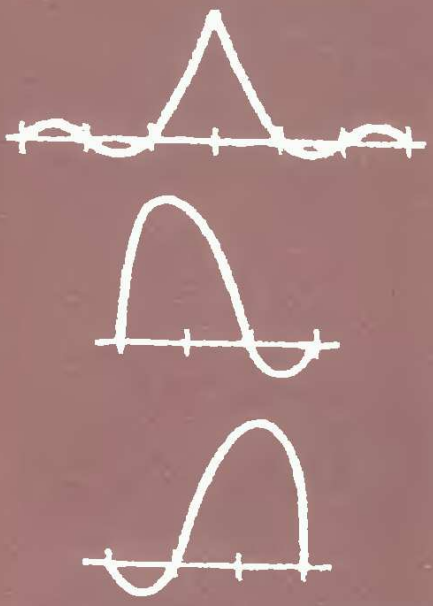
Г. СТРЕНГ
Дж. ФИКС



ТЕОРИЯ МЕТОДА КОНЕЧНЫХ ЭЛЕМЕНТОВ

ИЗДАТЕЛЬСТВО
«МИР»

AN ANALYSIS OF THE FINITE ELEMENT METHOD

Gilbert Strang
Massachusetts Institute of Technology

George J. Fix
University of Maryland

PRENTICE-HALL, INC.
Englewood Cliffs, N. J.
1973

Г. СТРЕНГ,
ДЖ. ФИКС

ТЕОРИЯ МЕТОДА КОНЕЧНЫХ ЭЛЕМЕНТОВ

Перевод с английского
В. И. АГОШКОВА,
В. А. ВАСИЛЕНКО,
В. В. ШАЙДУРОВА

Под редакцией
Г. И. МАРЧУКА

ИЗДАТЕЛЬСТВО „МИР"
МОСКВА 1977

УДК 519.3

Метод конечных элементов получил в последнее время широкое распространение как один из современных и самых эффективных методов решения краевых задач математической физики. В монографии известных американских специалистов излагаются теоретические основы метода конечных элементов — интерполяция данных, выбор аппроксимирующих функций, модификация краевых условий, точность вычислений. Обсуждаются возможности применения в различных областях физики и техники, приводятся простые примеры для иллюстрации теоретических положений.

Книга доступна студентам и аспирантам университетов и втузов. Специалисты по численным методам найдут в ней большой фактический материал по практическому применению метода конечных элементов.

Редакция литературы по математическим наукам

С $\frac{20203-027}{041(01)-77}$ 27 — 77

ОТ РЕДАКТОРА ПЕРЕВОДА

В течение последних десяти лет метод конечных элементов превратился в мощную математическую основу для создания пакетов программ решения задач математической физики, позволяющих полностью автоматизировать процесс построения и решения систем вариационно-разностных уравнений.

Книга, предлагаемая Вашему вниманию, написана двумя известными американскими математиками — Гилбертом Стренгом и Джорджем Фиксом. Их работа — попытка наметить связь между инженерной теорией конечных элементов и математической основой метода. Этим объясняются особенности стиля книги, а также простота примеров, на которых рассматривается метод и иллюстрируются основные идеи. Вместе с тем в книге присутствуют все математические доказательства, необходимые для четкого обоснования оценок сходимости. Круг рассматриваемых задач весьма широк — эллиптические и параболические задачи, задачи на собственные значения. При этом не только обсуждаются теоретические вопросы, но и даются практические вычислительные рекомендации. Книга снабжена обширной библиографией.

Для чтения книги достаточно владеть математическим анализом в рамках программы технического вуза. Все дополнительные математические понятия вводятся непосредстенно в тексте или в конце книги.

Книга будет полезна как специалистам, работающим в области прикладной и вычислительной математики, так и инженерам и студентам вузов, желающим ознакомиться с этим популярным сейчас вычислительным методом решения различных научных и технических проблем — методом конечных элементов. Эта книга будет способствовать развитию и применению современных вычислительных методов.

Г. И. Марчук

Май 1976

ПРЕДИСЛОВИЕ К РУССКОМУ ИЗДАНИЮ

Я очень рад и очень горжусь тем, что эта книга переведена на русский язык. Вы, конечно, знаете и без нашей книги, что метод конечных элементов продолжает свое стремительное развитие не только как «красивая теория», но и как очень практический вычислительный метод решения прикладных задач. Расширяются его применения в строительной механике и гидромеханике, рождаются новые области применений. Вероятно, конечные элементы стали наиболее употребительным средством вычислительной математики во всем мире; это хорошо, но будет еще лучше, если мы научимся решать те же задачи с меньшими затратами.

Я уверен, что метод будет развиваться дальше, и надеюсь, что Вы поможете этому и настоящая книга также. Ее цель — разъяснить основные идеи метода настолько просто и ясно, чтобы инженеры смогли строить и применять конечные элементы, а специалисты по численному анализу увидели вопросы, которые их касаются, и чтобы те и другие нашли общий язык!

Предмет книги очень многим обязан русским математикам и прежде всего Галёркину и Михлину. Идея использования кусочно полиномиальных функций — это новый шаг, но шаг прямо по пути, намеченному их фундаментальными работами. Благодаря этому шагу метод перестал быть чисто теоретическим и сделался чрезвычайно практическим.

Сейчас появились другие книги по конечным элементам, некоторые из них хороши для инженеров, но эта остается для меня самой любимой. Надеюсь, она Вам понравится.

Гилберт Стренг

Октябрь 1975

Джилл и Линде

Cherche la f. e. m. [1])

ПРЕДИСЛОВИЕ

Метод конечных элементов удивительно успешно применяется в самых различных задачах. Он был создан для решения сложных уравнений теории упругости и строительной механики и оказался гораздо эффективнее метода конечных разностей. Сейчас активно разрабатываются и другие применения метода конечных элементов. Этот метод незаменим, если нужно учитывать геометрические особенности областей — тогда ЭВМ используется не только для *решения* системы уравнений, но в первую очередь для *формулирования* и *построения* дискретных аппроксимаций.

С математической точки зрения метод представляет собой обобщение метода Рэлея — Ритца — Галёркина. Поэтому он применим к широкому классу уравнений в частных производных. В методе Ритца, однако, не решается непосредственно дифференциальное уравнение; вместо этого исходная задача представляется в эквивалентной вариационной формулировке, а затем ищется приближенное решение последней в виде комбинации $\sum q_j \varphi_j$ заданных пробных функций $\varphi_j(x)$. При этом весовые коэффициенты q_j вычисляются из вариационного принципа, соответствующего задаче. Это и есть та система дискретных уравнений, которая решается с помощью ЭВМ.

Эта идея очень стара. Новым является лишь выбор пробных функций: в методе конечных элементов они *кусочно полиномиальны*. Именно этим выбором определяется успех метода. Каждая функция φ_j равна нулю на большей части области и отлична от нуля только в окрестности одного узла. В этой окрестности φ_j составлена из полиномов небольшой степени, и все вычисления становятся максимально простыми. Интересно, что преимущества кусочно полиномиальных функций одновременно и совершенно независимо были замечены в математической теории аппроксимации. Идея их применения оказалась весьма плодотворной, и она появилась как раз в нужное время.

Поскольку математическая основа метода построена, можно показать, почему он работает; этому и посвящена наша книга.

---

[1]) f. e. m. — сокращение для метода конечных элементов. — *Прим. перев.*

Ее цель — объяснить влияние различных факторов на вычислительную эффективность метода конечных элементов. Мы перечислим здесь основные факторы:

1) интерполяция физических данных,
2) выбор конечного числа полиномиальных пробных функций,
3) упрощение геометрии области,
4) модификация краевых условий,
5) численное интегрирование функционала в вариационном принципе,
6) ошибки округления при решении дискретной системы.

Эти вопросы в основном являются математическими, и авторы этой книги — математики. Тем не менее *не надо* думать, что эта книга предназначена исключительно для специалистов по численному анализу. Напротив, мы надеемся, что она поможет установлению более тесных связей между инженерами-математиками и математиками-аналитиками. Нам кажется, что метод конечных элементов обеспечивает благоприятную возможность укрепления таких связей: теория весьма привлекательна, приложений становится все больше и, что самое главное, метод еще достаточно молод и разрыв между теорией и практикой не стал еще непреодолимым.

Конечно, мы сознаем, что существуют помехи, которые нельзя устранить полностью. Одна из них — язык изложения: мы свели математические обозначения до минимума и вынесли их (вместе с определениями) в конец книги. Однако мы хорошо понимаем, что даже после интерпретации нормы как естественной меры энергии деформации, а гильбертова пространства как класса допустимых функций, в вариационной задаче физического происхождения остается самое трудное — свыкнуться с этими понятиями, сделать их своими собственными. Здесь наряду с совместными усилиями требуется настойчивость и терпимость с обеих сторон. Возможно, эта книга по меньшей мере выявит такие задачи, которые математик уже умеет решать, и такие, для которых он бесполезен.

В последние несколько лет очень многие специалисты по численному анализу стали заниматься методом конечных элементов, и мы им весьма благодарны. Это подтверждается явно на протяжении всей книги и неявно в библиографии, хотя мы не имели намерения создавать формальную историю метода. Здесь, в предисловии, мы хотим поблагодарить двух коллег — скорее инженеров, чем математиков — за помощь, особенно важную для нас. Первый — Айзек Фрид, влияние которого заставило нас отказаться от публикации законченной рукописи «Fourier Analysis of the Finite Element Method» и заняться этой книгой. Вто-

рой — Брюс Айронс, замечательные интуитивные соображения которого описаны (и доказаны строго, где мы смогли это сделать) в нашей книге.

Глава 1 существенно длиннее всех остальных. Она использовалась первым автором в качестве вводного курса в Массачусетском технологическом институте. В качестве домашнего задания предлагалось запрограммировать некоторые виды конечных элементов. Там, где такие программы имеются, следует совмещать численные эксперименты с теоретическим семинаром, основанным на этой книге.

Главы 2—5 написаны первым автором. Последние три главы написаны вторым автором, а затем редактировались и «унифицировались» первым. Весь текст был отпечатан миссис Ингрид Нааман, любезность которой заставила нас поверить, что она делала это с удовольствием.

Благодарим за внимание.

Гилберт Стренг
Джордж Дж. Фикс

Кембридж, Массачусетс

1 ВВЕДЕНИЕ В ТЕОРИЮ

1.1. ОСНОВНЫЕ ИДЕИ

Метод конечных элементов можно описать несколькими словами. Предположим, что задача, которую нужно решить, поставлена в вариационной форме: требуется найти функцию u, минимизирующую заданный функционал потенциальной энергии. Необходимость минимизации приводит к дифференциальному уравнению для u (уравнению Эйлера), которое обычно нельзя решить точно и приходится применять приближенные методы. Идея метода Рэлея — Ритца — Галёркина состоит в том, что выбирается конечное число пробных функций $\varphi_1, \varphi_2, \ldots, \varphi_N$ и среди всех линейных комбинаций вида $\Sigma q_j \varphi_j$ ищется комбинация, доставляющая минимум функционалу. Это аппроксимация Ритца. Неизвестные веса q_j определяются уже не из дифференциального уравнения, а из системы N дискретных алгебраических уравнений, для решения которой можно применить ЭВМ. Теоретическое обоснование этого метода очень простое: *процесс минимизации автоматически дает комбинацию, ближайшую к функции u.* Таким образом, цель состоит в том, чтобы выбрать пробные функции φ_j достаточно удобными для вычисления и минимизации потенциальной энергии и в то же время обеспечить хорошее приближение неизвестного решения u.

Наибольшая трудность при этом — достижение удобства и простоты вычислений. Теоретически всегда существует полный базис из пробных функций: их линейные комбинации при $N \to \infty$ приближают любой элемент пространства и потому аппроксимация Ритца сходится. Но можно ли будет численно работать с этими функциями — вот в чем вопрос. Именно этим вопросом и занимается теория конечных элементов.

Основополагающая идея весьма проста. Все начинается с разбиения исходной области на мелкие куски. Структура их должна быть проста для хранения и опознавания с помощью ЭВМ. Это могут быть треугольники или прямоугольники. Затем внутри каждого элемента разбиения задается пробная функция в максимально простой форме — обычно это полином, как правило, третьей или четвертой степени. Краевые условия гораздо проще поставить вдоль стороны треугольника или прямоугольника, чем сразу на всей границе области. Точность приближе-

ния повышается, если необходимо, не как в классическом методе Ритца за счет использования более сложных пробных функций, а за счет более мелкого разбиения области с сохранением тех же полиномов, что и прежде. ЭВМ при этом работает по той же программе, только дольше. При применении метода конечных элементов вычислительная машина помогает не только решать разностные уравнения, но и *строить* их, чего никогда прежде не было в случае сложных физических задач.

Метод конечных элементов придумали инженеры, и поначалу он не был понят как вариант метода Рэлея — Ритца. Разбиение области на простые части и составление уравнений равновесия и неразрывности для этих частей были выполнены на основе физических соображений. Построение более сложных конечных элементов проводилось так же; было замечено, что при возрастании степени полиномов значительно повышается точность, *однако неизвестные коэффициенты q_j, вычисляемые при дискретной аппроксимации, всегда имели некий физический смысл.* По этой причине результат было гораздо легче интерпретировать, чем весовые коэффициенты в классическом методе.

Вся эта процедура стала математически обоснованной, когда неизвестные коэффициенты q_j были отождествлены с коэффициентами в аппроксимации Ритца $u \approx \sum q_j \varphi_j$, а дискретные уравнения — с условиями минимума потенциальной энергии. Это поняли Аржирис в Германии и Англии, Мартин и Клаф в Америке; мы не знаем, кто из них первый. В результате появилась возможность заложить теоретическую основу метода. Процедуры построения более точных конечных элементов уже были разработаны, теория стала вырисовываться.

Основная задача состоит в исследовании точности, с которой кусочно полиномиальные функции могут аппроксимировать неизвестное решение u. Другими словами, надо определить, насколько хороши конечные элементы, построенные на основе вычислительной простоты, и дадут ли они хорошую аппроксимацию. Интуитивно ясно, что всякую достаточно хорошую функцию u можно с произвольной точностью приблизить кусочно линейными функциями. Математическая задача состоит в получении максимально точной оценки ошибки и определении скорости убывания ошибки при возрастании количества элементов разбиения (или степени полинома внутри каждого элемента). Разумеется, метод конечных элементов можно применять, не доказывая математические теоремы; так делали в течение более десяти лет. Однако мы считаем, что полезно, особенно для дальнейшего развития метода, понять и обобщить все, что уже сделано.

Мы попытаемся дать полный анализ *линейных задач* и *метода перемещений.* Подобной теории для нелинейных уравнений

пока не существует, хотя можно рассмотреть полулинейные уравнения, в которых нелинейность содержится в членах младшего порядка. Мы сделаем несколько предварительных замечаний о нелинейных уравнениях, но в основном оставим эти задачи на будущее. Почему мы выбрали именно метод перемещений, а не другую вариационную формулировку, мы поясним в гл. 2; здесь мы встали на сторону большинства. Это наиболее распространенный взгляд на метод конечных элементов. Разумеется, можно было бы построить теорию аппроксимации в других терминах, но переход от одной теории к другой был бы почти автоматическим.

Цель настоящей главы — иллюстрация основных этапов в методе конечных элементов:

1. Вариационная постановка задачи.
2. Построение кусочно полиномиальных пробных функций.
3. Вычисление матрицы и решение дискретной системы.
4. Оценка точности аппроксимации Ритца.

Мы ограничиваемся вариационной постановкой задачи, чтобы можно было использовать некоторые важные математические понятия, необходимые для нашей теории — гильбертовы пространства $\mathscr{H}^s$, оценки решения по исходным данным, энергетическое скалярное произведение, которое естественно связано со спецификой задачи. С помощью этого аппарата можно доказать сходимость метода конечных элементов даже в очень сложной геометрии.

Фактически простота вариационного подхода позволяет анализировать то, что уже недоступно методу конечных разностей.

1.2. ДВУХТОЧЕЧНАЯ КРАЕВАЯ ЗАДАЧА

Мы поясним метод конечных элементов и введем необходимый математический аппарат на хорошо известном примере. Возьмем одномерное пространство, чтобы конструкция элементов была проста и естественна, а математические преобразования вели прямо к цели — требуется всего лишь интегрирование по частям вместо использования общих формул Грина. Итак, мы выбираем уравнение

$$-\frac{d}{dx}\left(p(x)\frac{du}{dx}\right)+q(x)u=f(x). \tag{1}$$

Если поставить подходящие краевые условия в точках $x=0$ и $x=\pi$, то получим классическую задачу Штурма — Лиувилля. Это уравнение описывает ряд различных физических процессов, например распределение температуры в некотором стержне или амплитуду колебаний струны. В пространстве большей

размерности это соответствует *эллиптической краевой задаче*, например уравнению Лапласа.

Для того чтобы проиллюстрировать различные типы краевых условий, особенно при вариационной постановке задачи, мы фиксируем левый конец струны, а правый оставляем свободным. Таким образом, в точке $x = 0$ краевое условие является *главным (кинематическим, вынужденным, геометрическим)*, или *условием Дирихле*

$$u(0) = 0,$$

а в точке $x = \pi$, где струна не закреплена, возникает *естественное (динамическое)* краевое условие, или *условие Неймана*

$$u'(\pi) = 0.$$

Исследуем эту модельную задачу с четырех различных точек зрения:

1) чисто математической,
2) прикладной,
3) с точки зрения аппроксимации конечными разностями,
4) с точки зрения аппроксимации конечными элементами.

Важно увидеть в этих четырех аспектах одной и той же задачи общие черты: средства, применяемые математиком для доказательства существования и единственности решения, а вычислителем при численном анализе поведения решения, следует применить также при изучении численного алгоритма.

Сначала поступим, как математик, скомбинировав дифференциальное уравнение и краевые условия в одно целое:

$$Lu = f.$$

Здесь L — линейный оператор, действующий на определенном классе функций, а именно удовлетворяющих краевым условиям и дважды дифференцируемых. С математической точки зрения основная задача такова: *подобрать пространство функций u и класс правых частей f так, чтобы каждой функции f соответствовало единственное решение u.* Как только соответствие между f и u установлено, задача $Lu = f$ в абстрактном смысле «решена». Разумеется, это лишь первый шаг в определении решения u, соответствующего данной функции f. Эта задача составляет предмет всей книги. Однако мы считаем, что стóит потратить время на определение таких функциональных пространств. При использовании вариационных принципов и аппроксимации особенно важно знать точно, на каком функциональном пространстве они применимы. (Термин «пространство» предполагает линейность.)

Рассмотрим одно такое пространство, по-видимому, наиболее важное в теории, а именно пространство функций с *конечной энергией*, т. е. функций f, для которых

$$\int_0^\pi (f(x))^2\,dx < \infty. \tag{2}$$

Любая кусочно гладкая функция f принадлежит этому пространству, но δ-функция Дирака — уже нет. Позже мы вернемся к этому случаю «сосредоточенной нагрузки». Пространство функций, удовлетворяющих условию (2), часто обозначают L_2. Мы предпочитаем обозначение $\mathscr{H}^0$. Здесь верхний индекс указывает, сколько производных от функции f обладают конечной энергией (в нашем случае конечной энергией обладает только сама функция f).

Для простейшей задачи Штурма — Лиувилля — $u'' = f$ нетрудно построить соответствующее пространство решений. Такое пространство обозначается $\mathscr{H}^2_B$. Нижний индекс B означает, что выполнены краевые условия $u(0) = u'(\pi) = 0$, а верхний индекс 2 — что вторая производиая решения u обладает конечной энергией [1]). Можно показать, однако, что если предположить $p(x) \geqslant p_{\min} > 0$ и $q(x) \geqslant 0$, то пространство $\mathscr{H}^2_B$ будет к тому же пространством решений более сложного уравнения — $(pu')' + qu = f$. Итак, справедлива теорема:

Оператор L есть взаимно однозначное отображение из $\mathscr{H}^2_B$ в $\mathscr{H}^0$. Таким образом, для всякой функции $f \in \mathscr{H}^0$ дифференциальное уравнение (1) *имеет единственное решение и в $\mathscr{H}^2_B$. Более того, решение непрерывно зависит от f: если функция f мала, то мало и решение u.*

Последнее замечание требует разъяснения. Нам нужны нормы, с помощью которых можно измерять f и u. Нормы должны быть различными, так как различны пространства правых частей и решений. Естественно связать эти нормы с энергией:

$$\|f\|_0 = \left[\int (f(x))^2\,dx\right]^{1/2},$$

$$\|u\|_2 = \left[\int ((u''(x))^2 + (u'(x))^2 + (u(x))^2)\,dx\right]^{1/2}.$$

Пользуясь этими определениями, можно записать непрерывную зависимость решения от входных данных в количественной форме: существует такая постоянная C, что

$$\|u\|_2 \leqslant C\|f\|_0. \tag{3}$$

---

[1]) Эти пространства определены в указателе обозначений в конце книги.

Из этой оценки немедленно следует единственность решения: если $f = 0$, то и $u = 0$. Такие оценки лежат в основе современной теории уравнений в частных производных. Общая методика доказательства неравенства (3), которую можно применять для краевых задач в пространствах нескольких переменных, развита совсем недавно. В этой книге мы будем использовать такие оценки для эллиптических уравнений порядка $2m$:

$$\| u \|_{2m} \leqslant C \| f \|_0. \tag{4}$$

Перейдем к более практическому вопросу — явному построению решения. Если коэффициенты p и q постоянны, то решение можно построить в виде бесконечного ряда. Ключ к решению дает то обстоятельство, что собственные значения и собственные функции оператора L известны в явной форме:

$$u_n(x) = \sqrt{\frac{\pi}{2}} \sin\left(n - \frac{1}{2}\right)x, \quad \lambda_n = p\left(n - \frac{1}{2}\right)^2 + q. \tag{5}$$

Ясно, что $Lu_n = -pu_n'' + qu_n = \lambda_n u_n$, функции u_n удовлетворяют краевым условиям, лежат в $\mathscr{H}_B^2$ и ортонормальны:

$$\int_0^\pi u_n(x)\, u_m(x)\, dx = \delta_{nm}.$$

Предположим, что правая часть разложена в ряд по собственным функциям:

$$f(x) = \sum_{n=1}^{\infty} q_n \sqrt{\frac{\pi}{2}} \sin\left(n - \frac{1}{2}\right)x. \tag{6}$$

Формально интегрируя ряд и учитывая ортогональность u_n, получаем

$$\| f \|_0^2 = \int_0^\pi f^2\, dx = \sum_{n=1}^{\infty} a_n^2.$$

Функция f из $\mathscr{H}^0$ допускает точное гармоническое разложение в форме (6), и коэффициенты разложения удовлетворяют условию $\sum a_n^2 < \infty$. Здесь возникает небольшой парадокс, так как каждая функция f в форме (6) формально удовлетворяет условиям $f(0) = 0$, $f'(\pi) = 0$, хотя никаких краевых условий на f не налагается. От элементов из $\mathscr{H}^0$ требуется лишь конечная энергия, $\int f^2 < \infty$. Этот парадокс исчезает, если использовать полноту собственных функций u_n в $\mathscr{H}^0$. Независимо от того,

удовлетворяет ли f этим фальшивым краевым условиям, ее разложение сходится в среднем квадратичном:

$$\int_0^\pi \left(f(x) - \sum_{n=1}^{N} a_n \sqrt{\frac{\pi}{2}} \sin\left(n - \frac{1}{2}\right) x \right)^2 dx \to 0 \quad \text{при } N \to \infty.$$

Таким образом, эти краевые условия нестойки и при $N \to \infty$ «снимаются». На рис. 1.1 показано, как сходится последовательность функций f_n, лежащих в $\mathscr{H}_B^2$, к функции f, не лежащей в $\mathscr{H}_B^2$.

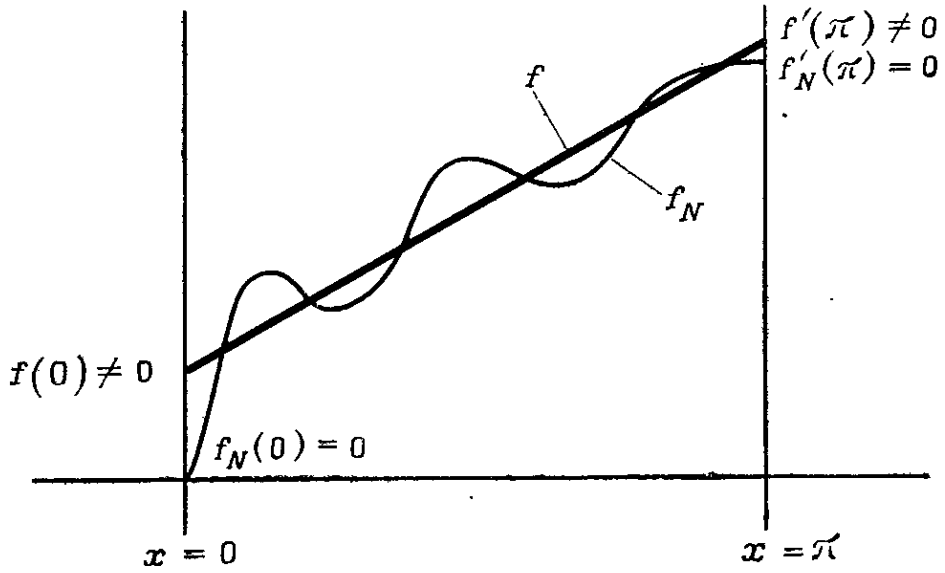


Рис. 1.1.
f_N в $\mathscr{H}_B^2$ аппроксимирует произвольную функцию f.

Теперь ясно, как решать дифференциальное уравнение Штурма — Лиувилля: если $f = \sum a_n u_n$, то решение u представимо в виде ряда

$$u = \sum \frac{a_n}{\lambda_n} u_n = \sqrt{\frac{\pi}{2}} \sum_1^\infty \frac{a_n \sin\left(n - \frac{1}{2}\right) x}{p\left(n - \frac{1}{2}\right)^2 + q}. \tag{7}$$

Используя это явное решение, можно непосредственно получить оценку $\| u \|_2 \leqslant C \| f \|_0$ и, таким образом, связь между пространством данных $\mathscr{H}^0$ и пространством решений $\mathscr{H}_B^2$.

Вопрос о краевых условиях сложнее и заслуживает дальнейшего рассмотрения. Мы уже видели, что хотя f приближается рядом по функциям u_n, удовлетворяющим краевым условиям, это еще не значит, что f удовлетворяет им. Почему же решение u подчинено краевым условиям? Дело в том, что разложение в ряд для функции u сходится в более сильном смысле, чем разложение для функции f: не только $\sum a_n u_n/\lambda_n$ сходится к u

в среднем квадратичном, но сходятся первые и вторые производные этого разложения. Точнее,

$$\left\| u - \sum_{1}^{N} \frac{a_n}{\lambda_n} u_n \right\|_2 \to 0 \text{ при } N \to \infty.$$

Смысл в том, что когда сходятся вторые производные, краевые условия *сохраняются* и предельная функция u также им удовлетворяет. (Заметим, что на рис. 1.1 вторые производные от f_N не сходятся к f'', поэтому f не удовлетворяет краевым условиям и не лежит в $\mathscr{H}_B^2$. Однако подобная ситуация невозможна для u.)

Общее правило таково: *краевые условия, включающие производные порядка менее s, сохраняются при предельном переходе в норме пространства* $\mathscr{H}^s$. Краевые условия с производными порядка s и выше нестойки, и их нельзя применить к функциям пространства $\mathscr{H}^s$. Теперь понятно различие между главными краевыми условиями, которые остаются, и естественными краевыми условиями, которые меняются. Это различие видно в вариационной задаче, так как она записывается в терминах *первых призводных*, т. е. $\mathscr{H}^1$-нормы. В аппроксимации по методу конечных элементов мы будем требовать удовлетворения всех краевых условий, содержащих производные порядка менее 1, т. е. условия типа $u(0) = 0$, но не будем требовать удовлетворения условия на первую производную. Это не помешает аппроксимации по методу конечных элементов сходиться в $\mathscr{H}^1$-норме к точному решению u, удовлетворяющему условию $u'(\pi) = 0$. Поэтому в следующем разделе мы сможем перейти от «чисто математической» постановки задачи к эквивалентной вариационной постановке.

1.3. ВАРИАЦИОННАЯ ПОСТАНОВКА ЗАДАЧИ

Линейное уравнение $Lu = f$ связано с квадратичным функционалом

$$I(v) = (Lv,\ v) - 2(f,\ u)$$

следующим образом: уравнение $Lu = f$ есть *уравнение Эйлера;* оно дает условие минимизации функционала I. Задачи обращения оператра L и минимизации функционала I эквивалентны и решением для них служит одна и та же функция u. По этой причине можно ставить задачи как в *операторной форме* — в терминах линейного оператора L, так и в *вариационной форме* — в терминах квадратичного функционала I. Цель этого раздела — найти точный вариационный эквивалент нашей двухточечной краевой задачи.

Эквивалентность дифференциальных уравнений и вариационных задач составляет основу выбора вычислительной схемы. Дифференциальное уравнение можно аппроксимировать дискретной системой, используя конечные разности, а вариационный функционал можно минимизировать на конечномерном пространстве функций, как в методе конечных элементов. В приложениях вариационная постановка часто бывает первичной и следует из физических соображений, а дифференциальное уравнение — результат такой постановки. Неудивительно поэтому, что нас интересует прежде всего, как приближенно минимизировать квадратичные функционалы.

Функционал мы называем квадратичным по аналогии со случаем, когда L, v и f — просто вещественные числа. Тогда $I(v) = Lv^2 - 2fv$ описывает параболу, и, *если число L положительно*, она достигает минимума в точке u, определяемой из уравнения

$$\left.\frac{dI}{dv}\right|_{v=u} = 2(Lu - f) = 0.$$

Если $L < 0$, то минимум равен $-\infty$, а если $L = 0$, то парабола вырождается в прямую.

Более интересен случай, когда v и f — это n-мерные векторы, а L — симметричная положительно определенная матрица порядка n. Функционал I имеет вид

$$I(v) = \sum_{j,\,k} L_{jk} v_k v_j - 2 \sum_j f_j v_j.$$

С учетом симметрии $L_{jk} = L_{kj}$ уравнения Эйлера запишутся так:

$$\left.\frac{\partial I}{\partial v_m}\right|_{v=u} = 2\left[\sum_k L_{mk} u_k - f_m\right] = 0, \quad m = 1, \ldots, n.$$

Эти уравнения дают вместе систему $Lu = f$. Минимум функционала I, достигаемый на векторе $u = L^{-1}f$, равен

$$I(L^{-1}f) = (f,\ L^{-1}f) - 2(f,\ L^{-1}f) = -(f,\ L^{-1}f).$$

(Запись $(,)$ означает обычное скалярное произведение векторов.) Так как матрица L положительно определена, то L^{-1} тоже положительно определена, и этот минимум отрицателен (или равен нулю, если $f = 0$). Геометрически $I(v)$ представляется выпуклой поверхностью, причем параболоид будет выпуклым вниз, если матрица L положительно определена.

Уравнение минимума $Lu = f$ получается при одновременном варьировании всех компонент.

Если u — точка минимума для I, то для всех v и ε

$$I(u) \leqslant I(u+\varepsilon v) = I(u) + 2\varepsilon[(Lu,\ v) - (f,\ v)] + \varepsilon^2 (Lv,\ v).$$

Так как число ε может быть произвольно мало и иметь любой знак, коэффициент при ε должен обратиться в нуль:

$$(Lu,\ v) = (f,\ v) \text{ для всех } v. \tag{8}$$

Отсюда следует, что $Lu = f$. Будем называть уравнение (8) уравнением в *слабой форме* или *уравнением Галёркина*. При составлении таких уравнений не надо требовать положительной определенности оператора L или его симметрии, так как уравнения Галёркина — это условия не минимума, а всего лишь стационарной точки. Такая постановка задачи приводит к методу Галёркина.

Разумеется, кроме функционала $(Lv,\ v) - 2(f,\ v)$, существуют и другие квадратичные функционалы, точкой минимума которых служит решение уравнения $Lu = f$. Очевидно, что функционал метода наименьших квадратов $Q(v) = (Lv - f,\ Lv - f)$ достигает минимума (равного нулю) в той же точке. Есть, однако, существенное различие: уравнения Эйлера $\partial Q/\partial v_m = 0$ приводят не к $Lu = f$, а к $L^T Lu = L^T f$. Теоретически эти уравнения эквивалентны, если оператор L обратим, но на практике появление $L^T L$ невыгодно.

Рассмотрим теперь наше дифференциальное уравнение

$$Lu = \left[-\frac{d}{dx}\left(p(x)\frac{d}{dx}\right) + q(x)\right]u = f,$$

$$u(0) = u'(\pi) = 0.$$

Построим $I(v) = (Lv, v) - 2(f, v)$. Скалярное произведение определено здесь для функций, заданных в интервале $0 \leqslant x \leqslant \pi$, но оно очень похоже на скалярное произведение векторов:

$$(f,\ v) = \int_0^\pi f(x)\, v(x)\, dx.$$

Функции f, v *вещественны*, как в большинстве приложений. Модификация на комплексный случай хорошо известна: над одним из сомножителей в интеграле надо поставить знак сопряжения.

Функционал I вычисляется интегрированием по частям (этот прием широко применяется в теории дифференциальных операторов):

$$(Lv,\ v) = \int_0^\pi [-(pv')' + qv]\, v\, dx = \int_0^\pi [p(v')^2 + qv^2]\, dx - pv'v\Big|_0^\pi. \tag{9}$$

Если v удовлетворяет краевым условиям $v(0) = v'(\pi) = 0$, то квадратичный функционал принимает вид

$$I(v) = \int_0^\pi [p(x)(v'(x))^2 + q(x)(v(x))^2 - 2f(x)v(x)]\,dx.$$

Этот функционал и нужно минимизировать.

Решение дифференциального уравнения $Lu = f$ соответствует функции u, минимизирующей I. На каком классе функций следует искать этот минимум? Мы знаем, что решение u лежит в $\mathscr{H}_B^2$. Значит, класс функций должен содержать $\mathscr{H}_B^2$. Минимум функционала $I(v)$ на классе $\mathscr{H}_B^2$ реализуется в нужной точке $v = u$. Заметим, однако, что *выражение для $I(v)$ не содержит вторых производных*, они исчезли при интегрировании по частям. Это значит, что $I(v)$ можно определить на функциях v, у которых первая производная, а не вторая, обладает конечной энергией. Следовательно, класс функций, на которых задача минимизации имеет смысл, шире, чем пространство $\mathscr{H}_B^2$.

Наш принцип таков: функцию v можно включить в класс, на котором мы решаем задачу минимизации, если только v — предел последовательности функций v_N из $\mathscr{H}_B^2$. Под словом «предел» мы понимаем предел в смысле квадратичных членов в выражении потенциальной энергии:

$$\int_0^\pi p(v' - v_N')^2 + q(v - v_N)^2 \to 0, \quad N \to \infty. \tag{10}$$

Отметим, что *такое расширение нашего пространства, действительно, не может уменьшить минимум функционала I*; каждое новое значение $I(v)$ есть предел старых значений $I(v_N)$. Таким образом, если минимум I уже достигался для какой-то функции u из $\mathscr{H}_B^2$, то она останется минимизирующей функцией. Это не вызывает сомнений. Однако у нас теперь огромное преимущество: минимум разрешается искать также и на функциях v, лежащих за пределом исходного класса $\mathscr{H}_B^2$. На практике это означает, что можно использовать непрерывные и всего лишь кусочно линейные функции — их легко построить и их первые производные обладают конечной энергией, но сами функции не принадлежат $\mathscr{H}_B^2$.

Наша задача состоит теперь в том, чтобы описать это новое, более приемлемое пространство. Другими словами, мы хотим найти свойство функций v, являющихся пределом в смысле (10), т. е. в норме пространства $\mathscr{H}^1$, последовательности

функций v_N, имеющих две производные и удовлетворяющих всем краевым условиям.

Свойств, которые мы хотим определить, два: гладкость допустимых функций и краевые условия, которым они должны удовлетворять. Первое свойство найти сравнительно просто: так как (10) влечет за собой только сходимость первых произ-

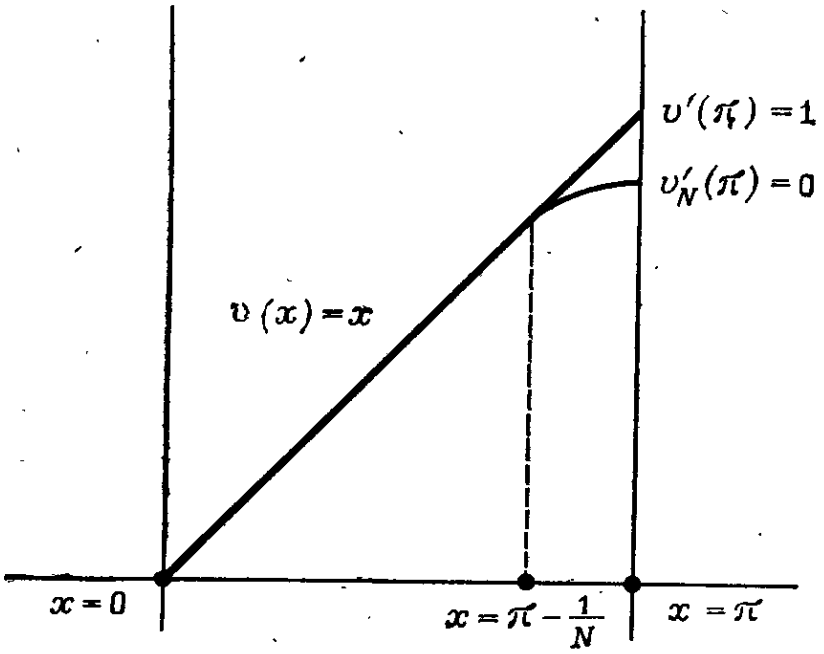


Рис. 1.2.
Сходимость в пространстве $\mathscr{H}^1$; $v'_N(\pi) = 0$, но $v'(\pi) \neq 0$.

