motion on an irregular grid free of non-linear instability // Monthly Weather Review. — 1966. — Т. 94, №1. [Рус. пер.: Численные методы решения задач динамики атмосферы и океана. — Л.: Гидрометиздат, 1968.]

Вакспресс (Wachspress E. L.). The numerical solution of boundary value problems // Mathematical Methods for Digital Computers. — N. Y.: 1960.

Валиуллин А. Н., Яненко Н. Н. Экономичные разностные схемы повышенной точности для полигармонического уравнения // Изв. Сиб. отд. АН СССР. Сер. тех. наук. — 1967. — Т. 13, № 3.

Валицкий Ю. Н. О сходимости разностных аппроксимаций собственных значений и собственных функций двумерного эллиптического оператора // ДАН СССР. — 1971. — Т. 198, № 2.

Волков Е. А. Исследование одного способа повышения точности метода сеток при решении уравнения Пуассона // Вычислительная математика. Вып. 1. — М.: ВЦ АН СССР, 1957.

Волков Е. А. Решение задачи Дирихле методом уточнений разностями высших порядков. Ч. I, II // Дифференциальные уравнения. — 1965. — Т. 1, № 7.

Волков Е. А. Метод неравномерных сеток для конечных и бесконечных областей с коническими точками // Дифференциальные уравнения. — 1966. — Т. 10, № 2.

Волков Е. А. Развитие метода сеток для уравнений Лапласа на конечных и бесконечных областях с кусочно-гладкой границей: автореф. дисс. д-ра. — М., 1967.

Годунов С. К., Забродин А. В. О разностных схемах второго порядка точности для многомерных задач// ЖВМ и МФ. — 1962. — Т. 2, № 4.

Годунов С. К., Прокопов Г. П. О расчетах конформных отображений и построении разностных сеток // ЖВМ и МФ. — 1967. — Т. 7, № 5.

Годунов С. К., Семендяев К. А. Разностные методы численного решения задач газовой динамики // ЖВМ и МФ. — 1962. — Т. 2, № 7.

Демьянович Ю. К. Метод сеток для некоторых задач математической физики // ДАН СССР. — 1964. — Т. 159, № 2.

Джойс (Joice D. C.). Survey of extrapolation processes in numerical analysis // SI AM Review. — 1971. — V. 13, №. 4.

Келлер (Keller H.). A new difference scheme for parabolic problems. // In: Numerical solution of partial differential equations. — II. SYNSPADE-1970. N. Y.; L.: Academic Press, 1971.

Келлог (Kellog R.). Singularities in interface problems. // In: Numerical solution of partial differential equations. — II. SYNSPADE-1970. N. Y.; L.: Academic Press, 1971.

Коновалов А. Н. Метод фиктивных областей в задачах кручения // Численные методы механики среды. — 1973. — Т. 4, № 2.

Копченов В. Д. Приближение решения задачи Дирихле методом фиктивных областей // Дифференциальные уравнения. — 1968. — Т. 4, № 1.

Кроули (Crowley W.). Second order numerical advection // J. Comp. Phys. — 1967 — V. 1, №. 4.

Кузнецов Ю. А., Мацокин А. М. Решение уравнения Гельмгольца методом фиктивных областей // Вычислительные методы линейной алгебры. — Новосибирск: ВЦ СО АН СССР, 1972.

Кузнецов Ю. А., Мацокин А. М. Матричный аналог метода фиктивных областей и его применения. — Новосибирск, 1977.

Кузнецов Ю. А., Шайдуров В. В. О равномерной сходимости разностных схем // Вычислительные методы линейной алгебры — Новосибирск: ВЦ СО АН СССР, 1972.

Курихара, Холловэй (Kurihara Y., Holloway I.). Numerical integration of a nine-level global primitive equations model formulated by the box method // Monthly Weather Review. — 1967. — V. 95, №. 8.

Куропатенко В. Ф. Метод построения разностных схем для численного интегрирования уравнений газодинамики // Изв. вузов. Математика. — 1962. — Т. 3, № 28.

Ландау Л. Д., Мейман Н. Н., Халатников И. М. Численные методы интегрирования уравнений в частных производных методом сеток // Труды III Всесоюзного математического съезда. Т. II. — М.: Изд-во АН СССР, 1956.

Лебедев В. И. Метод сеток для уравнений типа С. Л. Соболева // ДАН СССР. — 1957. — Т. 114, № 6.

Лебедев В. И. О методе сеток для одной системы уравнений в частных производных // Изв. АН СССР. Сер. матем. — 1958. — Т. 22, № 5.

Лебедев В. И. О задаче Дирихле и Неймана на треугольных и шестиугольных сетках // ДАН СССР. — 1961. — Т. 138, № 1.

Люстерник Л. А. О разностных аппроксимациях операторов Лапласа. — УМН, 1954. — Т. IX, № 2.

Марчук Г. И., Дымников В. П., Галин В. Я. и др. Гидродинамическая модель общей циркуляции атмосферы и океана. — Новосибирск, 1975.

Марчук Г. И., Ривин Г. С., Юдин М. И. Численные эксперименты с балансными схемами // Изв. АН СССР. Сер. ФАО. — 1973. — Т. II.

Марчук Г. И., Шайдуров В. В. О численном решении эволюционной задачи с ограниченным оператором // ДАН СССР. — 1974. — Т. 216.

Марчук Г. И., Шайдуров В. В. (Marchuk G. I., Shaydourov V. V.). Increasing of the accuracy of the projective-difference schemes. Lecture Notes in Computer Science, V. II. — Springer-Veriag, 1974.

Мацокин А. М. Автоматизация триангуляции областей с гладкой границей при решении уравнений эллиптического типа. // Вычислительные методы прикладной математики: семинар. — Препринт № 15 ВЦ СО АН СССР, 1975.

Мацокин А. М. К развитию фиктивных областей // Вычислительные методы линейной алгебры. — Новосибирск: ВЦ СО АН СССР, 1972.

Мацокин А. М. Вариационно-разностный метод решения эллиптических уравнений в трехмерных областях // Вариационно-разностные методы в математической физике. — Новосибирск: ВЦ СО АН СССР, 1976.

Мацокин А. М. О построении и методах решения систем вариационно-разностных уравнений: автореф. дис. ... канд. — Новосибирск: ВЦ СО АН СССР, 1975.

Пененко В. В. Вычислительные аспекты в задачах математического моделирования динамики атмосферных процессов: автореф. дис. ... д-ра. — Новосибирск, 1976.

Равьяр (Raviart P. A.). Sur 1’approximation de certaines equations d’evolution lineaires et non lineaires // J. de Mathem. Pures et Appl. — 1967. — V. 46, № 1.

Ривкинд В. Я. Приближенный метод решения задачи Дирихле и об оценках скорости сходимости разностных уравнений к решениям эллиптических уравнений с разрывными коэффициентами // Вестник Ленингр. ун-та. Сер. матем. — 1964. — Т. 3.

Ривкинд В. Я. Об оценке скорости сходимости однородных разностных схем для эллиптических и параболических уравнений с разрывными коэффициентами // Проблемы математического анализа. — Л.: Изд-во ЛГУ, 1966.

Ричардсон Л. Ф. (Richardson L. F.). The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stress in a masonry dam // Philos. Trans. Roy. Soc.: London, Ser. A, 1910.— V. 210.

Руховец Л. А. Замечание к методу фиктивных областей // Дифференциальные уравнения. — 1967. — Т. 3, № 4.

Самарский А. А. О монотонных разностных схемах для эллиптических и параболических уравнений в случае несамосопряженного эллиптического оператора // ЖВМ и МФ. — 1965. — Т. 5, № 3.

Самарский А. А. О точности метода сеток для задачи Дирихле в произвольной области // Appl., Math. — 1965. — Т. 10, № 3.

Самарский А. А. Некоторые вопросы теории разностных схем // ЖВМ и МФ. — 1966. — Т. 6, № 4.

Саульев В. К. Об одном методе автоматизации решения краевых задач на быстродействующих вычислительных машинах // ДАН СССР. — 1962. — Т. 142, № 3.

Саульев В. К. О решении некоторых краевых задач на быстродействующих краевых машинах методом фиктивных областей // Сибирск. матем. журнал. — 1963. — Т. 4, № 4.

Синяев В. Н. Об одном принципе построения конечно-разностных схем, основанных на законах сохранения полной энергии // Численные методы механики сплошной среды. — 1974. — Т. 5, № 2.

Тихонов А. Н., Самарский А. А. О разностных схемах для уравнений с разрывными коэффициентами // ДАН СССР. — 1956. — Т. 108, № 3.

Тихонов А. Н., Самарский А. А. Об однородных разностных схемах // ЖВМ и МФ. — 1961. — Т. 1, № 1.

Тихонов А. Н., Самарский А. А. Однородные разностные схемы на неравномерных сетках// ЖВМ и МФ. — 1962. — Т. 2, № 5.

Урванцев А. Л., Шайдуров В. В. Уточнение приближенного решения квазилинейного уравнения Пуассона // Вариационно-разностные методы в математической физике. — Новосибирск: ВЦ СО АН СССР, 1976.

Фикера (Fichera G.). Further development in the approximation theory of eigenvalues // Numerical solution of partial differential equations. — II. SYNSPADE 1970. — N. Y.; L.: Academic Press, 1971.

Фокс, Хенричи, Молер (Fox L., Henrici P., Moler C.). Approximations and bounds for eigenvalues of elliptic operators // SI AM J. Numer. Anal. — 1967. — V. 4, №. 1.

Фромм (Fromm J. E.). Numerical method for computing nonlinear, time dependent, buoyant circulation of air in rooms // JBM J. of Research and Development. — 1971. — V. 15, №. 3.

Чудов Л. А., Кудрявцев В. П. Об ошибках округления при решении разностными методами задач с начальными условиями для эллиптических

уравнений и систем // Численные методы в газовой динамике. — М.: Изд-во МГУ, 1963.

Шайдуров В. В. Об одном методе повышения точности разностных решений // Численные методы механики сплошной среды. — 1972. — Т. 3, № 2.

Шутяев В. П. Нестационарная задача для уравнения диффузии и параллельные алгоритмы ее решения // Сопряженные уравнения и алгоритмы возмущений в задачах математической физики. — М.: ОВМ АН СССР, 1989.

Яненко Н. Н., Сучков В. А., Погодин Ю. Я. О разностном решении уравнения теплопроводности в криволинейных координатах // ДАН СССР. — 1959. — Т. 28, №5.

[5] **Вариационно-разностные методы**

Агошков В. И. О вариационной форме интегрального тождества Г. И. Марчука. — Новосибирск: Препринт ВЦ СО АН СССР, 1977.

Агошков В. И. Обобщенная формулировка метода интегральных тождеств. — Новосибирск: Препринт ВЦ СО АН СССР, 1979.

Бабушка (Babuska I.). The finite element method for elliptic differential equations//Numerical solution of partial differential equations, — II. SYNSPADE-1970. — N. Y.; L.: Academic Press, 1971.

Бабушка (Babuska I.). The rate of convergence for finite element method // SIAM J. Numer. Anal. — 1971. — V. 8, №. 2.

Биркгоф, Шульц, Варга (Birkhoff G., Schultz M. H., Varga R. S.). Hermite interpolation in one and two variable with applications to partial differential equations // Numer. Math. — 1968. — V. 11, №. 3.

Брембл (Bramble J.). A second order finite difference analog of the first biharmonic boundary value problems // Numer. Math. — 1961. — V. 9, №. 3.

Брембл, Хаббард (Bramble J., Hubbard B.). On the formulation of fini difference analogues of the Dirichlet problem for Posiison's equation // Numer. Math. — 1962. — V. 4, №. 4.

Брембл, Шатц (Bramble J., Schatz A.). On the numerical solution of elliptic boundary value problems by least squares approximation of the data // Numerical solution of partial differential equations. — II. SYNSPADE-1970. — N. Y.; L.: Academic Press, 1971.

Гловински (GIovinski R.). Introduction to the Approximation of Elliptic Variational Inequalities. — Report 76006, Laboratoire d'AnaIyse Numerique de l'universite Paris, 1976. — V. 6.

Гловински, Лионс, Тремольер (GIovinski R., Lions J. L., Tremolieres R.). Analyse Numerique des Inequations Variationnelles. V. 1. 2. — P.: Dunod, 1976.

Гловински, Марроко (Glowinski R., Marroco A.). Sur 1'Approximation. par Elements Fin is d'Ordre Un. et al Resolution, par Penalisation — Dualite, d'Une Classe de Problemes de Dirichlet Non Lineares. — Revue Francaise d'Automatique, Informatique et Recherche Operationelle, 1975. — R-2.

Дуглас, Дюпон (Douglas J., Dupont T.). Alternating-direction Galerkin methods on rectangles // Numerical solution of partial differential equations. — II. SYNSPADE-1970. — N. Y.; L.: Academic Press, 1971.

Дьяконов Е. Г. Некоторые классы операторов, эквивалентных по спектру, и их применения // Вариационно-разностные методы в математической физике. — Новосибирск: ВЦ СА АН СССР, 1976.

Дюво, Лионс (Duvaur G., Lions J. L.). Les Inequations en Mecanique et en Physique. — P.: Dunod, 1972. (English translation: Grundlehren der Math., Springer-Veriag, 1976. — V. 219.)

Зенкевич О. Метод конечных элементов в технике. — М.: Мир, 1975.

Зламал (Zlamal M.). On the finite element method // Numer. Math. — 1968. — V. 12, №. 5.

Зламал (Zlamal M.). On some finite element procedures for solving second order boundary value problems // Numer. Math. — 1969. — V. 14, №. 1.

Келдыш М. В. О методе Галеркина для решения краевых задач // Изв. АН СССР. Сер. матем. — 1942. — Т. 6.

Курант (Courant R.). Variational methods for the solutions of problems of equalibrium and variations // Bull. Amer. Math. Soc. — 1943. — V. 69.

Лебедев В. И. Разностные аналоги ортогональных разложений основных дифференциальных операторов и некоторых краевых задач математической физики // ЖВМ и МФ. — 1964. — Т. 4, № 3.

Лионс, Темам (Lions J. L., Temam R.). Une methode d'eclatement des operateurs et des contraintes en calcul des variations // C. R. A cad. Sci. Paris. — 1966. — V. 263.

Лионс, Стампакья (Lions J., Stampacchia). Variational inequalities // Con. Pure Applied Math. — 1967. — V. XX.

Марчук Г. И., Агошков В. И. О выборе координатных функций в обобщенном методе Бубнова —- Галеркина // ДАН СССР. — 1977. — Т. 232, № 6.

Марчук Г. И., Агошков В. И. Введение в проекционно-сеточные методы. — М.: Наука, 1981.

Обэн (Aubin J. P.). Behavior of the error of the approximate solutions of boundary value problems for linear elliptic equations by Galerkin's and finite difference methods // Ann. Scuola Norm. Super. — Pisa, 1967. — V. 21, № 4.

Обэн (Aubin J. P.). Best approximation of linear operators in Hilbert space // SIAM J. Numer. Anal. — 1968. — V. 5, №. 3.

Обэн (Aubin J. P.). Approximation des espaces des distributions et des operateurs differentiels // Bull. Soc. Math. France, Memoire. — 1967. — V. 12.

Обэн, Буршард (Aubin J. P., Burchard H. G.). Some aspects of the method of the hypercircle applied to elliptic variational problems. — Proceedings of SYNSPADE. Acad. Press, 1971.

Оганесян Л. П. Численный расчет плит // Решение инженерных задач на электронно-вычислительных машинах. — Л., 1963.

Оганесян Л. А. Вариационно-разностная схема на регулярной сетке для задачи Дирихле // ЖВМ и МФ. — 1971. — Т. 11, № 6.