водных, предельная функция v должна лежать лишь в $\mathscr{H}^1$. Это означает, что норма

$$\| v \|_1 = \left[\int_0^\pi (v^2 + (v')^2)\, dx \right]^{1/2}$$

должна быть конечной.

Проблема краевых условий — вещь более тонкая. Если некоторая последовательность элементов из $\mathscr{H}^2_B$ удовлетворяет условиям $v_N(0) = 0$ и $v'_N(\pi) = 0$, сохранит ли предельная функция оба эти свойства? Оказывается, *первое свойство сохранится, а второе будет потеряно.* Чтобы показать, что предел не обязан удовлетворять условию Неймана $v'(\pi) = 0$, положим, например, $v(x) = x$ и рассмотрим последовательность $\{v_N\}$ на рис. 1.2. Так как $v - v_N$ обращается в нуль всюду, кроме малого интервала на границе, причем $0 \leqslant v' - v'_N \leqslant 1$, то требование (10), очевидно, выполнено. Таким образом, $v(x) = x$ принадлежит предельному пространству допустимых функций, хотя $v'(\pi) \neq 0$.

С другой стороны, условие $v(0)=0$ продолжает сохраняться в пределе. Действительно, v_N сходится в каждой точке x, так как по неравенству Шварца

$$|v_N(x)-v_M(x)|^2=\left|\int_0^x (v_N'(y)-v_M'(y))\,dy\right|^2\leqslant$$

$$\leqslant\int_0^x 1^2\,dy\int_0^x (v_N'-v_M')\,dy\to 0.$$

Из анализа известно, что предельная функция v непрерывна и сходимость v_N к v равномерна по x. В частности, $v(0)=\lim v_N(0)=0$ в точке $x=0$. Другими словами, первые производные сходятся только в среднем квадратичном и нет гарантии, что $v'(\pi)=0$.

Итак, *пространством функций, допустимых при минимизации, будет* $\mathscr{H}_E^1$; его элементы имеют первые производные с конечной энергией и удовлетворяют главному краевому условию $v(0)=0$ (на это указывает нижний индекс E). Естественное краевое условие $v'(\pi)=0$ не обязательно. Отметим, что, если наши математические рассуждения последовательны, функция u из $\mathscr{H}_E^1$, минимизирующая I, *автоматически* удовлетворяет условию $u'(\pi)=0$. Это легко проверить, так как для некоторого числа ε и некоторой функции v из $\mathscr{H}_E^1$

$$I(u)\leqslant I(u+\varepsilon v)=I(u)+2\varepsilon\int_0^\pi pu'v'+quv-fv+$$

$$+\varepsilon^2\int_0^\pi p(v')^2+qv^2.$$

Так как число ε может быть любого знака, линейный член (первая вариация) должен отсутствовать:

$$0=\int_0^\pi pu'v'+quv-fv=$$

$$=\int_0^\pi [-(pu')'+qu-f]\,v+p(\pi)u'(\pi)v(\pi). \qquad (11)$$

Если минимизирующая функция u имеет две производные (так что можно провести последнее интегрирование по частям), правая часть равна нулю для всех $v \in \mathscr{H}_E^1$ только тогда, когда $-(pu')' + qu = f$ и выполнено естественное краевое условие $v'(\pi) = 0$ на границе. Точно так же $u(0) = 0$, поскольку, подобно всякой другой функции из $\mathscr{H}_E^1$, функция u удовлетворяет главному краевому условию. Это замыкает круг: минимизация функционала I на пространстве $\mathscr{H}_E^1$ эквивалентна решению уравнения $Lu = f$, и функцию u можно вычислить одним из указанных способов.

Процесс расширения $\mathscr{H}_B^2$ до $\mathscr{H}_E^1$ допускает простую геометрическую интерпретацию. Квадратичный функционал I представляется выпуклой поверхностью — параболоидом в бесконечномерном случае. Сначала, когда функционал I был определен только для функций v из $\mathscr{H}_B^2$, на этой поверхности были «дыры». Мы их заполняли. Поверхность менялась так, чтобы ни в коем случае не изменилась минимальность значения, и в результате были устранены мелкие дефекты, соответствующие функциям v, лежащим в $\mathscr{H}_E^1$, но не в $\mathscr{H}_B^2$.

В завершение этого раздела отметим две вырожденные задачи, весьма важные в приложениях. Заслуживает внимания то, что в обоих случаях вариационная форма, которую мы только что исследовали, остается прежней; $I(v)$ следует минимизировать на допустимом пространстве $\mathscr{H}_E^1$ и минимизирующая функция u дает нужное решение. Напротив, операторная форма $Lu = f$ становится сложнее, в особую точку x_0 входят специальные условия и *решение не лежит более в* $\mathscr{H}_B^2$.

С точки зрения прикладной математики специальное поведение вблизи особенности не следует игнорировать; возможно, это и правильно. Однако с точки зрения аппроксимации конечными элементами важно, что алгоритм можно осуществить, не зная всей информации об особенности. Такая информация весьма полезна для ускорения сходимости аппроксимаций, как в гл. 8, но алгоритм действует и без нее.

Замечание 1. В точке x_0, где упругие свойства струны (или коэффициент диффузии среды в тепловом потоке) изменяется скачкообразно, коэффициент $p(x)$ может быть разрывен. В такой точке появляется *внутренняя граница*. Решение u уже не имеет вторую производную. Чтобы найти «условие скачка» в точке x_0, обратимся к вариационной форме задачи; первая вариация $\int pu'v' + quv - fv$ должна равняться нулю для всех v, если u — точка минимума. Полагая, что, кроме x_0, нет дру-

гих особенностей, интегрируем по частям отдельно в интервалах $(0, x_0)$ и (x_0, π):

$$0 = \int_0^{x_0} [-(pu')' + qu - f]\, v + p_- u'_- v_- +$$

$$+ \int_{x_0}^{\pi} [-(pu')' + qu - f]\, v + p(\pi) u'(\pi) v(\pi) - p_+ u'_+ v_+.$$

Нижние индексы — и "+" обозначают пределы при $x \to x_0$ слева и справа соответственно. Напомним, что $v_- = v_+$ для всех v из $\mathscr{H}_E^1$, так как функция v непрерывна; в частности $u_- = u_+$. Изменяя v, получаем, что дифференциальное уравнение справедливо в каждом интервале, $u'(\pi) = 0$ и

$$p_- u'_- = p_+ u'_+.$$

Это естественное краевое условие в точке x_0, вытекающее непосредственно из вариационной формы: хотя u' имеет скачок, *комбинация pu' остается непрерывной.*

Так как u' имеет скачок, решение лежит в $\mathscr{H}_E^1$, но не в $\mathscr{H}_B^2$. Это тот случай, когда *одна из дыр на поверхности $I(v)$ расположена на самом дне.* Минимальное значение $I(v)$ могло бы быть тем же самым на исходном пространстве $\mathscr{H}_B^2$, но внутри этого пространства нет функции, реализующей минимум. Поверхность находится сколь угодно близко от дыры, но ее можно заполнить только функцией, удовлетворяющей условию скачка.

Стандартные оценки ошибок для конечных элементов, справедливые в предположении гладкости u, не пригодны в точке разрыва x_0. Их, однако, можно сохранить, если в узле x_0 пробные функции в аппроксимации будут удовлетворять условию скачка. Хотя это условие скорее естественное, чем главное, не каждая пробная функция должна ему удовлетворять. Так как не требуется непрерывность производных пробных функций и потому нарушается условие скачка, аппроксимация будет хорошей.

Замечание 2. До сих пор требовалось, чтобы неоднородный член f принадлежал классу $\mathscr{H}^0$, так что δ-функции исключались. Физически было бы очень интересно рассмотреть случай, когда δ-функция допускается и интерпретируется как точечная нагрузка или точечный источник; математически соответствующая функция u есть *фундаментальное решение.* Поэтому мы попробуем исследовать этот случай с помощью

функционала

$$I(v) = \int_0^\pi p\,(v')^2 + qv^2 - 2fv,$$

минимизируемого на $\mathscr{H}_E^1$.

Предположим, например, что $p \equiv 1$ и $q \equiv 0$. Если f обладает конечной энергией, то интеграл I конечен и минимизируется просто. Но и для $f = \delta(x_0)$, $0 < x_0 < 1$, интеграл I конечен. Он имеет вид

$$I(v) = \int_0^\pi (v')^2\,dx - 2v(x_0),$$

и минимум реализуется на «*ломаной*» $v = u$, изображенной на рис. 1.3. Снова решение не принадлежит $\mathscr{H}_B^2$, и дыра соответствующая этой функции, находится на «дне» бесконечномерного параболоида $I(v)$.

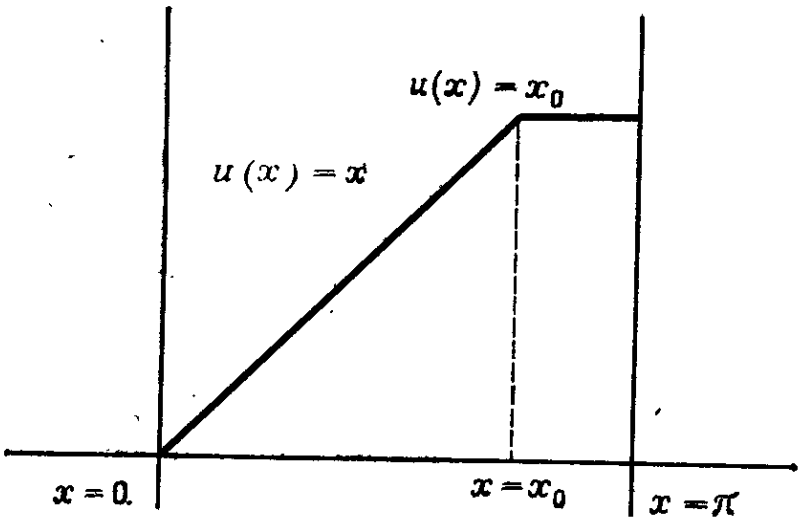


Рис. 1.3.
Фундаментальное решение для точечной нагрузки, f и u не принадлежат $\mathscr{H}^0$ и $\mathscr{H}^2$ соответственно.

Еще одна возможность: предположим, что точечная нагрузка приложена на конце $x_0 = \pi$, т. е. $f = \delta(\pi)$. Тогда решением будет $u(x) = x$ и у функции на рис. 1.3 *нет излома. Решение нарушает естественное краевое условие* $u'(\pi) = 0$. Возвращаясь к (11), видим, что это вполне возможно; при $p = 1$, $q = 0$ и $u(x) = x$ первая вариация функционала равна

$$\int_0^\pi (u'v' - fv)\,dx = \int_0^\pi (v' - \delta(\pi)\,v)\,dx \equiv 0$$

для любой функции v из $\mathscr{H}_E^1$. Таким образом, первая вариация становится равной нулю при $u(x) = x$, а значит, это минимизирующая функция. Единственное отличие состоит в том, что интегрирование по частям в (11), а значит, и последующий вывод естественного краевого условия $u'(\pi) = 0$ не проходят. Следовательно, если позволить функции f быть сингулярной в точке $x = \pi$, естественное краевое условие в этой точке не обязано выполняться.

Так как δ-функция представляет собой возможный выбор f, сразу возникает общий вопрос: какой класс неоднородных членов f мы можем себе позволить? Точнее, какое пространство данных соответствует пространству решений $\mathscr{H}_E^1$? Грубо говоря, пока функционал $I(v)$ остается конечным для всех v из $\mathscr{H}_E^1$, минимизация возможна. Это правило дает возможность выбирать в качестве f δ-функции и их линейные комбинации, но не их производные. Например, *дипольная функция* $f = \delta'(x_0)$ привела бы к

$$\int_0^\pi fv = -\int_0^\pi v'\delta = -v'(x_0);$$

интеграл в левой части может не быть конечным, так как v' может иметь неограниченный пик в точке x_0, хотя и обладает конечной энергией. Дипольная функция «слишком» сингулярна.

Итак, мы теперь допускаем, чтобы правая часть принадлежала пространству $\mathscr{H}^{-1}$ функций, производные порядка -1 которых, т. е. их неопределенные интегралы, принадлежат $\mathscr{H}^0$. Оператор второго порядка L переводит пространство $\mathscr{H}_E^1$ в $\mathscr{H}^{-1}$ так же, как он переводил $\mathscr{H}_B^2$ в $\mathscr{H}^0$. Соответствующая норма в $\mathscr{H}^{-1}$ определяется формулой

$$\|f\|_{-1} = \max_{v \in \mathscr{H}^1} \frac{\left|\int f(x)\, v(x)\, dx\right|}{\|v\|_1}. \tag{12}$$

Особый интерес представляет случай, когда f есть δ-функция, сосредоточенная в начале координат, т. е. $\int fv = 0$ для каждой функции v из допустимого пространства $\mathscr{H}_E^1$. Это означает, что f в вариационном смысле не отличается от нуля, и решением будет $u \equiv 0$. (Это функция, изображенная на рис. 1.3, при $x_0 = 0$.) Отсюда в свою очередь следует, что пространство решений $\mathscr{H}_E^1$ соответствует пространству данных $\mathscr{H}^{-1}$ при одной оговорке: две данные функции f_1 и f_2 считаются одинаковыми, если они отличаются на δ-функцию, сосредоточенную в нуле.

Наконец, решение должно быть непрерывным по f в смысле норм в пространстве решений для u и в пространстве данных для f. Доказательство опирается на равенство нулю первой вариации для каждой функции v из $\mathscr{H}_E^1$, в частности для $v = u$:

$$\int_0^\pi p(u')^2 + qu^2 = \int_0^\pi fu.$$

По определению нормы $\|f\|_{-1}$ правая часть ограничена величиной $\|f\|_{-1}\|u\|_1$, а левая часть, очевидно, превосходит $p_{\min}\|u'\|_0^2$ и; как легко показать, $\sigma\|u\|_1^2$ для некоторого положительного числа σ. Поэтому

$$\sigma\|u\|_1^2 \leqslant \|f\|_{-1}\|u\|_1, \quad \|u\|_1 \leqslant \frac{1}{\sigma}\|f\|_{-1}.$$

Это и доказывает непрерывную зависимость u от f.

1.4. АППРОКСИМАЦИЯ КОНЕЧНЫМИ РАЗНОСТЯМИ

Этот раздел проникнут верой в численный анализ, ибо все, что можно решить абстрактно, можно решить и с помощью конкретных численных расчетов. «Каждому непрерывному отображению соответствует сходящаяся последовательность дискретных аппроксимаций». В предыдущих разделах рассмотрены два непрерывных отображения, переводящих f в u: одно для дифференциального уравнения ($\|u\|_2 \leqslant C\|f\|_0$), другое для его вариационного эквивалента ($\|u\|_1 \leqslant \sigma^{-1}\|f\|_{-1}$). Обе задачи готовы для численного решения.

Начнем с дифференциального уравнения $Lu = f$ и заменим производные разностными отношениями. Результатом будет конечная линейная система $L^hU^h = f^h$, дискретная операторная форма. В теоретическом анализе решения разностного уравнения два основных шага:

1. Вычислить локальную ошибку «отсечения», или дискретизации, с помощью разложения в ряд Тейлора.
2. Обосновать общую устойчивость системы, т. е. показать, что U^h непрерывно зависит от f^h, если h стремится к нулю.

Взятые вместе, эти два шага дают скорость сходимости U^h к точному решению u при $h \to 0$. Наше обсуждение направлено как раз на то, чтобы противопоставить теорию сходимости для разностных уравнений технике, используемой в следующем разделе и во всей оставшейся части книги для доказательства

сходимости в вариационных задачах [1]). Прекрасно видно, в чем особенность двух описанных шагов, проделанных в вариационной форме: вычисление ошибки отсечения здесь заменяется проверкой аппроксимационных свойств (или полноты) системы пробных функций, а устойчивость вообще не требует специального доказательства — для конечных элементов она автоматически выполняется.

Пусть, как обычно, интервал $[0, \pi]$ разбит на равные части длины $h = \pi/N$ точками $x_i = ih$, $i = 0, 1, \ldots, N$. Производные в уравнении $-(pu')' + qu = f$ заменяются центральными разностными отношениями

$$u'(x) \to \Delta^h u(x) = \frac{u(x + h/2) - u(x - h/2)}{h},$$

в результате чего появляется дискретное уравнение $-\Delta^h(p\,\Delta^h U) + qU = f$, и мы требуем, чтобы оно удовлетворялось во внутренних точках сетки x_i:

$$\frac{1}{h^2}\left[-p\left(x_i + \frac{h}{2}\right)(U^h_{i+1} - U^h_i) + p\left(x_i - \frac{h}{2}\right)(U^h_i - U^h_{i-1})\right] + \\ + q(x_i)U^h_i = f(x_i). \quad (13)$$

Так как это уравнение второго порядка, нужно потребовать выполнения краевых условий на концах интервала. Очевидно, что на левом конце $U^h_0 = 0$. На другом конце нельзя найти разностную замену условия $u'(\pi) = 0$ однозначно, и мы исследуем две возможности: одностороннюю разность

$$\frac{U^h_N - U^h_{N-1}}{h} = 0 \quad (14a)$$

и центральную разность

$$\frac{U^h_{N+1} - U^h_{N-1}}{2h} = 0. \quad (14b)$$

В первом случае разностное уравнение (13) удовлетворяется при $0 < i < N$, и вместе с двумя краевыми условиями оно дает систему $N + 1$ уравнений с $N + 1$ неизвестными. Во втором случае разностное уравнение рассматривают также и в точке $x_N = \pi$, чтобы компенсировать лишнее неизвестное U^h_{N+1}. Эти краевые условия легко сравнить, когда $p = 1$ и $q = 0$: после исключения неизвестных на концах интервала разностные урав-

[1]) В этом разделе мы отступаем от нашей главной темы, но метод конечных элементов и метод конечных разностей так тесно связаны, что их необходимо сравнить. Читатель понимает, что́ именно предпочитаем мы, и, если он разделяет наше мнение, может пропустить этот раздел.

нения имеют вид соответственно

$$L^hU^h = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & . \\ -1 & 2 & . & 0 \\ 0 & . & 2 & -1 \\ . & 0 & -1 & 1 \end{pmatrix}\begin{pmatrix} U_1 \\ . \\ . \\ U_{N-1} \end{pmatrix} = \begin{pmatrix} f_1 \\ . \\ . \\ f_{N-1} \end{pmatrix} \quad (15a)$$

и

$$L^hU^h = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & . \\ -1 & 2 & . & 0 \\ 0 & . & 2 & -1 \\ . & 0 & -2 & 2 \end{pmatrix}\begin{pmatrix} U_1 \\ . \\ . \\ U_N \end{pmatrix} = \begin{pmatrix} f_1 \\ . \\ . \\ f_N \end{pmatrix}. \quad (15b)$$

Чтобы проанализировать разностные уравнения с переменными коэффициентами p и q, рассмотрим *локальную ошибку отсечения* $\tau^h(x)$, возникающую, если вместо U^h подставить точное решение u:

$$-\Delta^h(p\Delta^h u) + pu - f = \tau^h.$$

Это совершенно формальные вычисления; в них $u(x_i \pm h)$ и $p(x_i \pm h/2)$ разлагаются в ряд Тейлора вблизи центральной точки x_i. Так как u удовлетворяет дифференциальному уравнению, члена нулевого порядка не будет; такое сокращение выражает *согласованность* разностного и дифференциального уравнений. Останутся члены

$$\tau^h = -\frac{h^2}{24}[(pu')''' + (pu''')'] + O(h^4).$$

Тот же процесс можно применить и на границах. Для оператора односторонней разности

$$\frac{u(\pi) - u(\pi - h)}{h} = u'(\pi) - \frac{h}{2}u''(\pi) + \ldots =$$

$$= -\frac{h}{2}u''(\pi) + O(h^3).$$

Снова согласованность с точным условием $u'(\pi) = 0$ аннулирует член нулевого порядка. Центральная разность, разумеется, точнее:

$$\frac{u(\pi + h) - u(\pi - h)}{2h} = \frac{h^2}{6}u'''(\pi) + O(h^4).$$

Когда разностное уравнение и краевые условия рассматриваются вместе, а не отдельно, ошибки отсечения выглядят по-другому. Возьмем, например, матрицы L^h, описанные выше:

краевые условия использовались здесь для исключения последнего неизвестного. Разложение последней строки в ряд Тейлора для односторонней и центральной разностей дают соответственно

$$\frac{u(\pi-h)-u(\pi-2h)}{h^2}-f(\pi-h)=-\frac{1}{2}u''(\pi)+O(h)$$

и

$$\frac{2u(\pi)-2u(\pi-h)}{h^2}-f(\pi)=-\frac{h}{3}u'''(\pi)+O(h^2).$$

В такой форме возникающие полные ошибки составляют $O(1)$ и $O(h)$, а это не так.

Для оценки ошибки $E^h=U^h-u$ исследуем разностное уравнение, которому эта ошибка удовлетворяет:

$$-\Delta^h(p\,\Delta E^h)+qE^h=\tau^h.$$

При центральной разности в точке $x=\pi$ краевые условия для ошибки имеют вид

$$E_0^h=0,\quad \frac{E_{N+1}^h-E_{N-1}^h}{2h}=\frac{h^2}{6}u'''(\pi)+O(h^4).$$

Отметим, что это разностное уравнение аналогично исходному дифференциальному уравнению, только теперь *правыми частями являются локальные ошибки отсечения*. Поэтому мы считаем, что главный член в E^h, скажем h^2e_2, возникает из членов порядка h^2 в локальной ошибке:

$$-(pe_2')'+qe_2=-\frac{1}{24}[(pu')'''+(pu''')'],$$

$$e_2(0)=0,\quad e_2'(\pi)=\frac{1}{6}u'''(\pi).$$

Решение $e_2(x)$ этой задачи есть *главная функция ошибки*. Чтобы найти следующий член в ошибке, подставим $u+h^2e_2$ в разностную задачу. Это приводит к ошибке отсечения, начинающейся с членов порядка h^4, и коэффициент при этом члене есть правая часть в уравнении для e_4.

Короче, можно рекурсивно определить разложение

$$U^h=u+h^2e_2+h^4e_4+\ \ldots\ =\sum h^ne_n(x). \tag{16}$$

Вычисление члена e_n осуществляется точно так же; разложение U^h прекращается лишь тогда, когда нарушается гладкость решения или когда продолжить разложение не позволяют краевые условия. Одно из положительных качеств разложения (16) — изменение ошибки с изменением x; обычно же у нас только

далекий от истины максимум ошибки на всем интервале. Далее, разложение (16) позволяет обосновать *экстраполяцию Ричардсона* при h близких к 0; можно вычислить решения с двумя (или более) различными значениями h и выбрать их комбинацию так, чтобы уничтожить главные члены в разложении. Для задачи с центральной разностью на крае в нашем примере линейный член he_1 отсутствует, и комбинация U^h и U^{2h}, повышающая точность со второго порядка до четвертого, выглядит так:

$$\frac{4}{3}U^h - \frac{1}{3}U^{2h} = u + O(h^4).$$

Такая экстраполяционная техника обсуждалась много раз и проверялась численно, однако на практике еще не получила должного применения. Точность краевых условий составляет здесь наиболее серьезное препятствие, в частности в многомерных задачах, когда граница области пересекает сетку самым причудливым образом.

Применяя те же идеи для односторонней разности на крае, видим, что первый член разложения he_1 возникает из ошибки $O(h)$ на границе:

$$-(pe_1')' + qe_1 = 0,$$

$$e_1(0) = 0, \quad e_1'(\pi) = -\frac{1}{2}u''(\pi).$$

Естественно, центральная разность предпочтительнее.

Аппроксимация первого порядка получается также, если дифференциальное уравнение связано с вариационной задачей. Как всегда, положительно определенная симметричная система $L^h u^h = f^h$ представляет собой уравнение Эйлера, дающее равенство нулю первой вариации функционала

$$I_\Delta(V^h) = (L^h V^h, V^h) - 2(f^h, V^h)$$

на минимизирующей функции U^h.

С односторонней разностью на крае этот функционал имеет вид

$$I_\Delta = \sum_{i=1}^{N-1}\left[p_{i-1/2}\left(\frac{V_i^h - V_{i-1}^h}{h}\right)^2 + q_i(V_i^h)^2 - 2f_iV_i^h\right].$$

Очевидно, что I_Δ есть конечно-разностный аналог истинного квадратичного функционала

$$I(v) = \int_0^\pi p(v')^2 + qv^2 - 2fv$$

и, что почти очевидно, его точность только первого порядка. Вместо того, чтобы воспользоваться *формулой трапеций* (второй порядок точности), интеграл I заменили суммой односторонних разностей.

Описанная техника занимает промежуточное положение между техникой конечных разностей и методом конечных элементов — интеграл I аппроксимируется суммой I_Δ, включающей разностные отношения, а затем минимизируется. Этот простой путь получения аппроксимаций небольшого порядка точности заслуживает больше внимания, чем ему уделяется. Но он теряет свои преимущества, если требуется высокая точность.

До сих пор анализ разностных уравнений был достаточно формальным и мы получили разложение ошибки $\sum h^n e_n$. Теперь сделаем второй шаг: докажем, что U^h сходится к u и разложение асимптотически справедливо. (Само разложение не может сходиться для конечного значения h, так как для сходимости надо, чтобы функция U^h от h была аналитической и чтобы задача была корректно поставлена даже для комплексных h; мы надеемся доказать, что $U^h - \sum_0^M h^n e_n = O(h^{M+1})$ при $h \to 0$.) Второй шаг требует оценки того же вида, что и для дифференциального уравнения: решение U^h должно непрерывно зависеть от правой части f^h.

Вернемся теперь к разностному уравнению и выясним прежде всего, существует ли единственное решение U^h для каждой функции f^h. Другими словами: является ли матрица L^h невырожденной? Один из наиболее эффективных способов доказательства обратимости L^h приводит к дискретному принципу максимума.

Пусть $L^h U^h = 0$. Предположим, что наибольшая компонента $|U_i^h|$ имеет номер n, и выберем знак U^h так, чтобы $U_n^h \geqslant 0$. Запишем разностное уравнение в точке x_n:

$$p_{n+1/2}(U_n^h - U_{n+1}^h) + p_{n-1/2}(U_n^h - U_{n-1}^h) + h^2 q_n U_n^h = 0.$$

Так как все члены неотрицательны, они должны равняться нулю. Если коэффициент q_n положителен, то сразу же получаем $U_n^h = 0$. В любом случае $U_{n+1}^h = U_n^h = U_{n-1}^h$. Таким образом, U_{n-1}^h и U_{n+1}^h также максимальны, и те же рассуждения можно повторить для $n-1$ или для $n+1$. Отсюда $U_n^h = U_0^h = 0$. Итак, $L^h U^h = 0$ только при $U^h = 0$, и матрица L^h *обратима*.

Аналогично можно показать, что при неоднородных краевых условиях все компоненты U_i^h не превосходят U_0^h и U_N^h. Это и есть *дискретный принцип максимума*, из которого следует, что *дискретная функция Грина* $(L^h)^{-1}$ — *неотрицательная матрица*.

Подобное доказательство приводит к теореме Гершгорина в теории матриц: все собственные значения λ матрицы A лежат в объединении кругов

$$|\lambda - A_{ii}| \leqslant \sum_{j \neq i} |A_{ij}|.$$

Выбирая в качестве A матрицу h^2L^h из формулы (15), видим, что собственные значения удовлетворяют неравенству $|\lambda - 2| \leqslant 2$. Таким образом, теорема Гершгорина не исключает возможности $\lambda = 0$, т. е. вырожденности матрицы L^h; здесь нужно повторить рассуждения для $i = n-1, n-2, \ldots, 1$.

Теорема Гершгорина становится совершенно бесполезной для задач четвертого порядка. Коэффициенты в простейшем случае равны: $A_{ii} = 6$, $A_{i, i\pm1} = -4$, $A_{i, i\pm2} = 1$, а круги Гершгорина таковы: $|\lambda - 6| \leqslant 10$. Так как точка $\lambda = 0$ лежит внутри этих кругов, доказательство не проходит даже для полуопределенной матрицы A. Эта трудность связана с тем, что в случае задач четвертого порядка не действует принцип максимума. Если сравнить два уравнения $u'' = 0$ и $u^{(IV)} = 0$, то очевидно, что прямая имеет экстремумы на концах интервала, а кубическая кривая нет.

Принцип максимума, если он работает, позволяет сделать доказательство сходимости простым. Но мы хотим сохранить полную аналогию между дифференциальным и разностным уравнениями, вводя дискретное неравенство, соответствующее $\|u\|_2 \leqslant C\|f\|_0$. Прежде всего зададим нормы, которые можно применять к сеточным функциям. Для дискретной энергии, очевидно, положим

$$\|f^h\|_0 = \left(\sum h |f_j^h|^2\right)^{1/2}.$$

Для квадрата второй нормы сложим энергию функции с энергией ее первых и вторых разностных отношений «вперед»:

$$\|U^h\|_2^2 = \|U^h\|_0^2 + \|\Delta_+ U^h\|_0^2 + \|\Delta_+^2 U^h\|_0^2.$$

В этих суммах используются только узловые точки: разностное отношение «вперед» определено как $\Delta_+ f_i = (f_{i+1} - f_i)/h$ и не берется в последней узловой точке.

Следует учесть еще одно обстоятельство, а именно неоднородность краевых условий. Для двухточечной краевой задачи непрерывная зависимость решения u от f *и краевых данных* выражается неравенством

$$\|u\|_2 \leqslant C(\|f\|_0 + |u(0)| + |u'(\pi)|). \qquad (17)$$

Для конечно-разностного уравнения, в котором Δ_π означает краевой оператор, примененный на правом конце интервала, со-

ответствующее неравенство будет

$$\|U^h\|_2 \leqslant C(\|f^h\|_0 + |U_0^h| + |\Delta_\pi U^h|). \tag{18}$$

При любом выборе разностных уравнений это основная оценка, и ее надо доказать. Она не вытекает автоматически из неравенства (17), которое тем не менее должно быть выполнено: *выполнение непрерывного неравенства необходимо, но не достаточно для выполнения дискретного неравенства.* В этом смысле теория разностных уравнений сложнее; существует множество разностных схем и каждая требует более или менее нового доказательства неравенства (18). Как и в непрерывной задаче, мы будем предполагать, что неравенство справедливо, и займемся его применением; техника доказательства таких неравенств в одномерном случае подробно рассмотрена Крайссом [К 7]. Неравенство (18) означает *устойчивость разностного уравнения*: дискретное решение U^h непрерывно зависит от данной функции f^h и равномерно по h.

Для обоснования сходимости U^h к u нам понадобится одна из самых важных теорем численного анализа: *аппроксимация* [1]) *и устойчивость влекут за собой сходимость.* Эта теорема доказывается в два этапа.

1. Ошибка E^h удовлетворяет тому же разностному уравнению (13), что и U^h, и потому по устойчивости непрерывно зависит от *своих* данных, представляющих собой не что иное, как локальную ошибку отсечения:

$$\|E^h\|_2 \leqslant C(\|\tau^h\|_0 + |E^h(0)| + |\Delta_\pi E^h|).$$

2. Так как существует аппроксимация, доказываемая с помощью разложений Тейлора функций τ^h, $E^h(0)$ и $\Delta_\pi E^h$, то правая часть стремится к нулю при $h \to 0$. Сходимость, таким образом, доказана.

Остановимся еще на одном вопросе. Так как τ^h включает четвертую производную от u, оценка ошибки при центральных разностях имеет вид $\|E^h\|_2 \leqslant Ch^2\|u\|_4$. Это неравенство после небольших преобразований сводится к

$$\|E^h\|_0 \leqslant C'h^2\|u\|_2.$$

Таким образом, порядок сходимости равен h^2, если u лежит в $\mathscr{H}^2$, или, другими словами, если f лежит в $\mathscr{H}^0$. Скорость сходимости здесь та же, что и для простейшего метода конечных элементов в разд. 1.6, но доказательство гораздо проще.

---

[1]) Когда говорят, что дифференциальное уравнение аппроксимируется разностным, подразумевают, что ошибка отсечения τ^h стремится к нулю при $h \to 0$. — *Прим. перев.*

Подводя итог, отметим, что для одномерных разностных уравнений удовлетворительная теория возможна, но она не тривиальна. Для многомерных задач, таких, как уравнения в частных производных, доказательство сходимости в литературе по численному анализу почти всегда основано на принципе максимума. Когда этот принцип не работает, для задач специального вида можно использовать некоторые соображения, но общая теория сейчас развивается в таком направлении, что, по-видимому, ее чрезвычайно трудно применить. Вся трудность в том, что одно дифференциальное уравнение допускает множество различных разностных аппроксимаций, особенно при криволинейных границах. В противовес этому вариационные методы подчиняются более строгим правилам, и именно эти ограничения позволяют сделать теорию более полной.

Мы будем заниматься исключительно этой теорией — построением и сходимостью конечных элементов.

1.5. МЕТОД РИТЦА И ЛИНЕЙНЫЕ ЭЛЕМЕНТЫ

Приступим, наконец, к изложению собственно метода конечных элементов. Общие рамки вопроса определены, и, как мы видим, есть выбор: можно либо аппроксимировать отдельные члены дифференциального уравнения, либо использовать присущий задаче вариационный принцип. Метод конечных элементов выбирает последнее. В то же время дискретные уравнения, возникающие при вариационной аппроксимации, являются разностными.

В вариационной форме задача заключается в минимизации квадратичного функционала

$$I(v)=\int_0^{\pi}[p(x)(v'(x))^2+q(x)(v(x))^2-2f(x)v(x)]\,dx$$

на бесконечномерном пространстве $\mathscr{H}_E^1$. *Метод Ритца состоит в замене* $\mathscr{H}_E^1$ *в вариационной задаче конечномерным подпространством* S, *или, точнее, последовательностью конечномерных подпространств* S^h, *содержащихся в* $\mathscr{H}_E^1$. Элементы v^h из S^h называются *пробными функциями*. Так как они принадлежат $\mathscr{H}_E^1$, они удовлетворяют главному краевому условию $v^h(0)=0$. На каждом подпространстве S^h минимизация функционала I приводит к решению системы линейных уравнений; число уравнений совпадает с размерностью подпространства S^h. *Аппроксимация Ритца — это функция* u^h, *минимизирующая* I *на подпространстве* S^h:

$$I(u^h)\leqslant I(v^h) \quad \text{для всех} \quad v^h\in S^h.$$

Основные задачи: 1) определить u^h, 2) оценить расстояние между u^h и истинным решением u. Этот раздел посвящен задаче 1), а следующий — задаче 2).

Начнем с двух примеров, иллюстрирующих классический метод Ритца. Подпространства S^h не будут иметь специального вида, связанного с методом конечных элементов; вместо этого каждое подпространство в последовательности $\{S^h\}$ будет содержать предыдущее.

Пусть сначала коэффициенты p и q постоянны, а S^h — подпространство, натянутое на первые $N = 1/h$ собственных функций дифференциальной задачи. Пробные функции, т. е. элементы из S^h, будут тогда линейными комбинациями вида

$$v^h(x) = \sum_1^N q_j \varphi_j(x) = \sum_1^N q_j \sqrt{\frac{\pi}{2}} \sin(j - {}^1/_2)x.$$

Ясно, что эти функции принадлежат $\mathscr{H}_E^1$; они удовлетворяют даже естественному краевому условию $(v^h)'(\pi) = 0$, что вовсе и не обязательно.

Весовые коэффициенты q_j должны быть определены из условия минимума I. Так как собственные функции ортогональны, интеграл имеет вид

$$I(v^h) = \sum_1^N \left[q_j^2 \lambda_j - 2 \int_0^\pi f q_j \varphi_j \, dx \right],$$

где $\lambda_j = p(j - {}^1/_2)^2 + q$. Условия минимума $\partial I/\partial q_j = 0$ сразу приводят к оптимальным значениям весов:

$$Q_j = \frac{1}{\lambda_j} \int_0^\pi f \varphi_j \, dx = \frac{(f, \varphi_j)}{\lambda_j}, \quad j = 1, \ldots, N.$$

Поэтому аппроксимацией Ритца будет функция

$$u^h = \sum_1^N \frac{(f, \varphi_j)\varphi_j}{\lambda_j}.$$

В этом случае система линейных уравнений $\partial I/\partial q_j = 0$, определяющая оптимальные координаты Q_j, решается очень просто; матрица системы оказалась диагональной, так как собственные функции ортогональны. Более того, *u^h представляет собой проекцию точного решения*

$$u = \sum_1^\infty \frac{(f, \varphi_j)\varphi_j}{\lambda_j}$$

на подпространство, натянутое на первые N собственных функций.

Легко подсчитать, какой получается выигрыш на каждом шаге в последовательности аппроксимаций Ритца. Когда N-я пробная функция φ_N включается в рассмотрение, u^h улучшается добавлением члена $\lambda_N^{-1}(f, \varphi_N)\varphi_N$. Эта добавка при $N \to \infty$ очень быстро стремится к нулю для гладких функций f, но даже для произвольной функции f она не превышает $\lambda_N^{-1}\|f\|_0$. В реальной задаче собственные функции точно не известны, но если геометрия области остается простой, все еще огромное значение имеет использование синусов и косинусов в качестве пробных функций. (Орсзаг и другие авторы показали, как с помощью быстрого преобразования Фурье сохранить в разумных пределах объем вычислений.) Для областей со сложной геометрией мы предпочитаем конечные элементы.