Оганесян Л. А., Ривкинд В. Я., Руховец Л. А. Вариационно-разностные методы решения эллиптических уравнений I, II // Дифференциальные уравнения и их применение. Вып. 5, 8. — Вильнюс, 1974.

Оганесян Л. А., Руховец Л. А. О вариационно-разностных схемах для линейных эллиптических уравнений второго порядка в двумерной области с кусочно-гладкой границей // ЖВМ и МФ. — 1968. — Т. 8, № 1.

Оганесян Л. А., Руховец Л. А. Исследование скорости вариационно-разностных схем для эллиптических уравнений второго порядка в двумерной области с гладкой границей // ЖВМ и МФ. — 1969. — Т. 9.

Руховец Л. А. Исследование скорости сходимости вариационно-разностных схем для двумерных эллиптических уравнений второго порядка: автореф. дис. ... канд. — Л., 1970.

Сеа (Cea J.). Approximation opirationnelle des problcmes aux limites // Ann. Inst. Fourier, Grenoble. — 1964. — V. 14, №. 2.

Сеа, Гловински (Cea J., Glowinski R.). Sur des mithodes d'optimisation par relaxation.— Revue Francaise d'Automatique, Informatique et Recherche Operationnelle, 1973. — R-3. — S-32.

Сеа, Гловински (Cea J., Glowinski R.). Methodes numeriques pour 1'ecoulement laminaire d'un fluide vigide viscoplastique incompressible // Inf. J. of Сотр. Math. Ser. B. — 1974. — V. 3.

Смелов В. В. Аппроксимация кусочно-гладких функций тригонометрическими многочленами и использование последних в вариационных методах. — Новосибирск, 1975.

Стрэнг (Strang G.). The finite element method and approximation theory. In: Numerical solution of partial differential equations. — II. SYNSPADE-1970. — N. Y.; L.: Academic Press, 1971.

Стрэнг, Фикс (Strang G., Fix G.). A Fourier analysis of the finite element variational method. — Preprint, 1970.

Федорова О. А. Вариационно-разностная схема для однородного уравнения диффузии // Матем. заметки. — 1975. — Т. 17, № 6.

Хаббард (Hubbard B.). Remarks on the convergence in the discrete Dirichlet problem // Numerical solution of partial differential equations / Ed. by James H. Bramble. — N. Y.; L.: Academic Press, 1965.

Шайдуров В. В. Экстраполяция Ричардсона для проекционно-разностной задачи Штурма-Лиувилля // Вариационно — разностные методы в математической физике. — Новосибирск: ВЦ СО АН СССР, 1974.

[6] **Теория устойчивости разностных схем**

Ильин А. М. Устойчивость разностных схем задачи Коши для систем дифференциальных уравнений в частных производных // ДАН СССР. — 1965. — Т. 164, № 3.

Келлер, Тома (Keller H. B., Thomes V.). Unconditionally stable difference method for mixed problems for quasilinear hyperbolic systems in two dimensions // Comm. Pure Appl. Math. — 1962. — V. 15, №. 1.

Крайс (Kreiss H. O.). Uber die Stabilitatsdefinition fur Differenzengleichungen die partielle Differentialgleichungen approzimieren // Nordisk. Tidskr. Informations Behandlung. — 1962. — V. 2, №. 2.

Крайс (Kreiss H. O.). On difference approximations of the dissipative type for hyperbolic differential equations // Comm. Pure Appl. Math. — 1964. — V. 17, №. 3.

Крайс (Kreiss H. O.). Initial boundary value problems for partial differential and difference equations in one space dimensions // Numerical solution of partial differential equations. — II. SYNSPADE-1970. — N. Y.; L.: Academic Press, 1971.

Лакс П. Об устойчивости конечно-разностных аппроксимаций решений гиперболических уравнений с переменными коэффициентами // Математика (сб. перев.). — 1962. — Т. 6, № 3.

Лакс, Вендроф (Lax P. D., Wendroff B.). On the stability of difference schemes with variable coefficients // Comm. Pure Appl. Math. — 1962. — V. 15, №. 4.

Лакс П., Ниренберг Г. Об устойчивости разностных схем; точная форма неравенства Гординга // Математика (сб. перев.). — 1967. — Т. 11, № 6.

Рихтмайер Р. Д. О нелинейной неустойчивости разностных схем // Некоторые вопросы вычислительной и прикладной математики. — Новосибирск: Наука, 1966.

Рябенький В. С., Филиппов А. Ф. Об устойчивости разностных уравнений. — М.: Гостехиздат, 1956.

Самарский А. А. Необходимые и достаточные условия устойчивости двухслойных разностных схем // ДАН СССР. — 1968. — Т. 181, № 4.

Сердюкова С. И. Исследование устойчивости в C явных разностных схем с постоянными действительными коэффициентами, устойчивых в l_2 // ЖВМ и МФ. — 1963. — Т. 3, № 2.

Стрэнг (Strang G.). Difference methods for mixed boundary value problem // Duke Math. J. — I960. — V. 27, №. 2.

Томэ (Thomee V.). Generally unconditionally stable difference operators // SIAM J. Numer. Anal. — 1962. — V. 4. №. 1.

Федорюк М. В. Об устойчивости в C задачи Коши для разностных уравнений и уравнений с частными производными // ЖВМ и МФ. — 1967. — Т. 7, № 3.

Филиппов А. Ф. Об устойчивости разностных уравнений // ДАН СССР. — 1955. — Т. 100, № 6.

[7] **Устойчивость и сходимость**

Андреев В. Б. О сходимости разностных схем, аппроксимирующих вторую и третью краевые задачи для эллиптических уравнений II // ЖВМ и МФ. — 1968. — Т. 8, №6.

О'Брайен, Хайман, Каплан (O'Brien G. G., Hyman M. A., Kaplan S.). A study of the numerical solution of partial differential equations // J. of Math. and Phys, — 1951. — V. 29, №. 4.

Вендроф (Wendroff B.). Well-posed and stable difference operators // SIAM J. Numer. Anal. — 1968. — V. 5, №. 1.

Видлунд (Widlund O. B.). Stability of parabolic difference schemes in the maximum norm // Numer. Math. — 1968. — V. 8, №2.

Годунов С. К., Рябенький В. С. Канонические виды систем линейных обыкновенных разностных уравнений с постоянными коэффициентами // ЖВМ и МФ. — 1963. — Т. 3, № 2.

Годунов С. К., Рябенький В. С. Спектральные признаки устойчивости краевых задач для несамосопряженных разностных уравнений // УМН. — 1963. — Т. XVIII, № 3.

Джон (John F.). On the integration of parabolic equations by difference methods. I. Linear and quasilinear equations for the infinite interval // Comm. Pure Appl. Math. — 1952. — V. 5, №. 2.

Дюфорт, Франкел (Du Fort E. C., Frankel S. P.). Stability condi- tions in the numerical treatment of parabolic differential equations // Math. Tables and Other Aids Comput. — 1953. — V. 7, №. 3.

Курант, Фридрикс, Леви (Courant R., Friedrichs K., Lewy H.). Uber die partiellen Differenzengleichungen der mathematischen Physik // Math. Ann. — 1928. — V. 100, №. 32. (Рус. пер.: О разностных уравнениях математической физики // УМН. — 1940. — Т. VIII.)

Ладыженская О. А. Метод конечных разностей в теории уравнений с частными производными // УМН. — 1957. — Т. XII, № 5.

Лакс, Вендроф (Lax P. D., Wendroff B.). System of conservations laws // Comm. Pure Appl. Math. — I960. — V. 13, №. 2.

Лакс, Рихтмайер (Lax P. D., Richtmyer R. D.). Survey of the stability of linear finite difference equations // Comm. Pure Appl. Math. — 1956. — V. 9, №. 2.

Лиз (Lees M.). A priori estimate for the solution of difference approximations to parabolic partial differential equations // Duke Math. J. — I960. — V. 27, №. 3.

Лиз (Lees M.). Energy inequalities for the solution of differential equations // Trans. Amer. Math. Soc. — I960. — V. 94, №. 1.

Лионс (Lions J.). Equations differentielles operationnelles dans les espaces de Hilberth. — Centro Int. Mat. Estivo, Varenna (1963). (Equazioni differenziali astratte. Gremonese, Roma, 1963.)

Нейман, Рихтмайер (Neuman J., Richtmyer R. D.). A method for the numerical calculation of hydrodynamic shocks // J. Appl. Phys. — 1950. — V. 21, №. 3.

Рябенький В. С. Структура спектров свойств несамосопряженных разностных операторов // Материалы к совместному советско-американскому симпозиуму по уравнениям с частными производными. — Новосибирск, 1963.

Рябенький В. С. Спектр семейств разностных операторов над функциями на сеточном графе // ЖВМ и МФ. — 1967. — Т. 7, № 6.

Самарский А. А. Некоторые вопросы общей теории разностных схем // Дифференциальные уравнения с частными производными (труды симпозиума, посвященного 60-летию академика С. Л. Соболева). — М.: Наука, 1970.

Соболев С. Л. Некоторые замечания о численном решении интегральных уравнений // Изв. АН СССР. Сер. матем. — 1956. — Т. 20, № 4.

Стрэнг (Strang G.). Implicite difference methods for initialboundary value problems // J. Math. Anal. Appl. — 1966. — V. 16, №. 1.

Стрэнг (Strang G.). Accurate partial difference methods. I. Linear Cauchy problems // Arch. Rational Mech. Anal. — 1963. — V. 12, №. 5.

Томэ (Thomee V.). On the rate of convergence of difference schemes for hyperbolic equations // Numerical solution of partial differential equations. — II. SYNSPADE-1970. — N. Y.; L.: Academic Press, 1971.

Филлипс (Phillips N. A.). The atmosphere and the sea in motion. // Scientific Contributions to the Rossby Memorial Volume. The Rossby Memorial Volume. — The Rockfeller Institute, 1959. (Рус. пер.: Пример нелинейной вычислительной неустойчивости // Атмосфера и океан в движении. — М.: ИЛ, 1963.)

Яненко Н. Н., Бояринцев Ю. Е. О сходимости разностных схем для уравнения теплопроводности с переменными коэффициентами // ДАН СССР. — 1961. — Т. 139, № 6.

Яненко Н. Н., Шокин Ю. И. О связи корректности первых дифференциальных приближений и устойчивости разностных схем для гиперболических систем уравнений // Матем. заметки. — 1968. — Т. 4, № 5.

Яненко Н. Н., Шокин Ю. И. О корректности первых дифференциальных приближений разностных схем // ДАН СССР. — 1968. — Т. 182, № 4.

[8] **Вычислительные методы линейной алгебры**

Абрамов А. А. Идеи теории возмущений в некоторых алгоритмах линейной алгебры // Вычислительные методы линейной алгебры. Вып. 1. — М.: ВЦ АН СССР, 1968.

Бауэр, Фике (Bauer F. L., Pike C. T.). Norms and exclusion theorems // Numer. Math. — 1960. V. 2, № 3.

Бахвалов Н. С. К вопросу о гипотезе независимости ошибок определения при численном интегрировании // ЖВМ и МФ. — 1964. — Т. 4, № 3. — С. 339—404.

Бахвалов Н. С. Основы вычислительной математики: курс лекций. — М.: Изд-во МГУ, 1970.

Беллман Р. Введение в теорию матриц. — М.: Наука, 1969.

Воеводин В. В. Ошибки округления и устойчивость в прямых методах линейной алгебры. — М.: Изд-во МГУ, 1969.

Воеводин В. В. Численные методы алгебры. Теория и алгорифмы. — М.: Наука, 1966.

Воеводин В. В., Кузнецов Ю. А. Матрицы и вычисления. — М.: Наука, 1984.

Воеводин В. В., Тыртышников Е. Е. Вычислительные процессы с тѣплицевыми матрицами. — М.: Наука, 1987.

Гантмахер Ф. Р. Теория матриц. — М.: Наука, 1967.

Дородницын А. А. К задаче вычисления собственных чисел и собственных векторов матриц // ДАН СССР. — 1959. — Т. 126, № 6.

Дьяконов Е. Г. (D'yakonov E. G.). On the solution of some elliptic difference equations // J. Inst. Math. Appl. — 1971. — V. 7.

Икрамов Х. Д. Матричные нормы и методы типа Якоби. — М.: Изд-во МГУ, 1969.

Ильин В. П. О некоторых оценках для методов сопряженных градиентов // ЖВМ и МФ. — 1976. — Т. 16, № 4.

Келлог, Нодерер (Kellog R., Noderer L.). Sealed iterations and linear equations // SIAM J. — 1960. — V. 8, №. 4.

Ким Г. О распределении ошибок округления итерационных методов решения систем алгебраических уравнений.— М.: Изд-во МГУ, 1969.

Кублановская В. Н. Применение ортогональных преобразований для решения задач алгебры: автореф. дис. ... д-ра. — Л., 1972.

Кузнецов Ю. А. Итерационные методы решения несовместимых систем линейных уравнений. Некоторые проблемы вычислительной и прикладной математики. — Новосибирск: Наука, 1975.

Кузнецов Ю. А. Iterative Methods for Solution of Noncompatible Systems of Linear Equations // Lecture Notes in Economics and Mathematical Systems, Springer-Verlag. — 1976. — V. 134.

Кузнецов Ю. А., Марчук Г. И. Итерационные методы решения систем линейных уравнении с собственными матрицами. — Acta Universitatis Carobinse — Mathematica et Physica, Praha, 1974.

Кузнецов Ю. А., Марчук Г. И. Stationary Iterative Methods for the Solution of Systems of Linear Equations with Singular Matrices. — Munich, Germany: Gatlinburg VI, Symposium on Numerical Algebra, Conference Manuscripts, 1974.

Марек (Магес I.). Итерация линейных органических операторов и процесс Келлога: диссертация. — Прага, 1962.

Марек (Магес I.). On iteration of linear bounded operators and the convergence of Kellog's iteration process // Cech. Mat. J. — 1962. — V. 12.

Марчук Г. И., Кузнецов Ю. А. Итерационные методы и квадратичные функционалы // Методы вычислительной математики.— Новосибирск: Наука, 1975.

Немчинов С. В., Либов С. Л. Прямой метод повышенной точности решения краевых задач для уравнения Гельмгольца на сетке точек в прямоугольнике // ЖВМ и МФ. — 1964. — Т. 4, № 4.

Парлетт Б. Симметричная проблема собственных значений. — М.: Мир, 1983.

Самарский А. А., Николаев Е. С. Методы решения сеточных уравнений. — М.: Наука, 1978.

Стрэнг (Strang Ct.). Linear Algebra and its Applications. — N. Y.; L.: Acad. Press, 1976.

Трауб (Traub J.). Iterative Methods for the Solution of Equations. — Englewood Cliffs.: Prentice-Hall, 1964.

Уилкинсон Дж. Х. Алгебраическая проблема собственных значений. — М.: Наука, 1970.

Уилкинсон, Рейнш (Wilkinson J. H., Reinsch C.). Linear Algebra. — Berlin — Heidelberg, Springer-Verlag, 1971.

Фаддеев Д. К. О некоторых последовательностях полиномов, полезных для построения итерационных методов системы линейных алгебраических уравнений // Вестник ЛГУ. Сер. матем. — 1958. — Т. 7, № 2.

Фаддеев Д. К., Фаддеева В. Н. Вычислительные методы линейной алгебры. — М.: Физматгиз, 1963.

Фаддеев Д. К., Фаддеева В. Н. Вычислительные методы линейной алгебры // Зап. науч. сем. ЛОМИ, 54. — Л.: Наука, 1975.

Фаддеев Д. К., Фаддеева В. Н., Кублановская В. Н. Линейные алгебраические системы с прямоугольными матрицами // Вычислительные методы линейной алгебры. — М.: Наука, 1968.

Фландерс, Шортли (Flanders D. A., Shortley Ct.). Numerical determination of fundamental modes // J. Appl. Phys. — 1950. — V. 21.

Форсайт Дж., Молер К. Численное решение систем линейных алгебраических уравнений. — М.: Мир, 1969.

Френсис (Frencis J.). The QR-transformation. Part I, II // Computer J. — 1961. — V. 4.

Штифель (Stiefel E.). Kernel polynomials in linear algebra and their numerical applications // N. B. S. Appl. Math. Ser. 49. — 1958. — V. 1.

Эберлейн (Eberlein P.). A Jacobi-like method fot the automatic computation of eigenvalues of an arbitrary matrix // J. Soc. Industr. Appl. Math. — 1962. — V. 10.

Янг (Young D. M.). Iterative Solution of Large Linear Systems. — L.: Acad. Press, 1971.

[9] **Спектральные методы оптимизации итерационных процессов**

Абрамов А.А. Об одном способе ускорения итерационных процессов // ДАН СССР. — 1950. — Т. 74, № 6.

Бахвалов Н. С. О сходимости одного релаксационного метода при естественных ограничениях на эллиптический оператор // ЖВМ и МФ. — 1966. — Т. 6, № 5.

Воробьев Ю. В. Случайный итерационный процесс // ЖВМ и МФ. — 1964. — Т. 4, № 6; 1965. — Т. 5, № 5.

Гавурин М. К. Применение полиномов наилучшего приближения для улучшения сходимости итерационных процессов // УМН. — 1950. — Т. V, № 3.

Гавурин М. К. Нелинейные функциональные уравнения и непрерывные аналоги итерационных методов // Изв. вузов. Математика. — 1958. — Т. 5, № 6.

Голуб, Варга (Golub G. H., Varga R. S.). Chebysheve semi-iterative methods, successive over-relaxation iterative methods and second order Richardson iterative methods. Parts I, II // Numer. Math. — 1961. — V. 3, №. 2.

Дьяконов Е. Г. О построении итерационных методов на основе использования операторов, эквивалентных по спектру // ЖВМ и МФ. — 1966. — Т. 6, №1. — С. 4.

Золотарев Е. И. Приложение эллиптических функций к вопросам о функциях, наименее и наиболее отклоняющихся от нуля. — СПБ.: Записки Российской академии наук. 1877.

Иванов В. К. О сходимости итерационных процессов при решении систем линейных алгебраических уравнений // Изв. АН СССР. Сер. матем. — 1939. — Т. 4.

Коллатц (Collatz L.). Fehlerabschatzung fur das Iterationsverfahren zur Auflosung linear Gleichungssysteme // Z. Angew. Math. Mech. — 1942. — V. 22.

Ланцош (Lanczos C.). An iteration methods for the solution for the eigenvalue problem of linear, differential and integral operators // J. Res. Nat. Bus. Stand. — 1950. — V. 45, №. 1.

Ланцош (Lanczos C.). Chebyshev polionomials in the solution of largescale linear systems. — Proc. Assoc. Comput. Math. Toronto Meeting, September, 1952.

Лебедев В. И. Об итерационных методах решения операторных уравнений со спектром, лежащим на нескольких отрезках // ЖВМ и МФ. — 1969. — Т. 9, № 6.

Лебедев В. И. О построении операторов Р в КР-методе II // ЖВМ и МФ. — 1969. — Т. 9, № 4.

Лебедев В. И. Итерационный метод с чебышевскими параметрами для определения наибольшего собственного значения и соответствующей собственной функции // ЖВМ и МФ. — 1977. — Т. 17, № 1.

Лебедев В. И., Финогенов С. А. О порядке выбора итерационных параметров в чебышевском циклическом итерационном методе // ЖВМ и МФ. — 1971. — Т. 11, № 2.

Лебедев В. И., Финогенов С. А. Решение проблемы упорядочения параметров в чебышевских итерационных методах // ЖВМ и МФ. — 1973. — Т. 13, № 1.

Лебедев В. И., Финогенов С. А. Об использовании упорядоченных чебышевских параметров в итерационных методах // ЖВМ и МФ. — 1976. — Т. 16, № 4.

Марчук Г. И., Сарбасов К. Е. Об одном методе решения стационарной задачи // ДАН СССР. — 1968. — Т. 182, № 1.

Островский А. М. (Ostrowski A. M.). On the linear iteration procedures for symmetric matrices.— Univ. Roma, Inst. Naz. Alta Mat. Rend. Mat. e Appl. — 1954. — V. 14, №. 1-2.

Петришин (Petryshyn W.). On a general iterative method for the approximate solution of linear operator equations // Math. Comput. — 1963. — V. 17, №. 1.

Рейк (Reich E.). On the convergence of the classical iterative method of solving linear simultaneous equations // Ann. Math. Statist. — 1949. — V. 20, №. 3.

Федоренко Р. П. Релаксационный метод решения разностных эллиптических уравнений // ЖВМ и МФ. — 1961. — Т. 1, № 5.

Федоренко Р. П. О скорости сходимости одного итерационного процесса // ЖВМ и МФ. — 1964. — Т. 4, № 3.

Юнкоза, Милликен (Juncosa M. L., Milliken T. M.). On the increase of convergence rates of relaxation procedures for elliptic partial of difference equations // J. Assoc. Comput. Math. — I960. — V. 7, №. 1.

[10] **Метод верхней релаксации**

Бройден (Broyden C. G.). Some generalizations of the theory of succesive over-relaxation // Numer. Math.— 1964.— V. 6, №. 4.

Бройден (Broyden C. G.). On convergence criteria for the method of successive over-relaxation // Math. Comut. — 1964. — V. 18, №. 85.

Варга (Varga R. S.). .P-cyclic matrices: a generalization of the Young — Francel successive over-relaxation scheme // Pacific J. Math. — 1959. — V. 9.

Варга (Varga R. S.). Orderings of the successive over-relaxation scheme // Pacific. J. Math.— 1959.— V. 9.

Гарабедян (Garabedian P.). Estimation of the relaxation factor for small mesh size // Math. Tables and Other Aids Comput. — 1956. — V. 10, №. 56.

Гастинол (Gastinel N.). Sur le meilleur choix des parameters do surrelaxation (Procede de Peaceman — Rachford) // Chilfres. — 1962. — V. 5, №. 2.

Голуб (Golub G. H.). The use of Chebyshev matrix polinomials in the iterative solution of linear equations comared with the method of successive over-relaxation Doct. Thesis. — L'niv. of Illinois, 1959. 133 p.

Ивенс (Evans D. J.). Note on the line over-rolaxation factor for small mesh size // Comput. J. — 1962. — V. 5, №. 1.

Ивенс, Форингтон (Evans D. J., Forington C. B.). An iterative process for optimizing summetric successive over-relaxation // Comput. J. — 1963. — V. 6, №. 3.

Климова Е. Г., Ривин Г. С. О методах решения уравнения Булеева — Марчука // Изв. АН СССР. Сер. ФАО. — 1979. — Т. 15, №. 4.

Линн (Linn M. S.). On the round — off error in the method of successive over-relaxation // Math. Comput. — 1964. — V. 18, №. 85.

Островский (Ostrowski A. M.). On over- and under-relaxation in the theory of the cyclic single step iteration // Math. Tables and Other Aids Comput. — 1953. — V. 7, №. 43.

Петришин (Petryshyn W.). The generalized over — relaxation method for the approximate solution of operator equations in Hilbert-space // SIAM J. — 1962. — V. 10, №. 4.

Петришин (Petryshyn W.). On the extrapolated Jacobi or simultaneous displacements method in the solution of matrix and operator equations // Math. Comput. — 1965. — V. 19, №. 89.

Фаддеев Д. К. К вопросу о верхней релаксации при решении систем линейных уравнений // Изв. вузов. Математика. — 1958. — Т. 5.

Хагеман, Келлог (Hageman L. A., Kellogg R. B.). Estimating optimum over-relaxation parameters // Math. Comput. — 1968. — V. 22, №. 101.

Шелдон (Sheldon J.). On the numerical solution of elliptic difference equations // Math. Tables Aids Comput. — 1955. — V. 9.

Янг (Young D. M.). Iterative methods for solving partial difference equations of elliptic type // Trans. Amer. Math. Soc. — 1954. — V. 76.

Янг (Young D. M.). A bound for the optimum relaxation factor for the successive over-relaxation method // Numer. Math. — 1971. — V. 16, №. 5.

Янг (Young D. M.). Convergence properties of the symmetric and unsymmetric successive over-relaxation methods and related methods // Math Comput. — 1971. — V. 24, №. 112.

[11] **Градиентные методы**

Бирман М. Ш. Некоторые оценки для метода наискорейшего спуска // УМН. — 1950. — Т. V, №. 3.

Годунов С. К., Прокопов Г. П. Вариационный подход к решению больших систем линейных уравнений, возникающих в сильно эллиптических задачах. — М.: ИПМ АН СССР, 1968.

Годунов С. К., Прокопов Г. П. О решении разностного уравнения Лапласа // ЖВМ и МФ. — 1969. — Т. 9, №. 2.

Горбенко Н. И., Ильин В. П. О градиентных методах переменных направлений // Некоторые проблемы вычислительной и прикладной математики. — Новосибирск: Наука, 1975.

Даниель (Daniel J. W.). The conjugate gradient method for linear and nonlinear operator equations // SIAM J. Numer. Anal. — 1957. — V. 4, №. 1.

Даниель (Daniel J. W.). Convergence of the conjugate gradient method with computationally conveneient modifications // Numer. Math. — 1967. — V. 10, №2.

Канторович Л. В. О методе наискорейшего спуска // ДАН СССР. — 1947.— Т. 56, №. 3.

Красносельский М. А., Крейн С. Г. Итеративный процесс с минимальными невязками // Матем. сб. — 1952. — Т. 31.

Кузнецов Ю. А. К теории итерационных процессов // ДАН СССР. — 1969.— Т. 184, №. 2.

Кузнецов Ю. А. О симметризации итерационных процессов // Вычислительные методы линейной алгебры. — Новосибирск: ВЦ СО АН СССР, 1969.

Кузнецов Ю. А. Некоторые вопросы теории и приложений итерационных методов: автореф. канд. дисс.— Новосибирск, 1969.

Ланцош (Lanczos C.). Solution of the system of linear equations by minimized iterations // J. Res. Nat. Stand. — 1952. — V. 49, №. 1.

Марчук Г. П., Кузнецов Ю. А. К вопросу об оптимальных итерационных процессах // ДАН СССР. — 1968. — Т. 181, №. 6.

Марчук Г. И., Кузнецов Ю. А. Некоторые вопросы теории многошаговых итерационных процессов // Вычислительные методы линейной алгебры. — Новосибирск: ВЦ СО АН СССР, 1969.

Марчук Г. И., Кузнецов Ю. А. К решению систем линейных уравнений итерационными методами // Вопросы точности и эффективности вычислительных алгоритмов. Вып. 1. — Киев: Изд-во Ин-та кибернетики АН УССР, 1969.

Самокиш Б. А. Исследование быстроты сходимости метода наискорейшего спуска // УМН. — 1957. — Т: XII, №. 1.

Форсайт (Forsythe G. E.). On the asymptotic directions of the s-dimensional optimum gradient method // Numer. Math. — 1968. — V. 11, №. 1.

Форсайт, Мотцкин (Forsythe G. E., Motzkin T. S.). Asymptotic properties of [the optimum gradient method // Bull. Amor. Math. Soc. — 1951. — V. 57, №. 2.

Форсайт, Мотцкин (Forsythe G. E., Motzkin T. S.). Acceleration of the optimum gradient method // Bull. Amer. Math. Soc. — 1951. — V. 57, №. 4.

Форсайт, Форсайт (Forsythe A. I., Forsythe G. E.). Punchedcard experiments with accelerated gradient methods for linear equations. Contributions to the solution of linear aquations and the determination of eigenvalues // N. B. S. Appl. Math., Ser. 39. — 1954.

Фридман В. М. Некоторые методы решения линейного операторного уравнения // ДАН СССР. — 1959. — Т. 128, №. 3.

Хестенс, Штифель (Hestenes M. R., Stielel E.). Method of conjugate gradients for solving linear systems // J. Res. Nat. Bur. Stand. — 1952. — V. 49.

[12] **Методы факторизации (прогонки)**

Абрамов А. А., Андреев В. Б. О применении метода прогонки к нахождению периодических решений дифференциальных и разностных уравнений // ЖВМ и МФ. — 1963. — Т. 3, №. 2.

Аинс Э. Обыкновенные дифференциальные уравнения. — М.: ОНТИ, 1939.

Бахвалов Н. С. О накоплении вычислительной погрешности при численном решении дифференциальных уравнений. — Сб. ВЦ МГУ. — 1962. Т. 1.

Бужби, Голуб, Нильсон (Buzbee B., Golub G., Nilson E.). On direct methods for solving Poinsson’s equations // SIAM J. Numer. Anal. — 1970. — V. 7, №. 4.

Булеев Н. И. Численный метод решения двумерных и трехмерных уравнений диффузии // Матем. сб. — I960. — Т. 5, №. 2.

Булеев Н. И. Метод неполной факторизации для решения двумерных и трехмерных уравнений типа диффузии // ЖВМ и МФ. — 1970. — Т. 10, №. 4.

Владимиров В. С. Приближенное решение одной краевой задачи для дифференциального уравнения второго порядка // Прикл. матем. и мех. — 1955. — Т. 19, №. 3.

Гельфанд И. М., Локуциевский О. В. Метод прогонки для решения разностных уравнений // Введение в теорию разностных схем / Годунов С. К., Рябенький В. С. — М.: Физматгиз, 1962.

Годунов С. К. Метод ортогональной прогонки для решения систем разностных уравнений // ЖВМ и МФ. — 1962. — Т. 2, №. 6.

Дегтярев Л. М., Фаворский А. П. Потоковый вариант метода прогонки // ЖМВ и МФ. — 1968. — Т. 8. №. 3.

Дегтярев Л. М., Фаворский А. П. Потоковый вариант метода прогонки для разностных задач с сильно меняющимися коэффициентами // ЖВМ и МФ. — 1969. — Т. 9, №. 1.

Огнева В. В. Метод прогонки для решения разностных уравнений // ЖВМ и МФ. — 1967. — Т. 7, №. 4.

Олифант (Oliphant T. A.). An implicit, numerical method for solving two-dimentional time dependent diffusion problems // Quart. Appl. Math. — 1961. — V. XIX, №. 3.

Русанов В. В. Об устойчивости метода матричной прогонки // Вычислительная математика. М., 1960.

Сафронов И. Д. О методе для решения краевых задач для разностных уравнений // ЖВМ и МФ. — 1964. — Т. 4, №. 2.

Сафронов И. Д. Разностная схема с диагональными направлениями прогонок для решения уравнения теплопроводности // ЖВМ и МФ. — 1965. — Т. 5, №. 2.

Фаге М. К. О методе прогонки // ДАН СССР. — 1970. — Т. 191, №. 2.

[13] **Быстрое преобразование Фурье**

Бингем, Годфри, Таки (Bingham C., Godfrye M. D., Tukey J.). Modern techniques of power spectrum estimation // IEEE Trans., Audio and Electroacoustics. — AU, 1967. — V. 15.

Голд, Радер (Gold B., Rader C. M.). Digital processing of signal. N. Y.: McGraw-Hill, 1969.

Канеко Т., Лю Б. (Kaneko T., Liu B.). Accumulation of roundoff error in fast Fourier transforms // J. Assoc. Comput. Mach. — 1970. — V. 17.

Клаудер, Прайс, Дарлингтон, Элберзгайм (Klauder J. R., Price A. C., Darlington S., Albersheim W. J.). The theory and design of chirp radars // Bell System Tech. J. — 1960. — V. 39. (См. также: Клаудер, Прайс, Дарлингтон, Элберзгайм. Теория и расчет импульсных радиолокационных станций с частотной модуляцией. — Зарубежная радиоэлектроника. 1961. Вып. 1.)