Пусть p и q остаются постоянными, а в качестве пробных функций теперь возьмем полиномы

$$v^h(x) = q_1 x + q_2 x^2 + \ \dots\ + q_N x^N.$$

Такие функции опять удовлетворяют главному краевому условию $v^h(0) = 0$, но естественное краевое условие нарушается. В этом случае

$$I(v^h) = \int_0^\pi \left[p\left(\sum q_j j x^{j-1}\right)^2 + q\left(\sum q_j x^j\right)^2 - 2f\sum q_j x^j\right] dx.$$

Дифференцируя I по параметрам q_j, находим систему N линейных уравнений относительно оптимальных параметров $Q_1, \dots, Q_N$:

$$KQ = F.$$

Неизвестный вектор здесь $Q = (Q_1, \dots, Q_N)$, компоненты вектора F являются «моментами» $F_j = \int f x^j\, dx$, а элементы матрицы коэффициентов K таковы:

$$K_{ij} = \int_0^\pi [p(ix^{i-1})(jx^{j-1}) + qx^i x^j]\, dx =$$

$$= \frac{pij\pi^{i+j-1}}{i+j-1} + \frac{q\pi^{i+j+1}}{i+j+1}.$$

Матрица, элементы которой имеют вид $(i+j+1)^{-1}$, известна как матрица Гильберта; с ней производить вычисления совершенно невозможно. Ее собственные значения настолько несоиз-

меримы — она так плохо обусловлена, — что уже при $N \approx 6$ ошибки округления «забивают» правую часть f. Если в K включены другие члены, ситуация еще хуже. Трудность в том, что степени x^j почти линейно зависимы; в окрестности точки $x = \pi$ все они имеют нулевые веса. Численная устойчивость зависит от выбора «более независимого» базиса в подпространстве.

Для получения такого базиса надо *сначала* ортогонализовать исходные пробные функции $\varphi_j = x^j$. В большинстве случаев можно в качестве нового базиса для пространства полиномов взять полиномы Лежандра или Чебышева. Такой базис удобен на интервале и даже на областях большей размерности, если только геометрия области очень проста. Для областей общего вида, однако, эти ортогональные полиномы опять становятся неработоспособными.

Вернемся к построению подпространств S^h в методе конечных элементов. Область — в нашем случае интервал $[0, \pi]$ — разбивается на отрезки, и на каждом из них в качестве пробных функций v^h берутся полиномы. В узлах между отрезками требуется некоторая степень непрерывности, и обычно эта непрерывность не более чем требуют следующие условия:

1. Функции v^h должны быть допустимы для вариационного принципа задачи.

2. Величины, представляющие физический интерес, такие, как перемещения, нагрузки или моменты, должны удобно вычисляться из приближенного решения u^h.

Очень трудно построить кусочно полиномиальные функции, если требовать слишком большую гладкость в узловых точках между отрезками.

В нашем примере допустимым пространством $\mathscr{H}_E^1$ будет пространство непрерывных функций. Кусочно постоянные функции сразу отбрасываются. Поэтому проще всего взять в качестве S^h *множество функций, линейных на каждом интервале* $[(j-1)h, jh]$, *непрерывных в узлах* $x = jh$ *и равных нулю при* $x = 0$. Производная от такой функции кусочно постоянна и обладает, очевидно, конечной энергией: таким образом, S^h — подпространство пространства $\mathscr{H}_E^1$. Такие пробные функции мы будем называть *линейными элементами*.

Пусть функция φ_j^h принадлежит S^h, равна 1 в узле $x = jh$ и 0 во всех остальных узлах, $j = 1, \ldots, N$ (рис. 1.4). Такие *функции-крышки* образуют базис в подпространстве S^h, так как любой элемент из S^h можно записать в виде

$$v^h(x) = \sum_1^N q_j \varphi_j^h(x).$$

Отметим важный факт, касающийся коэффициента q_j: он совпадает со значением v в j-м узле $x = jh$. Поскольку координаты q_j — это не что иное, как узловые значения функции, оптимальные координаты Q_j будут иметь непосредственный физический смысл: они будут аппроксимациями Ритца для перемещения струны в узлах. Отметим также, что φ_j^h образуют *локальный базис*, так как каждая функция φ_j^h отлична от нуля только в области диаметра $2h$. В самом деле, очевидно, что функции

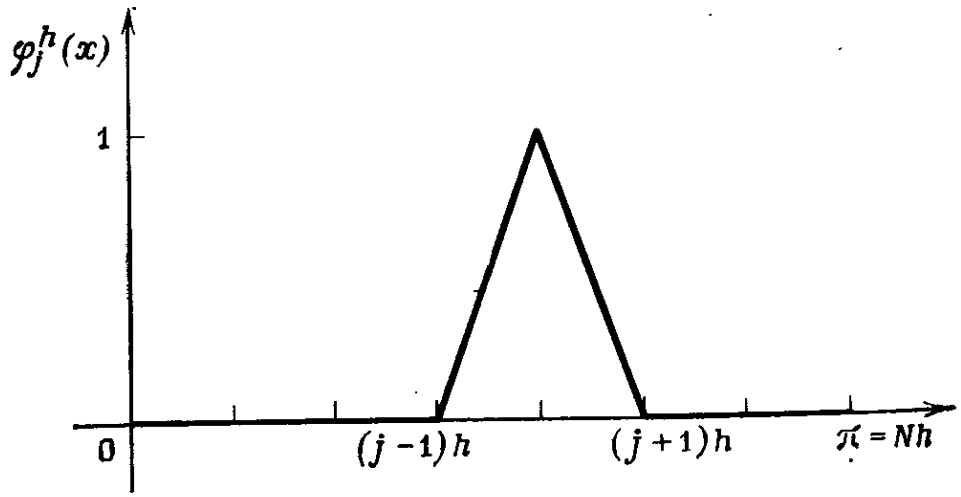


Рис. 1.4.
Кусочно линейные базисные функции.

φ_{j-1}^h и φ_{j+1}^h *ортогональны*, так как там, где одна из них отлична от нуля, равна нулю другая. Таким образом, хотя базис не совсем ортогонален, только соседние элементы в произведении дают не нуль.

С нормализованными коэффициентами $p = q = 1$ задача состоит в минимизации функционала

$$I(v^h) = \int_0^\pi [((v^h)')^2 + (v^h)^2 - 2fv^h]\, dx.$$

При $v^h = \sum q_j \varphi_j^h$ этот интеграл является квадратичной функцией координат $q_1, q_2, \ldots, q_N$ и его можно вычислять на одном каком-нибудь подынтервале. На j-м подынтервале функция v^h линейна, $v^h((j-1)h) = q_{j-1}$, $v^h(jh) = q_j$ и $(v^h)' = (q_j - q_{j-1})/h$ (будем считать, что $q_0 = 0$). Поэтому

$$\int_{(j-1)h}^{jh} ((v^h)')^2\, dx = \frac{(q_j - q_{j-1})^2}{h}.$$

Чуть более длинное вычисление дает

$$\int_{(j-1)h}^{jh} (v^h)^2\,dx = \frac{h}{3}\left(q_j^2 + q_j q_{j-1} + q_{j-1}^2\right).$$

Эти члены сответствуют отдельному куску струны с линейным изменением перемещения. Для всей струны член второго порядка в $I(v^h)$ равен

$$\int_0^\pi p\,((v^h)')^2 + q\,(v^h)^2 = \sum_1^N \frac{p\left(q_j - q_{j-1}\right)^2}{h} + \frac{qh\left(q_j^2 + q_j q_{j-1} + q_{j-1}^2\right)}{3}.$$

Это не слишком удобная форма записи результата. Предпочтительнее получить результат в матричном виде q^TKq (или (Kq, q)), так как матрица K нам еще понадобится. Причина в том, что выражение для $I(v^h)$ квадратично относительно параметров $q = (q_1, q_2, \ldots, q_N)$ и имеет вид

$$I(v^h) = q^TKq - 2F^Tq.$$

Минимум такого выражения достигается (как мы знаем из разд. 1.3, где мы положили $\partial I/\partial q_m = 0$) на векторе $Q = (Q_1, \ldots, Q_N)$, определяемом системой

$$KQ = F.$$

Именно эту систему и нужно решить, поэтому все, что нам нужно знать, — это матрицу K и вектор F.

Наилучший способ — найти вклад в матрицу K каждого «элемента», т. е. каждого куска струны. Для этого вернемся к формуле

$$\int_{(j-1)h}^{jh} ((v^h)')^2 = \frac{(q_j - q_{j-1})^2}{h}$$

и запишем правую часть (величину интеграла) в матричной форме

$$(q_{j-1}q_j)\,\frac{1}{h}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} q_{j-1} \\ q_j \end{pmatrix} = (q_{j-1}q_j)\,k_1\begin{pmatrix} q_{j-1} \\ q_j \end{pmatrix}.$$

Матрица k_1 называется *матрицей жесткости элемента*. Ее достаточно вычислить один раз, так как она не зависит от дифференциального уравнения. Точно так же вычисление члена $\int (v^h)^2\,dx$

нулевого порядка на отдельном элементе проводится один раз и в результате получается *матрица массы элемента*:

$$\frac{h}{3}(q_{j-1}^2 + q_{j-1}q_j + q_j^2) = (q_{j-1}q_j)\frac{h}{6}\begin{pmatrix}2 & 1\\ 1 & 2\end{pmatrix}\begin{pmatrix}q_{j-1}\\ q_j\end{pmatrix} =$$

$$= (q_{j-1}q_j)\,k_0\begin{pmatrix}q_{j-1}\\ q_j\end{pmatrix}.$$

Далее суммирование по всем элементам $j = 1, \ldots, N$ заменяется *построением глобальной матрицы жесткости* K. Это означает, что после соответствующего расположения в позициях глобального массива матрицы элементов складываются. Матрица, связанная с $\int_0^\pi ((v^h)')^2\,dx$, при условии, что неизвестное q_0 отброшено (так как оно определяется главным краевым условием), имеет вид

$$K_1 = \frac{1}{h}\begin{pmatrix}1 & & & \\ & & & \\ & & & \\ & & & \end{pmatrix} + \frac{1}{h}\begin{pmatrix}1 & -1 & & \\ -1 & 1 & & \\ & & & \\ & & & \end{pmatrix} + \ldots + \frac{1}{h}\begin{pmatrix} & & & \\ & 1 & -1 & \\ & -1 & 1 & \\ & & & \end{pmatrix} =$$

$$= \frac{1}{h}\begin{pmatrix}2 & -1 & 0 & 0 & .\\ -1 & 2 & -1 & . & 0\\ 0 & -1 & . & -1 & 0\\ 0 & . & -1 & 2 & -1\\ . & 0 & 0 & -1 & 1\end{pmatrix}.$$

Снова матрица и интеграл связаны формулой

$$\int_0^\pi ((v_h)')^2\,dx = (q_1 \ldots q_N)\,K_1\begin{pmatrix}q_1\\ .\\ .\\ .\\ q_N\end{pmatrix} = q^T K_1 q.$$

Интеграл от $(v^h)^2$ равен $q^T K_0 q$, где матрица массы K_0 (в дальнейшем обозначаемая через M) строится точно так же:

$$K_0 = \frac{h}{6}\begin{pmatrix}4 & 1 & 0 & 0 & .\\ 1 & 4 & 1 & . & 0\\ 0 & 1 & . & 1 & 0\\ 0 & . & 1 & 4 & 1\\ . & 0 & 0 & 1 & 2\end{pmatrix}.$$

Требуемая матрица K будет суммой $K_1 + K_0$. Эти матрицы не нужно все сразу строить и хранить; вместо них можно вычислить матрицы элементов, когда они понадобятся при решении итоговой системы $KQ = F$.

Осталось вычислить член $\int fv^h$, с которым неоднородная нагрузка f входит в аппроксимацию. Этот интеграл линеен относительно координат q_j:

$$\int_0^\pi fv^h\,dx = \sum_1^N q_j \int_0^\pi f\varphi_j^h\,dx = F^T q;$$

здесь вектор нагрузок F имеет координаты

$$F_j = \int_0^\pi f\varphi_j^h\,dx.$$

На практике эти величины вычисляются так же, как матрицы жесткости и массы, интегрированием сразу только на одном элементе. Пусть на j-м интервале

$$\int_{(j-1)h}^{jh} fv^h = \alpha_j q_{j-1} + \beta_j q_j.$$

Когда такие интегралы суммируются, коэффициент при q_k равен $F_k = \beta_k + \alpha_{k+1}$. Отсюда ясно, какова простейшая форма хранения в памяти ЭВМ. Заданный узел kh является правым концом в k-м подынтервале — и это дает $\beta_k q_k$ в интеграле, а также левым концом в следующем подынтервале — и это дает $\alpha_{k+1} q_k$. Подпрограмма, строящая вектор, должна учитывать, что оба подынтервала стыкуются в k-м узле, и комбинировать результаты. (Нет сомнений, что такое построение метода конечных элементов сделает программу длиннее программы метода конечных разностей для двухточечной краевой задачи; преимущества метода конечных элементов опять проявляются только на задачах размерности > 1.)

Для произвольной функции f интегралы нельзя подсчитать точно, и нужно использовать численные квадратурные формулы. Одна возможность состоит в *аппроксимации f линейной интерполяцией по узлам.* Другими словами, f заменяется кусочно линейным интерполянтом $f_I = \sum f_k \varphi_k^h(x)$, где f_k — значение f в узле $x = kh$. Тогда при вычислении интеграла $\int f_I v^h$ возникают *те же*

самые скалярные произведения $\int \varphi_k^h \varphi_j^h$, *что и ранее при построении матрицы массы* k_0. На j-м интервале

$$\int_{(j-1)h}^{jh} f_I v^h \, dx = (f_{I-1} f_I)\, k_0 \begin{pmatrix} q_{I-1} \\ q_I \end{pmatrix},$$

так как единственное различие между вычислением этого интеграла и $\int (v^h)^2 \, dx$ заключается в замене пары коэффициентов q_j и q_{j-1} узловыми значениями f.

Снова, суммируя на ЭВМ эти результаты от $j = 1$ до $j = N$, получаем матрицу массы

$$\int_0^\pi f_I v^h dx = (f_0 \ldots f_N)\, \tilde{K}_0 \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ \cdot \\ q_N \end{pmatrix}.$$

$\tilde{K}_0$ отличается от K_0 только наличием нулевой строки, так как, вообще говоря, $f(0) \neq 0$.

Этот интеграл приближает точный линейный член $F^T q$, который можно было бы получить, выполняя точное интегрирование. Обозначим приближение через $\tilde{F}^T q$. Матрица массы K_0 дает коэффициенты в аппроксимации вектора нагрузок $\tilde{F}$:

$$\tilde{F}_I = \frac{h}{6} [f((j-1)h) + 4f(jh) + f((j+1)h)].$$

Разность между $\tilde{F}_I$ и точным значением $F_I = \int f \varphi_I^h$ нетрудно оценить. Если функция f линейна, коэффициенты $\tilde{F}_I$ и F_I совпадают, поскольку f_I совпадает с f. Для квадратичной функции f это не так. Если $f(x) = x^2$, то

$$F_I = \int_{(j-1)h}^{jh} x^2 \left(-j + 1 + \frac{x}{h}\right) dx + \int_{jh}^{(j+1)h} x^2 \left(j + 1 - \frac{x}{h}\right) dx = j^2 h^3 + \frac{h^3}{6},$$

в то время как

$$\tilde{F}_I = \frac{h^3}{6} [(j-1)^2 + 4j^2 + (j+1)^2] = j^2 h^3 + \frac{h^3}{3}.$$

Поэтому ошибка численного интегрирования равна $h^3/6 = h^3 f''(jh)/12$. Для произвольной гладкой функции f — это главный член в $\tilde{F}_j - F_j$.

Существует формула интегрирования, столь же простая, как и формула для $\bar{F}_j$, точная на квадратичных (и даже кубических) членах:

$$\bar{F}_j = \frac{h(f_{j-1} + 10f_j + f_{j+1})}{12}.$$

Она выведена Коллатцем (см. [Г2]). Ее можно получить с помощью *квадратичной* интерполяции f на двойном интервале $[(j-1)h,\ (j+1)h]$ и последующего точного интегрирования $\bar{F}_j = \int f_I \varphi_j^h\, dx$. Однако для программ метода конечных элементов этот вывод неестествен. Вместо того, чтобы один раз вычислить интеграл на $[jh,\ (j+1)h]$, надо пройти по каждому интервалу дважды: сначала использовать квадратичную интерполяцию по узлам x_{j-1}, x_j, x_{j+1} и вычислить $\bar{F}_j$, а затем — по узлам x_j, x_{j+1}, x_{j+2} и вычислить $\bar{F}_{j+1}$. Это типичная ситуация: наиболее эффективная формула в определенном классе не обнаруживается при применении метода конечных элементов; если в каждом специальном случае предоставить полную свободу выбора наилучшей формулы, то могут оказаться предпочтительнее конечные разности. Важно, что при решении сложных задач формула метода конечных элементов «почти» оптимальна и просто и дешево реализуется на ЭВМ.

На практике вместо замены функции f ее интерполянтом f_I часто проводится *прямое численное интегрирование*. На каждом подынтервале fv^h интегрируется по стандартной квадратурной формуле

$$\int fv^h = \sum w_i f(\xi_i) v^h(\xi_i).$$

Чаще всего выбирается весьма эффективная квадратура Гаусса, например с равными весами w_i и двумя симметрично расположенными узлами $\xi_i = (j + 1/2 \pm 1/\sqrt{3})h$. Эта формула точна для кубических функций f и обеспечивает автоматически ту же точность, что и формула Коллатца, описанная выше. *Одноточечная* формула Гаусса с точкой ξ_i, взятой в середине каждого интервала, уже обеспечивает точность, присущую линейным пробным функциям:

$$\tilde{F}_j = \frac{h(f_{j-1/2} + f_{j+1/2})}{2}.$$

(Формула трапеций с равноотстоящими узлами дает в точности правую часть $\bar{F}_j = hf(jh)$ обычного трехточечного разностного уравнения, полученного в разд. 1.4.)

В результате любого численного интегрирования точные линейные члены $F^T q$ заменяются некоторым приближенным выра-

жением $\tilde{F}^T q$, все еще линейным относительно неизвестных $q_1, \ldots, q_N$.

Те же идеи применяются к квадратичным членам, если коэффициенты $p(x)$ и $q(x)$ в дифференциальном уравнении действительно зависят от x. Интегралы от $p(x)((v^h)')^2$ и $q(x)(v^h)^2$ опять вычисляются на каждом подынтервале с помощью численной квадратуры. Итоговые результаты хранятся в виде приближенных матриц, квадратичные формы которых близки к точным интегралам $q^T K_1 q$ и $q^T K_0 q$. Мы будем предполагать, что все интегралы вычислены точно, а в разд. 4.3 изучим влияние ошибок, возникающих при численном интегрировании.

Теперь попытаемся суммировать наши идеи. Обозначая через K сумму $K_1 + K_0$ и производя описанные вычисления, получаем

$$I(v^h) = I\left(\sum q_j \varphi_j^h\right) = q^T K q - 2F^T q.$$

Это дискретное выражение, подлежащее минимизации в методе Ритца. Сразу видно, что это стандартная вариационная форма; минимизирующий вектор Q определяется линейным уравнением

$$KQ = F.$$

Назовем его *уравнением метода конечных элементов.* Решение этого уравнения — центральная часть всех вычислений, и если число h мало, порядок системы будет большим. Матрица K с гарантией положительно определена и поэтому обратима; так как $p > 0$, то величина

$$q^T K q = \int p(x)\left(\sum q_j \varphi_j'\right)^2 + q(x)\left(\sum q_j \varphi_j\right)^2$$

может равняться нулю только при $\sum q_j \varphi_j' \equiv 0$, а это возможно только тогда, когда $q_j = 0$ при всех j.

Глобальная матрица жесткости K похожа на матрицу метода конечных разностей L^h, вернее на hL^h из предыдущего раздела. При постоянных коэффициентах главные члены у них совпадают, обе пропорциональны вторым разностям с весами $-1, 2, -1$. Член нулевого порядка qu входит только в диагональные элементы матрицы L^h. А вот в матрице K этот член проявляется в связях между соседними неизвестными и «сглажен» с весами 1, 4, 1, возникающими из формулы Симпсона. Подчеркнем еще раз, что *как только выбрано аппроксимирующее подпространство S^h, дискретная форма каждого члена уравнения полностью определена.* Метод Ритца действует сразу на все уравнение и не требует от пользователя принятия неза-

висимых решений при аппроксимации различных членов уравнения.

В частности, упрощается работа с краевыми условиями, и, учитывая различные возможности, возникающие в разностных уравнениях, можно только удивляться тому, что порядок точности некоторых краевых условий выбирается «автоматически» с помощью уравнения метода конечных элементов. Используя последнюю строку матрицы K, запишем уравнение в граничной точке $x = \pi$ в случае $p = q = 1$:

$$\frac{-q_{N-1}+q_N}{h}+\frac{h}{6}(q_{N-1}+2q_N)=\int f\varphi_N^h=\int_{\pi-h}^{\pi} f(y)\left(1+\frac{y-\pi}{h}\right)dy.$$

Подставляя точное значение $u(x_j)$ вместо q_j и разлагая u и f в ряд Тейлора в точке $x = \pi$, находим ошибку отсечения

$$u'-\frac{h}{2}u''+\frac{h^2}{6}u'''-\frac{h^3}{24}u^{(\mathrm{IV})}+\frac{h}{6}\left(3u-hu'+\frac{h^2}{2}u''\right)-$$
$$-\frac{h}{2}f+\frac{h^2}{6}f'-\frac{h^3}{24}f''+\ldots\approx\frac{h^3}{24}u''(\pi)+\ldots.$$

(Мы воспользовались дифференциальным уравнением $-u''+u=f$ и краевым условием $u'(\pi)=0$.) В терминах разностных уравнений это означает, что краевое условие имеет третий порядок точности.

Последний шаг при вычислении аппроксимации u^h метода конечных элементов — решение линейной системы $KQ = F$. Мы собираемся обсудить здесь только прямые методы исключения, так как в подавляющем большинстве программ по методу конечных элементов они предпочтительнее итерационных методов. (Было бы интересно обсудить возникновение и упадок итерационных методов за последние десятилетия. Очень много сложных и математически интересных работ было посвящено методам верхней релаксации и переменных направлений; они составляли основную тему численного анализа. Теперь они вытеснены методом исключений и методами типа быстрого преобразования Фурье. Методы типа метода Фурье, безусловно, эффективнее, если геометрия области и уравнения подходящие, а методы исключения особенно хороши, когда нужно решать одну систему с многими правыми частями, как в задачах проектирования.)

Обсудим кратко теорию исключения Гаусса. При применении этого знакомого алгоритма к матрице K общего вида сначала исключается Q_1 из последних $N-1$ уравнений, затем Q_2 из последних $N-2$ уравнений и т. д.; наконец, исключается Q_{N-1} из последнего уравнения. Система $KQ = F$ преобразуется

в эквивалентную систему

$$UQ = \begin{Vmatrix} U_{11} & \cdot & \cdot & \cdot & U_{1N} \\ & U_{22} & \cdot & & U_{2N} \\ & & \cdot & \cdot & \cdot \\ & & & & U_{NN} \end{Vmatrix} \begin{Vmatrix} Q_1 \\ Q_2 \\ \cdot \\ \cdot \\ \cdot \\ Q_N \end{Vmatrix} = F'.$$

После этого неизвестные Q_j определяются с помощью *обратной подстановки*: последнее уравнение решается относительно Q_N, после подстановки значения Q_N решается предпоследнее уравнение относительно Q_{N-1} и т. д.

Важно понять, что происходит, в терминах матриц. Предположим, что мы ведем процесс, обратный исключению, прибавляя к последнему уравнению $(N-1)$-е с множителем $l_{N,\,N-1}$ — в прямом процессе исключения Q_{N-1} оно вычиталось. Далее прибавим к последним двум уравнениям $(N-2)$-е с множителями $l_{N-1,\,N-2}$, $l_{N,\,N-2}$ — при исключении Q_{N-2} оно вычиталось. В конце концов восстанавливается исходная система $KQ = F$; к i-му уравнению прибавляется предыдущее, умноженное на l_{ij}, $j = 1, \ldots, i-1$, — они вычитались при исключении неизвестных $Q_1, \ldots, Q_{i-1}$. В терминах матриц *система $KQ = F$ восстанавливается умножением преобразованной системы $UQ = F'$ на матрицу*

$$L = \begin{Vmatrix} 1 & & & & \\ l_{21} & 1 & & & \\ \cdot & \cdot & 1 & & \\ \cdot & \cdot & \cdot & & \\ \cdot & \cdot & l_{N-1,\,N-2} & 1 & \\ l_{N1} & \cdot & l_{N,\,N-2} & l_{N,\,N-1} & 1 \end{Vmatrix}.$$

Это означает, что $LUQ = LF'$ равносильно $KQ = F$. *Исключение Гаусса есть не что иное, как разложение K в произведение*

$$K = LU$$

нижней треугольной и верхней треугольной матриц. Таким образом, решение $K^{-1}F$, которое мы ищем, есть не что иное, как $U^{-1}L^{-1}F$, и треугольные матрицы L и U легко обращаются. Действительно, $L^{-1}F = F'$ — правая часть системы после исключений и $Q = U^{-1}F'$ — результат обратной подстановки. (Если нужно решить много систем $KQ = F_n$ с различными правыми частями, но с одной и той же матрицей жесткости K, сомножители L и U следует хранить.)

Рассмотрим процесс исключения, когда известно, как в нашем случае, что матрица K симметрична, положительно определена и трехдиагональна. Прежде всего заметим, что процесс достигает цели: исключения можно осуществить и разложение $K = LU$ существует. Процесс нельзя провести, например, в случае матрицы $K = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, поскольку Q_1 нельзя исключить из второго уравнения, используя первое. Условие возможности провести процесс исключения таково: каждая матрица в левом верхнем углу матрицы K, т. е.

$$K^{(1)} = (K_{11}), \quad K^{(2)} = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix}, \ldots,$$

должна иметь ненулевой определитель. Для положительно определенной матрицы все эти определители положительны, и потому процесс осуществим без перестановки строк. Действительно, определитель матрицы $K^{(j)}$ равен произведению $U_{11}U_{22}\ldots U_{jj}$ и все элементы U_{jj}, лежащие на главной диагонали матрицы U, положительны.

Для численной устойчивости алгоритма исключения требуется еще, чтобы элемент U_{jj} был не только ненулевым, но и достаточно большим. Если алгоритм неустойчив, то можно потерять всю информацию об исходных коэффициентах K_{ij}. Мы осуществляем только частичный контроль за размером элементов U_{ij}, так что неизбежны некоторые ошибки округления. Внутренняя чувствительность матрицы K к малым возмущениям определяется ее *числом обусловленности*, примерно равным отношению наибольшего собственного значения к наименьшему. Этот вопрос обсуждается в гл. 5. Число обусловленности зависит от размера шага h и от порядка дифференциального уравнения. Вычислительные трудности иногда возникают не из-за плохой обусловленности матрицы K, а из-за неудачно выбранного алгоритма. В матрице типа $K = \begin{pmatrix} \varepsilon & 1 \\ 1 & 0 \end{pmatrix}$, например, следует сначала поменять уравнения, или что то же самое, выбрать главный элемент. Если каждый раз менять строки так, чтобы элемент U_{jj} был максимальным, алгоритм исключения Гаусса станет настолько устойчивым, насколько позволяет число обусловленности. «Неприятная» матрица K в нашем примере, разумеется, не является положительно определенной; такая ситуация типична для матриц, возникающих в методе «смешанного типа» (разд. 2.3).

Прямой метод для матриц жесткости (в котором неизвестные представляют собой перемещения u, как в большей части

этой книги) автоматически приводит к положительно определенной матрице K. В этом случае *исключение Гаусса не только возможно без перестановки строк, но и численно устойчиво.* Чтобы понять почему, приведем разложение $K = LU$ к более симметричной форме, выделяя диагональ матрицы U:

$$D = \begin{pmatrix} U_{11} & & & \\ & U_{22} & & \\ & & \ddots & \\ & & & U_{NN} \end{pmatrix}.$$

Ясно, что $K = LD(D^{-1}U)$, причем все три сомножителя определены однозначно: L — нижняя треугольная матрица с единичной диагональю, $D^{-1}U$ — верхняя треугольная матрица с единичной диагональю и D — диагональная матрица с положительными элементами. По симметрии матрица $D^{-1}U$ должна быть транспонированной к L. Итак, $K = LDL^T$ — получили симметричную форму разложения. Можно даже пойти дальше и ввести новую нижнюю треугольную матрицу $\tilde{L} = LD^{1/2}$; это даст разложение Холесского $K = \tilde{L}\tilde{L}^T$. Теперь можно объяснить, почему положительно определенная симметричная матрица не требует выбора главного элемента: множитель $\tilde{L}$ в разложении ведет себя как корень квадратный из K и число обусловленности в точности равно числу обусловленности этого корня. Это можно сравнить с разложением

$$\begin{pmatrix} \varepsilon & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \varepsilon^{-1} & 1 \end{pmatrix} \begin{pmatrix} \varepsilon & 0 \\ 0 & -\varepsilon^{-1} \end{pmatrix} \begin{pmatrix} 1 & \varepsilon^{-1} \\ 0 & 1 \end{pmatrix},$$

множители в правой части которого велики, даже когда матрица в левой части не велика.

Наконец, трехдиагональность матрицы K приводит к серьезному уменьшению объема вычислений. Так как первое неизвестное Q_1 исключается лишь из второго уравнения (в других его нет), все множители $l_{3,1}, \ldots, l_{N-1,1}$, необходимые для исключения Q_1, равны нулю. Неизвестное Q_2 исключается лишь из третьего уравнения и т. д. Таким образом, ненулевые элементы нижней треугольной матрицы L расположены лишь на главной диагонали и на первой поддиагонали; трехдиагональная матрица разлагается на двухдиагональные.

Так как разложение Холесского требует извлечения квадратных корней из главных элементов и приводит к некоторым дополнительным проблемам, связанным с необходимостью избежать операций с нулями, наиболее популярный численный

алгоритм основан на разложении $K = LDL^T$. Разложение Холесского и такое разложение эквивалентны как математически, так и с точки зрения устойчивости. Для трехдиагональной матрицы элементы матриц L и D удовлетворяют простой рекуррентной формуле

$$d_j = K_{j,j} - d_{j-1} l_{j,j-1}^2, \quad d_0 = 0,$$

$$l_{j+1,j} = \frac{K_{j+1,j}}{d_j}.$$

Соответственно вектор $F' = L^{-1}F$ и решение Q, вычисляемое из обратной подстановки, равны

$$F'_j = F_j - F'_{j-1} l_{j,j-1}. \quad F_0 = 0,$$

$$Q_j = \frac{F'_j}{d_j} - Q_{j+1} l_{j+1,j}. \quad Q_{N+1} = 0.$$

Важно, что число арифметических операций здесь пропорционально N.

В многомерных задачах матрица жесткости K также будет разреженной и симметричной положительно определенной. Однако здесь уже нет такого очевидного упорядочения неизвестных, как в одномерном случае. *Способ упорядочения узлов приобретает основное значение*, и разрабатываются подпрограммы упорядочения, приводящего к минимальным затратам при исключении Гаусса.

Наиболее простым и популярным критерием выбора порядка узлов является ширина ленты матрицы. Пусть известно, что только первые w поддиагоналей и первые w наддиагоналей в матрице K ненулевые (для трехдиагональных матриц по этому определению $w = 1$). На каждом шаге процесса исключения эта информация должна использоваться наилучшим образом. Ниже каждого главного элемента может быть только w ненулевых элементов, подлежащих исключению; более того, каждое исключение — это вычитание из одной строки другой с некоторым множителем, а мы знаем, что в строке не более $2w + 1$ ненулевых элементов. Ленточная структура в процессе исключения сохраняется и переносится в сомножители L и U. Для симметричной матрицы число операций равно примерно $Nw^2/2$ вместо $N^3/3$ для плотной матрицы.

В качестве простейшей иллюстрации хорошего и плохого упорядочения узлов, с точки зрения ширины ленты, рассмотрим двумерный случай при прямоугольном расположении узлов. Если по горизонтали узлов меньше, чем по вертикали, то неизвестные нумеруем последовательно вдоль строк, а не столбцов.

Ширина ленты равна примерно длине строки, так как для данного узла связь с узлом, лежащим выше него (с которым данный узел связан ненулевым элементом в матрице жесткости), при таком упорядочении проявится быстрее. В общем случае матрицы в методе конечных элементов куда менее систематичны, чем в случае разностных уравнений, и выбор оптимального упорядочения совсем не очевиден.

Существует другой критерий, учитывающий разреженность матриц; он чуть точнее, чем ширина ленты. Он основан на *профиле* (очертании) матрицы. Возьмем первый ненулевой элемент в i-й строке. Если он оказался в j-м столбце и об этом «известно» ЭВМ, то не обязательно вычитать строки с номерами $1, 2, \ldots, j-1$ из i-й строки. Соответствующие множители $l_{i,1}$, $l_{i,2}, \ldots, l_{i,j-1}$ будут равны нулю, так как неизвестные $Q_1, Q_2, \ldots$ $\ldots, Q_{j-1}$ не надо исключать из i-го уравнения: их уже нет. Профиль матрицы формируется при определении этих первых ненулевых элементов в каждой строке, и, подобно ленточной структуре, профиль сохраняется в процессе исключения Гаусса и переходит без изменения в множитель L. Профиль как бы вклинивается в ленту, так что длина многих строк гораздо меньше $2w+1$, и очень полезно хранить профиль в ЭВМ и даже упорядочивать неизвестные в соответствии с алгоритмом, ориентированным на профиль.

Отметим, что число арифметических операций — не единственный критерий выбора алгоритма; по крайней мере столь же важной может оказаться потребность в оперативной памяти. Для ленточной матрицы стандартная процедура требует хранения диагоналей матрицы; эта ситуация близка к оптимальной, и для нее надо порядка Nw ячеек. Для линейных или билинейных элементов на прямоугольной сетке 50×50 число N равно 2500 и $w \approx 50$. Современная большая ЭВМ позволяет хранить информацию за пределами оперативной памяти, но программирование и обмен данными становятся гораздо сложнее. Поэтому большее внимание следует уделять алгоритмам, учитывающим и использующим, где возможно, разреженность матрицы даже внутри ленты или профиля. В крайнем случае можно даже запоминать положение каждого ненулевого элемента матрицы A и порядок неизвестных, как в «алгоритме для разреженной матрицы», чтобы минимизировать число ненулевых элементов в нижней треугольной матрице L. Нам кажется, правда, что для матриц метода конечных элементов это слишком дорого; в нем иногда трудно учесть систематическую структуру матриц.

Если задача так велика, что стандартный алгоритм, основанный на ленте матрицы или ее профиле, не помещается в оперативную память, мы предпочитаем следовать циклу статей

Алана Джорджа. Он задался такой целью: для конечных элементов с N неизвестными параметрами на плоскости добиться $O(N^{3/2})$ арифметических операций при хранении $O(N \log N)$ ненулевых элементов матрицы L. Это было бы действительно хорошо. (Существует несколько специальных прямых методов, аналогичных быстрому преобразованию Фурье, требующих только $O(N \log N)$ арифметических операций при общем объеме памяти $O(N)$. Однако применение этих методов ограничивается простыми задачами на прямоугольниках.) Такие цифры достигаются при упорядочении [Д7], напоминающем алгоритм минимальных степеней: на каждом шаге неизвестное, которое исключается, должно быть связано только с несколькими неизвестными. Метод значительно отличается от прямого метода Айронса и связан с большими затратами при составлении программы, возможно, даже слишком большими. Последние предложения по этому методу даны в статье Джорджа «An efficient band-oriented scheme for solving n by n grid problems». Он разбивает область на узкие полосы и применяет ленточный алгоритм для соответствующих подматриц (неизвестные внутри полосы занумерованы так, чтобы уменьшить ширину ленты). Между двумя полосами остаются неизвестные, расположенные на прямой, и в упорядочении Джорджа они нумеруются только *после неизвестных в полосах, выделенных прямыми.* На этих прямых сравнительно мало неизвестных, которые вносят вклад в расширение ленты матрицы. Большие подматрицы, соответствующие связи одной полосы с другой, пусты. С количеством полос, пропорциональным $h^{-1/2}$, потребности в памяти составляют $O(N^{5/4})$ — не оптимальная величина, но это на $N^{1/4}$ лучше, чем в прямом ленточном алгоритме, требующем $Nw = N^{3/2}$ ячеек. Для трехмерных задач ситуация сохраняется.

Любой из этих алгоритмов (очевидно, остается еще обширное поле для дальнейших исследований) в процессе исключения дает решение системы $KQ = F$. После этого аппроксимация метода конечных элементов найдена.

1.6. ОШИБКИ АППРОКСИМАЦИИ ЛИНЕЙНЫМИ ЭЛЕМЕНТАМИ

Как близка аппроксимация Ритца u^h к точному решению u? В соответствии с приведенной теоремой, утверждающей, что *энергия ошибки* $u - u^h$ *минимальна*, эта аппроксимация близка насколько возможно. Таким образом, метод Ритца оптимален при условии, что энергия измеряется естественным образом. Измерение должно быть связано с особенностями задачи, т. е. с функционалом $I(v)$: энергия v — член второго порядка в $I(v)$. (Наше определение отличается от физически корректного множителем $^1/_2$, но нам удобно игнорировать этот множитель.)

Итак, если функционал записан в виде

$$I(v) = a(v, v) - 2(f, v), \tag{19}$$

то *энергия функции v задается величиной $a(v, v)$*.

Энергия соответствует члену, который до сих пор имел вид (Lv, v) и интегрировался по частям. Это интегрирование приводит к более симметричному выражению — симметрия подчеркивается видом $a(v, v)$. В частности, если (Lv, w) проинтегрировать по частям, получим симметричное выражение

$$a(v, w) = \int_0^\pi (p(x) v'(x) w'(x) + q(x) v(x) w(x)) \, dx.$$

Это *энергетическое скалярное произведение*. Оно определено для всех v и w в допустимом пространстве $\mathscr{H}_E^1$ и представляет собой скалярное произведение, «внутреннее» для данной задачи.

Наша цель в этом разделе состоит в доказательстве теоремы, утверждающей, как сказано выше, что энергия ошибки в методе Ритца минимальна, и в применении ее для установления границ ошибки при аппроксимации линейными элементами.

Теорема 1.1. *Предположим, что u минимизирует $I(v)$ на всем допустимом пространстве $\mathscr{H}_E^1$, а S^h — его замкнутое подпространство. Тогда*

а) *минимум $I(v^h)$ и минимум $a(u - v^h, u - v^h)$, где v^h пробегает подпространство S^h, достигается на одной и той же функции u^h, так что*

$$a(u - u^h, u - u^h) = \min_{v^h \in S^h} a(u - v^h, u - v^h); \tag{20}$$

б) *по отношению к энергетическому скалярному произведению u^h есть проекция u на S^h, или, что то же самое, ошибка $u - u^h$ ортогональна S^h:*

$$a(u - u^h, v^h) = 0 \text{ для всех } v^h \in S^h; \tag{21}$$

в) *функция u^h, на которой достигается минимум, удовлетворяет условию*

$$a(u^h, v^h) = (f, v^h) \text{ для всех } v^h \in S^h; \tag{22}$$

в частности, если S^h — все пространство $\mathscr{H}_E^1$,

$$a(u, v) = (f, v) \text{ для всех } v \in \mathscr{H}_E^1. \tag{23}$$

Следствие. *Из (21) следует, что $a(u - u^h, u^h) = 0$, или $a(u, u^h) = a(u^h, u^h)$, и в силу теоремы Пифагора энергия ошибки равна ошибке в энергии:*

$$a(u - u^h, u - u^h) = a(u, u) - a(u^h, u^h).$$

Далее, так как левая часть неотрицательна, энергия деформации в u^h всегда мажорируется энергией деформации в u:

$$a(u^h, u^h) \leqslant a(u, u). \tag{24}$$

Это теорема основная в теории метода Ритца, и три ее части тесно связаны. Утверждение (б) непосредственно вытекает из (в): если равенство (23) справедливо для всех v, то оно справедливо и для $v^h \in S^h$; вычитая из него (22), получаем (21).

Утверждение (б) вытекает из (а): в пространстве со скалярным произведением функция из подпространства S^h, ближайшая к заданной функции u, всегда является ее проекцией на S^h. Обратно, покажем, что (а) вытекает из (б):

$$a(u - u^h - v^h, u - u^h - v^h) =$$
$$= a(u - u^h, u - u^h) - 2a(u - u^h, v^h) + a(v^h, v^h).$$

Если справедливо равенство (21), то

$$a(u - u^h, u - u^h) \leqslant a(u - u^h - v^h, u - u^h - v^h).$$

Равенство возможно, только когда $a(v^h, v^h) = 0$, т. е. когда $v^h = 0$. Таким образом, u^h — единственная функция в (20), и утверждение (а) доказано.

Осталось доказать утверждение (в) — из него вытекает (б), откуда в свою очередь следует (а). Если u^h минимизирует I на S^h, то для всех ε и v^h

$$I(u^h) \leqslant I(u^h + \varepsilon v^h).$$

Правая часть есть

$$a(u^h + \varepsilon v^h, u^h + \varepsilon v^h) - 2(f, u^h + \varepsilon v^h) =$$
$$= I(u^h) + 2\varepsilon [a(u^h, v^h) - (f, v^h)] + \varepsilon^2 a(v^h, v^h).$$

Поэтому

$$0 \leqslant 2\varepsilon [a(u^h, v^h) - (f, v^h)] + \varepsilon^2 a(v^h, v^h).$$

Так как это верно для сколь угодно малого числа ε любого знака, то $a(u^h, v^h) = (f, v^h)$. Последнее уравнение выражает *равенство нулю первой вариации функционала I в точке u^h в направлении v^h*. В частности, $a(u, v) = (f, v)$, и первая вариация в u равна нулю в любом направлении v. Мы получили уравнение (11), выведенное ранее. Таким образом, утверждение (в) доказано; соотношение (23) дает *уравнение виртуальной работы*.

Если в этом уравнении положить $v = u$, получим интересный результат: *в точке минимума энергия деформации равна потенциальной энергии с обратным знаком*:

$$I(u) = a(u, u) - 2(f, u) = -a(u, u). \tag{25}$$

Аналогично $I(u^h) = -a(u^h, u^h)$. В любом случае $I(u) \leqslant I(u^h)$, так как u доставляет минимум на более широком классе функций, и потому изменение знака приводит к результату, сформулированному в следствии: энергия деформации всегда оценивается сверху,

$$a(u^h, u^h) \leqslant a(u, u).$$

Теорема теперь доказана, за исключением одного пункта: ни существование, ни единственность u^h (когда S^h — все пространство $\mathscr{H}_E^1$, самого решения u) не обоснованы. Для специалиста по функциональному анализу это означает, что доказательство только начинается. Попытаемся его успокоить, указав основное предположение теоремы: подпространство S^h должно быть *замкнуто*, т. е. оно должно содержать все предельные функции. Если последовательность v_N в S^h такова, что

$$a(v_N - v_M, v_N - v_M) \to 0 \text{ при } N, M \to \infty,$$

то в S^h найдется функция v, для которой

$$a(v_N - v, v_N - v) \to 0 \text{ при } N \to \infty.$$

Это всегда верно, если S^h конечномерно; именно этот случай рассматривается в методе Ритца. Вообще говоря, нельзя гарантировать существование функции $u^h \in S^h$, ближайшей к u, не предполагая замкнутости подпространства. Приведем в качестве примера случай $S^h = \mathscr{H}_B^2$, подпространство $\mathscr{H}_B^2$ не замкнуто. Оно содержит функции, сколь угодно близкие к $u(x) = x$, но ближайшей нет, проекции $u(x)$ на $\mathscr{H}_B^2$ не существует.

Для того чтобы доказать существование функции u, определенной как минимизирующая функция на всем пространстве $\mathscr{H}_E^1$, надо считать пространство $\mathscr{H}_E^1$ замкнутым. Это как раз то, чего мы добились, пополняя $\mathscr{H}_B^2$ до $\mathscr{H}_E^1$. В частности, допустимое пространство становилось полным (или замкнутым), когда отбрасывалось естественное краевое условие $u'(\pi) = 0$. Была, правда, одна техническая деталь, которую мы использовали: пространство было пополнено в естественной энергетической норме, $a(v - v_N, v - v_N) \to 0$, как в (10), а чуть раньше мы описали пополненное пространство в терминах $\mathscr{H}^1$-нормы. Эти два подхода оправданы эквивалентностью двух норм: существуют такие постоянные σ и K, что

$$a(v, v) \leqslant K \|v\|_1^2, \tag{26a}$$

$$a(v, v) \geqslant \sigma \|v\|_1^2. \tag{26b}$$

Последнее неравенство также дает единственность u и u^h, так как оно означает, что энергия положительно определена:

$a(v, v) = 0$ тогда и только тогда, когда $v = 0$. Поверхность $I(v)$ строго выпукла и имеет лишь одну стационарную точку — точку минимума.

Первое неравенство очевидно, поскольку

$$\int [p\,(v')^2 + qv^2]\,dx \leqslant \max(p\,(x),\, q\,(x)) \int [(v')^2 + v^2]\,dx.$$

Таким образом, в качестве K можно взять $\max(p, q)$.

Доказательство неравенства (26b) в другую сторону начинается так же, поскольку величина p ограничена снизу положительной постоянной $p_{\min}$:

$$\int p\,(v')^2\,dx \geqslant p\min \int (v')^2\,dx. \tag{27}$$

Трудности появляются в членах нулевого порядка, так как величина q не обязательно отделена от нуля; в самом деле, может оказаться, что $q \equiv 0$. Поэтому нам нужно неравенство *типа Пуанкаре*, с оценкой v через v'. С учетом краевого условия $v(0) = 0$ естественно написать

$$v\,(x_0) = \int_0^{x_0} v'\,(x)\,dx$$

и применить неравенство Шварца

$$|\,v\,(x_0)\,|^2 \leqslant \left(\int_0^{x_0} 1^2\right)\left(\int_0^{x_0} (v')^2\right) \leqslant \pi \int_0^{x_0} (v')^2.$$

Интегрируя по отрезку $0 \leqslant x_0 \leqslant \pi$, приходим к неравенству Пуанкаре

$$\int_0^{\pi} v^2 \leqslant \pi^2 \int_0^{\pi} (v')^2.$$

Теперь, учитывая правую часть в (27), находим, что

$$\int p\,(v')^2 + qv^2 \geqslant p_{\min} \int (v')^2 \geqslant \frac{1}{2\pi^2}\, p_{\min} \int (v')^2 + v^2.$$

Это и есть требуемое неравенство $a(v, v) \geqslant \sigma \| v \|_1^2$, доказывающее *эллиптичность* нашей задачи. Вместе с (26a) оно означает, что пополнения пространства $\mathscr{H}_B^2$ в обычной норме $\| v \|_1$ и в энергетической $\sqrt{a(v, v)}$ совпадают; этим общим пространством является $\mathscr{H}_E^1$.

Заметим, что при естественных краевых условиях на *обоих* концах интервала ситуация совсем другая. Неравенство Пуан-

каре зависит от условия $v(0)=0$ и нарушается для каждой постоянной функции. При $v=\text{const}$ и $q\equiv 0$ энергия $a(v, v)=\int p\,(v')^2$ может равняться нулю и без требования $v=0$. Соответственно решение дифференциального уравнения

$$-(pu')'=f, \quad u'(0)=u'(\pi)=0,$$

не единственно; u определяется с точностью до постоянной. Таким образом, задача Неймана в чистом виде — физически это задача, где возможны жесткие перемещения тела — может привести к техническим трудностям: квадратичная форма $a(v, v)$ будет не определена.

Теперь об общей теории. Наша цель — продемонстрировать ее на примере конечных элементов, когда подпространство S^h составлено из кусочно линейных функций, и оценить ошибку $e^h=u-u^h$. Ключ к решению дает свойство минимизации (20):

$$a(e^h, e^h)\leqslant a(u-v^h, u-v^h) \text{ для всех } v^h\in S^h.$$

Разумеется, функция u неизвестна. Можно только утверждать, что если f принадлежит $\mathscr{H}^0$, то u принадлежит $\mathscr{H}^2_B$. Поэтому вопрос таков: *насколько хорошо можно аппроксимировать произвольную функцию* $u\in\mathscr{H}^2_B$ *элементами из* S^h? Нам не обязательно работать здесь с аппроксимацией Ритца u^h, достаточно найти в S^h хорошую аппроксимацию функции u, а u^h будет еще лучше. Таким образом, оценка ошибки e^h приводит непосредственно к задаче аппроксимации: насколько далеки от S^h функции из пространства $\mathscr{H}^2_B$ в естественной норме $\sqrt{a(v, v)}$?

Удобнее всего взять в качестве функции из S^h, близкой к u, ее интерполянт u_I. Обе функции, u и u_I, совпадают во всех узлах $x=jh$, и u_I линейна между узлами. Ее можно записать в виде комбинаций функций-крышек:

$$u_I(x)=\sum_1^N u(ih)\,\varphi_j^h(x).$$

В каждом узле только одна соответствующая базисная функция φ_j^h отлична от нуля.

Сравним u и u_I сначала на основе простых разложений в ряд Тейлора, дающих поточечную оценку разности функций.

Теорема 1.2. *Если вторая производная* u'' *непрерывна, то*

$$\max|u(x)-u_I(x)|\leqslant\frac{1}{8}h^2\max|u''(x)| \tag{28}$$

и

$$\max|u'(x)-u_I'(x)|\leqslant h\max|u''(x)|. \tag{29}$$

Доказательство. Рассмотрим разность $\Delta(x) = u(x) - u_I(x)$ на интервале $(j-1)h \leqslant x \leqslant jh$. Так как на концах интервала Δ обращается в нуль, найдется по крайней мере одна точка z, для которой $\Delta'(z) = 0$. Тогда

$$\Delta'(x) = \int_z^x \Delta''(y)\, dy \text{ для всех } x.$$

Но $\Delta'' = u''$, поскольку функция u_I линейна; отсюда сразу получаем (29):

$$|\Delta'(x)| = \left| \int_z^x u''(y)\, dy \right| \leqslant h \max |u''(x)|.$$

Максимум величины $|\Delta(x)|$ может достигаться только в точке, где производная равна нулю, $\Delta'(z) = 0$. Посмотрим, в какой

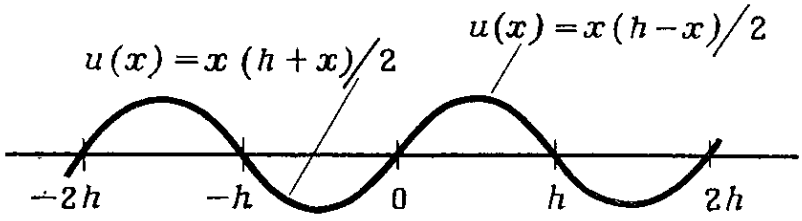


Рис. 1.5.
Экстремальный случай при кусочно линейной аппроксимации.

половине интервала лежит z; пусть z, например, ближе к правому концу интервала, т. е. $jh - z \leqslant h/2$. Разложим в ряд Тейлора в точке z:

$$\Delta(jh) = \Delta(z) + (jh - z)\Delta'(z) + \frac{1}{2}(jh - z)^2 \Delta''(w),$$

где $z < w < jh$. Так как $\Delta = 0$ на концах интервала jh и $\Delta'' = u''$, то

$$|\Delta(z)| = \left| \frac{1}{2}(jh - z)^2 \Delta''(w) \right| \leqslant \frac{1}{8} h^2 \max |u''|.$$

Постоянная 1/8 неулучшаема, и не только для ошибки *линейной интерполяции*, но и для произвольной кусочно линейной аппроксимации. Вторая производная u'' от функции, где аппроксимация хуже всех, меняется от $+1$ до -1 на соседних интервалах (рис. 1.5). Наилучшая кусочно линейная аппроксимация в этом экстремальном случае — тождественный нуль, и ошибка составляет $h^2/8$.

Более подробное доказательство неравенства (29) улучшило бы оценку до $\max|\Delta'| \leqslant (1/2)h \max|u''|$, и экстремальная функ-

ция, изображенная на рис. 1.5, также показывает, что постоянная $1/2$ — наилучшая из возможных.

Из теоремы следует, что если вторая производная u'' непрерывна, то

$$a(u-u_I, u-u_I)=\int_0^\pi p\,(\Delta')^2+q\Delta^2 \leqslant Ch^2 \max|u''|^2.$$

Так как аппроксимация Ритца u^h приближает u не хуже, чем u_I, *ошибка в энергии при аппроксимации линейными элементами удовлетворяет неравенству*

$$a(u-u^h, u-u^h) \leqslant Ch^2 \max|u''|^2.$$

Это уже почти тот результат, какой мы хотим получить. Множитель h^2 совершенно правильный, он отражает скорость убывания ошибки, когда сетка сгущается. Неточность оценки отражает другой множитель, $\max|u''|^2$. Он не удовлетворителен, так как нужно предполагать непрерывность второй производной u'' или даже ее ограниченность, а нам достаточно было бы работать в предположении, что u'' обладает конечной энергией в $\mathscr{H}^0$-норме, т. е. $\int (u'')^2 dx < \infty$. Доказательство этого более точного результата в примере с линейной интерполяцией основано уже на разложении Фурье, а не Тейлора. Оценка ошибки дана ниже в (34).

Теорема 1.3. *Если u'' принадлежит $\mathscr{H}^0$, то*

$$\|u-u_I\|_0 \leqslant \frac{1}{\pi^2} h^2 \|u''\|_0, \tag{30}$$

$$\|u'-u_I'\|_0 \leqslant \frac{1}{\pi} h \|u''\|_0, \tag{31}$$

$$a(u-u_I, u-u_I) \leqslant \left(\frac{h^2}{\pi^2} p_{\max} + \frac{h^4}{\pi^4} q_{\max}\right) \|u''\|_0^2. \tag{32}$$

Доказательство. Рассмотрим какой-нибудь интервал длины h, например первый: $0 \leqslant x \leqslant h$. Разность $\Delta(x) = u(x) - u_I(x)$ равна нулю на концах интервала; представим ее в виде разложения Фурье по синусам:

$$\Delta(x) = \sum_1^\infty a_n \sin\frac{n\pi x}{h}.$$

Непосредственные вычисления дают

$$\int_0^h (\Delta')^2\,dx = \frac{h}{2}\sum\left(\frac{n\pi}{h}\right)^2 a_n^2,$$

$$\int_0^h (\Delta'')^2\,dx = \frac{h}{2}\sum\left(\frac{n\pi}{h}\right)^4 a_n^2.$$

Так как $n \geqslant 1$, то

$$\left(\frac{n\pi}{h}\right)^2 a_n^2 \leqslant \frac{h^2}{\pi^2}\left(\frac{n\pi}{h}\right)^4 a_n^2.$$

Поэтому, суммируя по n, получаем

$$\int_0^h (\Delta')^2 \leqslant \frac{h^2}{\pi^2}\int_0^h (\Delta'')^2 = \frac{h^2}{\pi^2}\int_0^h (u'')^2. \tag{33}$$

Здесь $\Delta'' = u''$, поскольку функция u_I линейна. Равенство достигается тогда и только тогда, когда все коэффициенты a_n, кроме первого, равны нулю, т. е. разность Δ должна иметь вид $\sin \pi x/h$.

Неравенство (33) справедливо на каждом подынтервале, скажем на $(j-1)h \leqslant x \leqslant jh$, так что можно просуммировать по всем подынтервалам:

$$\sum_1^N \int_{(j-1)h}^{jh} (\Delta')^2 \leqslant \frac{h^2}{\pi^2}\sum_1^N \int_{(j-1)h}^{jh} (u'')^2.$$

Мы хотим упростить это неравенство до

$$\int_0^\pi (\Delta')^2 \leqslant \frac{h^2}{\pi^2}\int_0^\pi (u'')^2.$$

Этот шаг, который выглядит вполне очевидным, оправдан только тем, что нет никаких неприятностей в точках стыковки подынтервалов. Отметим, что если бы функция Δ'' стояла справа, как в (33), то равенство

$$\sum \int_{(j-1)h}^{jh} (\Delta'')^2 = \int_0^\pi (\Delta'')^2$$

было бы неверным, так как правая часть в действительности может оказаться бесконечной (Δ'' является δ-функцией в узлах). Эта ситуация возникнет снова, когда мы будем анализировать разницу между *согласованными* и *несогласованными*

элементами; если пробные функции не принадлежат допустимому пространству, то $I(v)$ нельзя вычислить на каждом элементе.

По тем же соображениям

$$\int_0^\pi \Delta^2\,dx = \frac{h}{2}\sum a_n^2 \leqslant \frac{h}{\pi^4}\int_0^\pi (u'')^2,$$

откуда

$$a(\Delta,\ \Delta) = \int_0^\pi p(\Delta')^2 + q\Delta^2 \leqslant$$

$$\leqslant \left(\frac{h^2}{\pi^2}\,p_{\max} + \frac{h^4}{\pi^4}\,q_{\max}\right)\int_0^\pi (u'')^2 \leqslant C^2h^2\,\|u''\|_0^2.$$

Теорема доказана.

Другой вывод неравенства (33) (без разложения в ряд Фурье) основан на вариационной задаче: найти максимум $\int(\Delta')^2$ при ограничениях $\Delta(0) = \Delta(h) = 0$, $\int(\Delta'')^2 = 1$. Стационарными точками будут $\Delta = \sin n\pi x/h$ и экстремум достигается на функциях $\sin \pi x/h$.

Интерполянт u_I не следует путать с аппроксимацией Ритца u^h. Обе функции кусочно линейны, но u^h определяется вариационно, в то время как u_I — всего лишь удобно выбранная близкая к u функция. Теорема 1.1, утверждающая, что u^h лежит к u еще ближе, дает первое из неравенств (34).

Следствие. *Ошибка $e^h = u - u^h$ метода конечных элементов удовлетворяет неравенствам*

$$a(e^h,\ e^h) \leqslant C_1h^2\,\|u''\|_0^2 \leqslant C_2h^2\,\|f\|_0^2. \qquad (34)$$

Второе неравенство вытекает из (3), где решение оценивается через правую часть уравнения. В постоянных C_1 и C_2 основные члены $p_{\max}/\pi^2$ и $p_{\max}/\pi^2p_{\min}^2$ соответственно.

Окончательный результат — оценка порядка h^2 для ошибки в энергии. Практические вычисления показывают, что эта оценка подтверждается, и ошибка $a(u - u^h, u - u^h)$ почти *пропорциональна* h^2, начиная уже с очень грубых сеток ($h = 1/2$ или $1/4$?). Такую регулярность можно объяснить асимптотическим поведением ошибки, которое было введено для разностных уравнений в (16).

Теорема сходимости доказывалась в предположении, что u обладает двумя производными, поэтому исключался случай,

когда f есть δ-функция, а решение u имеет излом. Простые выкладки показывают, что порядок окончательной ошибки в энергии в этом случае равен h, если узел не расположен в точке разрыва функции f. (В случае двух переменных при разрыве вдоль прямой сходимость будет также порядка $O(h)$.) Вообще, если функция u принадлежит только $\mathscr{H}_E^1$, о *скорости* сходимости сказать ничего нельзя, она может произвольно уменьшаться при $h \to 0$. Однако сходимость все-таки есть, это легко доказать.

Теорема 1.4. *Для решения u из пространства $\mathscr{H}_E^1$, т. е. для соответствующих правых частей f, метод конечных элементов сходится в энергетической норме*

$$a(e^h, e^h) \to 0 \text{ при } h \to 0.$$

Доказательство. Так как пространство $\mathscr{H}_E^1$ было построено как пополнение пространства $\mathscr{H}_B^2$, найдется последовательность v_N из $\mathscr{H}_B^2$, сходящаяся в энергетической норме к u. Для каждого фиксированного индекса N аппроксимация по методу конечных элементов v_N^h сходится к v_N при $h \to 0$ (теорема 1.3). Поэтому, если выбрать N достаточно большим, а h достаточно малым, функция v_N^h из S^h будет произвольно близка к u. Так как проекция u^h будет еще ближе, последовательность u^h должна сходиться к u.

Это доказательство применяется без всяких изменений ко всем таким задачам минимизации, и нет нужды повторять его в каждом случае. Необходимое и достаточное условие для сходимости метода Ритца очевидно: *для всякой допустимой функции u ее расстояние до пространства пробных функций S^h* (измеренное по энергии) *должно стремиться к нулю при $h \to 0$*. Из доказательства предыдущей теоремы видно, что эту сходимость можно проверять на плотном подпространстве, т. е. таком, пополнение которого в энергетической норме включает все допустимые функции; сходимость тогда будет автоматически следовать для каждой функции u. Однако интересно установить *скорость сходимости* в энергетической норме в случае, когда u — достаточно гладкая функция.

Интересно также, но несколько труднее, найти скорость сходимости в другой норме. Согласно следствию, напряжения, т. е. первые производные $(u^h)'$, имеют ошибку порядка $O(h)$. Какова *ошибка перемещения*? Насколько быстро убывает $e^h = u - u^h$ по норме $\| e^h \|_0$?

Приблизительно на этот вопрос можно ответить, вспомнив неравенство Пуанкаре $|e^h(x_0)| \leqslant \sqrt{\pi}\, \| e^h \|_1$, выведенное выше. Границы ошибки в каждой точке x_0 будут равномерно иметь

порядок $O(h)$. Можно, однако, улучшить эту оценку до $O(h^2)$. Как это сделать, хорошо видно из (30), где ошибка при интерполировании элементом u_I второго порядка точности. Это по крайней мере доказывает, что S^h содержит функции, отличающиеся от u на $O(h^2)$ по перемещению. Трудность состоит в том, что *в $\mathscr{H}^0$-норме аппроксимация Ритца u^h уже не обладает минимизирующим свойством*, и нет уверенности, что u^h ближе к u, чем u_I. В дальнейшем мы рассмотрим задачу четвертого порядка, в которой ошибка по перемещению не лучше, чем по наклону.

В примере, который мы рассматриваем сейчас, ошибка по перемещению действительно составляет $O(h^2)$. Одно из возможных доказательств — забыть о вариационном происхождении уравнений $KQ = F$ метода конечных элементов и вычислить из них ошибку отсечения как из разностных уравнений (на границе $x = \pi$ это уже сделано). Применяя принцип максимума, мы действительно получаем поточечную оценку $|e^h(x)| = O(h^2)$, которая оптимальна. Но этот подход не полностью удовлетворителен, так как распространение его на нерегулярные конечные элементы в задачах с двумя переменными вызывает огромные трудности. Поэтому важно найти соображения, позволяющие установить *вариационно* скорость сходимости ошибки по перемещению $\| e^h \|_0$.

Следующий прием приводит к успеху: пусть z — решение исходной вариационной задачи на $\mathscr{H}_E^1$, в которой ошибка $e^h = u - u^h$ выбрана в качестве правой части. Приравняем нулю первую вариацию:

$$a(z, v) = (e^h, v) \quad \text{для всех} \quad v \in \mathscr{H}_E^1. \tag{35}$$

В частности, можно положить $v = e^h$; тогда

$$a(z, e^h) = \| e^h \|_0^2. \tag{36}$$

С другой стороны, в теореме 1.1 утверждается, что $a(v^h, e^h) = 0$ для всех $v^h \in S^h$. Вычитая из (36), получаем

$$a(z - v^h, e^h) = \| e^h \|_0^2. \tag{37}$$

К левой части применим неравенство Шварца в энергетической норме:

$$|a(v, w)| \leqslant (a(v, v))^{1/2} (a(w, w))^{1/2}$$

при $v = z - v^h$, $w = e^h$. Согласно следствию из теоремы 1.2,

$$(a(e^h, e^h))^{1/2} \leqslant Ch \| u'' \|_0.$$

Если выбрать v^h как аппроксимацию Ритца функции z, то, согласно тому же следствию,

$$(a(z-v^h,\ z-v^h))^{1/2} \leqslant Ch \| z'' \|_0.$$

Таким образом, применение к (37) неравенства Шварца дает

$$\| e^h \|_0^2 \leqslant C^2 h^2 \| u'' \|_0 \| z'' \|_0.$$

Наконец, можно оценить решение z через правую часть e^h; в силу неравенства (3)

$$\| z'' \|_0 \leqslant \| z \|_2 \leqslant \rho \| e^h \|_0.$$

Это ключевой момент: чтобы оценить ошибку e^h метода Ритца в $\mathscr{H}^0$-норме, что в вариационном смысле неестественно, потребовалось оценить решение в $\mathscr{H}^2$-норме, что также неестественно. Последняя оценка, однако, совершенно естественна с точки зрения дифференциальных уравнений: действительно, основной результат теории состоял в том, чтобы оценить решение в терминах пространства $\mathscr{H}^2$ по правой части из $\mathscr{H}^0$. Подставляя эту оценку в предыдущее неравенство и сокращая на общий множитель $\| e^h \|_0$, приходим к оценке h^2 для ошибки по перемещению (такой подход в литературе по численному анализу известен как *прием Нитше*):

Теорема 1.5. *Кусочно линейная аппроксимация u^h по методу конечных элементов, полученная из теории Ритца, удовлетворяет неравенствам*

$$\| u - u^h \|_0 \leqslant \rho C^2 h^2 \| u'' \|_0 \leqslant \rho^2 C^2 h^2 \| f \|_0. \tag{38}$$

Интересно, что для вывода этой оценки не применялся непосредственно факт возможности аппроксимации порядка h^2 в $\mathscr{H}^0$-норме. Этот факт, таким образом, можно считать следствием теоремы: если S^h дает аппроксимацию порядка $O(h)$ в $\mathscr{H}^1$, то в $\mathscr{H}^0$ достигается аппроксимация $O(h^2)$.

Отметим, что скорости убывания ошибок — h^2 для перемещения и h для напряжения — опять-таки подтверждаются численным экспериментом. Многие экспериментаторы подсчитывали ошибки только в отдельных узлах сетки вместо среднеквадратичных ошибок на интервале и получили те же самые скорости сходимости. (Чтобы предсказать поточечные ошибки, мы должны вернуться к принципу максимума или предположить бо́льшую гладкость данных в среднем и улучшить вариационную оценку. В некоторых важных задачах решение по методу Ритца действительно точнее всего в узловых точках; например, для $-u'' = f$, $u(0) = u(\pi) = 0$ функция u^h совпадает

в узлах с u и точность бесконечна.) Бывают, однако, случаи, в которых ожидаемая сходимость *не* подтверждается из-за простого просчета в эксперименте: ЭВМ работает с величиной

$$E^h = \max_j |e^h(jh)|.$$

При убывании h количество точек сетки, входящих в эту формулу, растет. В частности, появляются точки ближе к границе, где ошибка часто наибольшая, и эти точки начинают определять численную величину E^h. Конечно, нет никаких оснований считать, что эта ошибка будет убывать к нулю с оптимальной скоростью h^2.

Наконец, исследуем ошибку, возникающую при замене функции нагрузок f ее линейным интерполянтом f_I. В результате этой замены, производимой для упрощения интегрирования $\int f\varphi_j^h dx$, вектор нагрузок F заменяется на $\tilde{F}$. Это приводит в свою очередь к приближенному решению по методу конечных элементов $\tilde{Q} = K^{-1}\tilde{F}$, $\tilde{u}^h = \sum \tilde{Q}_j \varphi_j^h$, *представляющему собой точную аппроксимацию по методу конечных элементов задачи с правой частью* f_I. Поэтому мы рассмотрим сейчас лишь изменения в решении по методу Ритца при изменении правой части.

Теорема 1.6. *Ошибка* $u^h - \tilde{u}^h$ *в решении по методу конечных элементов, возникающая при замене функции* f *ее линейным интерполянтом* f_I, *удовлетворяет неравенству*

$$a(u^h - \tilde{u}^h,\ u^h - \tilde{u}^h) \leqslant \frac{K\rho^2}{\pi^4} h^4 \|f''\|_0^2.$$

Доказательство. Точное решение $u - \tilde{u}$ задачи с правой частью $f - f_I$ удовлетворяет неравенствам

$$\|u - \tilde{u}\|_2 \leqslant \rho \|f - f_I\|_0 \leqslant \frac{\rho}{\pi^2} h^2 \|f''\|_0. \tag{39}$$

В последнем неравенстве учтена оценка ошибки линейной интерполяции, взятая из теоремы 1.3. Далее,

$$a(u - \tilde{u},\ u - \tilde{u}) \leqslant K\|u - \tilde{u}\|_1^2 \leqslant K\|u - \tilde{u}\|_2^2 \leqslant \frac{K\rho^2}{\pi^4} h^4 \|f''\|_0^2.$$

(Мы не смогли использовать самую сильную часть неравенств (39), а именно, что даже вторая производная от $u - \tilde{u}$ имеет порядок $O(h^2)$.)

Доказательство заканчивается применением следствия из теоремы 1.1: $u^h - \tilde{u}^h$ есть проекция функции $u - \tilde{u}$ на S^h, а про-

ектирование на S^h не может увеличить энергию:

$$a(u^h - \tilde{u}^h,\ u^h - \tilde{u}^h) = a(u - \tilde{u},\ u - \tilde{u}) -$$
$$- a((u - \tilde{u}) - (u^h - \tilde{u}^h),\ (u - \tilde{u}) - (u^h - \tilde{u}^h)) \leqslant$$
$$\leqslant a(u - \tilde{u},\ u - \tilde{u}) \leqslant \frac{K\rho^2}{\pi^4} h^4 \|f''\|_0^2.$$

Таким образом, ошибка аппроксимации Ритца при интерполировании правой части меньше (h^4 в смысле энергии), чем ошибка h^2 в методе Ритца при аппроксимации линейными элементами.

1.7. МЕТОД КОНЕЧНЫХ ЭЛЕМЕНТОВ В ОДНОМЕРНОМ СЛУЧАЕ

Этот раздел обобщает предыдущий в трех направлениях: здесь вводятся неоднородные краевые условия, рассматриваются квадратичные и даже кубические элементы, а не линейные, и решаются дифференциальные уравнения четвертого порядка, а не только второго. Оценки ошибок для различных конечных элементов часто приводятся без доказательств, так как они вытекают из теории, которая будет развита далее в этой книге. Этапы метода конечных элементов те же, что и прежде: вариационная постановка задачи, выделение кусочно полиномиальных подпространств в некотором допустимом пространстве, построение и решение линейных уравнений $KQ = F$. Эта схема в одномерном случае более или менее закончена.

Начнем с изучения того же дифференциального уравнения $-(pu')' - qu = f$, но с краевыми условиями более общего вида

$$u(0) = g, \quad u'(\pi) + \alpha u(\pi) = b.$$

Первое из условий по-прежнему главное, и ему должна удовлетворять каждая функция v из допустимого пространства $\mathscr{H}_E^1$. Поэтому *разность между двумя допустимыми функциями* $v_0 = v_1 - v_2$ *будет удовлетворять однородному условию* $v_0(0) = 0$. Обозначим через V_0 пространство таких разностей v_0; это допустимое пространство в случае однородного главного условия.

Краевое условие на другом конце выглядит теперь по-новому; оно содержит u' и u, и потому функционал $I(v)$ надо вычислить заново. Физически система представляет собой струну, которая не фиксирована и не свободна полностью в точке

$x = \pi$. Новый функционал имеет вид

$$I(v) = \int_0^\pi (p(v')^2 + qv^2)\,dx + \alpha p(\pi)\,v^2(\pi) - \\ - 2\int_0^\pi fv\,dx - 2bp(\pi)\,v(\pi).$$

Таким образом, новые краевые условия входят как в линейную часть функционала, так и в энергию

$$a(v,\ v) = \int (p(v')^2 + qv^2)\,dx + \alpha p(\pi)\,v^2(\pi).$$

Последний член представляет собой энергию струны.

Проверим, что равенство нулю первой вариации в каждом направлении v_0 приводит к тем же условиям на точку минимума u, что и дифференциальное уравнение вместе с краевыми условиями. (Заметим, что u изменяется функцией v_0 из V_0, обеспечивающей выполнение главного краевого условия $(u + \varepsilon v_0)(0) = g$. Измененная функция не принадлежит $\mathscr{H}_E^1$). Коэффициент при 2ε в $I(u + \varepsilon v_0)$ есть

$$\int [pu'v_0' + quv_0] + \alpha p(\pi)\,u(\pi)\,v_0(\pi) - \int fv_0 - bp(\pi)\,v_0(\pi) = \\ = \int [-(pu')' + qu - f]\,v_0 + p(\pi)\,[u'(\pi) + \alpha u(\pi) - b]\,v_0(\pi).$$

Это выражение равно нулю для всех v_0 тогда и только тогда, когда u удовлетворяет дифференциальному уравнению и новым краевым условиям. *Поэтому краевое условие — естественное для измененного функционала* $I(v)$.

В общем методе Ритца больше нет смысла задавать себе вопрос, является ли S^h подпространством в $\mathscr{H}_E^1$, так как $\mathscr{H}_E^1$ — это уже не векторное пространство, оно сдвинуто относительно нуля. По этой причине пусть S^h имеет тот же вид, что и прежде. *Пробные функции* v^h *не должны лежать в допустимом пространстве* $\mathscr{H}_E^1$, *но их разности обязаны лежать в пространстве* V_0 *функций с однородным условием.* Эти разности $v_0^h = v_1^h - v_2^h$ образуют конечномерное пространство S_0^h, которое должно быть подпространством в V_0.

Для линейных конечных элементов все очень просто. В точке $x = \pi$ никаких ограничений нет, там краевое условие естественное. Пространство пробных функций S^h будет состоять поэтому *из всех кусочно линейных функций, удовлетворяющих*

условию $v^h(0) = g$. (Главному краевому условию можно удовлетворить точно в одномерном случае, так как v^h «зажата» только в точке. В случае двух и более переменных краевое условие $v^h(x, y) = g(x, y)$ не удовлетворяется полиномами и S^h не содержится в $\mathscr{H}_E^1$.) S_0^h есть пространство кусочно-линейных пробных функций, равных нулю в точке $x = 0$, и каждую функцию v^h можно записать в виде

$$v^h(x) = g\varphi_0^h(x) + \sum_1^N q_j \varphi_j^h(x).$$

Коэффициент при φ_0^h фиксирован: $q_0 = g$.

Потенциальная энергия $I(v^h)$ представляет собой квадратичный функционал неизвестных $q_1, q_2, \ldots, q_N$ и его минимизация приводит опять к линейной системе $KQ = F$. Внутри интервала, т. е. для всех строк матрицы, кроме первой и последней, эта система такая же, как в предыдущем разделе. Первая строка матрицы, соответствующая левому концу интервала, похожа на все другие строки, за исключением того, что $q_0 = g$ [1]). Поэтому первое уравнение системы (с коэффициентами $p = q = 1$) выглядит так:

$$\frac{-g + 2Q_1 - Q_2}{h} + \frac{h}{6}(g + 4Q_1 + Q_2) = \int_0^{2h} f\varphi_1^h \, dx.$$

Если перенести члены с g в другую часть уравнения, первая строка матрицы K будет в точности той же, что и раньше. Неоднородное условие изменяет только первую компоненту F_1 вектора нагрузок на величину $g/h - gh/6$.

Для другого конца вычисления почти так же просты. Новые члены в $I(v^h)$ таковы:

$$\alpha p(\pi)(v^h(\pi))^2 - 2bp(\pi)v^h(\pi) = \alpha p(\pi) q_N^2 - 2bp(\pi) q_N.$$

Поэтому в последнем уравнении $\partial I/\partial q_N = 0$ после сокращения на 2 к компоненте F_N вектора нагрузок добавится $bp(\pi)$, а в элемент K_{NN} матрицы жесткости войдет $\alpha p(\pi)$. Такие малые изменения объясняются локальностью базиса: только одна базисная функция (последняя) не равна нулю в точке $x = \pi$ и именно она связана с краевым условием.

---

[1]) Это соответствует способу, которым главное краевое условие вводится на практике. Его игнорируют при построении матрицы, а потом неизвестной (в нашем случае q_0) приписывают значение, которое берется из краевого условия.

Оценка ошибки для этой задачи та же, что и в предыдущем разделе:

$$a(u-u^h,\ u-u^h)=O(h^2)$$

и

$$\|u-u^h\|_0=O(h^2).$$

Первая оценка опять связана с вариационной теоремой: u ближе к u^h, чем к u_I. В свою очередь это зависит от того, принадлежит ли линейный интерполянт u_I пространству пробных функций S^h. Поэтому можно опять все свести к теореме 1.3, дающей оценку расстояния между функцией и ее интерполянтом.

Второе обобщение метода состоит во введении более «точных», чем кусочно линейные функции, элементов. Заданная функция $u(x)$ лучше аппроксимируется квадратичными или кубическими интерполянтами, чем линейными, и это приводит к соответствующему улучшению точности u^h. Поэтому естественно строить пространство пробных функций S^h с помощью полиномов высших степеней.