Кузнецов Ю. А., Мацокин А. М. Решение уравнения Гельмгольца методом фиктивных областей // Вычислительные методы линейной алгебры. — Новосибирск: ВЦ СО АН СССР, 1972.

Кули, Льюис, Уэлч (Cooley J. W., Lewis P. A., Welch P. D.). The fast Fourier transform algorithm and its applications.— IBM Research Paper RC — 1743, Feb. 1967.

Кули, Таки (Cooley J. W., Tukey J. W.). An algorithm for the machine calculation of complexes Fourier series // Math. Comput. — 1965. — V. 19, №. 90.

Немчинов С. В. О применении метода сеток к решению краевых задач для уравнений в частных производных с периодическими краевыми условиями // Динамическая метеорология. — Ташкент: Наука, 1965.

Сегет К. (Segeth K.). Roundoff errors in the fast computation of discrete convolutions.— Praha: Math. Ustav CSAV, 1979.

Таки (Tukey J. W.). An introduction to the calculations of numerica: spectrum analysis // Spectral Analysis in Time Series, Bern Harris, L. D. Wiley. — N. Y., 1967.

Хелмс (Helms R. D.). Fast Fourier transform method for computing diffrence equation and simulating filters. — IEEE Trans., Audio and Electroacoustics. — AU, 1967. — V. 15.

Хокни (Hockney R. W.). A fast direct solution of Poissons equation using Fourier analysis // J. Assoc. Comp. Mech. — 1965. — V. 12, №. 1.

[14] **Интерполяция с помощью сплайнов**

Алберг, Нильсон, Уолш (Alberg J. H., Nilson E. N., Walsh J. L.). Extremal orthogonalines // J. Math. Appl. — 1965. — V. 12, №. 1.

Алберг Дж., Нильсон Э., Уолш Дж. Теория сплайнов и ее приложения. — М.: Мир, 1972.

Ананьин А. З., Смелов В. В., Василенко В. А. Эффективный способ преобразования вариационной задачи сглаживания к линейной алгебраической системе.— Новосибирск: Препринт ВЦ СО АН СССР, 1976. — Вып. 28.

Анселон, Лоран (Anselon P. M., Laurent P. J.). A general method for construction of interpolating or smoothing spline – functions // Numer. Math. — 1968. — V. 12, №. 1.

Атья (Atteia M.). Generalisation de la definition et des proprietes des «spline fonctions» // C. R. Acad. Sci., P. — 1965. — V. 260.

Бежаев А. Ю. Ошибки сплайн-интерполяции в многомерных ограниченных областях. — Новосибирск: Препринт ВЦ СО АН СССР, 1984.

Бежаев А. Ю., Василенко В. А. (Bezhaev A. Yu., Vasilenko V. A.). Splines in Hilbert spaces and their finite element approximations // SNAMM. — 1987. — V. 2, №. 3. — Pp. 191—202.

Белоносов А. С., Цецохо В. А. Вычислительный алгоритм и процедуры сглаживания функций, заданных приближенно в узлах нерегулярной сетки на плоскости // Некорректные задачи математической физики и проблемы интерпретации геофизических наблюдений (Математические проблемы геофизики). — Новосибирск: ВЦ СО АН СССР, 1976.

Биркгоф, Гарабедян (Birkhoff G., Garabedian P.). Smooth surface interpolation // J. Math. Phys. — 1960. — V. 39, №. 3.

Бур (Bde Boor C.). Bicubic spline interpolation // J. Math. Phys. — 1962. — V. 41, №. 2.

Василенко В. А. Сплайн-функции: теория, алгоритмы, программы. — Новосибирск: Наука, 1983.

Василенко В. А. Сходимость операторных интерполирующих сплайнов // Вариационно-разностные методы в математической физике. — Новосибирск: ВЦ СО АН СССР, 1973.

Василенко В. А. Сглаживающие сплайны на подпространствах и теоремы компактности // Численные методы механики сплошной среды. — 1974. — Т. 5, №. 5.

Василенко В. А. Конечномерная аппроксимация в методе наименьших квадратов // Вариационно-разностные методы в математической физике. Вып 2. — Новосибирск: ВЦ СО АН СССР, 1975.

Василенко В. А. Сходимость сплайнов в гильбертовом пространстве // Численные методы механики сплошной среды. — 1972. — Т. 3, №. 3.

Василенко В. А., Зюзин М. В., Ковалков А. В. Сплайн-функции и цифровые фильтры. — Новосибирск: ВЦ АН СССР, 1984.

Василенко В. А., Переломов Е. М. Сплайн-интерполяция в прямоугольной области с хаотически расположенными узлами // Машинная графика и ее применение. — Новосибирск: ВЦ СО АН СССР, 1973.

Гребенников А. И. Метод сплайнов в численном анализе. — М.: Изд-во МГУ, 1979.

Гребенников А. И. Метод сплайнов и решение некорректных задач теории приближений. — М.: МГУ, 1983.

Завьялов Ю. С. Интерполирование кубическими многозвенниками // Вычислительные системы. Вып. 38. — Новосибирск, 1970.

Завьялов Ю. С. Экстремальное свойство кубических многозвенников и задача сглаживания // Вычислительные системы. Вып. 42. — Новосибирск, 1970.

Завьялов Ю. С. Интерполирование мультикубическими сплайнами // Вычислительные системы. Вып. 65. — Новосибирск, 1975.

Завьялов Ю. С., Квасов Б. И., Мирошниченко В. Л. Методы сплайн-функций. — М.: Наука, 1980.

Лебедев В. И. Об одном способе интерполяции в n-мерном пространстве по произвольным узлам и некоторых квадратурных формулах. — Новосибирск: Препринт ВЦ СО АН СССР, 1975. — Вып. 10.

Михалевич Ю. И., Омельченко О. К. Процедуры кусочно-полиномиальной интерполяции функции одной и двух переменных. — Новосибирск: ВЦ СО АН СССР, 1970.

Морозов В. А. О выборе параметра при решении функциональных урав-

нений методом регуляризации // ДАН СССР. — 1967. — Т. 175, №. 6.
Морозов В. А. Теория сплайнов и задачи устойчивого вычисления значений неограниченного оператора // ЖВМ и МФ. — 1971. — Т. 11, №. 3.
Пивоварова Н. Б., Пухначева Т. П. Сглаживание экспериментальных данных локальными сплайнами.— Новосибирск: Препринт ВЦ СО АН СССР, 1975. — Вып. 9.
Реинш (Reinsch C. H.). Smoothing by spline functions // Numer. Math. — 1967. — V. 10, №. 4.
Рябенький В. С. Локальные формулы гладкого восполнения и гладкой интерполяции функций по их значениям в узлах неравномерной прямоугольной сетки. — М.: ИПМ АН СССР, 1974.
Стечкян С. Б., Субботин Ю. Н. Сплайны в вычислительной математике. — М.: Наука, 1976.
Уолш, Алберг, Нильсон (Waksh J. L., Ahlberg J. H., Nilson E. N.). Best approximation properties of the spline fit // J. Math. Mech. — 1962. —V. 11, №. 2.
Холидеи (Holladey J. C. D. Smoothest curve approximation // Math. Tables Aids Computation. — 1957. — V. 11, №. 60.
Цецохо В. А., Белоносов А. С., Белоносова А. В. Об одном методе гладкого приближения функций многих переменных.— Новосибирск: Препринт ВЦ СО АН СССР, 1974. Вып. 8.
Шенберг (Schoenberg I. J.). Contibutions to the problem of approximation of equidistant data by analitic functions. Parts A and B // Quart. Appl. Math. — 1946. — V. 4.
Шумахер (Schumaker L. L.). Approximation by splines: Theory and applications of spline functions.— N. Y.; L.: Academic Press, 1969.
Яненко Н.Н., Квасов Б.И. Итерационный метод построения поликубических сплайн-функций // Численные методы механики сплошной среды, 1970.— Т. 1, №3.

[15] **Методы расщепления**

Андреев В. Б. О разностных схемах с расщепляющимся оператором для общих p-мерных параболических уравнений второго порядка со смешанными производными // ЖВМ и МФ. — 1967. — Т. 7, № 2.
Багриновский К. А., Годунов С. К. Разностные методы для многомерных задач // ДАН СССР. — 1957. — Т. 115, № 3.
Бейкер (Baker G. A.). An implicit numerical method for solving the n-dimensional heat equation // Quart. Appl. Math. — 1960. — V. 17, №. 4.
Бейкер, Олифант (Baker G. A., Oliphant T. A.). An implicit numerical

method for solving the two-dimensional heat equation // Quart. Appl. Math. — I960. — V. 17, №. 4.

Бенсусан (Bensoussan A.). Pure decentralization for interrelated payofss. // In: Symposium on Optimization. — Los Angeles, 1971.

Бенсусан А., Лионе Ж.-Л., Темам Р. Методы декомпозиции, децентрализации, координации и их приложения // Методы вычислительной математики. — Новосибирск: Наука, 1975.

Биркгоф, Варга (Birkhof G., Varga R.). Implicitc alternating direction methods // Trans. Amer. Math. Soc. — 1959. — V. 92, №. 1.

Биркгоф, Варга, Янг (Birkhof G., Varga R., Young D.). Alternating direction implicit methods // Advances in Comp. — N. Y.; L.: Academic Press. — 1962. — V. 3.

Булеев Н. И. Численный метод решения двумерных и трехмерных уравнений диффузии // Матем. сб. — 1960. — Т. 51, №$2.

Вакспресс (Wachspress E. L.). Optimum alternating-direction-implicit iteration parameters for a model problem // SIAM J. — 1962. — V. 10, №. 2.

Вакспресс (Wachspress E. L.). Extended application of alternating-direction-implicit iteration model problem theory // SIAM J — 1963 — V. 11, №. 3.

Вакспресс (Wachspress E. L.). Iterative Solution of Elliptic Systems and Applications to the Neutron Diffusion Equations of Reactor Physics. — Englewood Cliffs; Prentice-Hall, 1966.

Вакспресс, Хабетлер (Wachspress E. L., Habetler G. J.). An alternating-direction-implicit iteration technique // SIAM J. — 1960. — V. 8, №. 2.

Варга (Varga B.). Some results in approximation theory with applications to numerical analysis // Numerical solution of partial differential equations. — II. SYNSPADE-1970. — N. Y.; L.: Academic Press, 1971.

Видлунд (Widlund O.). On the rate of an alternating-direction-implicit method in a non-commutative case // Math. Comput. — 1966. — V. 20, №. 96.

Видлунд (Widlund O.). On the effects of scalling of the Peaceman-Rachford method // Math. Comput. — 1971. — V. 25, №. 113.

Воробьев Ю.В. Случайный итерационный процесс в методе переменных направлений // ЖВМ и МФ. — 1968. — Т. 8, № 3.

Ганн (Gunn J. E.). The solution of elliptic difference equations by semiexplicit iterative techniques // SIAM J. Numer. Anal. — 1965. — V. 2, №. 1.

Дуглас, Ган (Douglas J., Gunn J. E.). Two high-order correct difference aniogues for the equation of multi-dimensional heat flow // Math. Comput. — 1963. — V. 17, №. 81.

Дуглас, Ганн (Douglas J., Gunn J. E.). A general formulation of alternating direction methods.— Part I. Parabolic and hyperbolic problems // Numer. Math. — 1964. — V. 6, №. 5.

Дуглас, Джонс (Douglas J., Jones B. F.). On predictor-corrector methods for nonlinear parabolic differential equations // J. Soc. Industr. Appl. Math. — 1963. — V. 11, № 1.

Дуглас, Келлот, Варга (Douglas J., Kellogg R. B., Varga R. S.). Alternating direction methods for n-space variables // Math. Comput. — 1963. — V. 17, №. 83.

Дуглас, Пирси (Douglas J., Pearcy C. M.). On convergence of alternating direction procedures in the presence of singular operators // Numer. Math. — 1963. — V. 5, №. 2.

Дуглас, Рэчфорд (Douglas J., Bachford H.). On the numerical solution of heat conduction problems in two and three space variables // Trans. Amer. Math. Soc. — 1956. — V. 82, №. 2.

Дьяконов Е. Г. Метод переменных направлений решения систем конечно-разностных уравнений // ДАН СССР. — 1961. — Т. 138, № 2.

Дьяконов Е. Г. О некоторых разностных схемах для решения краевых задач // ЖВМ и МФ. — 1962. — Т. 2, № 1.

Дьяконов Е. Г. Разностные схемы с расщепляющимся оператором для многомерных стационарных задач // ЖВМ и МФ. — 1962. — Т. 2, № 4.

Дьяконов Е. Г. Решение некоторых многомерных задач математической физики при помощи сеток: автореф. дис. ... канд.— М., 1962.

Дьяконов Е. Г., Лебедев В. И. Метод расщепления для третьей краевой задачи // Вычислительные методы и программирование. Вып. IV. — М.: Изд-во МГУ, 1967.

Дюпон (Dupont T.). A factorization procedure for the solution of elliptic difference equations // SIAM J. Numer. Anal. — 1968. — V. 5, №. 4.

Ильин В. П. О расщеплении разностных уравнений параболического и эллиптического типов // Сиб. матем. ж. — 1965. — Т. VI, № 1.

Ильин В. П. О явных схемах переменных направлений // Изв. СО АН СССР. Сер. техн. наук. — 1967. — Т. 13, № 3.

Келлог (Kellog R. B.). Another alternating-direction-implicit method // J. Soc. Industr. Appl. Math. — 1963. — V. 11, №. 4.

Келлог (Kellog R. B.). An alternating direction method for operations // J. Soc. Industr. Appl. Math. — 1964. — V. 12, №. 4.

Келлог, Сианьер (Kellogg R. B., Spanier J.). On optimal alternating direction parametei's for singular matrices // Math. Comput. — 1965. — V. 19, №. 91.

Коновалов А. Н. Метод дробных шагов решения задачи Коши для многомерного уравнения колебаний // ДАН СССР. — 1962. — Т. 147, № 1.

Коновалов А. Н. Применение метода расщепления к численному решению динамических задач теории упругости // ЖВМ и МФ. — 1964. — Т. 4, № 4.

Коновалов А. Н. Задачи фильтрации многофазной несжимаемой жидкости. — Новосибирск: Изд-во НГУ, 1972.
Кузнецов Б. Г. (Kuznetsov B. G.). Numerical methods for solving some problems of viscous liquid // Fluid Dynamics Transactions, 1969. — Т. 4.
Лиз (Lees M.). Alternating direction methods for hyperbolic differential equations// J. Soc. Industr. Appl. Math. — 1962. — V. 10, №. 4.
Лиз (Lees M.). Alternating direction and semi-explicit difference methods for parabolic partial differential equations // Numer. Math. — 1961. — V. 3, №. 5.
Лионс П. Л., Мерсье Б. (Lions P. L., Mercier B.). Splitting algorithms for the sum of two nonlinear operators. — P. Centre de Mathematiques appliquees; Janvier 1978, Rapport interne № 29.
Марчук Г. И. Методы расщепления.— М.: Наука, 1988.
Марчук Г. И. (Marchuk G. I.). On the theory of the spliting-up method. // In: Numerical solution of partial differential equations. — II. SYNSPADE-1970. — N. Y.; L.; Academic Press, 1971.
Марчук Г. И., Султангазин У. М. К обоснованию метода расщепления для уравнения переноса излучения // ЖВМ и МФ. — 1965. — Т. 5, № 5.
Марчук Г. И., Яненко Н. Н. Применение метода расщепления (дробных шагов) для решения задач математической физики // Некоторые вопросы вычислительной и прикладной математики. — Новосибирск: Наука, 1966.
Писман, Рэчфорд (Peacoman D. W., Rachford H. H.). The numerical solution of parabolic and elliptic differential equations // SIAM J. — 1955. — V. 3, №. 1.
Самарский А. А. Об одном экономическом разностном методе многомерного параболического уравнения в произвольной области // ЖВМ и МФ. — 1962. — Т. 2, № 5.
Самарский А. А. О сходимости метода дробных шагов для уравнения теплопроводности // ЖВМ и МФ. — 1962. — Т. 2, № 6.
Самарский А. А. Локально одномерные разностные схемы на неравномерных сетках // ЖВМ и МФ. — 1963. — Т. 3, № 3.
Самарский А. А. Об одном экономическом алгоритме численного решения систем дифференциальных и алгебраических уравнений // ЖВМ и МФ. — 1964. — Т. 4, № 3.
Самарский А. А. Экономические разностные схемы для гиперболической системы уравнений со смешанными производными и их применение для уравнений теории упругости // ЖВМ и МФ. — 1965. — Т. 5, № 1.
Самарский А. А. Аддитивные схемы // Тезисы докладов на Международном съезде математиков в Москве. — М., 1966.