Мы начнем с предположения, что S^h состоит из всех кусочно квадратичных функций, непрерывных в узлах $x=jh$ и удовлетворяющих условию $v^h(0)=g$. Прежде всего вычислим размерность пространства S^h (число свободных параметров q_j) и определим его базис. Заметим, что если x попадает в узел, непрерывность налагает только одно ограничение на параболу, начинающуюся в этом узле; два параметра параболы остаются свободными. Поэтому размерность должна быть вдвое больше числа парабол, т. е. $2N$.

Базис можно построить, вводя в дополнение к узловым точкам $x=jh$ средние между узлами точки $x=(j-1/2)h$. Тогда узлов будет $2N$, поскольку $x=0$ исключается и $Nh=\pi$; будем обозначать эти узлы z_j, $j=1, 2, \ldots, 2N$. Каждому узлу соответствует непрерывная кусочно квадратичная функция, равная 1 в z_j и 0 в z_i, $i\neq j$:

$$\varphi_j(z_i)=\delta_{ij}. \tag{40}$$

Эти функции будут двух типов в зависимости от того, была ли точка z_j узлом (рис. 1.6, *а*) или средней между узлами (рис. 1.6, *б*). Отметим, что обе функции непрерывны и принадлежат $\mathscr{H}_E^1$. Функция φ_2, не равная нулю только на одном подынтервале, не определяется внутренней структурой подпространства, промежуточный узел можно выбрать где-нибудь на интервале. Выбор средней точки влияет на базис, но не на само пространство.

Матрица жесткости элемента k_1, соответствующая интегралу от $(v')^2$ по отрезку $0\leqslant x\leqslant h$, вычисляется по трем значениям

параболы: q_0 при $x=0$, $q_{1/2}$ в средней точке $x=h/2$ и q_1 при $x=h$. Такая парабола имеет вид

$$v^h(x)=q_0+\frac{x}{h}(4q_{1/2}-q_1-3q_0)+\left(\frac{x}{h}\right)^2(2q_1+2q_0-4q_{1/2}).$$

Чтобы выразить ее через базисные функции, найдем коэффициенты при q_0, $q_{1/2}$ и q_1:

$$v^h(x)=q_0\left(1-3\frac{x}{h}+2\frac{x^2}{h^2}\right)+$$
$$+q_{1/2}\left(4\frac{x}{h}-4\frac{x^2}{h^2}\right)+q_1\left(-\frac{x}{h}+2\frac{x^2}{h^2}\right).$$

Эти три коэффициента точно описывают три параболы на рис. 1.6: коэффициент при q_0 — правую половину графика функции φ_1 на рис. 1.6, *а* (равной 1 при $x=0$ и 0 при $x=h/2$, $x=h$),

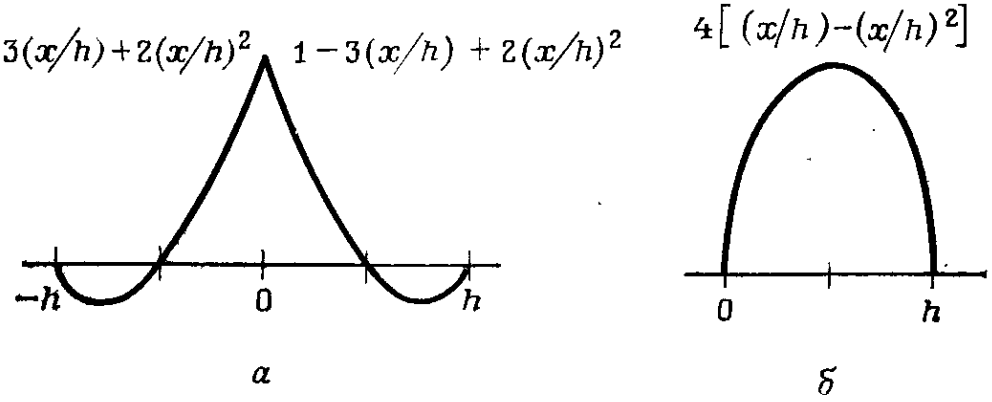


Рис. 1.6.
Базисные функции кусочно квадратичных элементов.

коэффициент при $q_{1/2}$ — параболу φ_2, изображенную на рис. 1.6, *б*, и коэффициент при q_1 — левую половину графика функции φ_1.

Для вычисления матрицы k_1 надо проинтегрировать $(dv^h/dx)^2$ и затем выписать результат в виде $(q_0q_{1/2}q_1)^T k_1 (q_0q_{1/2}q_1)$. Отметим, что k_1 имеет размер 3×3, так как на каждом фиксированном интервале появляется 3 параметра q; парабола определяется тремя условиями. Опуская вычисления, которые каждый может проделать сам, дадим только окончательный результат:

$$k_1=\frac{1}{3h}\begin{pmatrix}7 & -8 & 1\\ -8 & 16 & -8\\ 1 & -8 & 7\end{pmatrix}.$$

Заметим, что матрица k_1 вырождена; после умножения ее на вектор (1, 1, 1) получается нуль. Вектор $q_0=1$, $q_{1/2}=1$, $q_1=1$ соответствует параболе v^h, вырожденной в горизонтальную прямую $v^h\equiv 1$, так что ее производная равна нулю. Вырождение

матрицы k_1 — это незначительное препятствие: не должна быть вырождена матрица k_0.

Описанные идеи непосредственно распространяются на кубические элементы. Непрерывность пробных функций налагает одно ограничение в каждом узле $x = jh$, оставляя 3 параметра кубического элемента свободными, поэтому размерность пространства S^h равна $3N$. Для построения базиса введем два узла внутри каждого интервала, скажем, на расстоянии $h/3$ от его концов. Вместе со старой сеткой это дает $3N$ узлов $z_j = jh/3$. Функции φ_j, удовлетворяющие условию $\varphi_j(z_i) = \delta_{ij}$, образуют базис и могут быть трех типов (рис. 1.7). Матрицы элементов будут иметь порядок 4.

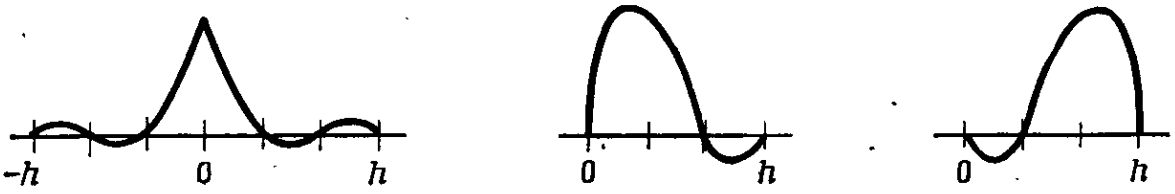


Рис. 1.7.
Кубические элементы, только непрерывные в узлах.

Есть и другие кубические элементы, лучшие почти во всех отношениях. Для их построения требуется непрерывность не только самих функций v^h, но и их первых производных. Это означает, что пробное пространство в данном случае действительно является *подпространством рассмотренного выше пространства пробных функций*: налагается по одному новому ограничению в каждом из $N-1$ внутренних узлов $x = h, 2h, \ldots, \pi - h$. Поэтому размерность нового пространства будет $3N - (N-1) = 2N + 1$. Число параметров, которые нужно вычислять, уменьшилось на одну треть. Единственный случай, когда введение нового пространства кубических элементов не улучшает аппроксимацию решения u, это случай, когда u *не* имеет непрерывную производную. В разд. 1.3 мы видели, что это происходит — u принадлежит $\mathscr{H}_E^1$, но не $\mathscr{H}_B^2$ — в случае точечной нагрузки или разрывного коэффициента $p(x)$ в дифференциальном уравнении. В такой ситуации существенно не требовать гладкости пробных функций. Особую точку x_0 надо поместить в узел, и в нем кубический элемент должен быть всегда лишь непрерывным. Это позволит сохранить порядок сходимости.

Расположение узлов для более гладких кубических элементов интереснее. В каждой точке $x = jh$ находится *двойной узел*. Вместо того, чтобы определять кубический элемент по его значениям в четырех различных точках 0, $h/3$, $2h/3$ и h, будем теперь определять его по значениям самого элемента и его

первых производных в обоих концах, т. е. по v_0, v_0', v_1, v_1'. Значения v_1 и v_1' будут совпадать со значениями в следующем подынтервале, так как v и v' непрерывны. Базисные функции будут двух типов (рис. 1.8). Эти функции имеют нули второго порядка на концах $(j \pm 1)h$. Их называют *эрмитовыми кубическими функциями*; они интерполируют значения функции и ее производной. Случай, рассмотренный ранее, связывают с именем Лагранжа.

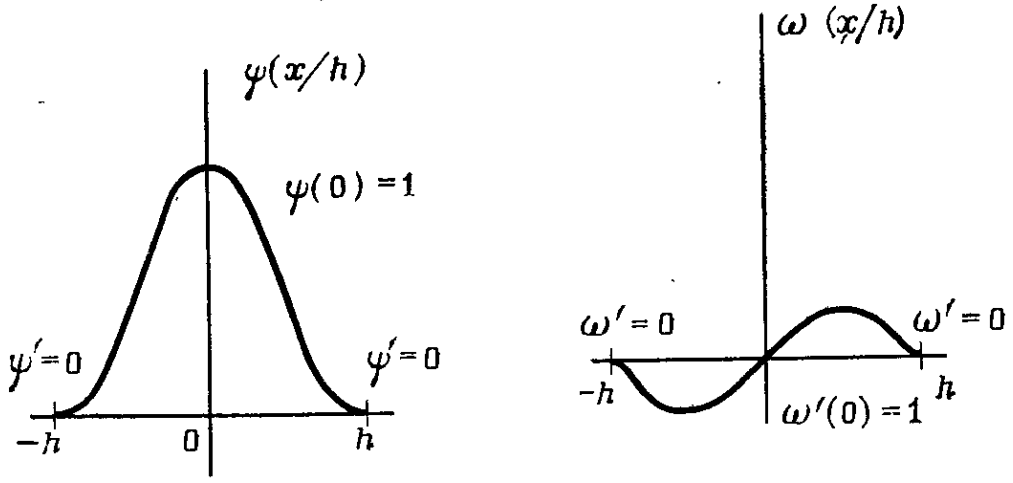


Рис. 1.8.

Эрмитовы кубические функции: v и v' непрерывны.
$\psi(x) = (|x|-1)^2(2|x|+1)$, $\omega(x) = x(|x|-1)^2$.

Кубический полином на отрезке $[0, h]$, принимающий четыре наперед заданных значения v_0, v_0', v_1 и v_1', имеет вид

$$v^h(x) = v_0\psi\left(\frac{x}{h}\right) + hv_0'\omega\left(\frac{x}{h}\right) + v_1\psi\left(\frac{x-h}{h}\right) + hv_1'\omega\left(\frac{x-h}{h}\right) =$$

$$= v_0 + v_0'x + (3v_1 + 3v_0 - v_1'h - 2v_0'h)\frac{x^2}{h^2} +$$

$$+ (2v_0 - 2v_1 + hv_1' + hv_0')\frac{x^3}{h^3}. \qquad (41)$$

Матрицы элементов (четвертого порядка) вычисляются интегрированием, причем q — вектор-столбец $(v_0, v_0', v_1, v_1')^T$:

$$\text{матрица массы } k_0\text{: } \int_0^h (v^h)^2 = q^T k_0 q,$$

$$\text{матрица жесткости } k_1\text{: } \int_0^h ((v^h)')^2 = q^T k_1 q,$$

$$\text{матрица изгиба } k_2\text{: } \int_0^h ((v^h)'')^2 = q^T k_2 q.$$

Заметим, что так как v^h принадлежит $\mathscr{H}^2$, эту технику можно использовать для задач четвертого порядка; здесь нужна матрица изгиба k_2.

Существует несколько способов организации вычисления матрицы массы k_0. Один из лучших состоит в том, чтобы составить матрицу H, связывающую четыре узловых параметра вектора q с четырьмя коэффициентами $A=(a_0, a_1, a_2, a_3)$ кубического полинома v^h: $A=Hq$, или, с учетом (41),

$$\begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -\frac{3}{h^2} & -\frac{2}{h} & \frac{3}{h^2} & -\frac{1}{h} \\ \frac{2}{h^3} & \frac{1}{h^2} & -\frac{2}{h^3} & \frac{1}{h^2} \end{pmatrix} \begin{pmatrix} v_0 \\ v_0' \\ v_1 \\ v_1' \end{pmatrix}.$$

Интегрирование функции $(v^h)^2=(a_0+a_1x+a_2x^2+a_3x^3)^2$ совсем тривиально:

$$\int_0^h (v^h)^2\,dx = (a_0a_1a_2a_3) \begin{pmatrix} h & \frac{h^2}{2} & \frac{h^3}{3} & \frac{h^4}{4} \\ \frac{h^2}{2} & \frac{h^3}{3} & \frac{h^4}{4} & \frac{h^5}{5} \\ \frac{h^3}{3} & \frac{h^4}{4} & \frac{h^5}{5} & \frac{h^6}{6} \\ \frac{h^4}{4} & \frac{h^5}{5} & \frac{h^6}{6} & \frac{h^7}{7} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix}.$$

Обозначая матрицу в правой части через N_0, получаем

$$\int_0^h (v^h)^2 dx = A^T N_0 A = q^T H^T N_0 H q.$$

Поэтому матрица массы элемента равна

$$k_0 = H^T N_0 H.$$

Все это легко программируется на ЭВМ.

Для матрицы жесткости матрица связи H между узловыми параметрами q и вектором коэффициентов A остается той же.

Единственное различие состоит в том, что

$$\int_0^h ((v^h)')^2\,dx = \int_0^h (a_1 + 2a_2x + 3a_3x^2)^2\,dx = A^T N_1 A =$$

$$= A^T \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & h & h^2 & h^3 \\ 0 & h^2 & \frac{4h^3}{3} & \frac{3h^4}{2} \\ 0 & h^3 & \frac{3h^4}{2} & \frac{9h^5}{5} \end{pmatrix} A.$$

Матрица жесткости элемента есть $k_1 = H^T N_1 H$. В результате этих вычислений и аналогичных для матрицы k_2 получаются матрицы (их надо дополнить по симметрии)

$$k_0 = \frac{h}{420}\begin{pmatrix} 156 & 22h & 54 & -13h \\ & 4h^2 & 13h & -3h^2 \\ & & 156 & -22h \\ & & & 4h^2 \end{pmatrix},$$

$$k_1 = \frac{1}{30h}\begin{pmatrix} 36 & 3h & -36 & 3h \\ & 4h^2 & -3h & -h^2 \\ & & 36 & -3h \\ & & & 4h^2 \end{pmatrix},$$

$$k_2 = \frac{1}{h^3}\begin{pmatrix} 12 & 6h & -12 & 6h \\ & 4h^2 & -6h & 2h^2 \\ & & 12 & -6h \\ & & & 4h^2 \end{pmatrix}.$$

Матрица k_0 положительно определена, а k_1 имеет нулевое собственное значение, соответствующее функции $v^h \equiv 1$, т. е. $q = (1, 0, 1, 0)$. Матрица k_2 обнуляет два линейно независимых вектора, так как $(v^h)'' \equiv 0$ для каждой линейной функции v^h. Новый вектор из ядра матрицы соответствует функции $v^h(x) \equiv x$, т. е. $q = (0, 1, h, 1)$.

Иногда полезно заменить *вырожденные блоки* k_1 и k_2 *естественными невырожденными матрицами*. Матрицы становятся невырожденными в результате *жестких движений тела*, т. е. при $(v^h)' \equiv 0$ и $(v^h)'' \equiv 0$. Порядки матриц снижаются до 3 и 2 соответственно, и теперь они не несут избыточной информации. Обнуление вектора (1, 0, 1, 0) матрицей k_1 естественно, поскольку

ку соответствует положению струны без напряжения — это принимается на веру и не доказывается. Оказывается, что эти «естественные» матрицы могут упростить вычисления для отдельной программы, допуская огромное разнообразие элементов. В этой книге мы сохраним матрицы жесткости в их вырожденной форме, так как в таком виде они яснее показывают роль всех четырех узловых параметров v_0, v_0', v_1 и v_1'.

Для того чтобы можно было применить матрицы элементов к задаче $-(pu')' + qu = f$, они должны составлять глобальную матрицу жесткости K. Если предположить, что коэффициенты постоянны, то типичной строкой (или, вернее, парой строк, поскольку каждой узловой точке $x_j = jh$ соответствует два неизвестных u_j и u_j') в построенной матрице K будет

$$\frac{p}{30h}(-36u_{j-1} - 3hu'_{j-1} + 72u_j - 36u_{j+1} + 3hu'_{j+1}) +$$
$$+ \frac{qh}{420}(54u_{j-1} + 13hu'_{j-1} + 312u_j + 54u_{j-1} - 13hu'_{j-1}) =$$
$$= F_j = \int f(x)\,\psi\left(\frac{x}{h} - j\right), \tag{42a}$$

$$\frac{p}{30}(3u_{j-1} - hu'_{j-1} + 8hu'_j - 3u_{j+1} - hu'_{j+1}) +$$
$$+ \frac{qh^2}{420}(-13u_{j-1} - 3hu'_{j-1} + 8hu'_j + 13u_{j+1} - 3hu'_{j+1}) =$$
$$= F'_j = h\int f(x)\,\omega\left(\frac{x}{h} - j\right). \tag{42b}$$

Интересно рассмотреть это как разностное уравнение. Предположим, что $p = 1$, $q = 0$ и $f = 1$, так что решается дифференциальное уравнение $-u'' = 1$. Уравнениями метода конечных элементов будут

$$-\frac{6}{5}\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + \frac{1}{5}\frac{u'_{j+1} - u'_{j-1}}{2h} = 1,$$

$$-\frac{1}{5}\frac{u_{j+1} - u_{j-1}}{2h} + \frac{1}{5}u'_j - \frac{1}{30}(u'_{j+1} - 2u'_j + u'_{j-1}) = 0.$$

Разлагая в ряд Тейлора, видим, что первое уравнение согласуется с $-u'' = 1$, а второе с $-hu'''/15 = 0$; это можно вывести дифференцированием первого уравнения. *Именно здесь проявляется, что метод конечных элементов способствует новой и полезной идее обоснования техники конечных разностей.* Вместо того, чтобы действовать только с неизвестными u_j и получать по одному уравнению в каждой точке сетки, разностные уравнения метода конечных элементов позволяют брать в качестве неизвестных перемещения и наклоны, так как уравнение для

наклона формально согласуется с продифференцированным исходным уравнением. При этом можно достичь высокой точности аппроксимации не только самой функции, но и производных высших порядков, *не отказываясь от локального характера разностного уравнения*. Это нововведение годится для неравномерных сеток и криволинейных границ. Его можно было бы серьезно взять на вооружение при исследовании разностных схем, так как здесь — без ограничения на то, что дискретный аналог возник из метода Ритца с полиномиальной аппроксимацией, — можно добиться даже большей эффективности.

Пусть для проверки порядка точности эрмитовых разностных уравнений (42) применяется разложение Тейлора. Прежде всего предположим, что

$$v_j = u(jh) + \sum h^n e_n(jh), \quad v_j' = u'(jh) + \sum h^n \varepsilon_n(jh), \qquad (43)$$

и разложим $v_{j\pm1}$ и $v'_{j\pm1}$ в центральной точке jh; с помощью исходного дифференциального уравнения и результатов его дифференцирований можно проверить, что $\varepsilon_n = e_n'$ и что эти члены исчезают при $n = 1, 2, 3$. Другими словами, эрмитово разностное уравнение имеет *четвертый порядок точности*. Эта оценка точно совпадает с оценкой, найденной вариационно; и разложение Тейлора, и вариационная оценка разности $u - u_I$ давали $O(h^2)$ в линейном случае. В работе [M7] описан один непредвиденный и довольно печальный случай: на границах асимптотические разложения (43) портятся и ошибка метода конечных элементов не описывается простым степенным рядом по h при $h \to 0$. Однако она имеет порядок h^4.

Рассмотрим еще один важный случай пространства кубических элементов, образованного функциями, у которых даже *вторая* производная непрерывна в узлах. Кусочно кубические функции с непрерывными вторыми производными называются *кубическими сплайнами*. Это пространство кубических сплайнов представляет собой подпространство эрмитовых кубических функций с новым ограничением в каждом из $N - 1$ внутренних узлов. Поэтому размерность подпространства сплайнов равна $3N - 2(N-1) = N + 2$. Это означает, что *каждому узлу соответствует одно неизвестное*, включая крайние точки $x_0 = 0$ (где $v_0 = 0$, а наклон v_0' можно считать свободным параметром) и $x_{N+1} = \pi + h$ (можно в качестве последнего параметра взять v_N'). Во внутренних точках сетки неизвестными являются перемещения v_j, и уравнения метода конечных элементов будут опять выглядеть в точности как уравнения в конечных разностях.

Теперь уже не очевидно, какие четыре узловых параметра определяют поведение кубического сплайна на заданном подын-

тервале, скажем на $(j-1)h \leqslant x \leqslant jh$. Значения в узлах x_{j-1} и x_j, принадлежащих подынтервалу, дают только два условия, а другие два надо откуда-то взять. Поэтому *в кубических сплайнах не существует простейшего локального базиса*, и поведение v^h внутри элемента определяется перемещениями за пределами этого элемента. Действительно, сплайн, равный нулю во всех узлах, кроме одного, не локален: он отличен от нуля во *всех* подынтервалах между узлами.

Для того чтобы вычислять с помощью сплайнов, нужно построить один сплайн, равный 1 в начале координат и отличный от нуля на как можно меньшем отрезке (рис. 1.9). Эта функция

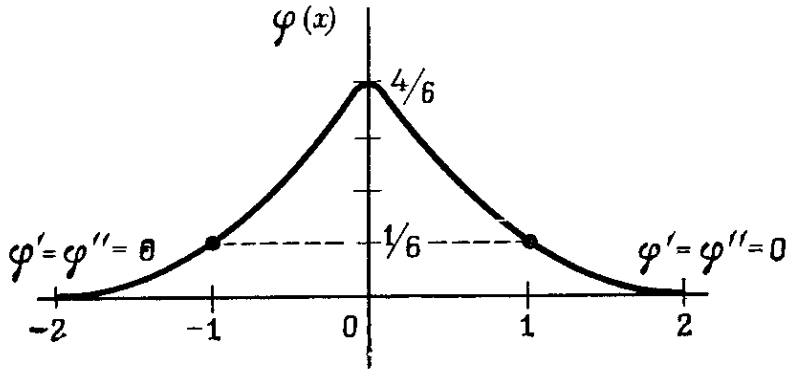


Рис. 1.9.

Кубический B-сплайн: непрерывные вторые производные в узлах.

известна под названием базисного сплайна или *B-сплайна*. Она очень важна в теории сплайнов и относится к одному из многих замечательных открытий Шёнберга. В частности, Шёнберг доказал, что каждый кубический сплайн на отрезке $[0, \pi]$ можно записать в виде линейной комбинации B-сплайнов

$$v^h(x) = \sum_{-1}^{N+1} q_j \varphi_j^h(x).$$

Базисные функции φ_j^h образованы из B-сплайна, изображенного на рис. 1.9, путем замены переменной x на x/h и переноса начала координат в точку jh:

$$\varphi_j^h(x) = \varphi\left(\frac{x}{h} - j\right).$$

В этом выражении для v^h не учтено, что v^h обращается в нуль в начале координат. Если φ_*^h и φ_{**}^h — комбинации функций φ_1^h и φ_0^h, удовлетворяющие этому главному условию, то соответствующими кубическими сплайнами будут

$$v^h(x) = q_* \varphi_*^h(x) + q_{**} \varphi_{**}^h(x) + \sum_{2}^{N+1} q_j \varphi_j^h(x).$$

В такой записи система $N+2$ уравнений метода конечных элементов $KQ = F$ имеет в качестве неизвестных $q_*, \ldots, q_{N+1}$. Неизвестное q_{N+1} появляется потому, что естественное краевое условие не налагает никакого ограничения в вариационной формулировке; было бы интересно выяснить, как действует это условие на u^h, и таким способом избавиться от последней неизвестной. Вероятно, действие это убывает по экспоненциальному закону.

Последние два замечания о сплайнах:

1. Видимо, наибольший интерес сплайны представляют *в теории приближений, а не при минимизации функционалов.* Заданное отношение между данными очень удобно аппроксимировать с помощью сплайнов, а вот отыскать *неизвестный* сплайн путем минимизации куда менее удобно.

2. Если узлы расположены неравномерно, сплайны будут ненулевыми на четырех интервалах (это минимум для кубического сплайна), и если каждый узел x_{2N-1} приближается к x_{2N}, B-сплайны превращаются в эрмитовы базисные функции ψ и ω.

Аппроксимационные свойства всех этих пространств кусочно полиномиальных функций легко обобщить одной формулой. Если полиномы имеют степень $k-1$ ($k=3$ для квадратичных элементов, $k=4$ для кубических), то гладкая функция u отличается от своего интерполянта u_I на величину

$$\| u - u_I \|_0 \leqslant Ch^k \| u^k \|_0.$$

Порядок ошибки аппроксимации производных уменьшается на единицу при каждом дифференцировании:

$$\| u - u_I \|_s \leqslant C_s h^{k-s} \| u^{(k)} \|_0. \tag{44}$$

Оценка (44) имеет смысл, если известно, что u_I обладает s производными, т. е. кусочно полиномиальная функция принадлежит $\mathscr{H}^s$. Поэтому $s \leqslant q$ в неравенстве (44): $q=1$ для непрерывных квадратичных и кубических элементов, принадлежащих $\mathscr{H}^1$, $q=2$ для эрмитовых кубических элементов и $q=3$ для сплайнов. При $s > q$ оценку еще можно получать между узлами, а в узлах появляются δ-функции.

Эти результаты по аппроксимации приводят к ожидаемым скоростям сходимости метода конечных элементов при условии, что производная $u^{(k)}$ обладает конечной энергией: наклоны аппроксимируются с ошибкой $O(h^{k-1})$, энергия деформации с ошибкой $O(h^{2(k-1)})$ и перемещение $u - u^h$ с ошибкой $O(h^k)$. Так как при расчетах эти скорости подтверждаются, есть все основания решать задачи с помощью хороших конечных элементов

Последний вопрос, рассматриваемый в этом разделе: решение уравнения четвертого порядка

$$Lu = (ru'')'' - (pu')' + qu = f. \tag{45}$$

Оператор L формально самосопряжен, поскольку u''' и u' здесь не встречаются, и положительно определен, если $r \geqslant r_{\min} > 0$, $p \geqslant 0$, $q \geqslant 0$.

Энергетическое скалярное произведение, соответствующее задаче, равно

$$a(u, v) = \int (ru''v'' + pu'v' + quv)\, dx,$$

а уравнением Эйлера для минимизации функционала $I(v) = a(v, v) - 2(f, v)$ служит $Lu = f$. Для того чтобы применить метод Ритца, нужно взять пробные функции v^h с конечной энергией, а это означает, что $v^h \in \mathscr{H}^2$. Эрмитовы и сплайновые кубические элементы здесь применимы, а просто непрерывные полиномиальные функции применять нельзя. Они здесь не подходят, и использовать их, игнорируя тот факт, что их вторые производные в узлах есть δ-функции, значит уже в одномерном случае нарваться на неприятность.

Уравнение (45) описывает изгиб балки. Если она закреплена в точке $x = 0$, краевыми условиями будут

$$u(0) = u'(0) = 0;$$

это *главные* условия: каждая пробная функция должна иметь нуль второго порядка в начале координат. Чтобы выяснить, каковы естественные условия при $x = \pi$, проинтегрируем по частям уравнение $a(u, v) = (f, v)$ и приравняем нулю первую вариацию. В результате для каждой функции v из допустимого пространства $\mathscr{H}^2_E$ получим

$$\int [(ru'')'' - (pu')' + qu - f]\, v + ru''v' \Big|_\pi + (pu' - (ru'')')\, v \Big|_\pi = 0. \tag{46}$$

Таким образом, если на v не налагаются условия в точке $x = \pi$, то естественными краевыми условиями на u будут условия, соответствующие физически свободному концу:

$$u''(\pi) = 0, \quad (pu' - ru''')(\pi) = 0.$$

Еще один очень важный случай — балка имеет опору: $u(\pi) = 0$. Это главное условие, которому должны удовлетворять пробные функции v. По этой причине последний член в первой вариации (46) автоматически равен нулю — без всякого условия на u в точке π. Однако другой проинтегрированный член равен нулю для всех v только тогда, когда $u''(\pi) = 0$, и это естественное краевое условие остается. Таким образом, концу балки с опорой

соответствует комбинация главного и естественного краевых условий:

$$u(\pi) = u''(\pi) = 0.$$

Можно представить себе и другую комбинацию краевых условий, например $u'(\pi) = u''(\pi) = 0$. Но их трудно понять как физически, так и вариационно.

1.8. ДВУМЕРНЫЕ КРАЕВЫЕ ЗАДАЧИ

Здесь мы рассмотрим несколько задач на плоскости, или, вернее, в области Ω на плоскости, ограниченной гладкой кривой Γ. Нашей целью в первую очередь будет сопоставление с дифференциальным видом этих задач, содержащих оператор Лапласа Δ и бигармонический оператор Δ^2, эквивалентной вариационной формулировки. Это означает, что в вариационной постановке мы должны подобрать допустимые пространства, в которых ищется решение. Естественно, что эти пространства зависят от краевых условий, и, как и в случае одномерной краевой задачи, условия Дирихле (главные условия) будут отличаться от условий Неймана (естественных условий). Примеры привести очень легко, но они представляют собой простейшие модели плоского напряженного состояния и изгиба пластины, так что полезнее еще раз проиллюстрировать основные идеи:

1) Эквивалентность дифференциальной и вариационной задач, допустимое пространство, в дальнейшем пополняемое в энергетической норме.

2) Равенство нулю первой вариации, дающее уравнение в слабой форме $a(u, v) = (f, v)$ и приводящее к методу Галёркина.

3) Процесс минимизации Ритца на подпространстве.

В следующем разделе подробно излагается метод конечных элементов, рассматриваются многие из наиболее важных способов выбора кусочно полиномиальных «элементов».

Требование гладкости граничной кривой Γ создает одну трудность, но избавляет от других. С одной стороны, ясно, что внутренность Ω нельзя разбить на многоугольники, скажем на треугольники, без потери точности около границы. Эта трудность в рамках теории аппроксимации обсуждается в гл. 3. С другой стороны, гладкость границы позволяет предположить гладкость самого решения. Это свойство следует из теории эллиптических краевых задач, если коэффициенты уравнения и правая часть также гладкие.

Рассмотрим для сравнения задачу $u_{xx} + u_{yy} = 1$ на многоугольнике с условием $u = 0$ на границе. Для единичного квадрата решение ведет себя, как $r^2 \log r$ вблизи угла, а вторые про-

изводные рвутся. (Разрывность очевидна, поскольку в угловой точке u_{xx} и u_{yy} равны нулю, а их сумма — единице.) Здесь u имеет вторые производные в среднем квадратичном; u принадлежит $\mathscr{H}^2$, но не $\mathscr{H}^3$. Это можно проверить непосредственно, разлагая u в ряд Фурье. Поэтому к задаче применима оценка ошибки кусочно линейной аппроксимации, но точность, связанную с элементами более высокого порядка, нельзя увеличить без специального построения сетки в углах или без введения специальной пробной функции с той же особенностью, что и у решения u. Для невыпуклого многоугольника, например для L-образной области, решение u не обладает вторыми производными даже в среднем квадратичном. Решение должно принадлежать пространству $\mathscr{H}^1$, представляющему собой допустимое пространство (вернее, $\mathscr{H}^1$ *содержит* допустимое пространство, зависящее от краевых условий), и так как u вблизи тупого угла L-образной области ведет себя как $r^{2/3}$, то u не принадлежит $\mathscr{H}^2$. Особенности решения, возникающие из-за нарушения гладкости Г, изучаются в гл. 8.

Начнем с задачи Дирихле для уравнения Пуассона

$$-\Delta u = f \quad \text{в} \quad \Omega,$$

краевое условие — типа Дирихле:

$$u = 0 \quad \text{на} \quad \Gamma.$$

Знак «минус» в дифференциальном уравнении выбран потому, что для оператора

$$L = -\Delta = -\frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial y^2}$$

соответствующая квадратичная форма (Lu, u) положительна.

Как и в одномерном случае, нужно сначала ввести норму для неоднородного члена, т. е. выбрать множество правых частей f, для которых задача Дирихле решается. Выберем норму, как прежде:

$$\|f\|_0 = \left(\iint_\Omega |f(x, y)|^2\, dx\, dy\right)^{1/2}.$$

Если она конечна, то f принадлежит пространству $\mathscr{H}^0(\Omega)$. Как и в одномерном случае, это пространство содержит все кусочно непрерывные функции f и не содержит δ-функций.

В качестве возможных решений дифференциального уравнения исследуем все функции u, равные нулю на границе Г и имеющие в Ω производные до второго порядка включительно.

Естественной нормой для такого пространства решений будет

$$\| u \|_2 = \left[\iint\limits_{\Omega} (u^2 + u_x^2 + u_y^2 + u_{xx}^2 + u_{xy}^2 + u_{yy}^2)\, dx\, dy \right]^{1/2}.$$

Пространством функций, для которых *эта* норма конечна (т. е. функций, вторые производные которых обладают конечной энергией), будет $\mathscr{H}^2(\Omega)$. Пространство решений задачи Дирихле $\mathscr{H}_B^2$ — подпространство в $\mathscr{H}^2$, определяемое краевым условием $u = 0$ на Γ.

Из определения норм ясно, что оператор Лапласа $L = -\Delta$ является ограниченным оператором из $\mathscr{H}_B^2$ в $\mathscr{H}^0$:

$$\| Lu \|_0 \leqslant K \| u \|_2.$$

В теории Дирихле решающую роль играет обратное в некотором смысле утверждение: оператор, обратный к L, задаваемый функцией Грина задачи, дает решение u, непрерывно зависящее от правой части f. Это означает, что *для каждой функции f существует единственное решение u и для некоторой постоянной* ρ

$$\| u \|_2 \leqslant \rho \| f \|_0. \tag{47}$$

Это решение можно построить непосредственно методом конечных разностей; чаще всего используют *пятиточечное* разностное уравнение

$$\frac{-U_{i+1,j} + 2U_{i,j} - U_{i-1,j}}{(\Delta x)^2} + \frac{-U_{i,j+1} + 2U_{i,j} - U_{i,j-1}}{(\Delta y)^2} = f_{i,j}. \tag{48}$$

Около границы это уравнение надо изменить, но здесь удается сохранить и второй порядок точности схемы, и дискретный принцип максимума, очевидный из (48): если $f = 0$, то $U_{i,j}$ не может превышать ни одного из четырех значений $U_{i\pm1,\,j\pm1}$. (Мы не знаем, существует ли теоретический предел порядка точности схем, удовлетворяющих принципу максимума. Ясно одно: если точность возрастает, то краевые условия для разностного уравнения становятся чрезвычайно сложными.)

Не входя в детали, отметим, что уравнение, конечно, требует непрерывности f, чтобы были корректно определены значения $f_{i,j} = f(i\Delta x, j\Delta y)$. Если же функция f недостаточно гладка, можно применить некоторый осредняющий процесс (который опять будет исходить из вариационных методов!). Наилучшая оценка для пятиточечной схемы на квадрате дает

$$\max_{i,\,j} | u_{i,j} - U_{i,j} | \leqslant Ch^2 | \ln h | \max | f |.$$

Здесь $h = \Delta x = \Delta y$. Дополнительный множитель $|\ln h|$, ненужный в одномерном случае, здесь необходим из-за характера поведения $r^2 \ln r$ в угле.

Для нас важна формулировка задачи Дирихле в вариационной форме: *среди всех допустимых функций v, равных нулю на границе* Γ, *только решение и минимизирует квадратичный функционал*

$$I(v) = \iint_\Omega (v_x^2 + v_y^2 - 2fv)\, dx\, dy.$$

Сначала нужно показать, что решение u дифференциальной задачи действительно минимизирует I. Изменяя u в направлении v, получаем

$$I(u + \varepsilon v) = I(u) + 2\varepsilon \iint (u_x v_x + u_y v_y - fv) + \varepsilon^2 \iint (v_x^2 + v_y^2).$$

Достаточно убедиться, что коэффициент при ε равен нулю. Тогда, так как коэффициент при ε^2 положителен, если $v \neq \text{const}$ (если $v = \text{const}$, то $v = 0$ из краевых условий), u будет единственной функцией, минимизирующей I. Равенство нулю коэффициента при ε, или первой вариации, означает, что

$$\iint u_x v_x + u_y v_y = \iint fv \tag{49}$$

для всех допустимых v. Докажем с помощью теоремы Грина, что решение задачи Дирихле удовлетворяет (49). Проинтегрируем по частям в области Ω:

$$-\iint_\Omega (u_{xx} + u_{yy} + f)v\, dx\, dy + \int_\Gamma u_n v\, ds = 0. \tag{50}$$

Здесь u_n — производная от u в направлении внешней нормали. Так как $v = 0$ на Γ, а u удовлетворяет уравнению Пуассона в Ω, то первая вариация равна нулю. Следовательно, u — минимизирующая функция.

Конечно, все это можно сделать и в обратном направлении. Уравнение Пуассона $-\Delta u = f$ приводится к своей *слабой форме* умножением на функцию v, равную нулю на границе, а затем интегрируется по Ω с преобразованием левой части по формуле Грина. Слабая форма дает в точности уравнение (49). К нему применим *метод Галёркина*; это уравнение удовлетворяется в подпространстве S^h и решение u^h ищется тоже в S^h. Как легко видеть, если квадратичная форма $a(v, v)$ самосопряжена и положительно определена, *минимизация Ритца и метод Галёркина приравнивания нулю первой вариации эквивалентны.* Уравнение в слабой форме $a(u, v) = (f, v)$ сохраняет смысл и без требования самосопряженности.

Все эти рассуждения основаны на теории эллиптических уравнений. Предположим теперь, что все начинается с квадратичного функционала $I(v)$, и попытаемся непосредственно мини-

мизировать его. Первая проблема — выяснить точно класс функций v, допустимых в процессе минимизации.

Соображения здесь очень похожи на одномерный случай разд. 1.5. Всякая гладкая функция v, равная нулю на Г, допустима, так как, если $f = -\Delta v$, функция v будет той самой минимизирующей функцией, которую мы хотим получить. С другой стороны, требование гладкости v, например $v \in \mathscr{H}_B^2$, не обязательно для определения $I(v)$. Поэтому пополняем допустимое пространство: если $a(v - v_N, v - v_N) \to 0$ для некоторой последовательности $\{v_N\}$ из $\mathscr{H}_B^2$, то функция v также будет допустима. Процесс пополнения приводит к классу функций, который интуитивно представляется естественным: функция v должна равняться нулю на Г (условие Дирихле — главное), но она обладает только *первыми* производными в среднем квадратичном. Другими словами, квадратичный член $a(v, v)$ в функционале $I(v)$ должен иметь смысл,

$$a(v, v) = \iint (v_x^2 + v_y^2)\, dx\, dy < \infty. \tag{51}$$

Функции, удовлетворяющие (51), принадлежат $\mathscr{H}^1(\Omega)$. Подпространство функций, удовлетворяющих к тому же условию Дирихле $v = 0$ на Г, обозначается $\mathscr{H}_0^1(\Omega)$. (В наших обозначениях это было $\mathscr{H}_E^1$, но в специальном случае задачи Дирихле будем употреблять $\mathscr{H}_0^1$; в литературе также встречается обозначение $\mathscr{H}^1$.) *Это и есть пространство допустимых функций для задачи Дирихле.*

Важно подчеркнуть, что краевое условие $v = 0$ сохраняется для всего допустимого пространства, но не потому, что нам так хочется, а потому, что оно настолько сильное, что справедливо в пределе. Другими словами, условие Дирихле $v = 0$ на Г *устойчиво* в $\mathscr{H}^1$-норме: если последовательность функций v_N, равных нулю на границе, сходится в естественной энергетической норме к v, то предел v также равен нулю на границе.

Такой устойчивости краевого условия нет для уравнения

$$-\Delta u + qu = f \quad \text{в} \quad \Omega \tag{52}$$

с естественным краевым условием

$$u_n = 0 \quad \text{на} \quad \Gamma.$$

Функционал для этой *задачи Неймана* равен

$$I(v) = a(v, v) - 2(f, v) = \iint (v_x^2 + v_y^2 + qv^2 - 2fv)\, dx\, dy.$$

Для дифференциального уравнения решение u ищется в $\mathscr{H}_B^2$, т. е. оно обладает двумя производными и удовлетворяет усло-

вию Неймана $u_n = 0$. Однако каждая функция v из $\mathscr{H}^1$ есть предел последовательности функций из $\mathscr{H}_B^2$: пространство $\mathscr{H}_B^2$ *плотно* в $\mathscr{H}^1$. Поэтому после пополнения *допустимым пространством для вариационной задачи Неймана будет все пространство* $\mathscr{H}^1(\Omega)$. В результате на пробные функции v^h в методе Ритца не налагается никаких краевых условий, приемлемо любое подпространство $S^h \subset \mathscr{H}^1$. При практическом применении метода конечных элементов это означает, что значения v^h в граничных точках не подчинены никаким ограничениям — это несколько упрощает дело по сравнению с задачей Дирихле. (В действительности естественные краевые условия приводят к возрастанию численной неустойчивости, а это заставляет нас сомневаться в преимуществах задач типа задачи Неймана.)

Конечно, минимизирующая функция u (но не ее приближение u^h) должна автоматически удовлетворять условию Неймана, если она достаточно гладкая. Это подтверждается уравнением (50) для первой вариации, равной нулю для всех v из допустимого пространства $\mathscr{H}^1(\Omega)$. Прежде всего функция $w = u_{xx} + u_{yy} + f$ должна равняться нулю всюду в Ω, и функцию v можно взять равной w в некоторой малой окрестности Γ и нулю в остальных точках. Это дает $\iint w^2 = 0$. Так как $w = 0$, ясно, что $u_n = 0$ на границе. Поэтому условие Неймана справедливо для u, даже если оно не выполняется для всех допустимых v.

Если в (52) $q = 0$, то очевидно, что решение u не единственно: $u + c$ будет решением для любой постоянной c. Такая свобода выбора решения, если учесть альтернативу Фредгольма, наводит на мысль, что должно быть ограничение на правую часть f. После интегрирования обеих частей дифференциального уравнения $-\Delta u = f$ по Ω левая часть равна нулю по формуле Грина (50) с $v = 1$. Поэтому ограничение на f таково: задача Неймана с $q = 0$ не может иметь решения при $\iint f\,dx\,dy \neq 0$. Факт отсутствия решения непосредственно проверяется на одномерной задаче Неймана

$$u'' = 2, \quad u'(0) = u'(1) = 0.$$

Парабола $u = x^2 + Ax + B$, т. е. общее решение уравнения $u'' = 2$, не может удовлетворять краевым условиям. Соответственно

$$\int u'' = u'(1) - u'(0) = 0 \neq \int 2.$$

Напротив, решение задачи Дирихле в этом случае единственно.

Различие между условиями Дирихле и Неймана должно проявляться также в теории, когда проверяется эквивалентность

энергетической нормы $\sqrt{a(v, v)}$ и обычной нормы $\|v\|_1$ в допустимом пространстве. Проблема существования и единственности упирается в вопрос: является ли задача *эллиптической*, т. е. существует ли такая постоянная $\sigma > 0$, что

$$a(v, v) \geqslant \sigma \|v\|_1^2 \tag{53}$$

для всех допустимых v?

Для уравнения $-\Delta u + qu = f$ это то же самое, что

$$\iint v_x^2 + v_y^2 + qv^2 \geqslant \sigma \iint v_x^2 + v_y^2 + v^2.$$

При $q > 0$ эллиптичность очевидна, так что существует единственная минимизирующая функция u (или u^h в методе Ритца). При $q = 0$ для задачи Неймана эллиптичности нет: если $v = 1$, левая часть равна нулю, а правая нет. В задаче Дирихле жесткое перемещение тела $v = 1$ недопустимо, и выполняется неравенство типа Пуанкаре:

$$\iint v_x^2 + v_y^2 \geqslant \sigma' \iint v^2 \quad \text{для} \quad v \in \mathscr{H}_0^1.$$

Таким образом, задача Дирихле — эллиптическая даже при $q = 0$. Действительно, эллиптичность просто означает, что q превышает наибольшее собственное значение $\lambda_{\max}$ оператора Лапласа Δ. В задаче Неймана $\lambda_{\max} = 0$, $v = 1$ — соответствующая собственная функция и при $q = 0$ эллиптичность отсутствует. В задаче Дирихле $\lambda_{\max} < 0$, и она остается эллиптической даже для некоторых отрицательных значений q.

Конечно, можно потребовать, чтобы $u = 0$ только на части границы, скажем на Γ_1, и положить $u_n = 0$ на $\Gamma_2 = \Gamma - \Gamma_1$. Допустимое пространство для такой *смешанной задачи* состоит из функций $v \in \mathscr{H}^1$, равных нулю на Γ_1, и решение будет иметь особенность на стыке Γ_1 и Γ_2.

Другая возможность выбора краевого условия — равенство нулю на границе *косой производной*:

$$u_n + c(x, y) u_s = 0 \quad \text{на} \quad \Gamma,$$

где u_s — производная по касательной. Это *естественное* краевое условие, связанное со скалярным произведением

$$a(u, v) = \iint (u_x v_x + u_y v_y + c u_x v_y - c u_y v_x + c_y u_x v - c_x u_y v).$$

Это скалярное произведение дает уравнение Пуассона как интеграл от $u_x v_x + u_y v_y$. Действительно, тождество Грина преобразует равенство $a(u, v) = (f, v)$ в равенство

$$-\iint (\Delta u + f) v + \int (u_n + c u_s) v \, ds = 0.$$

Изложенное хорошо иллюстрирует тот факт, что *не существует единого способа интегрирования* (Lv, v) *по частям для получения энергетической нормы* $a(v, v)$. Различные действия над (Lv, v) приводят к различным формам $a(v, v)$ и соответственно к различным естественным краевым условиям. Мы увидим это еще раз чуть позже в примере, где коэффициент Пуассона входит в краевые условия и в энергетическую норму $a(v, v)$, но не в оператор $L = \Delta^2$.

Неоднородные краевые условия для уравнения $-\Delta u = f$ бывают двух типов: либо закрепление на границе

$$u = g(x, y) \quad \text{на} \quad \Gamma,$$

либо задание нагрузок на границе, что можно включить в условие Ньютона общего типа

$$u_n + d(x, y)\, u = b(x, y) \quad \text{на} \quad \Gamma.$$

С вариационной точки зрения эти условия совершенно различны. Первое представляет собой неоднородное условие Дирихле, и ему должны удовлетворять пробные функции; решение u минимизирует $\iint v_x^2 + v_y^2 - 2fv$ на классе функций $v \in \mathscr{H}^1$, удовлетворяющих условию $v = g$ на границе. Заметим, что допустимый класс $\mathscr{H}_E^1$ здесь не будет пространством: сумма двух допустимых функций равна $2g$ на границе, так что не будет допустимой. Однако *разность* двух допустимых функций равна нулю на границе и принадлежит пространству $\mathscr{H}_0^1$. Простейшее описание допустимого класса $\mathscr{H}_E^1$ таково: возьмем в нем какую-нибудь функцию, скажем $G(x, y)$, совпадающую с g на границе; тогда каждая допустимая функция v представима в виде $G + v_0$, где v_0 принадлежит $\mathscr{H}_0^1$. Короче

$$\mathscr{H}_E^1 = G + \mathscr{H}_0^1.$$

В методе Ритца *не требуется, чтобы пробные функции в точности совпадали с* g *на границе*. Достаточно, чтобы пробные функции имели вид $v^h = G^h(x, y) + \sum q_j \varphi_j^h(x, y)$, где $\varphi_j^h \in \mathscr{H}_0^1$, а G^h принимает на Γ значения, *близкие* к g. Это означает, что класс пробных функций есть

$$S^h = G^h(x, y) + S_0^h,$$

где S_0^h — подпространство в $\mathscr{H}_0^1$. В разд. 4.4 проверяется, что в этом случае применима основная теорема 1.1 метода Ритца (можно даже отказаться от предположения, что $S_0^h \subset \mathscr{H}_0^1$, но это уже выходит за рамки теории Ритца).

Наконец, рассмотрим краевое условие $u_n + du = b$. Это условие будет естественным, если основной функционал $\iint v_x^2 + v_y^2 - 2fv$ изменить, вводя функции d и b. Изменения касаются только граничных членов:

$$I(v) = \iint_\Omega v_x^2 + v_y^2 - 2fv + \int_\Gamma (dv^2 - 2bv)\, ds.$$

Подчеркнем, что первый новый член входит в энергию

$$a(v, v) = \iint (v_x^2 + v_y^2) + \int dv^2$$

и вносит, таким образом, вклад в матрицу жесткости K метода Ритца. Новый линейный член $-2bv$ вносит вклад в граничные компоненты вектора нагрузок F. Допустимым пространством будет все пространство $\mathcal{H}^1$, и потому любую непрерывную между элементами кусочно полиномиальную функцию можно использовать как пробную.

Мы хотим исследовать три задачи четвертого порядка, относящиеся к бигармоническому уравнению

$$\Delta^2 u = u_{xxxx} + 2u_{xxyy} + u_{yyyy} = f \quad \text{в} \quad \Omega. \tag{54}$$

Это уравнение описывает поперечное перемещение u тонкой пластины под действием силы $f(x, y)$ с нормализованным коэффициентом жесткости $D = 1$. Как обычно, число m краевых условий равно половине порядка уравнения, т. е. $m = 2$.

Первая возможность — условия Дирихле, означающие физически *закрепленную пластину*:

$$u = 0, \quad u_n = 0 \quad \text{на} \quad \Gamma.$$

Большой интерес представляет также *задача о свободно опертой пластине*, в которой одно краевое условие главное, а другое естественное. Как и в задаче для уравнения Пуассона с косой производной, *вид естественного краевого условия будет зависеть от вида вариационного интеграла* $I(v)$. В теории упругости естественное краевое условие включает коэффициент Пуассона ν, определяющий изменение ширины при растяжении материала в длину; обычно выбирают $\nu = 0{,}3$. Краевые условия, определяемые физическими соображениями, таковы:

$$u = 0, \quad \nu \Delta u + (1 - \nu) u_{nn} = 0 \quad \text{на} \quad \Gamma. \tag{55}$$

Сюда входят, конечно, случаи $\nu = 0$ и $\nu = 1$, которые могут возникнуть в других приложениях. Заметим, что на прямолинейной границе все производные по касательной равны нулю, $\Delta u = u_{nn}$, и ν *исчезает из краевого условия*. В разд. 4.4 показано, к каким замечательным и парадоксальным следствиям

приводит это исчезновение при аппроксимации окружности многоугольником.

Наконец, исследуем чистую задачу Неймана, соответствующую *свободной границе*. Второе краевое условие (55) сохраняется; его часто записывают в виде

$$-\frac{M_{nn}}{D} = u_{nn} + \nu(\varphi_s u_n + u_{ss}) = 0,$$

где φ — угол между нормалью и осью x. Условие $u = 0$, фиксирующее край, отбрасывается. Чтобы почувствовать, зачем это условие нужно, попробуйте вычислить коэффициент при δu, когда меняется энергетический функционал в (56). Этот коэффициент найден Ландау и Лифшицем в терминах u и φ: он обычно записывается в сжатой форме Кирхгофа $Q_n + \partial M_{ns}/\partial s = 0$. В практических задачах вся граница редко бывает свободной.

Все приведенные условия, разумеется, можно переписать, непосредственно заменяя производные конечными разностями, однако построить хорошие уравнения вблизи границы становится необычайно трудно, и лучше уж сразу начинать с вариационной формулировки задачи.

С вариационной точки зрения, задача состоит в минимизации функционала

$$I(v) = a(v, v) - 2(f, v) =$$
$$= \iint_{\Omega} (v_{xx}^2 + v_{yy}^2 + 2\nu v_{xx} v_{yy} + 2(1-\nu) v_{xy}^2 - 2fv)\, dx\, dy \qquad (56)$$

при соответствующих краевых условиях. В задаче Неймана не налагается ограничений на границе и пространство допустимых функций v есть в точности $\mathscr{H}^2(\Omega)$. В задаче Дирихле требуется $v = 0$ и $v_n = 0$; подпространство, удовлетворяющее этим ограничениям, есть $\mathscr{H}_0^2(\Omega)$. В промежуточном случае свободно опертой пластины главным является только условие $v = 0$; обозначим соответствующее подпространство через $\mathscr{H}_{ss}^2$ и заметим, что $\mathscr{H}_0^2 \subset \mathscr{H}_{ss}^2 \subset \mathscr{H}^2$. Теорема Грина дает эквивалентность этих вариационных и дифференциальных задач.

Подчеркнем, что четвертые производные, появляющиеся в члене $\Delta^2 u$ в теореме Грина, не требуются для справедливости *теоремы о минимуме потенциальной энергии* в вариационной формулировке. Обратное тоже верно. Предел функций, имеющих непрерывные четвертые производные и стремящихся к решению, может оказаться функцией другого типа, а идея пополнения состоит в получении допустимого (условиями минимума) пространства функций, удовлетворяющих лишь главным крае-

вым условиям и обладающих конечной энергией $a(v, v)$. Решение u тогда удовлетворяет уравнению в слабой форме $a(u, v) = (f, v)$.

Для теории метода Ритца ключевой момент — эквивалентность энергетической нормы $\sqrt{a(v, v)}$ обычной норме $\| v \|_2$. Это опять условие эллиптичности: существует такая постоянная $\sigma > 0$, что для всех допустимых v

$$a(v, v) \geqslant \sigma \| v \|_2^2,$$

или

$$\iint v_{xx}^2 + v_{yy}^2 + 2\nu v_{xx} v_{yy} + 2(1-\nu) v_{xy}^2 \geqslant$$
$$\geqslant \sigma \iint v^2 + v_x^2 + v_y^2 + v_{xx}^2 + v_{xy}^2 + v_{yy}^2.$$

При полностью свободной границе эллиптичность нарушается, так как решение единственно только с точностью до линейной функции $a + bx + cy$. Если дифференциальную задачу заменить задачей $\Delta^2 u + qu = f$, соответствующей пластине с постоянным вращением $q = \rho\omega^2 > 0$, она снова станет эллиптической.

Эллиптичность трудно доказать, если — как в задаче линейной упругости — неизвестных u_j два или три и энергия деформации содержит только определенные комбинации $\varepsilon_{ij} = (u_{i,j} + u_{j,i})/2$ их производных. Существует, правда, *неравенство Корна*, утверждающее, что энергия деформации превосходит $\mathscr{H}^1$-норму, т. е. $\sum \int \varepsilon_{ij}\varepsilon_{ij} \geqslant \sigma \sum \int u_{i,j} u_{i,j}$.

Прежде чем строить конечные элементы для решения таких задач, обсудим кратко функциональные пространства $\mathscr{H}^s(\Omega)$. В одномерном случае они описываются очень просто: функция v принадлежит $\mathscr{H}^s[0, 1]$, если она является первообразной порядка s от f, $\int f^2 dy < \infty$. Отсюда следует, что функция v и ее первые $s - 1$ производных непрерывны; только s-я производная, т. е. исходная функция f, быть может, имеет скачки или что-нибудь похуже.

На плоскости возможна ситуация, когда функция v *разрывна и в то же время дифференцируема*. Такова, например, функция $v = \log\log 1/r \in \mathscr{H}^1$ в круге $r \leqslant 1/2$:

$$\iint v_x^2 + v_y^2 = \iint \left[v_r^2 + \left(\frac{v_\theta}{r}\right)^2\right] r\, dr\, d\theta =$$
$$= \iint (\log r)^{-2}\, d(\log r)\, d\theta = \frac{2\pi}{\log 2}.$$

Таким образом, производные от функции v, как и она сама, обладают конечной энергией, но функция разрывна в начале

координат. В n-мерном пространстве действует общее правило: *если* $v \in \mathscr{H}^s$ *и* $s > n/2$, *то функция* v *непрерывна и*

$$\max |v(x_1, \ldots, x_n)| \leqslant C \|v\|_s. \tag{57}$$

В этом смысл замечательного *неравенства Соболева*, связывающего два свойства: непрерывность и конечность энергии производных. Если $v \in \mathscr{H}^s$ и $s \leqslant n/2$, то гарантировать непрерывность v нельзя.

По соображениям двойственности с помощью теоремы Соболева можно выяснить, когда n-мерная δ-функция принадлежит $\mathscr{H}^{-s}$. Конечно, это возможно, только если $-s < 0$; лишь после достаточного количества интегрирований δ-функция может обладать конечной энергией. Норма в $\mathscr{H}^{-s}$ определяется аналогично (12):

$$\|w\|_{-s} = \max_v \frac{|(v, w)|}{\|v\|_s} = \max_v \frac{\left|\int wv\, dx_1 \ldots dx_n\right|}{\|v\|_s}. \tag{58}$$

Если w есть δ-функция, скажем в начале координат, то

$$\|\delta\|_{-s} = \max_v \frac{|v(0)|}{\|v\|_s}.$$

Согласно неравенству Соболева (57), эта величина конечна и δ-функция принадлежит $\mathscr{H}^{-s}$ тогда и только тогда, когда $s > n/2$.

В частности, δ-функция на плоскости не принадлежит $\mathscr{H}^{-1}$ и соответственно фундаментальное решение уравнения Лапласа $u = \log r$ не принадлежит $\mathscr{H}^{-1}$:

$$\iint u_x^2 + u_y^2 = \iint \left[u_r^2 + \left(\frac{u_\theta}{r}\right)^2\right] r\, dr\, d\theta = \iint r^{-1}\, dr\, d\theta = \infty.$$

Поэтому *точечно нагруженная мембрана, строго говоря, недопустима в вариационной задаче.*

Для пластины ситуация другая, так как дифференциальное уравнение имеет четвертый порядок. Пространством решений будет $\mathscr{H}^2$ (с краевыми условиями), а у пространства правых частей гладкость на 4 ниже, т. е. $\mathscr{H}^{-2}$. На плоскости δ-функция принадлежит $\mathscr{H}^{-2}$, и возможна ситуация точечно нагруженной пластины.

Еще один вопрос относительно функциональных пространств $\mathscr{H}^s$, очень важный для конечных элементов: *при каком условии элемент является согласованным*? Другими словами, если задано дифференциальное уравнение порядка $2m$ в пространстве n независимых переменных, то *какие кусочно полиномиальные функции принадлежат допустимому пространству* $\mathscr{H}_E^m$? Очень

легко проверить выполнение главных краевых условий; единственный вопрос — насколько гладким должен быть элемент для того, чтобы он принадлежал $\mathscr{H}^m$?

Стандартное условие согласованности хорошо известно: пробная функция и ее первые $m-1$ производных должны непрерывно продолжаться за границы элемента. Это условие, очевидно, достаточно для допустимости, так как m-е производные могут в худшем случае иметь скачок между элементами, а их энергия конечна. С другой стороны, пример $\log\log(1/r)$ показывает, что вряд ли это необходимое условие согласованности; существуют функции, не обладающие $m-1$ непрерывными производными, но принадлежащие $\mathscr{H}^m$ и являющиеся допустимыми. К счастью, такие «нехорошие» функции не могут быть кусочно полиномиальными. Если v — полином (или отношение полиномов) на каждой стороне границы элемента, то *v принадлежит $\mathscr{H}^m$ тогда и только тогда, когда производные порядка, меньшего m, непрерывно продолжены за границу элемента.* Залог успеха метода конечных элементов состоит в построении таких элементов, чтобы обеспечить удобный базис и одновременно высокую степень аппроксимации.

1.9. ТРЕУГОЛЬНЫЕ И ПРЯМОУГОЛЬНЫЕ ЭЛЕМЕНТЫ

В этом разделе описываются наиболее важные конечные элементы на плоскости. История их построения насчитывает примерно 30 лет, если вспомнить раннюю работу Куранта о кусочно линейных элементах; это одна из излюбленных тем в прикладной математике. Она требует знания алгебры лишь в рамках средней школы, а результаты ее очень важны — редкая и счастливая комбинация. Цель состоит в выборе кусочно полиномиальных функций, определяемых небольшим и удобным набором узловых значений, и достижении нужной степени непрерывности и аппроксимации.

Существует много элементов, конкурирующих между собой, и пока не ясно, что эффективнее — разбить область на треугольники или на четырехугольники. Очевидно, что треугольники лучше при аппроксимации криволинейной границы, а четырехугольники (особенно прямоугольники) имеют преимущество внутри области: их меньше и они позволяют строить простые элементы высших степеней. Уже эти замечания наводят на мысль, что лучше всего использовать обе возможности при условии, что их можно объединить с помощью узловых точек, обеспечив при этом требуемую непрерывность при стыковке.

Начнем с разбиения исходной области Ω на треугольники (рис. 1.10). Объединение этих треугольников даст многоугольник Ω^h, и, вообще говоря, если граница Γ криволинейна, при-

граничное множество $\Omega - \Omega^h$ будет непустым. Длину наибольшей стороны j-го треугольника обозначим через h_j, $h = \max h_j$. Ради удобства предположим, что Ω^h — подмножество в Ω и ни одна вершина треугольника не лежит на стороне другого треугольника. На практике, поскольку положение каждой вершины $z_j = (x_j, y_j)$ должно быть известно ЭВМ, триангуляция проводится так, как это допускают возможности ЭВМ.

Полностью автоматизированная подпрограмма триангуляции начинается с покрытия Ω регулярной треугольной сеткой, а затем вносятся необходимые исправления вблизи границы.

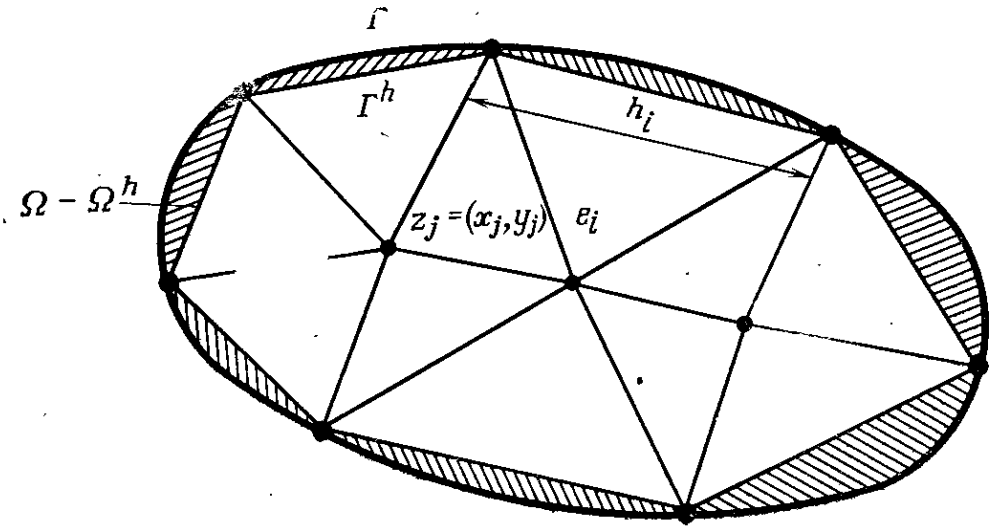


Рис. 1.10.

Разбиение многоугольника Ω^h на нерегулярные треугольники.

Если сетка оказывается слишком грубой в одной части области Ω и более мелкой в другой или область имеет углы и другие особенности, триангуляцию нужно производить вручную. В некоторых случаях, как показал Джордж [Д6], *сгущение* грубой сетки достигается разбиением каждого треугольника на четыре подобных; для такой механической процедуры идеально подходит ЭВМ. Пользователь может начать с введения в ЭВМ грубой сетки с минимальными затратами времени и усилий, а затем уже работать с полученной хорошей сеткой. Более того, он может использовать в своей программе один небольшой прием, который рекомендуем всем: проверять, приводит ли изменение в h к приемлемому изменению в численном решении.

Опишем простейший и самый важный способ построения пробной функции, если задана триангуляция. *Такая функция линейна внутри каждого треугольника* ($v^h = a_1 + a_2 x + a_3 y$) *и непрерывно продолжается за его стороны.* График $v^h(x, y)$ представляет собой поверхность, состоящую из треугольных кусков,

стыкующихся вдоль сторон. Это очевидное обобщение кусочно линейной функции в одномерном случае. Подпространство S^h, составленное из таких кусочно линейных функций, предложено Курантом [К15] для решения вариационных задач. Оно образует подпространство в $\mathscr{H}^1$, так как первые производные кусочно постоянны. Такие функции в самом начале развития метода конечных элементов были независимо исследованы Тёрнером и другими авторами; они иногда называются *треугольниками Тёрнера.*

Непрерывность позволяет избавиться от δ-функций в первых производных на границах между элементами; без этого ограничения функции не будут допустимыми и их (бесконечную) энергию на Ω нельзя получить суммированием по внутренностям всех элементов.

Простота пространства Куранта связана с тем, что внутри каждого треугольника три коэффициента функции $v^h = a_1 + a_2x + a_3y$ однозначно определяются значениями v^h в трех вершинах. Это означает, что *функцию можно удобно описать, задавая узловые значения*, или, что то же самое, S^h имеет удобный базис. Более того, v^h вдоль каждой стороны оказывается линейной функцией одной переменной и эта функция очевидным образом определяется значениями на концах стороны. Значение v^h в третьей вершине не влияет на функцию вдоль этой стороны независимо от того, где расположена эта третья вершина. *Поэтому непрерывность v^h на стороне гарантируется непрерывностью в вершинах.*

В случае главного краевого условия, скажем $u = 0$ на Γ, простейшее подпространство $S^h \subset \mathscr{H}_0^1$ образовано функциями, от которых требуется равенство нулю на границе многоугольника Γ^h. Расширенные нулем в полосе $\Omega - \Omega^h$, эти функции v^h непрерывны во всей области Ω, принадлежат пространству $\mathscr{H}_0^1$ и допустимы для задачи Дирихле.

Размерность N пространства S^h, т. е. число свободных параметров в функциях v^h, *совпадает с числом незакрепленных узлов.* (Граничный узел, в котором требуется равенство v^h нулю или другому заданному перемещению, называется закрепленным; он не влияет на размерность подпространства.) Для доказательства обозначим через $\varphi_j(x, y)$ пробную функцию, равную 1 в j-м узле и нулю в остальных. *Такие пирамидальные функции φ_j образуют базис в пространстве пробных функций S^h.* Произвольную функцию $v^h \in S^h$ можно представить единственным образом в виде линейной комбинации

$$v^h(x, y) = \sum_{j=1}^{N} q_j \varphi_j(x, y).$$

В таком виде координата q_j имеет естественный физический смысл: перемещение v^h в j-м узле $z_j = (x_j, y_j)$. Это характерная черта элементов, построенных для инженерных расчетов: каждая координата q_j соответствует значению функции v^h или одной из ее производных в узловой точке области.

Оптимальные координаты Q_j определяются из условия минимизации функционала $I(v^h) = I\left(\sum q_j\varphi_j\right)$, квадратичного по переменным $q_1, \ldots, q_N$. Подчеркнем, что минимизирующая функция $u^h = \sum Q_j\varphi_j$ *не зависит от выбора базиса.* Базис выбирался только для того, чтобы привести задачу к виду $KQ = F$, удобному для вычислений. С другой стороны, выбор базиса очень влияет на вычисление решения. Матрица жесткости K', полученная при выборе другого базиса, подобна матрице K; $K' = SKS^T$, где S — некоторая матрица, F заменяется на $F' = SF$. Вопрос сводится к выбору такого базиса, чтобы матрица K была разреженной и хорошо обусловленной и при этом элементы матриц K и F по возможности упрощали вычисление.

Выбор конечных элементов φ_j, основанный на интерполировании узловых значений, весьма эффективен. Матрица K в разумных пределах хорошо обусловлена и разрежена, так как два узла связаны только тогда, когда они принадлежат одному и тому же элементу. Более того, скалярные произведения $K_{ij} = a(\varphi_i, \varphi_j)$ и $F_j = (f, \varphi_j)$ находятся очень быстро по стандартному алгоритму, причем можно не вычислять по очереди сами скалярные произведения, а лишь вклады в них по всем треугольникам. Это означает, что интегралы вычисляются на каждом треугольнике, в результате образуются матрицы жесткости элементов k_i. Каждая матрица k_i включает только узлы i-го треугольника, остальные ее элементы равны нулю. Затем из этих блоков k_i строится глобальная матрица жесткости K, содержащая все скалярные произведения. В разд. 1.5 этот процесс был описан на интервале, а в следующем разделе мы дадим его обобщение на треугольную сетку.

Треугольники Куранта приводят к очень интересной матрице жесткости K. Для уравнения Лапласа получается стандартная пятиточечная разностная схема, если треугольники строятся регулярным образом, т. е. разбиением квадратной сетки диагоналями в северо-восточном направлении. (Более точную девятиточечную схему можно аналогично получить с помощью билинейных элементов [Ф9], но это редко делают.) Такая простая и систематическая структура матрицы жесткости позволяет использовать для решения уравнения $KQ = F$ *быстрое преобразование Фурье.* Оно дает отличный результат на прямоугольнике; его применение на непрямоугольных областях успешно развивается в работах Дорра, Голуба и др. С математической

точки зрения основное свойство пятиточечной схемы заключается в *принципе максимума*: все внедиагональные элементы K_{ij} в матрице жесткости отрицательны, диагональные элементы K_{ii} доминируют над ними, так что *обратная матрица* K^{-1} *неотрицательна*. Простой подсчет показывает, что это справедливо для линейных элементов на любой триангуляции, в которой нет углов, превышающих $\pi/2$. (Точное условие таково: сумма двух углов, прилежащих к данной стороне, не превышает π.) То же верно и для n-мерных линейных элементов. Так как K^{-1} и все $\varphi_j(x)$ неотрицательны, то для аппроксимации u^h по методу конечных элементов выполняется тот же физический закон, что и для точного перемещения u: *если нагрузка* f *всюду положительна, то таково же и перемещение*. Фрид задает вопрос: справедливо ли это для элементов высших степеней? Среди внедиагональных элементов K_{ij} могут быть положительные, но ведь условие их отрицательности не является необходимым для положительности матрицы K^{-1}.

В задаче Неймана не налагается никаких ограничений на v^h в граничных узлах и размерность пространства S^h равна общему количеству внутренних и граничных узлов. Базисные функции φ_j опять равны 1 в одном узле и 0 в остальных. В этом случае, однако, нельзя доопределять пробные функции нулем в $\Omega - \Omega^h$, так как они станут разрывными. Вместо этого можно просто продолжить линейно в каждую часть полосы $\Omega - \Omega^h$ функцию из прилегающего треугольника.

Есть и другие возможности, мы упомянем об одной: можно игнорировать приграничную полосу и изучать задачу Неймана только внутри многоугольной области Ω^h. Конечно, приближенное решение может значительно измениться и точность уменьшится по сравнению с минимизацией интеграла на Ω. Такие ошибки, связанные с изменением краевой задачи, оцениваются в гл. 4.

Займемся теперь более точными элементами. В технике конечных элементов решающим шагом было обобщение основной идеи Куранта о простых пробных функциях. Вместо линейности функции v^h внутри каждого треугольника будем предполагать ее квадратичность:

$$v^h = a_1 + a_2x + a_3y + a_4x^2 + a_5xy + a_6y^2. \tag{59}$$

Для того чтобы функция v^h принадлежала $\mathscr{H}^1$, она должна быть непрерывна при переходе через сторону в соседний треугольник.

Для удобства работы с таким подпространством нужно построить его базис, т. е. такое множество непрерывных кусочно квадратичных функций φ_j, что любой элемент из S^h единственным образом представляется в виде

$$v^h(x, y) = \sum q_j\varphi_j(x, y).$$

Для построения базиса существует прекрасная конструкция. Добавим к вершинам треугольников узлы, помещая их в середины сторон треугольников (рис. 1.11, *а*). Каждому узлу, будь он вершиной или серединой стороны, поставим в соответствие функцию φ_j, равную 1 в этом узле и 0 в остальных. Функция φ задается в каждом треугольнике в шести точках — в трех вершинах и в трех серединах сторон, поэтому 6 коэффициентов a_i в формуле (59) определяются однозначно.

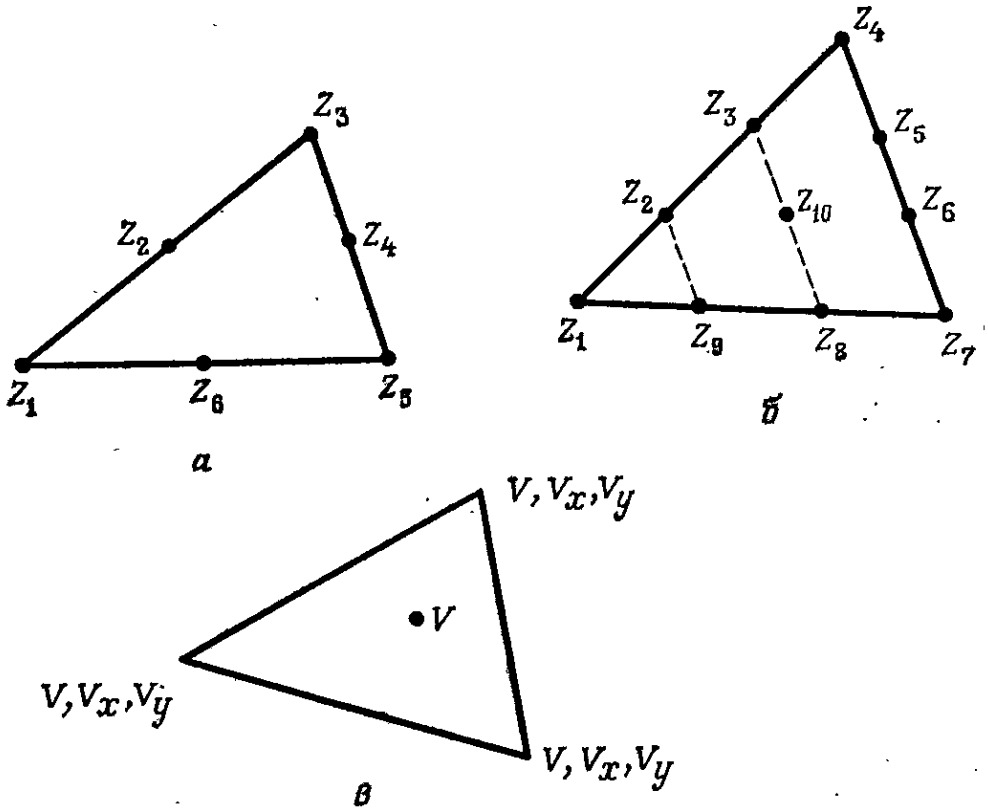


Рис. 1.11.

Расположение узлов для квадратичных и кубических элементов: *а* — непрерывные квадратичные элементы, *б* — иепрерывные кубические элементы, *в* — кубические элементы из пространства Z_3.

Докажем непрерывность построенных кусочно квадратичных элементов на сторонах треугольников. Доказательство очень простое. Вдоль каждой стороны v^h будет полиномом второй степени от одной переменной. На каждой стороне лежат 3 узла — две вершины и средняя точка. Полином второй степени тремя значениями в узлах определяется однозначно. Для двух соседних треугольников эти значения одинаковы, а значения в остальных узлах не влияют на v^h вдоль стороны, так что непрерывность доказана. Для стороны, лежащей на внешней границе Γ^h в задаче Дирихле, три узловые значения (а значит, и весь полином) равны нулю.