Темам (Temam R.). Sur la stabilite et la convergence de la Methode des pas fractionnaires // Annali di Mat. Pura ed Appl. — 1968. — V. IV, №. 79.
Темам (Temam R.). Quelques methodes de decomposition en analyse numerique // Acta du Congres Intern, des Math. — 1970. — V. 3.
Фейрвезер, Митчелл (Fairweather G., Mitchell A. R.). Some computational results of an improved A. D. I. method for the Dirichlet problem // Comput. J. — 1966. — V. 9, №. 3.
Фрязинов И. В. О разностных схемах для уравнения Пуассона в полярной цилиндрической и сферической системах координат // ЖВМ и МФ. — 1971. — Т. 11, №. 5.
Хабард (Hubbard B. E.). Alternating direction schemes for the heat equation in a general domain J. Numera. // SIA Anal. — 1966. — V. 2, №. 3.
Яненко Н. Н. Об одном разностном методе счета многомерного уравнения теплопроводности // ДАН СССР. — 1959. — Т. 125, № 6.
Яненко Н. Н. Об экономических неявных схемах (метод дробных шагов) // ДАН СССР. — 1960. — Т. 134, № 5.
Яненко Н. Н. О неявных разностных методах счета многомерного уравнения теплопроводности // Изв. вузов. Математика. — 1961. — Т. 4, № 23.
Яненко Н. Н. О сходимости метода расщепления для уравнения теплопроводности с переменными коэффициентами // ЖВМ и МФ. — 1962. — Т. 2, № 5.
Яненко Н. Н. О слабой аппроксимации систем дифференциальных уравнений // Сиб. матем. ж. — 1964. — Т. V, № 6.
Яненко Н. Н., Демидов Г. В. Метод слабой аппроксимации как конструктивный метод построения решения задачи Коши // Некоторые вопросы вычислительной и прикладной математики. — Новосибирск: Наука, 1966

[16] **Условно-корректные задачи и некоторые обратные задачи математической физики**

Аниконов Ю. Е. Некоторые методы исследования многомерных обратных задач для дифференциальных уравнений. — Новосибирск: Наука, 1978.
Березанский Ю. М. Об однозначности определения уравнения Шредингера по его спектральной функции // ДАН СССР. — 1953. — Т 93, № 4. — С. 591—594.
Бухгейм А. Л. Об одном классе операторных уравнений Вольтерра первого рода // Функц. анализ. — 1972. — Т. 6, № 1. — С. 1—9.
Бухгейм А. Л. Операторные уравнения Вольтерра в шкалах банаховых пространств // ДАН СССР. — 1978. — Т. 242, № 2. — С. 272—275.

Гончарский А. В., Черепащук А. М., Ягола А. Г. Численные методы решения обратных задач астрофизики. — М.: Наука, 1978.
Джон (John F.). Differential Equation with Approximate and Improper Data: Lectures. — New York Univ., 1955.
Дуглас (Douglas J.). On the relation between stability and convergence in the numerical solution of linear parabolic and hyperbolic differential equations // J. Soc. Indust. Appl. Math. — 1956. — V. 4, № 1.
Иванов В. К. О некорректно поставленных задачах // Матем. сб. — 1963. — Т. 61, № 2.
Иванов В. К., Васин В. В., Танана В. П. Теория линейных некорректных задач и ее приложения. — М.: Наука, 1978.
Кадомцев Б. Б. О функции влияния в теории переноса лучистой энергии // ДАН СССР. — 1957. — Т. 113, № 3.
Крейн С. Г. О классах корректности для некоторых задач // ДАН СССР. — 1957. — Т. 114, № 6.
Крейн С. Г., Прозоровская О. И. О приближенных методах решения некорректных задач // ЖВМ и МФ. — 1963. — Т. 3, № 1.
Лаврентьев М. А. Numerical solution of conditionally properly posed problems // Numerical solution of partial differential equations. — II. SYNSPADE-1970. — N. Y.; L.: Academic Press, 1971.
Лаврентьев М. М. О задаче Коши для уравнения Лапласа // ДАН СССР. — 1956. — Т. 102, № 2.
Лаврентьев М. М. О постановке некоторых некорректных задач математической физики // Некоторые вопросы вычислительной и прикладной математики. — Новосибирск: Наука, 1956.
Лаврентьев М. М., Васильев В. Г. О постановке некоторых некорректных задач математической физики // Сиб. матем. ж. — 1960. — Т. VII, № 3.
Лаврентьев М. М., Романов В. Г., Васильев В. Г. Многомерные обратные задачи для дифференциальных уравнений. — Новосибирск: Наука, 1969.
Лаврентьев М. М., Романов В. Г., Шишатский С. П. Некорректные задачи математической физики и анализа. — М.: Наука, 1980.
Лаврентьев М. М. Условно-корректные задачи для дифференциальных уравнений. — Новосибирск: Изд-во НГУ, 1973.
Ландис Е. М. О некоторых свойствах решений эллиптических уравнений // ДАН СССР. — 1956. — Т. 107, № 4.
Лионс, Латтес (Lions J., Lattes R.). Метод квазиобращения и его применения. — М.: Мир, 1970.
Магницкий Н. А. Об одном методе регуляризации уравнения Вольтерра 1-го рода // ЖВМ и МФ. — 1975. — Т. 15, № 5. — С. 1317—1323.
Марченко В. А. Операторы Штурма — Лиувилля и их приложения. —

Киев: Наукова думка, 1978.
Марчук А. Г. Оптимальные по точности методы решения задач восстановления. — Новосибирск: Препринт / ВЦ СО АН СССР, 1976. — № 10.
Марчук Г. И. Уравнения для ценности информации с метеорологических спутников и постановка обратных задач // Космические исследования. — 1964. — Т. 11, № 3.
Марчук Г. И., Атанбаев С. А. Некоторые вопросы глобальной регуляризации // ДАН СССР. — 1970. — Т. 190, № 3.
Марчук Г. И., Васильев В. Г. О приближенном решении операторных уравнений первого рода // ДАН СССР. — 1970. — Т. 199, № 4.
Марчук Г. И., Дробышев Ю. П. Некоторые вопросы линейной теории измерений // Автометрия. — 1967. — Т. 3.
Марчук Г. И., Орлов В. В. К теории сопряженных функций // Нейтронная физика. — М.: Атомиздат, 1961.
Мергелян С. Н. Гармоническая аппроксимация и приближенное решение задачи Коши для уравнения Лапласа // УМН. — 1956. — Т. XI.
Морозов В. А. Методы решения неустойчивых задач (тексты лекций). — М.: Изд-во МГУ, 1967.
Мухаметов Р. Г. К задаче восстановления анизотропной римановской метрики в n-мерной области. — Новосибирск: Препринт № 136 ВЦ СО АН СССР, 1978.
Прилепко А. И. Обратные задачи теории потенциала // Матем. заметки. — 1973. — Вып. 14. — № 5. — С. 755—765.
Романов В. Г. Некоторые обратные задачи для уравнений гиперболического типа. — Новосибирск: Наука, 1972.
Романов В. Г. Обратные задачи для дифференциальных уравнений (обратная кинематическая задача сейсмики). — Новосибирск: Изд-во НГУ, 1978.
Сергеев В. О. Регуляризация уравнения Вольтерра I рода // ДАН СССР. — 1971. — Т. 197, № 3. — С. 531—534.
Тихонов А. Н. Об устойчивости обратных задач // ДАН СССР. — 1943. — Т. 39, № 1.
Тихонов А. Н. О решении некорректно поставленных задач и методе регуляризации // ДАН СССР. — 1963. — Т. 151, № 3.
Тихонов А. Н. О регуляризации некорректно поставленных задач // ДАН СССР. — 1963. — Т. 153, № 1.
Тихонов А. Н., Арсенин В. Я. Методы решения некорректных задач. — М.: Наука, 1974.
Фаддеева В. Н. Сдвиг для систем с плохо обусловленными матрицами // ЖВМ и МФ. — 1965. — Т. 5, № 5.

Федотов А. М. О двух подходах к исследованию условно-корректных задач со случайными ошибками в исходных данных. — Новосибирск: Препринт № 80 ВЦ СО АН СССР, 1978.

Франк Л. С., Чудов Л. А. Разностные методы решения некорректной задачи Коши // Численные методы в газовой динамике. — М.: Изд-во МГУ, 1965.

Фукс (Fuks K.). Perturbation theory in neutron multiplication problems // Proc. Phys. Soc. — 1949. — V. 62, № 791.

Шишатский С. П. Об одном методе приближенного решения некорректной задачи Коши для эволюционного уравнения // Математические проблемы геофизики. Вып. 3. — Новосибирск: ВЦ СО АН СССР, 1972. — С. 216—228.

[17] **Вычислительные методы в теории переноса**

Агошков В. И. Вариационные методы в теории переноса: автореф. дис. ... канд. — Новосибирск: ВЦ СО АН СССР, 1975.

Агошков В. И. Выбор базисных функций при решении некоторых эллиптических уравнений // Вычислительная математика и программирование. — Новосибирск: ВЦ СО АН СССР, 1974.

Агошков В. И. Некоторые особенности решения уравнения переноса и учет их при построении базисных функций. — Новосибирск: Препринт № 58 ВЦ СО АН СССР, 1977.

Агошков В. И. Обобщенные решения уравнения переноса и свойства их гладкости. — М.: Наука, 1988.

Агошков В. И. О гладкости решений уравнения переноса и приближенных методах их построения. I, II // Дифференциальные и интегро-дифференциальные уравнения. — Новосибирск: ВЦ СО АН СССР, 1977.

Агошков В. И. Решение уравнения переноса в X — Y-геометрии методом интегральных тождеств. — Новосибирск: Препринт № 159 ВЦ СО АН СССР, 1979.

Бардос (Bardos P. G.). Equations du premier ordere a coefficients reels // Ann. Sci. EC. Norm. Sup. 4^e series. — 1970. — T. 3.

Боголюбов Н. Н. Проблемы динамической теории в статистической физике. — М.: Гостехиздат, 1946.

Владимиров В. С. Численное решение кинетического уравнения для сферы // Вычислительная математика, Т. 3. — М.: ВЦАН СССР, 1958.

Владимиров В. С. Математические задачи односкоростной теории переноса частиц. // Труды матем. ин-та АН СССР. — 1961. — Т. 61.

Владимиров В. С. О некоторых вариационных методах приближенного

решения уравнения переноса // Вычислительная математика. Т 7. — М.: ВЦ АН СССР, 1961.
Гермогенова Т. А. Локальные свойства решений уравнений переноса. — М.: Наука, 1986.
Гермогенова Т. А. О сходимости некоторых приближенных методов решения уравнения переноса // ДАН СССР. — 1968. — Т. 181, № 3.
Гермогенова Т. А. Обобщенные решения краевых задач для уравнения переноса // ЖВМ и МФ. — 1969. — Т. 9, № 3.
Годунов С. К. Использование интеграла энергии для оценки точности приближенных собственных значений // ЖВМ и МФ. — 1971. — Т. 11, № 5.
Годунов С. К., Султангазин У. М. О диссипативности граничных условий В. С. Владимирова для симметрической системы метода сферических гармоник // ЖВМ и МФ. — 1971. — Т. 11, № 3.
Гольдин В. Я. Квазидиффузионный метод решение кинетического уравнения // ЖВМ и МФ. — 1964. — Т. 4, № 6.
Иоргенс (Jorgens K.). An asymptotic expansion in the tneory of neutron transport // Comm. Pure Appl. Math. — 1958. — V. 11, №. 2.
Карлсон Б., Белл Дж. Решение транспортного уравнения S_n-методом // Физика ядерных реакторов. — М.: ИЛ, 1963.
Кузнецов Е. С., Марчук Г. И. Вычислительные методы в теории переноса излучения // Труды IV Всесоюзного математического съезда. Ленинград, 3—12 июля 1961 г. Т. II. Секционные доклады. — Л.: Наука, 1964.
Лебедев В. И. О нахождении решений кинетических задач теории переноса: автореф. дис. ... д-ра. — Новосибирск, 1967.
Лебедев В. И. О KP-методе и разностных схемах для кинетического уравнения // Вычислительные методы в теории переноса. — М.: Атомиздат, 1969.
Марек (Marec I.). On a problem of mathematical physics // Appl. Math. — 1966. — V. II, № 89.
Марчук Г. И. Численные методы расчета ядерных реакторов. — М.: Атомиздат, 1958.
Марчук Г. И., Кочергин В. П. Эффективный метод решения двумерного уравнения диффузии для ячеек квадратной и шестиугольной формы // Атомная энергия. — 1965. — Т. 18, № 6.
Марчук Г. И., Лебедев В. И. Численные методы в теории переноса нейтронов. — М.: Атомиздат, 1971.
Марчук Г. И., Султангазин У. М. О сходимости метода расщепления уравнений переноса излучений // ДАН СССР. — 1965. — Т. 161, № 1.
Марчук Г. И., Султангазин У. М. К вопросу о решении кинетического

уравнения переноса методом расщеплений // ДАН СССР. — 1965. — Т. 163, № 4.

Марчук Г. И., Яненко Н. Н. Решение многомерного кинетического уравнения методом расщепления // ДАН СССР. — 1964. — Т. 157, № 6.

Николайшвили Ш. С. Приближенное решение уравнения переноса методом моментов // Атомная энергия. — 1961. — Т. 9, № 2.

Николайшвили Ш. С. О решении односкоростного уравнения переноса с использованием приближения Ивона — Мартенса // Атомная энергия. — 1966. — Т. 20, № 4.

Смелов В. В. Лекции по теории переноса нейтронов. — М.: Атомиздат, 1972.

Султангазин У. М. Дифференциальные свойства решений смешанной задачи Коши для нестационарного кинетического уравнения. — Новосибирск: Препринт СО АН СССР, 1971.

Султангазин У. М. К обоснованию метода слабой аппроксимации для уравнения сферических гармоник. — Новосибирск: Препринт СО АН СССР, 1971.

Султангазин У. М. Слабая сходимость сферических гармоник.— Новосибирск: Препринт СО АН СССР, 1971.

Шихов С. Б. Некоторые вопросы математической теории критического состояния реактора // ЖВМ и МФ. — 1967. — Т. 7, № 1.

Шутяев В. П. Спектр разностной задачи для уравнения переноса в плоском слое и его асимптотическое поведение. — М.: Препринт № 38 ОВМ АН СССР, 1982.

[18] **Метод Монте-Карло**

Бахвалов Н. С. Об оптимальности оценок скорости сходимости квадратурных процессов и методов интегрирования типа Монте-Карло на классах функций // Численные методы решения дифференциальных и интегральных уравнений и квадратурные формулы. — М.: Наука, 1964. — С. 5—63.