Каждая функция v^h из S^h допускает единственное разложение вида $\sum q_j\varphi_j$, где коэффициент q_j есть значение v^h в j-м

внутреннем узле (в середине стороны или в вершине). Поэтому φ_j образуют базис в пространстве S^h, *размерность* которого, таким образом, *равна числу N незакрепленных узлов.*

Для непрерывных кусочно кубических элементов базис строится точно так же. Кубический полином двух переменных x и y определяется десятью коэффициентами и, в частности, значениями в десяти узлах (рис. 1.11, *б*). Опять-таки 4 узловых значения на каждой из сторон определяют однозначно кубический полином, заданный на этой стороне, и непрерывность обеспечена. Это треугольный аналог одномерных кубических элементов, построенных в разд. 1.6 (тех, которые только непрерывны и имеют 2 узла внутри каждого интервала).

Та же самая конструкция распространяется на полиномы степени $k-1$ нескольких переменных $x_1, \ldots, x_n$ при условии, что основные области разбиения — симплексы: интервалы при $n=1$, треугольники при $n=2$, тетраэдры при $n=3$. Можно получить дискретные аналоги произвольно высокой степени точности в n-мерном пространстве. К сожалению, с точки зрения практических приложений существует фатальное обстоятельство: размерность пространства S^h, равная общему числу внутренних узлов, растет чрезвычайно быстро при росте k и n. Главная проблема в методе конечных элементов — наложить дополнительные ограничения на пробные функции (тем самым уменьшая размерность пространства S^h) без нарушения свойств аппроксимации и простоты локального базиса.

Это можно сделать, увеличивая требования на гладкость. В случае кубической аппроксимации можно добавить *условие непрерывности первых производных функции v^h в каждой вершине.* Очевидно, что получится подпространство пространства просто непрерывных кубических элементов v^h. Если m треугольников имеют общую вершину внутри области, непрерывность v_x^h и v_y^h налагает новые ограничения на пробные функции и размерность N соответственно уменьшается. Чтобы построить базис, уберем средние точки на сторонах и образуем в вершинах «тройной узел». Другими словами, 10 коэффициентов кубического полинома определяются значениями v, v_x v_y в каждой вершине и значением v в центре тяжести; этот десятый узел нельзя передвигать. Кубический полином однозначно определяется этими десятью значениями, а вдоль каждой стороны — четырьмя значениями (функции v и ее производной по направлению стороны в обоих концах этой стороны), и непрерывная стыковка обеспечена. В результате получится очень полезное пространство кубических пробных функций, которое мы обозначим через Z_3 (рис. 1.11, *в*).

Кусочно полиномиальные элементы из Z_3 легко описать внутри области. О границе поговорим особо, здесь есть несколько

возможностей. В случае главного условия $u = 0$ ограничением будет $u^h = 0$. Если предполагать $v^h = 0$ вдоль всей границы Γ^h, то и производные вдоль обеих хорд должны быть равны нулю, и в такой вершине не будет свободных параметров. Более удовлетворительную аппроксимацию дает изопараметрический метод разд. 3.3, или, в терминах переменных x, y, условие равенства нулю производной по направлению, касательному к Γ. В последнем случае мы должны отказаться от условия Дирихле $v^h = 0$ на границе Γ^h и от его продолжения на истинную границу Γ. Тогда *функции v^h будут действительно нарушать главное краевое условие, обеспечивающее допустимость*, и в таком варианте пространство кубических функций Z_3 не будет подпространством пространства Дирихле $\mathscr{H}_0^1$. Тем не менее удается дать строгую оценку ошибки, возникающей при использовании таких недопустимых элементов; это один из основных результатов гл. 4. Можно ожидать, что функция v^h близка к нулю на Γ, и если вычислять по криволинейным треугольникам в Ω, а не по обычным треугольникам в Ω^h, численные результаты будут лучше.

Кубические элементы имеют еще одно важное дополнительное достоинство: неизвестные Q_c, соответствующие узлам в центре треугольников, можно сразу исключить из системы метода конечных элементов $KQ = F$ и там останутся только неизвестные, соответствующие трем узловым точкам сетки. Такое исключение известно под названием *статической конденсации*. Оно связано с тем, что соответствующие базисные функции φ_c отличны от нуля *только внутри одного треугольника*, и каждое неизвестное Q_c выражается только через другие девять параметров в этом треугольнике. Таким образом, уравнение с номером c можно разрешить относительно Q_c через 9 «ближайших» параметров Q_j и исключить Q_c из системы, не увеличивая ширины ленты. Число арифметических операций прямо пропорционально числу отброшенных центральных точек, что необычно: в двумерных задачах общего вида нельзя исключить n узлов, произведя лишь αn операций. Физически оптимальное перемещение Q_c в центре определяется полностью девятью параметрами, задающими перемещение вдоль границы элемента.

Математически этот процесс можно рассматривать как *ортогонализацию* базисных функций φ_j по отношению к каждой центральной базисной функции φ_c. Разумеется, с помощью процесса Грама — Шмидта всегда можно ортогонализовать весь базис и свести матрицу жесткости к тривиальному виду $K = I$, но это было бы безумием. Легче прямо решить систему $KQ = F$. Специальная ортогонализация по отношению к φ_c возможна, так как в ней участвуют только соседние 9 функций φ_j и только внутри данного треугольника. Общее исключение неизвестных, отличное от обычного метода Гаусса, приводит к увеличению

ширины ленты матрицы, кроме специальных случаев — например, *четно-нечетной редукции*, с помощью которой исключается каждый второй узел в пятиточечной аппроксимации уравнения Лапласа.

Приведем еще один важный способ построения пространства кубических функций Z_3, в котором не применяется статическая конденсация; центральные точки здесь не будут узловыми, а будет только 9 параметров — значения v, v_x и v_y в каждой вершине. Соответственно нужно избавиться от лишней степени свободы в кубическом полиноме. Потребуем, чтобы коэффициенты при x^2y и xy^2 в разложении

$$v^h = a_1 + a_2x + a_3y + a_4x^2 + a_5xy + a_6y^2 + a_7x^3 + a_8(x^2y + xy^2) + a_9y^3$$

были равны. Это ограничение портит точность элемента; в дальнейшем мы свяжем скорость сходимости со степенью полинома, аппроксимируемого точно в пространстве пробных функций, и увидим, что степень снижается в этом случае с 3 до 2. Тем не менее такие кубические функции с ограничением относятся к наиболее важным конечным элементам, и многие инженеры предпочитают работать с ними, а не со всем пространством Z_3. Узловые параметры таких функций очень хороши. (Однако, как будет показано в разд. 4.2, оставшиеся 9 степеней свободы не инвариантны относительно поворота в плоскости x, y, и для некоторых направлений они даже не определяются однозначно девятью узловыми параметрами. Андерхегген предложил в качестве ограничения $\iint v^h = 0$, а Зенкевич [7] рассмотрел еще одну, весьма привлекательную возможность.)

Ни одно из описанных пространств нельзя применить для бигармонического уравнения, так как они не являются подпространствами в $\mathscr{H}^2$. (Вернее, их использование было бы незаконно; Z_3 с ограничением часто используется для расчета оболочек. Элементы, несогласованные вдоль внутренних сторон треугольников, обсуждаются в разд. 4.2.) Поэтому, обозначив через $\mathscr{C}^k$ класс функций с непрерывными производными до порядка k включительно, мы хотим построить элементы, принадлежащие классу $\mathscr{C}^1$. Существенно новое условие заключается в том, что производная по нормали должна быть непрерывна на границах между элементами. Функция из $\mathscr{C}^1$ автоматически принадлежит $\mathscr{H}^2$ и потому допустима для задач четвертого порядка; интегралы по Ω можно вычислить на каждом элементе, если нет δ-функций на границах.

Существует несколько возможностей. Одна состоит в том, чтобы заменить кубические полиномы в Z_3 *рациональными функциями*, выбранными так, чтобы уничтожить разрывы на сторонах в нормальных производных, не затрагивая при этом

узловых значений v, v_x и v_y. Функция v^h становится кусочно рациональной вместо кусочно полиномиальной, и опять уменьшается точность; пространство не содержит произвольного кубического полинома. Рациональные функции имеют, кроме того, другой важный недостаток: их трудно интегрировать и даже численное интегрирование приводит к серьезным трудностям (разд. 4.3).

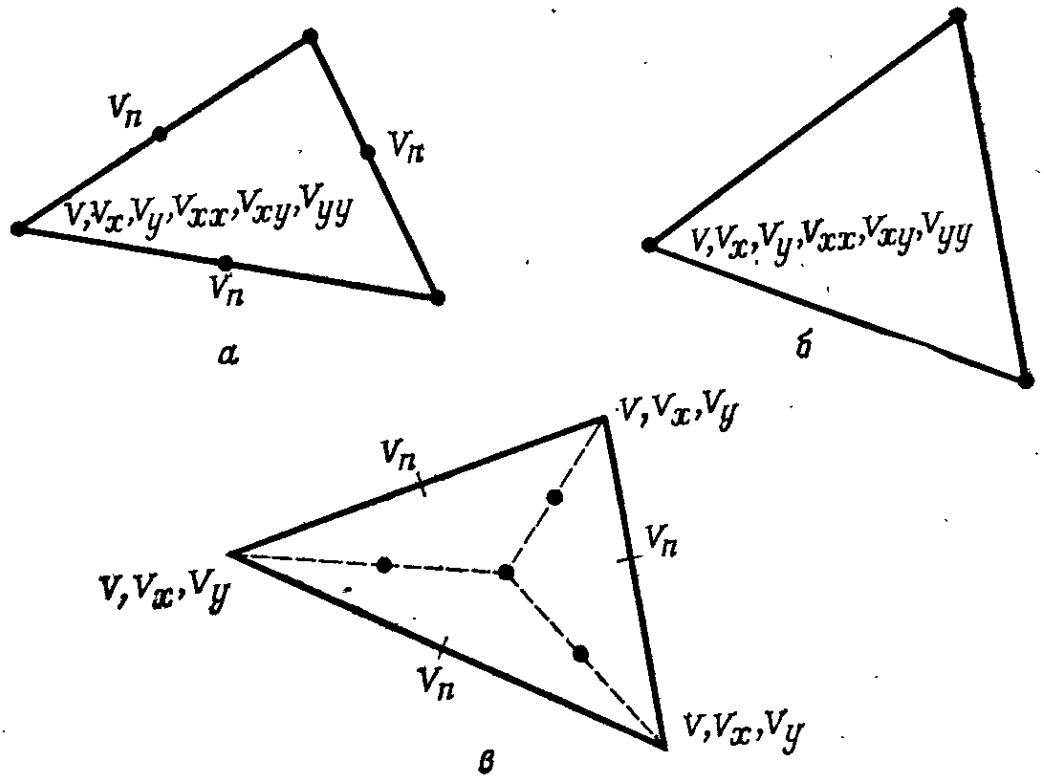


Рис. 1.12.

Два треугольника соответствуют аппроксимации пятой степени (*а* — без ограничений и *б* — с ограничением; функция v_n — кубическая на каждой стороне) и макротреугольник соответствует кубической аппроксимации (*в* — кубические полиномы класса $\mathscr{C}^1$; Клаф — Точер).

Если мы предпочитаем работать исключительно с полиномами, то приходим к наиболее важной и остроумной конструкции — элементам пятой степени (рис. 1.12). У полинома пятой степени по x и y надо определить 21 коэффициент, из которых 18 определяются значениями v, v_x, v_y, v_{xx}, v_{xy} и v_{yy} в вершинах. Вторые производные представляют собой *изгибающие моменты*, интересные с физической точки зрения; мы будем требовать их непрерывность в вершинах и получать их как результат процесса метода конечных элементов — они задаются непосредственно весовыми коэффициентами Q_j. Далее, при существовании вторых производных не возникает никаких трудностей на нерегулярной триангуляции. Полиномы пятой степени для двух треугольников на стороне между ними совпадают, так как три

условия в каждой вершине — значения функции v и ее производных v_s и v_{ss} по направлению стороны — однозначно определяют шесть его параметров в вершине.

Остается так задать три дополнительных ограничения, чтобы производная по нормали v_n была непрерывна между треугольниками. Один способ — добавить к узловым параметрам значения v_n в середине каждой стороны. Так как v_n — полином четвертой степени от s вдоль стороны, он однозначно определяется этим параметром вместе со значениями v_n и v_{ns} на концах отрезка (нормаль в средних точках сторон треугольника при переходе в соседний треугольник тоже надо учесть). Эта конструкция приводит к полному пространству полиномов пятой степени класса $\mathscr{C}^1$, построенному независимо по крайней мере в четырех работах. Оно обеспечивает точность по перемещению $O(h^6)$, если краевые условия удачно аппроксимированы. С помощью конечных разностей такая точность, по-видимому, никогда не достигалась для двумерных уравнений четвертого порядка.

Другой столь же эффективный способ введения трех дополнительных ограничений — потребовать превращения v_n в кубическую функцию вдоль каждой стороны, т. е. равенства нулю ведущего коэффициента в полиноме четвертой степени. При таком ограничении узловые значения v_n и v_{ns} определяют кубический полином вдоль стороны, и элемент принадлежит классу $\mathscr{C}^1$. Оценка ошибки для $u - u^h$ ухудшается с h^6 до h^5. В то же время размерность пространства S^h существенно уменьшается (в отношении 9 : 6, см. табл. 1.1), и это достаточная компенсация. Действительно, по результатам серии численных экспериментов [K14] можно отдать предпочтение этим замечательным элементам. По общему признанию, с ними не всегда просто работать (непрерывность вторых производных может нарушаться в углах области, где точное решение u менее гладкое), но точность все равно остается высокой [1]).

---

[1]) К такому классу относятся важные физические задачи с дополнительными условиями. Для несжимаемой жидкости, например, полезно наложить ограничения $\operatorname{div} u^h = 0$ для всех пробных функций. Представители фраицузской школы (Крузей, Фортен, Гловинский, Равьяр, Темам) обнаружили, что хотя треугольники Куранта не соответствуют условиям, сходимость можио установить при использовании

1) квадратичных полиномов на плоскости и кубических в трехмерием пространстве,

2) полиномов пятой степени из класса $\mathscr{C}^1$ как функций тока в пространстве двух переменных и полиномов четвертой степени, построеиных дифференцированием полиномов пятой степеии, для скорости поля,

3) несогласованных элементов, линейных иа треугольниках, но непрерывных только в *серединах* сторон (разд. 4.2).

Таблица 1.1

Треугольные элементы

Тип элемента	Гладкость	d	k	$N=Mn^2$
Линейный	$\mathscr{C}^0$	3	2	n^2
Квадратичный	$\mathscr{C}^0$	6	3	$4n^2$
Кубический	$\mathscr{C}^0$	10	4	$9n^2$
Кубический Z_3	$\mathscr{C}^0$, v_x, v_y непрерывны в вершинах	10	4	$5n^2$
Z_3 с ограничением	Как выше плюс равные коэффициенты при x^2y, xy^2	9	3	$3n^2$
5-й степеии	$\mathscr{C}^1$, v_{xx}, v_{xy}, v_{yy} непрерывны в вершинах	21	6	$9n^2$
5-й степени, иормальная производная кубическая на сторонах	$\mathscr{C}^1$, v_{xx}, v_{xy}, v_{yy} непрерывны в вершинах	18	5	$6n^2$
Кубический на макротреугольнике	$\mathscr{C}^1$	12	4	$M=6$

Заметим, что для достижения гладкости $\mathscr{C}^1$ элемента на треугольнике необходимо было задать даже вторые производные в вершинах. Женишек [Ж 1] доказал по этому поводу интересную теорему: для того *чтобы кусочно полиномиальная функция принадлежала классу* $\mathscr{C}^q$ *на произвольной триангуляции, узловые параметры должны включать все производные в узлах до порядка* $2q$ *включительно.* Он построил такие элементы в пространстве n переменных, используя полиномы степени 2^nq+1; эта степень представляется минимально возможной.

К счастью, существует способ обойти жесткое ограничение и все же построить согласованный элемент класса $\mathscr{C}^1$. Он состоит в образовании «макроэлемента» из нескольких стандартных элементов. Наиболее известен способ *треугольника Клафа — Точера*: комбинирование различных кубических полиномов в трех подтреугольниках (см. рис. 1.12). Окончательными узловыми параметрами в большом треугольнике будут значения v, v_x и v_y в вершинах и значения v_n в серединах сторон — всего 12 параметров. Гарантируется даже стыковка нормальной производной при переходе в соседний макротреугольник, так как эта производная есть квадратичная функция и потому она полностью определяется вдоль стороны параметрами на этой сто-

роне: значение v_n в средней точке задано непосредственно, а в двух вершинах — косвенно, как комбинация v_x и v_y. Так как каждый из трех кубических элементов имеет 10 степеней свободы, а макроэлемент — только 12 узловых параметров, нужно ввести 18 ограничений. Оказывается, это как раз и надо для достижения гладкости $\mathscr{C}^1$ внутри треугольника. Требование, чтобы v, v_x и v_y принимали одинаковые значения во всех внешних вершинах и в одной внутренней, и согласование v_n во всех средних точках сторон дают 18 ограничений.

Предполагая, что данная триангуляция позволяет объединять треугольники по три в макротреугольники, мы нашли таким образом базис пространства S^h всех кусочно кубических функций класса $\mathscr{C}^1$. Это один из случаев весьма важной и на вид очень трудной задачи — определить базис пространства кусочно полиномиальных функций $v(x_1, \ldots, x_n)$ степени $k-1$ класса гладкости $\mathscr{C}^q$ между симплексами. Мы не знаем даже, как определить (при $n > 1$) размерность такого пространства: в соответствии с табл. 1.1 кусочно кубические функции класса $\mathscr{C}^1$ имеют $M = 6$ параметров для каждой пары макротреугольников и потому в среднем одно неизвестное для каждого исходного треугольника [1]).

Чтобы подытожить свойства кусочно полиномиальных функций, описанных в этом разделе, сведем основные свойства в таблицу. В столбце d приведено число параметров, необходимое для определения полинома внутри каждой подобласти, т. е. число степеней свободы, если на соседние элементы не наложено ограничений. Целое число $k-1$ указывает на наивысшую степень полинома, аппроксимируемого точно в данном пространстве пробных функций; это означает, что полином степени k уже нельзя точно представить комбинацией пробных функций, и (как мы еще докажем) порядок ошибки $u - u^h$ равен $O(h^k)$. Наконец, N — размерность пространства пробных функций S^h в предположении, что Ω — квадрат, разбитый на $2n^2$ малых квадратов, разбитых на два треугольника диагональю с наклоном $+1$. В $N = Mn^2$ дается только основной член; будут, конечно, дополнительные члены, зависящие от условий на границе, но постоянная M самая важная. Коэффициент M для элемента, каждая вершина которого является p-кратным узлом, на каждой стороне лежит q узлов и внутри каждого треугольника содержится r узлов, равен $p + 3q + 2r$. В любой триангуляции в одной вершине сходятся два или более треугольников: сумма углов треугольника равна 180°, а вершине соответствует 360°. Далее, число сторон относится к числу треугольников, как

[1]) Добавлено в корректурах: сейчас мы уже догадываемся, какова будет размерность пространства кусочных полиномов степени $k-1$ класса гладкости $\mathscr{C}^q$, но ничего не представляем себе о конструкции базиса.

3:2, поскольку каждая сторона принадлежит двум треугольникам. Это объясняет веса 1, 2 и 3 при p, r и q. (*Второе доказательство*: если внутри уже существующего треугольника вводится одна новая вершина, это приводит к образованию дополнительно трех сторон и двух треугольников.) Правило 1:2:3 выполняется для любого множества треугольников, и оно означает, что в среднем на *треугольник приходится* $M/2$ *неизвестных*. Возрастание p приводит к умеренному возрастанию размерности и ширины ленты матрицы. Внутренние узлы также не страшны, так как их можно исключить с помощью статической конденсации, а узлы на сторонах приносят наибольший вред и возрастание их количества сильно отражается на времени счета.

Отметим, что при теоретическом сравнении двух конечных элементов эффективнее тот, который точно аппроксимирует многочлены более высокой степени $k-1$. Ошибка тогда убывает, как Ch^k, постоянная C зависит от конкретного элемента и от k-х производных от u. Поэтому в теоретическом плане единственным ограничением на скорость сходимости будет гладкость u, и даже его можно устранить (разд. 3.2), улучшая сетку.

Дело меняется, если, напротив, фиксировать требуемую точность и искать элементы, дающие такую точность с наименьшими затратами. Шаг сетки h здесь конечен, т. е. не является бесконечно малым. Возникает вопрос: достаточно ли провести вычисления лишь несколько раз, т. е. задача удобна для программирования, или же затраты на программирование и приготовления оправдываются только при длительном использовании программы? Мы считаем, что элементы фиксированного порядка, подобные элементам в табл. 1.1, или другие сходные конструкции (такие, как элементы на четырехугольниках, трехмерные элементы) будут обеспечивать достаточную свободу выбора при практическом применении метода конечных элементов.

Теперь мы хотим обсудить *прямоугольные элементы*, которые быстро завоевывают популярность. Они особенно хороши в трехмерных задачах, где один куб занимает тот же объем, что и 6 довольно сложных тетраэдров. (Нерегулярное разбиение на тетраэдры в трехмерном пространстве трудно осуществить даже с помощью ЭВМ.) Далее, на плоскости очень многие важные задачи решаются в прямоугольных областях или в областях, составленных из прямоугольников. Границу более сложной области нельзя удовлетворительно описать без использования треугольников, но очень часто появляется возможность комбинировать прямоугольные элементы внутри области с треугольными около границы.

Простейшая конструкция по аналогии с линейным элементом на треугольнике основана на *кусочно билинейных функциях* $v^h = a_1 + a_2x + a_3y + a_4xy$ в каждом прямоугольнике. Четыре

коэффициента такой функции определяются значениями v^h в вершинах. Вне четырехугольника пробная функция непрерывна, и ее можно применить к дифференциальным уравнениям второго порядка. Базис строится с помощью интерполяции: $\varphi_j = 1$ в j-м узле и $\varphi_j = 0$ в остальных узлах. Поверхность, соответствующая φ_j, напоминает пагоду или по крайней мере то, как мы ее себе представляем, так что функцию φ_j будем называть *функцией-пагодой*. Это произведение $\psi(x)\psi(y)$ основных кусочно линейных функций-крышек одной переменной; таким образом, пространство S^h — тензорное произведение двух более простых пространств. Это очень полезная конструкция.

Важно отметить, что для произвольного четырехугольника *такие кусочно билинейные функции не будут непрерывными при переходе от одного элемента к другому*. Предположим, что два четырехугольника прилежат к прямой $y = mx + b$. Вдоль этой общей стороны билинейная функция будет *квадратичной*; она линейна, только если сторона расположена горизонтально или вертикально. Квадратичный полином не определяется двумя узловыми значениями на концах стороны: на v^h влияют и другие узлы. Поэтому билинейные элементы можно использовать только на прямоугольниках. Правда, для общего случая четырехугольника можно изменить координаты так, чтобы он стал прямоугольником, и тогда допустимы билинейные функции. Более того, эту замену переменных можно также описать билинейной функцией, так что *в замене координат участвуют те же функции, что и в построении самого элемента*. Это простейшие из *изопараметрических элементов*, которые подробно обсуждаются в разд. 3.3.

Билинейный элемент на прямоугольниках можно легко соединить с линейным элементом Куранта на треугольниках, так как и тот и другой полностью определяются значениями v^h в узлах. Возможны и другие комбинации: билинейную функцию на треугольнике с узлом в середине одной из сторон можно соединить с квадратичной функцией на соседнем элементе. Вообще билинейные пробные функции лишь чуть-чуть «старше» линейных, так как они точно воспроизводят член второго порядка xy. Для лапласиана $L = -\Delta$ главная диагональ матрицы жесткости K, построенной с помощью билинейных элементов, пропорциональна 8, а остальные элементы пропорциональны -1 и соответствуют восьми соседним точкам на плоскости. В трехмерном пространстве, очевидно, нужно рассматривать *трилинейные* функции вида $a_1 + a_2x + a_3y + a_4z + a_5xy + a_6xz + a_7yz + a_8xyz$; такая функция опять определяется значениями в углах.

Так же, как билинейный элемент на прямоугольнике соответствует линейному элементу на треугольнике, биквадратичные и бикубические функции соответствуют квадратичным и куби-

ческим функциям класса $\mathscr{C}^0$ на треугольнике. (В многомерном случае это полиномы степени $k-1$ по каждой переменной $x_1, x_2, \ldots, x_n$, обладающие k^n степенями свободы внутри каждого параллелепипеда.) Биквадратичные элементы используются часто и описываются очень легко. На рис. 1.13 показано расположение узлов. Девять коэффициентов биквадратичного элемента определяются его узловыми значениями, и наличие трех узлов на каждой стороне обеспечивает непрерывность между прямоугольниками. Для бикубического элемента все аналогично.

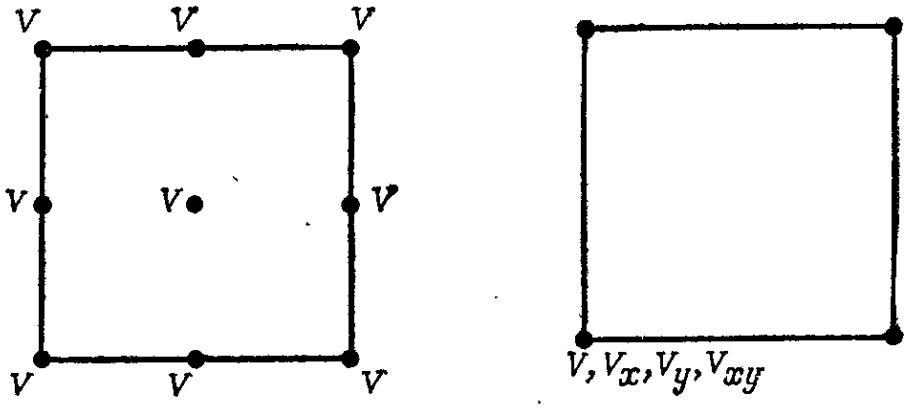


Рис. 1.13.
Расположение узлов при биквадратичной и эрмитовой бикубической интерполяции.

Простая модификация биквадратичного элемента состоит в исключении внутреннего узла и уменьшении числа параметров до восьми. Это делается за счет удаления из биквадратичного элемента члена x^2y^2, вклад которого в аппроксимацию u незначителен. Этот простой и удобный элемент принадлежит «сирендипову [1]) классу» элементов, описанному Зенкевичем. Практически он весьма полезен для четырехугольников с криволинейными сторонами, особенно после замены переменных, переводящей область в квадрат (разд. 3.3). Наряду с билинейным, этот элемент является самым полезным изопараметрическим элементом на плоскости.

Соответствующий трехмерный элемент определяется по 20 точкам и опять чрезвычайно удобен, так как не имеет ни внутренних, ни кратных узлов. Его узлы лежат в 8 вершинах и на серединах 12 ребер параллелепипеда. Допускается объединение членов вида $x^\alpha y^\beta z^\gamma$ второй степени по одной из переменных, например x^2yz, но не x^2y^2z или x^3.

---

[1]) Происхождение термина объясняется в книге Зенкевича О., Метод конечных элементов в технике, изд-во «Мир», М., 1975, стр. 124. — *Прим. перев.*

Элементы старшего порядка такого сирендипова типа на плоскости имеют $4p$ узлов, равномерно расположенных по периметру прямоугольника, включая четыре угловые точки. Функции такого типа содержат все члены $x^\alpha y^\beta$, в которых α и β не превышают p, а наименьший показатель степени равен 0 или 1. Поэтому первый недостающий член — это x^2y^2, и $k=4$. Эти функции опять особенно полезны при преобразованиях координат, переводящих границы прямоугольника в произвольные полиномиальные кривые степени p; при этом мы избавляемся от нежелательных внутренних узлов.

Существует несколько очень хороших элементов (предложенных Клафом, Фелиппа и др.), в которых четырехугольник получается как объединение двух или более треугольников, так что полином изменяется от одного треугольника к другому внутри четырехугольного макроэлемента. Расположение узлов и непрерывность между частями макроэлемента здесь довольно сложные. Мы опишем лишь один элемент другого типа, именно *эрмитов бикубический элемент*. Пробные функции снова будут кубическими по каждой переменной отдельно, $v^h=\sum a_{ij}x^iy^j$, $0\leqslant i,\ j\leqslant 3$, это дает 16 степеней свободы в каждом прямоугольнике. Параметры определяются значениями v, v_x, v_y и v_{xy} в четырех вершинах [1]). Таким образом, размерность пространства S^h намного меньше по сравнению с пространством обычных бикубических элементов, описанных выше и основанных на 16 различных узлах.

Такой элемент можно понимать как естественное обобщение фундаментального одномерного эрмитова кубического элемента, описанного в разд. 1.7. В этом случае функция $v=a_1+a_2x+a_3x^2+a_4x^3$ определялась на каждом подынтервале значениями v и v_x на его концах. Такая конструкция обеспечивала непрерывность v_x в узлах, а значит, и всюду, так что элемент принадлежал $\mathscr{C}^1$. Стандартный базис в одномерном случае состоял из функций двух типов, $\psi(x)$ и $\omega(x)$, интерполирующих значения функции и ее производных соответственно:

$$\psi_j=\begin{cases}1 & \text{в узле } x=x_j;\\ 0 & \text{в других узлах.}\end{cases}\qquad \omega_l=0 \text{ во всех узлах,}$$

$$\frac{d\psi_j}{dx}=0 \text{ во всех узлах,}\qquad \frac{d\omega_l}{dx}=\begin{cases}1 & \text{в узле } x=x_l,\\ 0 & \text{в других узлах.}\end{cases} \tag{60}$$

На эти функции натянуты все кусочно кубические функции класса $\mathscr{C}^1$.

---

[1]) Этот элемент был предложен в технической литературе Богнером, Фоксом и Шмитом.

Эрмитово бикубическое пространство есть произведение двух эрмитовых кубических пространств, и четыре параметра в обычных узлах $z = (x_j, y_l)$ приводят к четырем соответствующим базисным функциям:

$$\begin{aligned} \Phi_1 &= \psi_j(x)\,\psi_l(y), \quad & \Phi_2 &= \psi_j(x)\,\omega_l(y), \\ \Phi_3 &= \omega_j(x)\,\psi_l(y), \quad & \Phi_4 &= \omega_j(x)\,\omega_l(y). \end{aligned} \tag{61}$$

Подчеркнем, что *узлы должны лежать на прямоугольной решетке.* Прямые $x = x_j$ в одном направлении и прямые $y = y_l$ в другом могут быть произвольными, но их пересечения полностью определяют двумерный массив узлов. Поэтому бикубические элементы применяются только на прямоугольниках (или, после простого линейного преобразования плоскости, на параллелограммах).

На прямоугольной области эрмитов бикубический элемент — один из самых лучших. Его гладкость непосредственно следует из гладкости базиса (61); так как ψ и ω принадлежат $\mathscr{C}^1$, их произведения также обладают этим свойством. Поэтому *бикубические элементы можно употреблять для уравнений четвертого порядка*; пробные функции будут принадлежать $\mathscr{H}^2$. Даже смешанные производные $\partial^2 v/\partial x\,\partial y$ *все непрерывны.* (Пользуясь этим, можно охарактеризовать эрмитово пространство S^h, не прибегая к базису; оно состоит из всех непрерывных кусочно бикубических функций v, у которых v_x, v_y и v_{xy} непрерывны. Будем говорить в этом случае, что v принадлежит классу $\mathscr{C}^{1,1}$.) Замечательно то, что из обычных соображений не следует дополнительная гладкость функций. Функция v_{xy} квадратична вдоль каждой стороны и для двух соседних прямоугольников, однако совпадают только два значения v_{xy} на концах стороны, а по двум значениям нельзя определить квадратичный полином!

Идею эрмитовой конструкции можно распространить на элементы старшей степени $2q - 1$. В одномерном случае должно быть q разных функций, соответствующих двум функциям ψ и ω для кубического полинома; все их производные порядка меньше q будут равны нулю в узлах, за исключением p-й функции $\omega_p(x)$, у которой $(d/dx)^{p-1}\omega_p = 1$ в начале координат. Это наиболее естественный способ построения базиса для кусочно полиномиальных функций степени $2q - 1$ с $q - 1$ непрерывными производными. В двумерном случае нужно рассмотреть все возможные произведения $\omega_p(x)\,\omega_{p'}(y)$, а это значит, что каждому узлу соответствует q^2 неизвестных. Многие из них будут смешанными производными высокого порядка, так что конструкция слишком неэффективна при $q > 3$. Как всегда, число неизвестных можно уменьшить, предполагая дополнительную гладкость элементов. Предельный в этом направлении случай дают *сплайны*, обеспе-

чивающие наибольшую возможную гладкость. В этом случае каждому узлу соответствует только одно неизвестное, а базисными функциями служат B-сплайны $\varphi(x)$ на прямой и $\varphi(x)\varphi(y)$ на плоскости. Трудность, конечно, состоит в том, что в этой конструкции различные элементы связаны между собой; краевые условия становятся сложнее и изопараметрические преобразования (выводящие за прямоугольники) невозможны.

Так как все эти пространства натянуты на произведения одномерных базисных функций, матрицы жесткости K также могут допускать разложение на одномерные операторы. Грубо говоря, это происходит, когда в дифференциальном операторе L можно разделить переменные. На практике это встречается в параболических задачах, когда метод Галёркина приводит к неявной разностной схеме с двумерной матрицей массы M, или матрицей Грама, образованной из скалярных произведений функций φ_j, которую приходится обращать на каждом шаге по времени. Для разностных уравнений именно эта трудность породила *метод переменных направлений*, в котором обратная матрица приближалась с помощью обращений двух одномерных операторов. Для пространства, образованного как произведение одномерных, применима та же техника с обычной оговоркой, что если область не в точности прямоугольная, то метод переменных направлений дает хорошие результаты, но сходимость не доказана.

Таблица 1.2

Прямоугольные элементы

Тип элемента	Гладкость	d	k	$N=Mn^2$
Билинейный	$\mathscr{C}^0$	4	2	n^2
Биквадратичный	$\mathscr{C}^0$	9	3	$4n^2$
Биквадратичный с ограничением	$\mathscr{C}^0$	8	3	$3n^2$
Бикубический обычный	$\mathscr{C}^0$	16	4	$9n^2$
Бикубический эрмитов	$\mathscr{C}^{1,1}$	16	4	$4n^2$
Сплайны степеии $k-1$	$\mathscr{C}^{k-2,k-2}$	k^2	k	n^2
Эрмитов степени $k-1=2q-1$	$\mathscr{C}^{q-1,q-1}$	k^2	k	q^2n^2
Сиреидипов, $p>2$	$\mathscr{C}^0$	$4p$	4	$(2p-1)n^2$

В табл. 1.2 $M=p+2q+r$ для элемента с p параметрами в каждой вершине, q параметрами вдоль каждой стороны и r параметрами внутри каждого прямоугольника. Весовые коэффициенты равны 1, 2, 1 при любом разбиении области на четырехугольники: число вершин равно числу прямоугольников и равно половине числа сторон. В каждом координатном направлении берется n подынтервалов.

1.10. МАТРИЦЫ ЭЛЕМЕНТОВ В ДВУМЕРНЫХ ЗАДАЧАХ

В этом разделе мы укажем последовательность операций, производимых ЭВМ в процессе построения матрицы жесткости K, т. е. в процессе получения дискретной системы метода конечных элементов $KQ = F$. Кроме того, кратко опишем вычисление составляющих *вектора нагрузок* F. Мы не собираемся излагать частные детали, которые могут понадобиться программисту, а хотим прояснить решающий фактор успеха метода конечных элементов: при практическом применении метода Ритца чрезвычайно удобны полиномиальные элементы, подобные рассмотренным в предыдущем разделе, и, возможно, только они одни.

Напомним сначала, как появляется матрица K. Функционал $I(v)$, подлежащий минимизации, имеет в качестве основного члена квадратичное выражение $a(v, v)$, представляющее собой во многих случаях энергию деформации (или, строго говоря, удвоенную энергию деформации). В методе Ритца v отыскивается в конечномерном подпространстве S пробных функций вида $\Sigma q_j \varphi_j$. (Верхний индекс h в этом разделе будет опускаться, так как метод Ритца применяется к фиксированному набору элементов.) Подстановка пробной функции v в функционал энергии приводит к квадратичному выражению от координат q_j, описывающему энергию на подпространстве S:

$$a(v, v) = q^T K q. \tag{62}$$

Элементы матрицы K являются энергетическими скалярными произведениями $K_{jk} = a(\varphi_j, \varphi_k)$.

На практике эти скалярные произведения вычисляются косвенным образом. *Основное здесь — это вычисление энергетического интеграла* $a(v, v)$ *на каждом элементе* e, другими словами, на каждой подобласти разбиения Ω. Каждая такая часть всей энергии имеет вид

$$a_e(v, v) = q_e^T k_e q_e \tag{63}$$

по аналогии с (62). Вектор q_e содержит только те параметры q_j, которые вносят вклад в энергию на подобласти e. (Конечно, можно включить сюда все множество координат q_j, вводя в матрицу жесткости k_e элемента нулевые компоненты. Но гораздо лучше выбросить все функции φ_j, обращающиеся в нуль на подобласти e; тогда порядок матрицы k_e будет равен числу d степеней свободы полиномиальной функции внутри e.) Для линейных функций $v = a_1 + a_2 x + a_3 y$ на треугольниках, например, порядок матрицы k_e будет $d = 3$ [1]). При обычном выборе базиса

[1]) Девять элементов матрицы жесткости элемента для уравнения Лапласа возникают непосредственно из поточечных произведений сторон треугольника: $k_{ij} = s_i \cdot s_j / 2A$, где A — площадь элемента. Все внедиагональные элементы k_{ij} неположительны, если только треугольник не имеет тупого угла.

три компоненты вектора q_e будут равны значениям v в вершинах треугольника e. Для квадратичного элемента $d = 6$, а для элемента пятой степени будет 18 степеней свободы,

$$q_e^{18} = (v^1, v_x^1, v_y^1, v_{xy}^1, v_{xx}^1, v_{yy}^1, v^2, v_x^2, \ldots, v^3, v_x^3, \ldots).$$

Здесь, например, v_x^2 — производная по x во второй вершине треугольника; она равна весовому коэффициенту q_j при базисной функции φ_j, производная по x от которой в этой вершине равна 1, а все остальные узловые параметры равны 0. В дальнейшем мы будем рассматривать элементы пятой степени на треугольниках в качестве основного примера, так как они наиболее полно иллюстрируют возникающие трудности.