Бусленко Н. П., Голенко Д. И. и др. Метод статистических испытаний (метод Монте-Карло). — М.: Физматгиз, 1962.

Владимиров В. С. О применении метода Монте-Карло для отыскания наименьшего характеристического числа и соответствующей собственной функции линейного интегрального оператора // Теория вероятностей и ее применения. — 1956. — Т. 11. — С. 113—130.

Владимиров В. С., Соболь И. М. Расчет наименьшего характеристического числа уравнения Пайерлса методом Монте-Карло // Вычислительная

математика. — М.: ВЦ АН СССР, 1958.
Гельфанд И. М., Фролов А. С., Ченцов Н. Н. Вычисление континуальных интегралов методом Монте-Карло // Изв. вузов. Математика. — 1958. — Т. 5.
Ермаков С. М. Метод Монте-Карло и смежные вопросы. — М.: Наука, 1971.
Ермаков С. М., Золотухин В. Г. Полиномиальные приближения и метод Монте-Карло // Теория вероятностей и ее применения. — 1960. — Т. 5, № 4.
Ермаков С. М., Михайлов Г. А. Курс статистического моделирования. — М.; Наука, 1976.
Кертис Д. Методы Монте-Карло для итерации линейных операторов // УМН. — 1957. — Т. XII, № 5.
Марчук Г. И., Михайлов Г. А., Назаралиев М. А. и др. Метод Монте-Карло в атмосферной оптике. — Новосибирск: Наука, 1976.
Метрополис, Улан (Metropolis N., Ulan S.). The Monte-Carlo Method // J. Amer. Stat. Assoc. — 1949. — V. 44, №. 247.
Михайлов Г. А. Некоторые вопросы теории методов Монте-Карло. — Новосибирск: Наука, 1974.
Соболь И. М. Численные методы Монте-Карло. — М.: Наука, 1973.
Спанье Дж., Гелбард З. Метод Монте-Карло и задачи переноса нейтронов. — М.: Атомиздат, 1972.
Фано У., Спенсер Л., Бергер М. Перенос гамма-излучения. — М.: Госатомиздат, 1963.
Ченцов Н. Н. Статистические решающие правила и оптимальные выводы. — М.: Наука, 1972.

[19] **Метод крупных частиц**

Белоцерковский О. М., Давыдов Ю. М. Нестационарный метод «крупных частиц» для газодинамических расчетов // ЖВМ и МФ. — 1971. — Т. 11, № 1.
Ведешкина К. А., Левина З. Ф., Ломнев С. П. и др. Решение задач методом «крупных частиц». — М.: ВЦ АН СССР, 1970.
Дьяченко В. Ф. Об одном новом методе численного решения нестационарных задач газовой динамики с двумя пространственными переменными // ЖВМ и МФ. — 1965. — Т. 5, № 4.
Харлоу (Harlow F.). Численный метод частиц в ячейках для задач гидродинамики // Вычислительные методы в гидродинамике. — М.: Мир, 1967.

Яненко Н. Н., Анучина Н. Н., Петренко В. Е., Шокин Ю. И. О методах расчета задач газовой динамики с большими деформациями. — Новосибирск: ВЦ СО АН СССР, 1970. — Т. 1, № 1.

[20] **Методы оптимизации алгоритмов**

Бабушка И., Соболев Л. С. Оптимизация численных методов. — 1965. — Т. 10, № 2.

Бахвалов Н. С. Об оптимальных методах решения задач // Appl. Math. — 1968. — V. 13, №. 1.

Бахвалов Н. С. Об оценке количества вычислительной работы, необходимой при приближенном решении задач. // Дополнение IV. В кн.: Введение в теорию разностных схем. С. К. Годунов, В. С. Рябенький — М.: Физматгиз, 1962.

Виноградов И.М. К вопросу об оценке тригонометрических сумм // Изв. АН СССР. Сер. матем. — 1965. — Т. 29, № 3.

Дальквист (Dajiquist G.). Convergence and stability in the numerical integration of ordinary differential equations // Math. Scand. — 1956. — V. 4, №. 1.

Колмогоров А. Н. Дискретные автоматы и конечные алгоритмы // Труды IV Всесоюзного математического съезда. Т. I. — М.: Изд-во АН СССР, 1963.

Коробов Н. М. Вычисление кратных интегралов методом оптимальных коэффициентов // Вестник МГУ. Сер. матем. — 1959. — Т. 4.

Моисеев Н. Н. Численные методы, использующие варьирование в пространстве состояний и некоторые вопросы управления большими системами // Тезисы докладов Международного конгресса математиков. — М., 1966.

Моисеев Н. Н., Красовский Н. Н. Теория оптимальных управляемых систем // Изв. АН СССР. Техн. кибернетика. — 1967. — Т. 5.

Мур (Moor R.). Interval analysis. — Prentice-Hall, 1966.

Никел (Nickel K.). Uber die Notwendigkeit einer Fehlerschranken-Arithmetic fur Rechnenautomaten // Numer. Math. — 1966. — V. 9, № 1.

Никел (Nickel K.). Bericht uber neue Kalsruher Ergehnisse bei der Fhelererfassung von numerischen Prozessen // Appl. Math. — 1968. — V. 13, № 2.

Фролов К. К. О связи квадратурных формул и подрешеток решетки целых векторов // ДАН СССР. — 1977. — Т. 232, № 1. — С. 40—43.

Черноусько Ф. Л., Баничук Н. В., Петров В. М. Численные решения вариационных и краевых задач методом локальных вариаций // ЖВМ и

МФ. — 1966. — Т. 6, № 6.

[21] **Численные методы условий оптимизации**

Абади, Карпентье (Abadie J., Carpenter J.). Generalization of the reduced gradient method to the caze of nonlinear constraints // Optimization. — L.: Acad. Press, 1969. — Pp. 37—48.

Булавский В. А., Звягина Р. А., Яковлева М. А. Численные методы линейного программирования. — М.: Наука, 1977.

Булавский В. А., Рубинштейн Г. Ш. О решении задач выпуклого программирования с линейными ограничениями методом последовательного улучшения допустимого вектора // ДАН СССР. — 1963. — Т. 150, № 2. — С. 231—235.

Бут (Boot J. C. G.). Quadratic programming. — Amsterdam, North-Holland, 1964. — Т. 17.

Вулф Ф. Новые методы в нелинейном программировании // Применение математики в экономических исследованиях. Т. 3. — М.: Мысль, 1968. — С. 312—333.

Ганжела И. Ф. Об одном алгоритме спуска с ограничениями // ЖВМ и МФ. — 1970. — Т. 10, № 1. — С. 146—157.

Гасс С. Линейное программирование. Методы и приложения. — М.: Физматгиз, 1961. — 303. С.

Голдстайн (Goldstein A. A.). Convex programming in Hilbert Space // Bull. Amer. Math. Soc. — 1964. — V. 70, № 5. — Pp. 709—710.

Гловински (Glovinski R.). Introduction to fue Approximation of Elliptic Variational Inequalities. — Report 76006, Laboratoire d'Analyse Numerique de l'universite Paris, 1976. — Т. 6.

Гловински, Лионс, Тремольер (Glovinski R., Lions L , Tremolieres R.). Analyse Numerique des Inequations Variationneles. V. 1, 2. — P.: Uunod, 1976. [Рус. пер.: Численное исследование вариационных веществ. — М.: Мир, 1979.]

Данилин Ю. М. Минимизация нелинейных функционалов в задачах с ограничениями // Кибернетика. — 1970. — № 3. — С. 110—117.

Данциг Дж. Линейное программирование, его применение и обобщения. — М.: Прогресс, 1966. — С. 600.

Демьянов В. Ф. К минимизации функций на выпуклых ограниченных множествах // Кибернетика. — 1965. — № 6. — С. 65—74.

Демьянов В. Ф., Рубинов А. М. Приближенные методы решения экстремальных задач. — Л.: ЛГУ, 1968. — С. 180.

Джилберт (Gilbert E. G.). An iterative procedure for computing of a

quadratic form on a convex set // SI AM J. Control. — 966. — V. 4, № 1. — Pp. 61—81.

Дюво, Лионс (Duvaur G., Lions J. L.). Les Inequations en Mecanique et on Physique. — P.: Dunod, 1972. (Euqlistranslation. Grundlehren der Math., Springer-Verlag, 219, 1976.)

Еремин И. И. О методе штрафов в выпуклом программировании // Кибернетика. — 1967. — № 4. — С. 63—67.

Зангвилл У. Нелинейное программирование. Единый подход. — М.: Сов. радио, 1973. — 312. С.

Зойтендейк Г. Методы возможных направлений. — М.: ИЛ, 1963. — 176 С.

Канторович Л. В. Математические методы в организации и планировании производства. — Л.: ЛГУ, 1939.

Каплан А. А. К вопросу о реализации метода возможных направлений // Труды ин-та мат. СО АН СССР — 1972. — Т. 5, № 22. — С. 99—105.

Карманов В. Г. Лекции по математическому программированию. — М.: МГУ, 1971.

Карманов В. Г. Математическое программирование. — М.: Наука, 1975.

Кон (Conn R.). Constained optimization using a differentiable penalty functions // SIAM J. Numer. Anal. — 1973. — V. 10, № 4. — Pp. 760—784.

Кун, Такер (Kuhn H. W., Tucker A. W.). Nonlinear programming // Proceedings of the Second Berkley Symposium on Methematical Statistics an Probability. — Univ. of California Press, 1951. — Pp. 481—493.

Курант (Courant R.). Variational methods for the solution of problems of equilibrium and vibrations // Bull. Amer. Soc. — 1943. — V. 49, № 1. — Pp. 1—23.

Кюнци Г. П., Крелле В. Нелинейное программирование. — М.: Сов. радио, 1965.

Кэрролл (Carroll C. W.). The created response surface technique for optimizing nonlinear restained systems // Operat. Res. — 1961. — V. 9, №. 2.— Pp. 169—184.

Левитин Е. С., Поляк Б. Т. Методы минимизации при наличии ограничений // ЖВМ и МФ. — 1966. — Т. 6, № 5. — С. 787 —823.

Лионс, Темам (Lions J. L., Temam R.). Une methods d’eclatement des operateurs et des contraintes en calcul des variations. — C. R. Acad. Sci. P., 1966. — V. 263.

Лионс, Стампакья (Lions J. L., Stampacchia). Variational inequalities // Con. Pure Applied Math. — 1967. — V. XX.

Лутсма (Lootsma F. A.). Constrained optimization via penalty functions // Philips Res. Repts. — 1968. — V. 23, № 5. — Pp. 408—423.

Моисеев Н. Н., Иванилов Ю. П., Столярова Е. М. Методы оптимизации. — М.: Наука, 1978.

Морозов В. А. Линейные и нелинейные некорректные задачи // Мат. анализ. — М. 1973. — Т. 11.— С. 129—178.

Пауэлл (Pouell M. J. D.). A method for nonlinear constraints in minimization problems // Optimization— L.: Acad. Press, 1969.

Поляк Б. Т. Метод сопряженных градиентов в задачах на экстремум // ЖВМ и МФ. — 1969. — Т. 9, № 4. — С. 807—921.

Пшеничный Б. Н. Алгоритмы для общей задачи математического программирования // Кибернетика. — 1970. — № 5. — С. 120—125.

Пшеничный Б. Н., Данилин Ю. М. Численные методы в экстремальных задачах. — М.: Наука, 1975.

Розен (Rosen J. B.). The gradien projection method for nonlinear programming. I. Linear constraints. II. Nonlinear constraints // SIAM J. — 1960. — V. 8, №. 1. — Pp. 180—217; 1961. — V. 9, №. 4.— Pp. 514—532.

Розенблюм (Rosenbloom P.). The method of steepset descent // Proc. sympos. Appl. Math. — 1956. — V. 6. — Pp. 127—177.

Рубинштейн Г. Ш. Конечномерные модели оптимизации: курс лекций. — Новосибирск: Изд-во НГУ, 1970.

Саджент, Муртаг (Sargent R. W., Murtagh B. A. H.). Progestion methods for nonlinear programming // Math. Prog. — 1973. — V. 4. — Pp. 245—268.

Сеа Ж. Оптимизация, теория и алгоритмы. — М.: Мир, 1973.

Тихонов А. Н. О некорректных задачах оптимального планирования // ЖВМ и МФ. — 1966. — Т. 6, № 1. — С. 81—89.

Флетчер (Fletcher R.). A general quadratic programming algorithm // J. Inst. Maths Applies. — 1971. — V. 7. — Pp. 76—91.

Франк, Вульф (Frank M., Wolte R.). An algorithm for quadratic programming // Naval. Res. Legist. Quart. — 1956. — V. 3, №. 1.-2. — Pp. 95—110.

Хедли Д. Нелинейное и динамическое программирование. — М.: Мир, 1967.

Эрроу К. Дж., Гурвиц Л., Удзава Х. Исследования по линейному и нелинейному программированию. — М.: ИЛ, 1962.

Юдин Д. Б., Гольдштейн Е. Г. Линейное программирование. Теория и конечные методы. — М.: Физматгиз, 1963.

[22] **Теория оптимального управления (динамическое программирование и принцип максимума)**

Балакришнан А. Введение в теорию оптимизации в гильбертовом пространстве. — М., 1974.

Беллман Р. Динамическое программирование. — М.: ИЛ, 1963.
Беллман Р., Дрейфус С. Прикладные задачи динамического программирования. — М.: Наука, 1965.
Болтянский В. Г. Математические методы оптимального управления. — М.: Наука, 1969.
Болтянский В. Г. Оптимальное управление дискретными системами. — М.: Наука, 1973.
Бутковский А. Г. Методы управления системами с распределенными параметрами. — М.: Наука, 1975.
Габасов Р., Кириллова Ф. М. К вопросу о распространении принципа максимума Л. С. Понтрягина на дискретные системы // Автоматика и телемеханика. — 1966. — Т. 27, №11.
Красовский Н. Н. Теория управления движением. — М.: Наука, 1968.
Летов А. М. Динамика полета и управление. — М.: Наука, 1969.
Ли Э. Б., Маркус Л. Основы теории оптимального управления. — М.: Наука, 1972.
Лионс Ж.-Л. Оптимальное управление системами, описываемыми уравнениями с частными производными. — М.: Мир, 1972.
Лурье К. А. Оптимальное управление в задачах математической физики. — М.: Наука, 1977.
Моисеев Н. Н. Элементы теории оптимальных систем. — М.: Наука, 1975.
Понтрягин Л. С., Болтянский В. Г., Гамкрелидзе Р. В., Мищенко Е. Ф. Математическая теория оптимальных процессов. — М.: Наука, 1976.
Федоренко Р. П. Приближенное решение задач оптимального управления. — М.: Наука, 1978.
Фельдбаум А. А. Основы теории оптимальных автоматических систем. — М.: Наука, 1966.
Черноусько Ф. Л., Баничук В. П. Вариационные задачи механики и управления. — М.: Наука, 1973.
Энеев Т.М. О применении градиентного метода в задачах оптимального управления // Космические исследования. — 1966. — Т. 4, № 5.
Янг Л. Лекции по вариационному исчислению и теории оптимального управления. — М.: Мир, 1974.

[23] **Методы Шварца и разделения области**

Агошков В. И. Метод разделения области в задачах гидродинамики. I. Задача о плоской циркуляции в океане. — М. Препринт ОВМ АН СССР., 1985. — № 96.

Агошков В. И. Метод разделения области в задачах математической физики // Сопряженные уравнения и алгоритмы возмущений в задачах математической физики. — М.: ОВМ АН СССР, 1989.

Агошков В. И. (Agoshkov V. I.). Poincare-Stekiow's Operators and Domain Decomposition Methods in Finite Dimentional Spaces // First International Symposium on Domain Decomposition Methods for Partial Differential Equations. — SIAM, Philadelphia, USA, 1988.

Агошков (Agoshkov V. I.). Reflection operators and domain decomposition methods in transport theory problems // Sov. J. Numer. Anal. Math. Modelling. — 1987. — V. 2, №. 5. — Pp. 325—347.

Агошков В. И., Лебедев В. И. Операторы Пуанкаре — Стеклова и методы разделения области в вариационных задачах // Вычислительные процессы и системы. Т. 2. — М.: Наука, 1985.

Бремблл Дж., Пасьяк Дж., Шатц А. (Bramble J. H., Pasciak J. E., Schatz A. H.). The construction of preconditioners for elliptic problems by substructuring // J. Math. of Comp. — 1986. — V. 47. — Pp. 103—134.

Булеев С. Н., Агошков В. П. Исследование некоторых алгоритмов разделения области // Сопряженные уравнения и алгоритмы возмущений в задачах математической физики. — М.: ОВМ АН СССР, 1989.

Видлунд (Widlund O. B.). Iterative substructuring methods: algorithms and theory for elliptic problems in the plane // Domainn Decomposition Methods for Partial Differential Equations. — SIAM, Philadelphia, USA, 1988. — Pp. 113—128.

Волков Е. А. Асимптотически быстрый приближенный метод нахождения на сеточных отрезках решения разностного уравнения Лапласа. // Труды МИАН СССР. — 1986. — Т. 173. — С. 69—89.

Гловински (Glowinski R.). Domain decomposition methods for nonlinear problems in fluid dynamics // Rapports de Recherche, INRIA. — Paris, 1982.

Дмитриенко М. Е., Оганесян Л. А. Вариант метода Шварца для прилегающих сеточных областей // Вычисления с разреженными матрицами. — Новосибирск, 1981. — С. 36—44.

Дрыя (Dryja M.). A finite element-capacitance matrix method for the elliptic problem // SIAM. J. on Num. Anal. — 1983. — V. 20. — 671—680.

Дьяконов Е. Г. Асимптотическая минимизация вычислительной работы при решении сильно эллиптических краевых задач // Теория кубатурных формул и вычислительная математика. — Новосибирск: Наука, 1980.

Кацнельсон В. Э., Меньшиков В. В. Об одном аналоге альтернирующего метода Шварца // Теория функций, функциональный анализ и их приложения. — Харьков: ХГУ, 1973. — Вып. 17. — С. 206—215.

Кузнецов Ю. А. Вычислительные методы в подпространствах // Вычис-

лительные процессы и системы. Т. 2. — М.: Наука, 1985. — С. 265—350.
Кузнецов Ю. А. Новые алгоритмы приближенной реализации неявных разностных схем. — М.: Препринт ОВМ АН СССР, 1986. — № 142.
Лебедев В. И. Метод композиции. — М.: ОВМ АН СССР, 1986.
Лебедев В. И., Агошков В. И. Обобщенный алгоритм Щварца с переменными параметрами. — М.: Препринт ОВМ АН СССР, 1981.
Лебедев В. И., Агошков В. И. Операторы Пуанкаре — Стеклова и их приложения в анализе. — М.: ОВМ АН СССР, 1983.
Лионс (Lions P. L.). On the Schwarz Alternating Method // Proceedings of the First International Symposium on Domain Decomposition Methods for Partial Differential Equations. — SIAM, Philadelphia, 1988. — Pp. 1—42.
Марчук Г. И., Кузнецов Ю. А. (Marchuk G. I., Kuznetsov Yu. A.). Approximate algorithms for implicit difference schemes // Analyse mathematique et applications. — Gauthier-Villars, Paris, 1988. — Pp. 357—371.
Марчук Г. И., Кузнецов Ю. А., Мацокин А. М. (Marchuk G. I., Kuznetsov Yu. A., Matsokin A. M.). Fictitious domain and domain decomposition methods // Soviet J. of Num. Anal. and Math. Modelling. — 1986. — V. I, № 1. — Pp. 5—41.
Матеева Э. И., Пальцев Б. В. О разделении областей при решении краевых задач для уравнения Пуассона в областях сложной формы // ЖВМ и МФ. — 1973. — Т. 13, № 6. — С. 1441—1452.
Мацокин А. М. Критерий сходимости метода Шварца в гильбертовом пространстве // Вычислительные процессы и системы. Вып. 6. — М.: Наука, 1988.
Мацокин А. М. Метод фиктивных компонент и модифицированный разностный аналог метода Шварца // Вычислительные методы линейной алгебры. — Новосибирск: ВЦ СО АН СССР, 1980.
Мацокин А. М. Методы фиктивных компонент и альтернирования по подпространствам // Вычислительные алгоритмы в задачах математической физики. — Новосибирск: ВЦ СО АН СССР, 1985.
Мацокин А. М. Связь метода окаймления с методом фиктивных компонент и методом альтернирования по подпространствам // Дифференциальные уравнения с частными производными. — Новосибирск: Наука, 1986.
Мацокин А. М., Непомнящих С. В. Метод альтернирования Шварца в подпространстве // Известия вузов. Математика. — 1985. — № 10.
Михлин С. Г. Об алгоритме Шварца // Докл. АН СССР. — 1951. — Т. 77, № 4. — С. 569—571.
Осмоловский В. Г., Ривкинд В. Я. О методе разделения областей для

эллиптических уравнений с разрывными коэффициентами // ЖВМ и МФ. — 1981. — Т. 21, № 1. — С. 35—39.

Романова С. Е. Приближенные методы решения разностного уравнения Лапласа асимптотически за одно и два сложения на точку // ДАН СССР. — 1983. — Т. 273, № 1. — С. 49-54.

Самарский А. А., Капорин И. Е., Кучеров А. В., Николаев Е. С. Некоторые современные методы решения сеточных уравнений // Известия вузов. Математика. — 1983. — № 7. — С. 3—12.

Смелов В. В. Обоснование итерационного процесса по подобластям для задач теории переноса в нечетном P_{2+1} приближении. Новосибирск: Препринт ВЦ СО АН СССР, 1980. — № 71. — 27 с.

Смелов В. В. Принцип итерирования по подобластям в задачах с уравнением переноса // Методы решения систем вариационно-разностных уравнений. — Новосибирск, 1979. — С. 139—158.

Соболев С. Л. Алгоритм Шварца в теории упругости // ДАН СССР. — 1936. — Т. 4 (XIII), № 6. — С. 235—238.

Фунаро Д., Квартерони А., Занолли П. (Funaro D., Quarteroni A., Zanolli P.). An iterative procedure with interface relaxation for domain decomposition methods // SIAM. J. Numer. Anal. — 1988. — V. 25. — P. 69.

Цвик Л. Б. Обобщение алгоритма Шварца на случай областей, сопряженных без налегания // ДАН СССР. — 1975. — Т. 224, № 2. — С. 309—312.

[24] **Сопряженные уравнения и алгоритмы возмущений**

Агошков В. И. Оценка скорости сходимости некоторых алгоритмов теории возмущений.— М.: Препринт ОВМ АН СССР, 1982. — № 30.

Агошков В. И. Проекционно-сеточный метод в алгоритмах теории возмущений.— М.: Препринт ОВМ АН СССР, 1982.— № 38.

Агошков В. И. Сопряженные уравнения в алгоритмах возмущений N-го порядка точности // Сопряженные уравнения и теория возмущений в задачах математической физики. — М.: ОВМ АН СССР, 1985.

Агошков В. И., Попыкин А. П., Шихов С. Б. К теории малых возмущений для уравнения переноса // Сопряженные уравнения и теория возмущений в задачах математической физики. — М.: ОВМ АН СССР, 1985.

Беллман (Bellman B.). Perturbation Techniques in Mathematics, Physics and Engineering. — Holt, N. Y., 1964.

Боголюбов Н. Н., Митропольский Ю. А. Асимптотические методы в теории нелинейных колебаний. — М.: Физматгиз, 1958.

Вайнберг М. М., Треногин В. А. Теория ветвления решений нелинейных уравнений. — М.: Наука, 1969.

Ван-Дейк (Van-Dyke M. D). Perturbation methodc in fluid mechanics. — Academic Press, 1964. — №. 1.

Васильева А. Б., Бутузов В. Ф. Асимптотические разложения решений сингулярно возмущенных уравнений. — М.: Наука, 1973.

Вишик М. И., Люстерник Л. А. Решение некоторых задач о возмущении в случае матриц и самосопряженных и несамосопряженных дифференциальных уравнений. I // УМН. — 1960. — Т. XV, вып. 3. — С. 3—80.

Владимиров В. С., Волович И. В. Законы сохранения для нелинейных уравнений // ДАН СССР. — 1984. — Т. 279, № 4. — С. 843—847.

Като Т. Теория возмущений линейных операторов. — М.: Мир, 1972.

Като (Kato T.). Perturbation of continuous spectra by trace class operators // Proc. Japan Acad. — 1957. — V. 33. — Pp. 260—264.

Ладыженская О. А., Фаддеев Л. Д. О теории возмущения непрерывного спектра // ДАН СССР. — 1958. — Т. 120, № 6. — С. 1187—1190.

Ломов С. А. Введение в общую теорию сингулярных возмущений. — М.: Наука, 1981.

Льюинс Дж. Ценность. Сопряженная функция. — М.: Атомиздат, 1972.

Ляпунов А. М. Собр. соч. Т. 2. — М.; Л., 1956.

Марчук Г. И. Методы долгосрочного прогноза погоды на основе решения основных и сопряженных задач // Метеорология и гидрология. — 1974. — № 3. — С. 17—34.

Марчук Г. И. Применение сопряженных уравнений к решению задач математической физики // Успехи механики. — 1981. — Т. 4, вып. 1. — С. 3—27.

Марчук Г. И. Основные и сопряженные уравнения динамики атмосферы и океана // Метеорология и гидрология. — 1974. — № 2. — С. 9—37.

Марчук Г. И. Окружающая среда и проблема оптимизации размещения предприятий // ДАН СССР. — 1976. — Т. 227, № 5. — С. 1056—1059.

Марчук (Marchuk G. I.). Formulation of the theory of Perturbations for Complicated Models // Applied Math. and Optimization. — 1975. — V. 2, №. 1. — Pp. 1—33.

Марчук Г. И., Агошков В. И. Сопряженные операторы и алгоритмы возмущений в нелинейных задачах. Принципы построения сопряженных операторов. — М.: Препринт ОВМ АН СССР, 1986.

Марчук Г. И., Агошков В. И. Сопряженные операторы и алгоритмы возмущений в нелинейных задачах. Алгоритмы возмущений. — М.: Препринт ОВМ АН СССР, 1986.

Марчук Г. И., Агошков В. И. Симметризация нестационарного уравнения переноса и формулировка вариационного принципа. — Новосибирск: Препринт ОВМ АН СССР, 1980.

Марчук Г. И., Агошков В. И., Шутяев В. П. Сопряженные уравнения и алгоритмы возмущений. — М.: ОВМ АН СССР, 1986

Марчук Г. И., Кузин В. И., Скиба Ю. Н. Проекционно-разностный метод расчета сопряженных функций для модели переноса тепла в системе атмосфера - океан - почва // Актуальные проблемы вычислительной и прикладной математики. — Новосибирск, 1983. — С. 149—154.

Марчук Г. И., Орлов В. В. К теории сопряженных функций // Нейтронная физика. — М.: Госатомиздат, 1961. — С. 30—45.

Маслов В. П. Теория возмущений и асимптотические методы. — М.: Изд-во МГУ, 1965.

Михайлов Г. А. Использование приближенных решений сопряженной задачи для улучшения алгоритмов метода Монте-Карло // ЖВМ и МФ. — 1969. — Т. 9, № 5. — С. 1145—1152.

Моисеев Н. Н. Асимптотические методы нелинейной механики. — М.: Наука, 1981.

Найфэ А. Методы возмущений. — М.: Мир, 1976.

Пуанкаре А. Собр. соч. Т. I. — М.: Наука, 1971.

Пупко В. Я., Зродников А. В., Лихачев Ю. И. Метод сопряженных функций в инженерно - физических исследованиях. — М.: Энергоатомиздат, 1984.

Реллих (Rellich F.). Perturbation theory of eigenvalue problems. — N. Y.; L.; P.: Gordon and breach Science Publishers, 1969.

Реллих (Rellich P.). Storungstheorie der spektralzerlegung. I—V // Math. Ann. — 1936. — V. 113. — Pp. 600—619, 667—685; 1939. — V. 116, — Pp. 555—570; 1940. — V. 117. — Pp. 346—382; 1942. — V. 118. — Pp. 462—484.

Стумбур Э. А. Применение теории возмущений в физике ядерных реакторов. — М.: Атомиздат, 1976.

Теория ветвления и нелинейные задачи на собственные значения / под ред. Д. Б. Келлера, С. Актмана. — М.: Мир, 1974.

Усачев Л. Н. Уравнение для ценности нейтронов кинетического реактора и теория возмущений // Реакторостроение и теория реакторов. — М., Изд-во АН СССР, 1955. — 251 С.

Фаддеев Л. Д. О модели Фридрихса в теории возмущений непрерывного спектра. // Труды МИАН. — 1964. — Т. 73.

Фридрихс К. Возмущение спектра операторов в гильбертовом пространстве. — М.: Мир, 1969.

Шредингер (Schrodinger E.). Quantisierung als Eigenwertproblem // Ann. Phys. — 1926. — V. 80. — Pp. 437—490.

Шутяев В. П. Вопросы теории возмущений для решения задач переноса нейтронов: автореф. дис. ... канд. физ.-мат. наук. — М., 1983.

Шутяев В. П. Спектральные свойства условно-критической задачи переноса в дискретном приближении и алгоритмы теории возмущений // Сопряженные уравнения и теория возмущений в задачах математической физики. — М.: ОВМ АН СССР, 1985.

[25] **Вычислительные тензорные методы**

Баллани, Граседик, Клюге (Ballani J., Grasedyck L., Kluge M.). Black box approximation of tensors in hierarchical Tucker format // Linear Alg. Appl. — 2013. — V. 428. — Pp. 639—657.

Бейлкин, Моленкамп (Beylkin G., Mohlenkamp M. J.). Algorithms for numerical analysis in high dimensions // SIAM J. Sci. Comput. — 2005. — V. 26, №. 6. — Pp. 2133—2159.

Бейлкин, Моленкамп (Beylkin G., Mohlenkamp M. J.). Numerical operator calculus in higher dimensions // Proc. Nat. Acad. Sci. USA. — 2002. — V. 99, №. 16. — Pp. 10246—10251.

Бини (Bini D.). Relations between exact and approximate bilinear algorithms. Applications // Calcolo. — 1980. — V. 17, №. 1. — P. 87—97.

Бро (Bro R.). Multi-way analysis in the food industry: models, algorithms and applications: Ph. D. thesis // KøbenhavnsUniversitetKøbenhavns Universitet, Det Biovidenskabelige Fakultet for Fødevarer, VeteFaculty of Life Sciences, Institut for FødevarevidenskabDepartment of Food Science, Kvalitet og TeknologiQuality & Technology. — 1998.

Бро (Bro R.). PARAFAC: Tutorial and applications // Chemometrics and Intelligent Lab. Syst. — 1997. — V. 38, №. 2. — Pp. 149—171.

Гаврилюк, Хакбуш, Хоромский (Gavrilyuk I. P., Hackbush W., Khoromskij B. N.). Tensor-product approximation to the inverse and related operators in high-dimensional elliptic problems // Computing. — 2005. — №. 74. — Pp. 131—157.

Граседик (Grasedyck L.). Existence and computation of low Kronecker-rank approximations for large systems in tensor product structure // Computing. — 2004. — V. 72. — Pp. 247—265.

Граседик (Grasedyck L.). Hierarchical singular value decomposition of tensors // SIAM J. Matrix Anal. Appl. — 2010. — V. 31, №. 4. — Pp. 2029—2054.

Граседик (Grasedyck L.). Polynomial approximation in hierarchical Tucker format by vector-tensorization: DFG-SPP1324 Preprint 43. — Marburg: Philipps-Univ., 2010. — Availableat: http://www.dfg-spp1324.de/download/preprints/preprint043.pdf.

Граседик, Кресснер, Тоблер (Grasedyck L., Kressner D., Tobler C.). A

literature survey of low-rank tensor approximation techniques // GAMM-Mitteilungen. — 2013. — V. 36, №. 1. — Pp. 53—78.

Де Латаувер, де Мур, Вандевалле (de Lathauwer L., de Moor B., Vandewalle J.). A multilinear singular value decomposition // SIAM J. Matrix Anal. Appl. — 2000. — V. 21. — Pp. 1253—1278.

Де Латаувер, де Мур, Вандевалле (de Lathauwer L., de Moor B., Vandewalle J.). Computing of canonical decomposition by means of a simultaneous generalized Schur decomposition // SIAM J. Matrix Anal. Appl. — 2004. — V. 26. — Pp. 296—237.

Де Латаувер, де Мур, Вандевалле (de Lathauwer L., de Moor B., Vandewalle J.). On best rank-1 and rank-(R_1, R_2, ..., R_N) approximation of high-order tensors // SIAM J. Matrix Anal. Appl. — 2000. — V. 21. — Pp. 1324—1342.

Джекельман (Jeckelmann E.). Dynamical density-matrix renormalization-group method // Phys. Rev. B. — 2002. — V. 66. — Pp. 045114.

Долгов, Оселедец (Dolgov S. V., Oseledets I. V.). Solution of linear systems and matrix inversion in the TT-format // SIAM J. Sci. Comput. — 2012. — V. 34, №. 5. — Pp. A2718—A2739.

Долгов, Савостьянов (Dolgov S. V., Savostyanov D. V.). Alternating minimal energy methods for linear systems in higher dimensions. Part I: SPD systems: arXiv preprint 1301.6068, 2013. — Availableat: http://arxiv.org/abs/1301.6068.

Долгов, Савостьянов (Dolgov S. V., Savostyanov D. V.). Alternating minimal energy methods for linear systems in higher dimensions. Part II: Raster algorithm and application to nonsymmetric systems: arXiv Preprint 1304.1222, 2013. — Availableat: http://arxiv.org/abs/1304.1222.

Казеев, Хоромский (Kazeev V. A., Khoromskij B. N.). Low-rank explicit QTT representation of the Laplace operator and its inverse // SIAM J. Matrix Anal. Appl. — 2012. — V. 33, №. 3. — Pp. 742—758.

Каттелл (Cattell R. B.). «Parallell proportional profiles» and other principles for determining the choice of factors by rotation // Psychometrika. — 1944. — V. 9, №. 4. — Pp. 267—283.

Кольда, Бадер (Kolda T. G., Bader B. W.). Tensor decompositions and applications // SIAM Review. — 2009. — V. 51, №. 3. — Pp. 455—500.

Кресснер, Тоблер (Kressner D., Tobler C.). Preconditioned low-rank methods for high-dimensional elliptic PDE eigenvalue problems // Computational Methods in Applied Mathematics. — 2011. — Vol. 11, no. 3. — P. 363—381.

Кох, Любих (Koch O., Lubich C.). Dynamical low rank approximation // SIAM J. Matrix Anal. Appl. — 2007. — V. 29, №. 2. — Pp. 434—454.

Кох, Любих (Koch O., Lubich C.). Dynamical tensor approximation // SIAM

J. Matrix Anal. Appl. — 2010. — V. 31, №. 5. — Pp. 2360—2375.
Кэролл, Чанг (Caroll J. D., Chang J. J.). Analysis of individual differences in multidimensional scaling via n-way generalization of Eckart-Young decomposition // Psychometrika. — 1970. — V. 35. — Pp. 283—319.
Любих, Оселедец (Lubich C., Oseledets I. V.). A projector-splitting integrator for dynamical low-rank approximation // BIT. — 2014. — V. 54, №. 1. — Pp. 171—188.
Любих, Оселедец, Вандерейкен (Lubich C., Oseledets I. M., Vandereycken B.). Time integration of tensor trains: arXiv preprint 1407.2042, 2014. — Availableat: http://arxiv.org/abs/1407.2042.
Люгранс, Росс (Leurgans S., Ross R. T.). Multilinear models: applications in spectroscopy // Statistical Science. — 1992. — V. 7, №. 3. — Pp. 289—310.
Манте (Manthe U.). A multilayer multiconfigurational time-dependent Hartree approach for quantum dynamics on general potential energy surfaces. // J. Chem. Phys. — 2008. — V. 128, №. 16. — P. 164116.
Оселедец И. В. О новом тензорном разложении // ДАН. — 2009. — Т. 427, № 2. — С. 168—169.
Оселедец И. В. О приближении матриц логарифмическим числом параметров // ДАН. — 2009. — Т. 428, № 1. — С. 23—24.
Оселедец И. В., Тыртышников Е.Е. Рекурсивное приближение многомерных тензоров // ДАН. — 2009. — Т. 427, № 1. — С. 14—16.
Оселедец (Oseledets I. V.). Approximation of $2^d \times 2^d$ matrices using tensor decomposition // SIAM J. Matrix Anal. Appl. — 2010. — V. 31, №. 4. — Pp. 2130—2145.
Оселедец (Oseledets I. V.). Constructive representation of functions in low-rank tensor formats // Constr. Appr. — 2013. — V. 37, №. 1. — Pp. 1—18. — Availableat: http://pub.inm.ras.ru/pub/inmras2010-04.pdf.
Оселедец (Oseledets I. V.). DMRG approach to fast linear algebra in the TT-format // Comput. Meth. Appl. Math. — 2011. — V. 11, №. 3. — Pp. 382—393.
Оселедец (Oseledets I. V.). Tensor-train decomposition // SIAM J. Sci. Comput. — 2011. — V. 33, №. 5 — Pp. 2295—2317.
Оселедец, Савостьянов, Тыртышников (Oseledets I. V., Savostyanov D. V., Tyrtyshnikov E. E.). Fast simultaneous orthogonal reduction to triangular matrices // SIAM J. Matrix Anal. Appl. — 2009. — V. 31, №. 2. — Pp. 316—330.
Оселедец, Савостьянов, Тыртышников (Oseledets I. V., Savostianov D. V., Tyrtyshnikov E. E.). Tucker dimensionality reduction of three-dimensional arrays in linear time // SIAM J. Matrix Anal. Appl. — 2008. — V. 30, №. 3. — Pp. 939—956.
Оселедец, Тыртышников (Oseledets I. V., Tyrtyshnikov E. E.). Algebraic

wavelet transform via quantics tensor train decomposition // SIAM J. Sci. Comput. — 2011. — V. 33, №. 3. — Pp. 1315—1328.

Оселедец, Тыртышников (Oseledets I. V., Tyrtyshnikov E. E.). Breaking the curse of dimensionality, or how to use SVD in many dimensions // SIAM J. Sci. Comput. — 2009. — V. 31, №. 5. — Pp. 3755—3759.

Оселедец, Тыртышников (Oseledets I. V., Tyrtyshnikov E. E.). TT-cross approximation for multidimensional arrays // Linear Algebra Appl. — 2010. — V. 432, №. 1. – Pp. 70—88.

Остлунд, Роммер (Östlund S., Rommer S.). Thermodynamic limit of Density Matrix Renormalization // Phys. Rev. Lett. — 1995. — V. 75, №. 19. — Pp. 3537—3540.

Райт (White S. R.). Density matrix formulation for quantum renormalization groups // Phys. Rev. Lett. — 1992. — V. 69, №. 19. — Pp. 2863—2866.

Савостьянов (Savostyanov D. V.). Quasioptimality of maximum-volume cross interpolation of tensors // Linear Algebra Appl. — 2014. — V. 458. — Pp. 217—244.

Савостьянов, Оселедец (Savostyanov D. V., Oseledets I. V.). Fast adabtive interpolation of multi-dimensional arrays in tensor train format // Proceedings of 7^{th} International Workshop on Multidimensional Systems (nDS). — IEEE, 2011.

Тыртышников Е. Е. Тензорные аппроксимации матриц, порожденных асимптотически гладкими функциями // Матем. сб. — 2003. — Т. 194, № 5. — С. 147—160.

Таккер (Tucker L. R.). Some mathematical notes on three-mode factor analysis // Psychometrika. — 1966. — Vol. 31. — P. 279—311.

Ушмаев (Uschmajew A.). Local convergence of the alternating least squares algorithm for canonical tensor approximation // SIAM J. Matr. Anal. Appl. — 2012. — V. 33, №. 2. — Pp. 639—652.

Фаннес, Нахтергаеле, Вернер (Fannes M., Nachtergaele B., Werner R. F.). Finitely correlated states on quantum spin chains // Comm. Math. Phys. — 1992. — V. 144, №. 3. — Pp. 443—490.

Фаннес, Нахтергаеле, Вернер Fannes M., Nachtergaele B., Werner R. F.). Ground states of VBS models on Cayley trees // J. Stat. Phys. — 1992. — V. 66. — Pp. 939—973. — Availableat: http://dx.doi.org/10.1007/BF01055710.

Хакбуш (Hackbusch W.). Tensor spaces and numerical tensor calculus. — Springer-Verlag, Berlin, 2012.

Хакбуш, Хоромский (Hackbusch W., Khoromskij B. N.). Low-rank Kronecker-product approximation to multi-dimensional nonlocal operators. I. Separable approximation of multi-variate functions // Computing. — 2006. — Vo. 76, №. 3-4. — Pp. 177—202.

Хакбуш, Хоромский (Hackbusch W., Khoromskij B. N.). Low-rank Kronecker-product approximation to multi-dimensional nonlocal operators. II. HKT representation of certain operators // Computing. — 2006. — V. 76, №. 3-4. — Pp. 203—225.

Хакбуш, Хоромский, Тыртышников (Hackbusch W., Khoromskij B. N., Tyrtyshnikov E. E.). Hierarchical Kronecker tensor-product approximations // J. Numer. Math. — 2005. — V. 13. — Pp. 119—156.

Хакбуш, Кюн (Hackbusch W., Kühn S.). A new scheme for the tensor representation // J/ Fourier Anal. Appl. — 2009. — Vol. 15, №. 5. — Pp. 706—722.

Харшман (Harshman R. A.). Foundations of the PARAFAC procedure: models and conditions for an explanatory multimodal factor analysis // UCLA Working Papers in Phonetics. — 1970. — V. 16. — Pp. 1—84.

Хичкок (Hitchcock F. L.). Multiple invariants and generalized rank of a p-way matrix or tensor // J. Math. Phys. — 1927. — V. 7, №. 1. — Pp. 39—79.

Хичкок (Hitchcock F .L.). The expression of a tensor or a polyadic as a sum of products // J. Math. Phys. — 1927. — V. 6, №. 1. — Pp. 164—189.

Хольц, Роведдер, Шнайдер (Holtz S., Rohwedder T., Schneider R.). On manifolds of tensors of fixed TT-rank // Numer. Math. — 2012. — V. 120, №. 4. — Pp. 701—731.

Хольц, Роведдер, Шнайдер (Holtz S., Rohwedder T., Schneider R.). The alternating linear scheme for tensor optimization in the tensor train format // SIAM J. Sci. Comput. — 2012. — V. 34, №. 2. — Pp. A683—A713.

Хоромский (Khoromskij B. N.). $\mathcal{O}(d \log N)$–Quantics approximation of N–d tensors in high-dimensional numerical modeling // Constr. Appr. — 2011. — V. 34, №. 2. — Pp. 257—280.

Хоромский, Хоромская (Khoromskij B. N., Khoromskaia V.). Low rank Tucker-type tensor approximation to classical potentials // Central European journal of mathematics. — 2007. — V. 5, №. 3. — Pp. 523—550.

Хоромский, Оселедец (Khoromskij B. N., Oseledets I. V.). DMRG+QTT approach to computation of the ground state for the moleculara Schrödinger operator: Preprint 69. — Leipzig: MPI MIS, 2010. — Availableat: http://www.mis.mpg.de/preprints/2010/preprint2010_69.pdf.

Хоромский, Оселедец (Khoromskij B. N., Oseledets I. V.). QTT-approximation of elliptic solution operators in higher dimensions // Rus. J. Numer. Anal. Math. Model. — 2011. — V. 26, №. 3. — Pp. 303—322.

Хоромский, Оселедец (Khoromskij B. N., Oseledets I. V.). Quantics-TT collocation approximation of parameter-dependentd and stochastic elliptic PDEs // Comput. Meth. Appl. Math. — 2010. — V. 10, №. 4. — Pp. 376—394.

Хоромский (Khoromskij B. N.). Tensor-structured numerical methods in

scientific computing: Survey on recent advances // Chemometr. Intell. Lab. Syst. — 2012. — V. 110, №. 1. — Pp. 1—19.

Шоллвек (Schollwöck U.). The density-matrix renormalization group in the age of matrix product states // Annals of Physics. — 2011. — V. 326, №. 1. — Pp. 96—192.

Штрассен (Strassen V.). Gaussian elimination is not optimal // Numerishe Mathematik. — 1969. — V. 13, №. 4. — Pp. 354—356.

Элден, Савас (Eldén L., Savas B.). A Newton-Grassmann method for computing the best multilinear rank-(r_1, r_2, r_3) approximation of a tensor // SIAM J. Matrix Anal. Appl. — 2009. — V. 31, №. 2. — Pp. 248—271.

Dolgov S. V., Khoromskij B. N., Oseledets I. V., Savostyanov D. V. Computation of extreme eigenvalues in higher dimensions using block tensor train format // Computer Phys. Comm. — 2014. — V. 185, №. 4. — Pp. 1207—1216.

Ishteva M., de Lathauwer L., Absil P. A., van Huffel S. Differential-geometric newton method for the best rank-(r_1, r_2, r_3) approximation of tensors // Numerical Algorithms. — 2009. — V. 51, №. 2. — Pp. 179—194.

Lubich C., Rohwedder T., Schneider R., Vandereycken B. Dynamical approximaiton by Hierarchical Tucker and Tensor-Train Tensors // SIAM J. Matrix. Anal. Appl. — 2013. — V. 34, №. 2. — Pp. 470—494.

Affleck I., Kennedy T., Lieb E. H., Tasaki H. Rigorous results on valence-bond ground states in antiferromagnets // Phys. Rev. Lett. — 1987. — V. 59, №. 7. — Pp. 799—802.

Huegeman J., Cirac J. I., Osborne T. J. et al. Time-dependent variational principle for quantum lattices // Phys. Rev. Lett. — 2011. — V. 107, №. 7. — Pp. 070601.

Haegeman J., Lubich C., Oseledets I. et al. Unifying time evolution and optimization with matrix product states: arXiv preprint 1408.5056, 2014. — Available at: http://arxiv.org/abs1408.5056.

Издатель – Российская академия наук

Публикуется в авторской редакции

Издается по решению Научно-издательского совета
Российской академии наук (НИСО РАН)
и распространяется бесплатно

Оригинал–макет подготовлен в ООО «Амирит»

Подписано в печать 15.05.2018 г.
Формат 70×100 1/16. Гарнитура Times New Roman. Бумага офсетная.
Усл. печ. л. 61,75. Тираж 100 экз. Заказ № 10/15058.

Отпечатано в типографии ООО «Амирит»,
410004, г. Саратов, ул. Чернышевского, 88.
Тел.: 8-800-700-86-33 | (845-2) 24-86-33
E-mail: zakaz@amirit.ru
Сайт: amirit.ru