Теперь перед нами две задачи: вычислить матрицы k_e жесткости элементов и собрать их в общую энергию деформации

$$a(v, v) = q^T K q = \sum_e q_e^T k_e q_e.$$

Последний вопрос — это вопрос эффективного хранения, зависящий, в частности, от относительных параметров ЭВМ и задачи. Для очень больших задач одним из возможных способов организации служит *прямой метод*, в котором упорядочиваются элементы (подобласти), а не неизвестные. Матрицы для каждого элемента определяются по очереди, и по мере того как производятся вычисления для каждого элемента, содержащего некоторое неизвестное Q_n, в соответствующей строке матрицы K выполняются исключения и результаты запоминаются. Таким образом, в данный момент хранится только несколько неизвестных, принадлежащих нескольким элементам, некоторые из которых уже вычислены, а некоторые — нет.

В этом разделе мы уделим наибольшее внимание вычислению матриц элементов, распространяя на двумерные задачи технику, развитую в разд. 1.7 для эрмитовых кубических элементов. В конце раздела остановимся на численном интегрировании.

Существенный момент здесь состоит в том, что при использовании полиномиального базиса φ_j на многоугольных (не криволинейных) элементах энергия есть взвешенная сумма интегралов вида

$$P_{rs} = \iint_e x^r y^s \, dx \, dy.$$

Эти интегралы зависят от положения элемента e и от постоянных, которые можно протабулировать. Поэтому задача сводится к нахождению удобной системы координат, в которой легко описать геометрию области e, и связи между узловыми параметрами q_e и коэффициентами полинома v. Существует общее мнение, что «глобальные» координаты x, y не подходят,

и гораздо менее распространено, но все же принято считать, что локальная система координат наилучшая. Поэтому мы опишем пару возможностей.

Первая заключается в том, чтобы перенести начало координат в центр тяжести (x_0, y_0) треугольника e. Белл в [23] называет полученную систему координат *локально-глобальной*. Так как преобразование линейно, полином сохранит пятую степень по новым переменным $X = x - x_0$, $Y = y - y_0$:

$$v = a_1 + a_2X + a_3Y + a_4X^2 + \ldots + a_{21}Y^5.$$

Легко найти узловые параметры q_e^{18} в терминах этих коэффициентов a_i, а именно если новые координаты вершин обозначены (X_i, Y_i), то

$$\begin{aligned} v' &= a_1 + a_2X_1 + a_3Y_1 + a_4X_1^2 + \ldots + a_{21}Y_1^5, \\ v'_x &= v_X^1 = a_2 + 2a_4X_1 + \ldots, \\ v'_y &= v_Y^1 = a_3 + \ldots + 5a_{21}Y_1^4 \end{aligned} \tag{64}$$

и т. д. Для элемента с 21 степенью свободы нужно также вычислить три производных в серединах сторон; узловыми параметрами такого элемента будут $q_e^{21} = (q_e^{18}, v_n^1, v_n^2, v_n^3)$. Связь с вектором коэффициентов $A = (a_1, \ldots, a_{21})$ можно записать в матричной форме:

$$q_e^{21} = GA, \tag{65}$$

где первые 18 строк матрицы G задаются равенствами (64), а последние 3 включают не только координаты X_i, Y_i вершин, но и ориентацию треугольника. Обращая (65), получаем $A = G^{-1}q_e^{21} = Hq_e^{21}$.

Для элемента с 18 степенями свободы последние 3 строки матрицы G заменяются однородными ограничениями на коэффициенты a_i, такими, чтобы нормальная производная v_n вдоль каждой стороны была кубической функцией. Это приводит к соотношению

$$\begin{bmatrix} q_e^{18} \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix} = G_0A.$$

Обращая это уравнение, видим, что матрица H, связывающая q_e с A, теперь содержит 21 строку и 18 столбцов:

$$A = Hq_e^{18}.$$

Для всякого элемента первый шаг состоит в нахождении матрицы связи H между узловыми параметрами q и коэффициентами полинома a.

Вычислим теперь энергетический интеграл в терминах коэффициентов a_i. Например, для изогнутой пластины в разд. 1.8 этот интеграл имеет вид

$$a_e(v, v) = \iint_e \left(v_{XX}^2 + v_{YY}^2 + 2\nu v_{XX} v_{YY} + 2(1-\nu) v_{XY}^2\right) dX\, dY.$$

Подставляя сюда полином с v, получаем

$$a_e(v, v) = A^T N A, \tag{66}$$

где матрица N требует вычисления интегралов P_{rs}. Как только они вычислены, матрица жесткости элемента найдена. Из (66) при $A = Hq_e$ выводим

$$a_e(v, v) = q_e^T H^T N H q_e$$

и окончательно

$$k_e = H^T N H. \tag{67}$$

Таким образом, для нахождения матрицы жесткости k_e элемента e необходимо проделать вычисления двух типов: найти матрицу связи H между узловыми параметрами и коэффициентами полинома и матрицу N, состоящую из интегралов от полиномов.

Другой выбор системы локальных координат рекомендуют Купер, Коско, Линдберг и Олсон [К14]. В их системе геометрические величины a, b, c и θ на рис. 1.14 приобретают первостепенное значение; они легко вычисляются по координатам вершин (x_i, y_i). Пробные функции в повернутых координатах ξ и η остаются полиномами пятой степени. Поэтому вычисление новой матрицы N, задаваемой интегралами от полиномов на треугольнике, основано на удобной формуле

$$\iint_e \xi^r \eta^s\, d\xi\, d\eta = c^{s+1}\left(a^{r+1} - (-b)^{r+1}\right) \frac{r!\,s!}{(r+s+2)!}.$$

Нам нужна также новая матрица связи H, вычисляемая в два этапа. Сначала вычисляется матрица H', связывающая в плоскости ξ, η узловые параметры с коэффициентами полинома, а затем матрица вращения R, связывающая локальные и глобальные координаты; $H = H'R$. Детали полностью приводятся в [К14], и создается впечатление, что кубическое изменение нормальных производных у элементов пятой степени легко вносится в такую систему.

Интегралы $P_{rs} = \iint X^r Y^s \, dX \, dY$ можно вычислить без всяких вращений введением *координат, связанных с площадями.* Это наиболее естественные параметры, известные среди инженеров как *треугольные координаты*, а среди математиков как *барицентрические координаты.* Здесь каждая точка описывается тремя координатами, в сумме всегда дающими единицу: $\zeta_1 + \zeta_2 + \zeta_3 = 1$. Вершины треугольника e расположены в точках $(1, 0, 0)$, $(0, 1, 0)$ и $(0, 0, 1)$. Если эти вершины имели прямо-

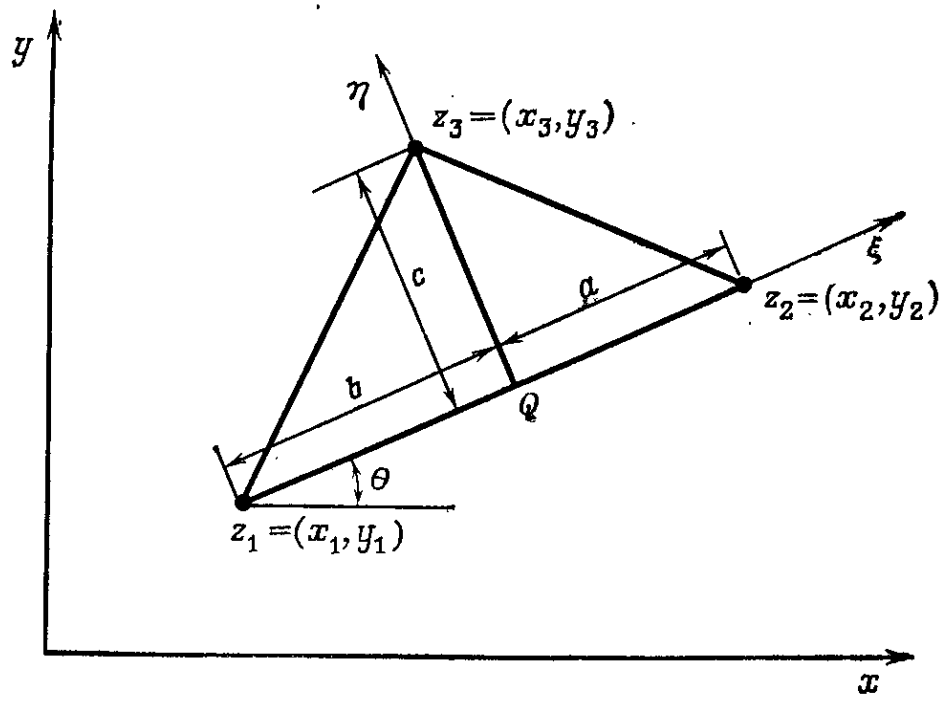


Рис. 1.14.
Другой выбор системы локальных координат.

угольные координаты (X_1, Y_1), (X_2, Y_2), (X_3, Y_3), то по линейности координаты (X, Y) и $(\zeta_1, \zeta_2, \zeta_3)$ произвольной точки P связаны соотношением

$$\begin{pmatrix} X \\ Y \\ 1 \end{pmatrix} = \begin{pmatrix} X_1 & X_2 & X_3 \\ Y_1 & Y_2 & Y_3 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_3 \end{pmatrix} = B \begin{pmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_3 \end{pmatrix}. \tag{68}$$

Геометрически ζ_i есть отношение площадей или расстояний до противоположной стороны (рис. 1.15). В этих координатах интеграл P_{rs}, согласно (68), принимает вид

$$P_{rs} = \iint (X_1\zeta_1 + X_2\zeta_2 + X_3\zeta_3)^r (Y_1\zeta_1 + Y_2\zeta_2 + Y_3\zeta_3)^s,$$

и теперь достаточно применить формулу интегрирования в барицентрических координатах (Холанд, Белл [23, стр. 84])

$$\iint \zeta_1^m \zeta_2^n \zeta_3^p = \frac{m!n!p!}{(m+n+p+2)!} \det B.$$

С помощью этой формулы можно вычислить якобиан преобразования координат: при $m = n = p = 0$ в знаменателе остается 2! и определитель матрицы B в (68) равен удвоенной площади A треугольного элемента.

Холанд и Белл [23] приводят явные формулы для интегралов P_{rs} в случае, когда начало координат расположено в центре тяжести треугольника. Замечательно, что при $r+s<6$ эти интегралы имеют очень простой вид:

$$P_{rs} = c_{r+s}A\left(X_1^rY_1^s + X_2^rY_2^s + X_3^rY_3^s\right),$$

причем $c_1 = 0$, $c_2 = 1/12$, $c_3 = c_4 = 1/30$, $c_5 = 2/105$. Из этих формул Белл получает матрицу жесткости k_e для кусочно по-

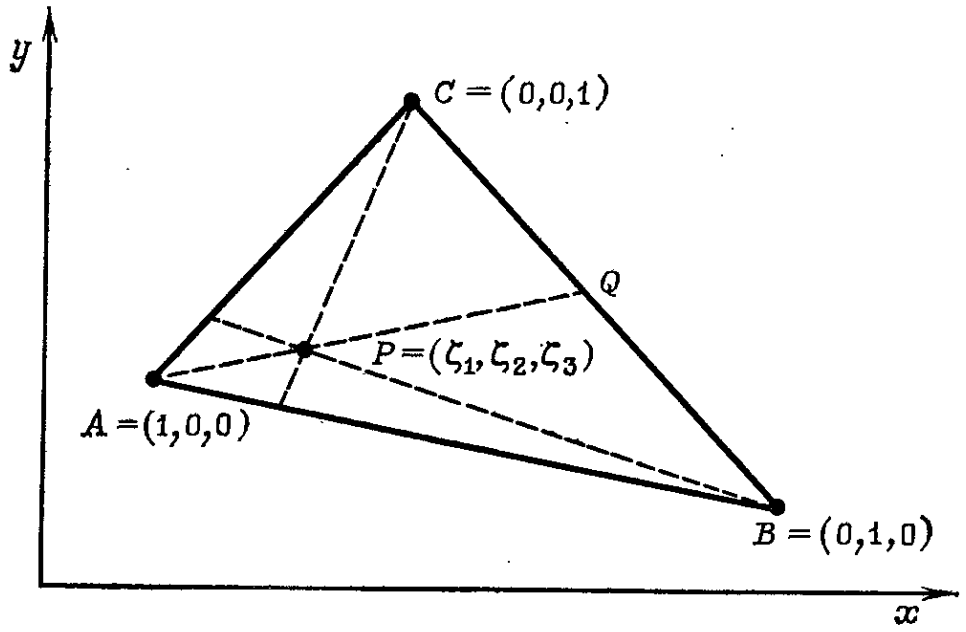


Рис. 1.15.
Координаты, связанные с площадями, для треугольника.

$$\zeta_1 = \frac{\text{длина } PQ}{\text{длина } AQ} = \frac{\text{площадь } BPC}{\text{площадь } BAC};$$

$$\zeta_1 + \zeta_2 + \zeta_3 = \frac{\text{площадь } (BPC + CPA + APB)}{\text{площадь } BAC} = 1.$$

линомиальных элементов пятой степени в применении к задаче об изгибе пластины.

Заметим, что другие локальные системы координат могут полностью основываться на естественных координатах ζ_j. В этом случае параметры q_e, включающие узловые производные вроде v_{xy}^1, можно заменить производными по ζ_j. Нам кажется, что алгебраические выкладки при использовании такой системы гребуют довольно много времени (и опыта!), однако Аржирис и его коллеги в этой системе успешно построили аналитически матрицу $H = G^{-1}$. Такая естественная система координат действительно оправдана численными квадратурами.

Теперь перейдем к вычислению *матрицы массы* M. Эта матрица появляется в связи с интегралом $\iint v^2\,dx\,dy$ в энергии. Другими словами, если записать v в виде $\Sigma q_j\varphi_j$, то M будет *матрицей Грама*[1]), образованной из скалярных произведений базисных функций φ_j:

$$M_{jk} = (\varphi_k, \varphi_j) = \int_\Omega \varphi_k\varphi_j\,dx\,dy.$$

Эта матрица играет основную роль при вычислении собственных значений.

Так же, как и матрицу K, матрицу массы можно построить из матриц массы элементов:

$$\iint_\Omega v^2 = q^T M q = \sum q_e^T m_e q_e = \sum \iint_e v^2. \tag{69}$$

Матрицу массы m_e элемента e можно, как раньше в случае k_e, записать в виде

$$m_e = H^T Z H.$$

Здесь H — та же самая матрица, связывающая параметры q_e с коэффициентами полинома

$$v = \sum_{i=1}^{d} a_i X^{m_i} Y^{n_i}.$$

Матрица Z сразу получается из скалярных произведений:

$$Z_{jk} = \iint_e X^{m_j} Y^{n_j} X^{m_k} Y^{n_k} = P_{m_j+m_k,\ n_j+n_k}.$$

Эти величины легко табулируются, и допустима любая локальная система координат. Таким образом, Z играет ту же роль для матрицы массы, что N для матрицы жесткости. В обоих случаях мы, конечно, считаем, что коэффициенты для элементов постоянны; если же материал пластины обладает переменными свойствами, которые нужно учитывать при расчетах, то следует применять численное интегрирование.

Наконец, обсудим вычисление вектора нагрузок F, который стоит в правой части уравнения метода конечных элементов $KQ = F$. Его вычисляют также поэлементно:

$$\iint_\Omega fv = q^T F = \sum q_e^T F_e = \sum \iint_e fv. \tag{70}$$

---

[1]) Мы хотели обсудить возникший каламбур «матрица Грама — матрица массы», но редактор сказал, что это успеется.

Вычисление опять начинается с соотношения $A = Hq_e$, связывающего узловые параметры элемента и коэффициенты полинома a_i. В терминах этих коэффициентов

$$\iint_e fv = \sum a_i \iint_e fX^{m_i}Y^{n_i} = A^T\sigma = q_e^T H^T \sigma, \qquad (71)$$

где компоненты d-мерного вектора σ определяются по формуле

$$\sigma_i = \iint_e fX^{m_i}Y^{n_i}.$$

Подставляя (71) в (70), получаем, что вектор нагрузок, соответствующий элементу e, равен

$$F_e = H^T\sigma.$$

В работе [K14] рассмотрены два специальных случая вычисления σ_i. Первый — это случай постоянной нагрузки $f(x, y) \equiv f_0$; здесь опять появляются затабулированные интегралы:

$$\sigma_i = \iint f_0 X^{m_i}Y^{n_i} = f_0 P_{m_i, n_i}.$$

Второй — случай нагрузки $f(x, y) = f_0\delta$, сосредоточенной в точке (X_0, Y_0). Если эта точка лежит вне элемента e, то, разумеется, $F_e = 0$; если же внутри, то

$$\sigma_i = f_0 X_0^{m_i} Y_0^{n_i}.$$

Все вычисления опять можно проводить в любой локальной системе координат.

При более общем виде вектора нагрузок f интегралы σ_i можно вычислить либо с помощью квадратурной формулы на каждом элементе, либо с помощью интерполирования f элементом из подпространства S^h. В последнем случае определяются узловые параметры f_j, т. е. значения функции f и ее производных в узлах, и строится *интерполяционный элемент*

$$f_I = \sum f_j \varphi_j. \qquad (72)$$

Замена f на f_I дает возмущенный вектор нагрузок $\tilde{F}$ с компонентами

$$\tilde{F}_k = \iint f_I \varphi_k = \iint \sum f_j \varphi_j \varphi_k = Mf';$$

M — матрица массы, описанная выше, f' — вектор, образованный из узловых параметров f_j. Эти вычисления проводятся сравнительно легко.

Итак, мы рассмотрели только аналитическое вычисление матриц элементов, основанное на точном интегрировании полиномов на многоугольниках. Конечно, *в подавляющем боль-*

шинстве случаев такие матрицы вычисляются приближенно, с помощью некоторых квадратур Гаусса. Для криволинейных элементов, возникающих при расчете оболочек или при решении плоских задач с криволинейными границами, численные квадратуры совершенно необходимы. В этом случае не только вектор F заменится на $\tilde{F}$, но и матрица K заменится на $\tilde{K}$. Это значит, что соответствующее решение $\tilde{Q}$ будет решением метода конечных элементов возмущенной задачи. В последней главе мы оценим ошибку $Q - \tilde{Q}$; она должна зависеть от точности квадратурной формулы.

Отметим, однако, что неточное численное интегрирование иногда может даже *улучшить* качество решения. Известен один пример (другой — для несогласованных элементов), в котором вычислительные эксперименты приводят к результатам, противоречащим положениям математического анализа, но с вычислительной точки зрения верным и важным. Улучшение для конечного шага h вытекает отчасти из следующего эффекта: точный метод Ритца всегда соответствует *слишком жесткой* аппроксимации и ошибки квадратурных формул уменьшают эту избыточную жесткость.

Жесткость — внутреннее свойство метода Ритца. Ограничивая перемещения v конечным числом величин $\varphi_1, \varphi_2, \ldots, \varphi_N$ вместо всех допустимых функций, мы получаем численную структуру, более ограниченную, чем реальная. В задаче на собственные значения такое ограничение выражается в том, что λ_j^h всегда больше истинного значения λ_j. В статических задачах потенциальная энергия $I(u^h)$ превышает $I(u)$, поскольку u^h получается из минимизации $I(v)$ на конечномерном подпространстве, натянутом на $\varphi_1, \varphi_2, \ldots, \varphi_N$. Такая верхняя оценка I соответствует *оценке снизу* энергии деформации a, как доказано в следствии из основной теоремы 1.1:

$$a(u^h, u^h) \leqslant a(u, u). \tag{73}$$

В частном случае точечной нагрузки $f = \delta(x_0)$, когда перемещение $u(x_0)$ пропорционально энергии деформации, оно также оценивается снизу в методе Ритца; более жесткая численная структура дает меньшие перемещения в нагруженной точке, чем истинная конструкция. Для распределенных нагрузок тенденция та же: перемещение $u^h(x)$ в методе конечных элементов обычно ниже истинного перемещения $u(x)$. Конечно, это не настоящая теорема, так как метод Ритца минимизирует энергию, а ее связь с перемещением не является строго монотонной. Другими словами, u^h может превышать u на некоторой части конструкции, но в то же время иметь меньшие производные в среднем квадратичном. Тем не менее односторонние

оценки и для перемещения, и для наклона те же, что и для конечных элементов. Это можно изменить либо фундаментальной перестройкой процесса Ритца (см. прямой, смешанный и гибридный методы, описанные в следующей главе), либо намеренным введением численных ошибок.

Последний эффект возникает в квадратурах Гаусса. Алгебраически он проявляется в уменьшении положительной определенности матрицы K; приближенная матрица $\tilde{K}$ удовлетворяет неравенствам

$$0 \leqslant q^T \tilde{K} q \leqslant q^T K q \quad \text{для всех} \quad q. \tag{74}$$

Может показаться парадоксальным, что это приводит к *увеличению* энергии деформации в решении по методу Ритца, но парадокс легко объяснить. Новое решение будет $\tilde{Q} = \tilde{K}^{-1} F$. Его энергия деформации равна $Q^T \tilde{K} \tilde{Q} = F^T \tilde{K}^{-1} F$, в то время как энергия невозмущенного решения была $Q^T K Q = F^T K^{-1} F$. Так как неравенство (74) для матрицы жесткости равносильно обратному неравенству для обратных матриц, то

$$F^T \tilde{K}^{-1} F \geqslant F^T K^{-1} F \geqslant 0 \quad \text{для всех} \quad F;$$

отсюда и следует возрастание энергии деформации.

В случае точечной нагрузки β в j-м узле это приводит, как и в непрерывном случае, к тем же выводам относительно перемещения. В векторе нагрузок F единственной ненулевой компонентой будет $F_j = \beta$ и перемещение в j-м узле будет равно $Q_j = (K^{-1}F)_j = \beta (K^{-1})_{jj}$. Если K уменьшится до $\tilde{K}$, то перемещение возрастет до $\beta (\tilde{K}^{-1})_{jj}$.

Остается выяснить, почему квадратура Гаусса приводит к описанному эффекту уменьшения K. Очевидно, K будет уменьшаться, если уменьшается каждая матрица элемента k_e. Строгое доказательство для одномерного случая можно найти в работе Айронса и Раззака [A7], где $(v^h)'$ разлагается в ряд по полиномам Лежандра. Энергия деформации на отрезке $[-1, 1]$ имеет вид

$$q^T k_e q = \int ((v^h)')^2 = \int (\alpha_0 + \alpha_1 P_1(x) + \ldots + \alpha_n P_n(x))^2 =$$

$$= 2\left(\alpha_0^2 + \frac{\alpha_1^2}{3} + \ldots + \frac{\alpha_n^2}{2n+1}\right);$$

n-точечная квадратура Гаусса сохраняет все члены до α_{n-1}^2 включительно, так как они возникают из полиномов степени меньше $2n$, и уничтожает член с α_n^2, поскольку $P_n = 0$ в узлах Гаусса (так определены полиномы Лежандра). Таким образом, в этом частном случае интеграл, очевидно, уменьшается и та

же самая тенденция сохраняется в более общих задачах. При сравнении различных конечных элементов полезен *критерий наименьшей матрицы жесткости.* Математически он должен отражать аппроксимационные свойства элементов и, в частности, числовые постоянные, которые входят в аппроксимации полиномов степени на единицу больше, чем у полиномов, точно аппроксимируемых конечными элементами. С вычислительной точки зрения это очевидно, и ряд теоретически возможных элементов сразу отбрасывается как слишком жесткие. Указанный критерий позволяет оптимизировать матрицу жесткости в предположении, что точно аппроксимируются полиномы степени $k-1$.

2 КРАТКОЕ ИЗЛОЖЕНИЕ ТЕОРИИ

2.1. БАЗИСНЫЕ ФУНКЦИИ ПОДПРОСТРАНСТВ S^h В МЕТОДЕ КОНЕЧНЫХ ЭЛЕМЕНТОВ

В настоящей главе мы собрали несколько основных результатов теории метода конечных элементов. Наша цель — описать общие рамки теории, в которой особое место отводится оценкам различных ошибок. В следующих главах мы будем рассматривать каждую из этих оценок подробно.

Сначала надо решить, какие подпространства S^h мы будем изучать. Для этого исследуем примеры, приведенные в гл. 1, и попытаемся выделить математически существенные свойства.

Дадим соответствующее этим примерам общее описание метода, называемого *методом узловых конечных элементов*. Оно будет фундаментом всей нашей теории. Каждая пробная функция v^h определяется своими узловыми параметрами — неизвестными q_j дискретной задачи. Каждый такой узловой параметр служит значением в заданном узле z_j либо самой функции, либо одной из ее производных. Таким образом, неизвестные можно представить в виде

$$q_j = D_j v^h(z_j),$$

где дифференциальный оператор D_j имеет нулевой порядок ($D_j v^h = v^h$), когда параметр служит значением функции, или равен одному из операторов $\partial/\partial x$, $\partial/\partial y$, $\partial/\partial n$, $\partial^2/\partial x\,\partial y$ и т. д.

Каждому из этих узловых параметров q_j поставим в соответствие пробную функцию φ_j, определяемую так: в точке z_j значение $D_j\varphi_j$ равно 1, а все остальные узловые параметры равны нулю. Заметим, что узлу z_j могут соответствовать несколько параметров, т. е. он может быть кратным узлом. Параметр q_j определяется единственным образом не точкой z_j, а парой (z_j, D_j). Таким образом, основное свойство заключается в том, что каждая функция φ_j удовлетворяет равенствам $D_j\varphi_j(z_j) = 1$ и $D_i\varphi_j(z_i) = 0$ при $i \neq j$:

$$D_i\varphi_j(z_i) = \delta_{ij}. \tag{1}$$

Функции φ_j образуют *интерполирующий базис* для пространства пробных функций, так как каждую пробную функцию

можно разложить по функциям φ_j:

$$v^h = \sum q_j \varphi_j.$$

Применяя к обеим частям равенства оператор D_i и вычисляя его в точке z_i, убеждаемся, что

$$q_i = D_i v^h(z_i).$$

Цель метода Ритца — минимизируя функционал $I(v^h)$, найти для этих параметров оптимальные значения Q_j. Тогда пробная функция с этими оптимальными параметрами будет аппроксимацией метода конечных элементов:

$$u^h = \sum Q_j \varphi_j.$$

Рассмотрим пару примеров в этой общей постановке. Для общеупотребимых кусочно линейных функций на треугольниках (треугольниках Тёрнера или Куранта) узлы z_j — это вершины триангуляции, а операторы D_j все нулевого порядка: $D_j v^h = v^h$. Неизвестные имеют вид $q_j = v^h(z_j)$ и базис образован пирамидальными функциями, определяемыми равенствами $\varphi_j(z_i) = \delta_{ij}$. То же справедливо для билинейных функций на четырехугольниках и для квадратичных на треугольниках, но здесь множество узлов z_j содержит и середины сторон. Для эрмитовых кубических функций одной переменной появляются производные: каждый узел z_j участвует в двух парах (z_j, I) и $(z_j, d/dx)$. Мы различали два вида базисных функций φ_j — функции ψ_j и ω_j. Для эрмитовых бикубических функций на каждый узел приходится четыре параметра, соответствующих v, v_x, v_y и v_{xy}, т. е. оператор D равен I, $\partial/\partial x$, $\partial/\partial y$ и $\partial^2/\partial x\,\partial y$. Для пространства кубических функций Z_3 на треугольниках узлы в вершинах тройные, а в центрах тяжести треугольников простые.

В заключение описания метода узловых конечных элементов остановимся на геометрии области. Область Ω разбивается на замкнутые подобласти e, пересекающиеся лишь по границам между элементами. *Каждый элемент e содержит не более d узлов z_j, и все базисные функции φ_j, кроме тех, которые соответствуют этим d узлам, равны нулю всюду на e.* Таким образом, функции φ_j образуют и *локальный базис*. Эта схема представляется достаточно общей, чтобы включать в себя большинство используемых пространств конечных элементов. Заметим тем не менее, что кубические сплайны не охватываются этой схемой, так как их базисные функции (B-сплайны) отличны от нуля на нескольких элементах.

Обратим внимание на одно дополнительное свойство функций φ_j, которое может оказаться полезным в теории метода. Это свойство появляется, лишь когда элементы построены гео-

метрически правильно, т. е. область покрывается равномерной сеткой с шагом h и в каждом квадрате сетки узлы попадают в одинаковые позиции. Свойство базиса тогда представляет собой свойство *инвариантности относительно переноса*: если базисная функция φ соответствует паре (z, D), а узел z сдвинут в новую точку $z^* = z + lh$, то *базисной функцией* φ^*, *соответствующей паре* (z^*, D), *будет перенос функции* φ:

$$\varphi^*(x) = \varphi(x - lh). \tag{2}$$

В n-мерном случае $l = (l_1, \ldots, l_n)$ — вектор с целочисленными координатами. Очевидно, что $D\varphi^*(z + lh) = D\varphi(z) = 1$; интерполирующая функция φ просто сдвинута и получена функция

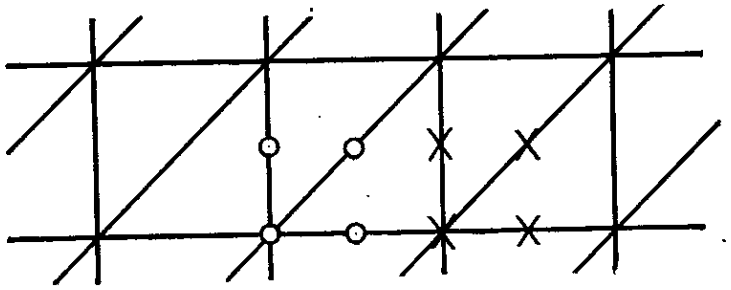


Рис. 2.1.
Узлы квадратичных функций из $\mathscr{C}^0$ на квадратной сетке.

φ^*, интерполирующая в точке z^*. Разумеется, у границы области Ω эта схема переносов нарушается, если только краевые условия не окажутся периодическими, в последнем случае можно считать, что базис обладает свойством периодичности.

Рассмотрим в качестве примера пространство непрерывных кусочно квадратичных функций, описанное в разд. 1.7. Узлы попадают на равномерную треугольную сетку, причем (рис. 2.1) шаг сетки равен 1. Заметим, что каждой точке сетки можно поставить в соответствие единственное множество из четырех узлов; кружками отмечены узлы, сопоставляемые с началом координат, а крестиками — с точкой $(1, 0)$. Отметим также, что число $M = 4$ базисных функций φ_j и неизвестных весовых коэффициентов q_j в одном квадрате сетки отличается от числа d степеней свободы; $d = 6$ в каждом треугольнике. *M — это коэффициент в последнем столбце таблицы в разд. 1.9, связанный с размерностью пространства* S^h. На единичном квадрате со свойством периодичности эта размерность равна M/h^2.

Обозначим через Φ_1, Φ_2, Φ_3 и Φ_4 базисные функции, интерполирующие в узлах z_1, z_2, z_3 и z_4 соответственно. Они кусочно квадратичны и равны нулю в каждом узле, кроме своего собственного, этим они полностью определяются. Базисные функции для четырех узлов, соответствующих точке сетки (l_1, l_2), получаются как раз описанным выше переносом; этими

сдвинутыми функциями будут $\Phi_i(x-l_1, y-l_2)$, $i=1, 2, 3, 4$. Уменьшение шага сетки в h раз приводит просто к изменению масштаба в независимых переменных: четырьмя базисными функциями, соответствующими точке сетки (l_1h, l_2h), будут $\Phi_i(x/h-l_1, y/h-l_2)$.

Эта схема переноса так полезна и важна потому, что мы положим ее в основу для второго общего описания метода конечных элементов. Эта форма метода применяется на равномерной сетке, назовем ее *абстрактным методом конечных элементов*. В n-мерном случае он начинается с выбора M функций $\Phi_1(x), \ldots, \Phi_M(x)$, которые в конечном счете приводят к M неизвестным на каждом кубе сетки, и уравнение метода конечных элементов $KQ=F$ принимает вид *системы M уравнений в конечных разностях*.

Для построения базисных функций, соответствующих определенной точке решетки, а именно началу координат, выберем просто масштаб переменной $x=(x_1, \ldots, x_n)$ так, чтобы получить $\Phi_1(x/h), \ldots, \Phi_M(x/h)$. Для построения базисных функций, соответствующих другой точке решетки $lh=(l_1, \ldots, l_n)h$, перенесем полученные только что функции. Таким образом, для обозначения всех базисных функций, построенных таким способом (выбором масштаба и переносом), нам понадобятся два индекса i и l:

$$\varphi^h_{i,l}(x)=\Phi_i(x/h-l), \quad i=1, \ldots, M.$$

Теперь осталось лишь потребовать, чтобы *исходные функции $\Phi_i(x)$ были равны нулю вне некоторого шара* $|x|\leqslant R$. Тогда $\varphi^h_{i,l}$ будут равны нулю вне локального шара $|x-lh|\leqslant Rh$, и у нас снова будет локальный (возможно, не строго локальный) базис.

Заметим, что в абстрактном методе конечных элементов *не* требуется, чтобы функции Φ_j интерполировали в некотором узле z_j и были кусочно полиномиальными. (Мы докажем, однако, что последние наиболее эффективны.) Интерполирующим свойством будет обладать любой базис, лежащий в пределах узлового метода конечных элементов, но наша абстрактная теория этим свойством не пользуется. В качестве примера рассмотрим кубические сплайны одной переменной. В этом случае $M=1$, и подходящий выбор для Φ_1 — это B-сплайн, приведенный на рис. 1.9. Пространство S^h всех комбинаций $\sum q_l\varphi^h_{1,l}$ тогда совпадает с пространством сплайнов, т. е. дважды непрерывно дифференцируемых кусочно кубических функций, стыкующихся в точках $x=0, \pm h, \pm 2h, \ldots$. Существенно, что B-сплайн Φ_1 не является интерполирующим, он отличается от нуля на *четырех* интервалах вместо двух, позволенных в узло-

вом методе, и, в частности, он не равен нулю в трех узлах вместо одного.

Подпространство сплайнов S^h, приемлемое для абстрактного метода конечных элементов, успешно применялось в одномерных приложениях. Построение уравнения $KQ = F$ и задание границы требуют изменений в технике, стандартной для узлового метода, но основные правила по-прежнему одинаковы для любой формы метода Ритца. Главное преимущество сплайнов в том, что дополнительная непрерывность уменьшает размерность пространства пробных функций без понижения степени аппроксимации. В узловом случае эрмитовы кубические полиномы определяются на каждом подынтервале значениями v и v' в его концах (это означает, что на каждую точку сетки приходится $M = 2$ параметров). Порождающие функции Φ_1 и Φ_2 изображены на рис. 1.8 (функции ψ и ω соответственно).

В двумерном случае также были проведены эксперименты со сплайнами, уменьшение числа M здесь еще значительнее, а выбор границы почти вынужден для обеспечения ее регулярности, сопоставимой с простотой узлового метода. Уменьшение ширины ленты матрицы жесткости K, разумеется, не следует из понижения числа M; у кубических сплайнов одной переменной в каждой строке матрицы K по семь ненулевых элементов, так как B-сплайн Φ_1 распространяется на четыре интервала. Фактически та же ширина ленты в случае эрмитовых полиномов, когда обычное упорядочение неизвестных дает матрицу вида

$$K = \begin{bmatrix} \cdot & \cdot & \cdot & \cdot & & \\ & \begin{pmatrix} x & x \\ x & x \end{pmatrix} & \begin{pmatrix} x & x \\ x & x \end{pmatrix} & \begin{pmatrix} x & x \\ x & x \end{pmatrix} & & \\ & & \begin{pmatrix} x & x \\ x & x \end{pmatrix} & \begin{pmatrix} x & x \\ x & x \end{pmatrix} & \begin{pmatrix} x & x \\ x & x \end{pmatrix} & \\ & & & \cdot & \cdot & \end{bmatrix}.$$

Число M непосредственно влияет на порядок матрицы K и сплайны становятся эффективнее (по крайней мере в смысле времени решения), когда область Ω — прямоугольник. Это подтверждается вычислениями в гл. 8.

Мы искренне верим, что узловой метод позволяет такую гибкость в отношении геометрических свойств, что в дальнейшем его применение будет предпочтительнее сплайнов. Несмотря на то что интерполирующее свойство несущественно для теории, удобнее иметь равномерную сетку, чтобы можно было использовать общее описание в терминах функций $\Phi_1, \ldots, \Phi_M$. Так как конструкция пространства S^h целиком зависит от этих функций, они должны содержать в себе ответы на все наши во-

просы об аппроксимации и численной устойчивости. Поэтому этот абстрактный подход сводит значительную часть теории метода конечных элементов к соответствующей задаче теории функций.

2.2. СКОРОСТИ СХОДИМОСТИ

Предположим, что u — решение n-мерной эллиптической вариационной задачи порядка m. Это означает, что u минимизирует $I(v)$ на допустимом классе $\mathscr{H}_E^m$, определяемом однородными или неоднородными главными краевыми условиями, и что (в силу эллиптичности) энергия деформации положительно определена: $a(v, v) \geqslant \sigma \| v \|_m^2$. Предположим также, что u^h — функция, минимизирующая $I(v)$ на пространстве пробных функций S^h, а $\tilde{u}^h$ — решение задачи, возмущенной ошибками численного интегрирования, координаты вектора $\tilde{u}^h$ удовлетворяют уравнению $\tilde{K}\tilde{Q} = \tilde{F}$. Предположим, наконец, что $\bar{u}^h$ представляет собой фактически вычисленное решение, отличающееся от $\tilde{u}^h$ из-за ошибок округления численного решения. Очевидно, что три приближения u^h, $\tilde{u}^h$, $\bar{u}^h$ содержат нарастающим образом источники ошибки. Мы хотим выяснить порядки величин этих ошибок для задач с гладкими решениями и для типичных конечных элементов.

Начало всегда одинаково: если S^h — подпространство в $\mathscr{H}_E^m$, то по основной теореме 1.1

$$a(u - u^h, u - u^h) = \min_{v^h \in S^h} a(u - v^h, u - v^h). \tag{3}$$

Теорема остается справедливой для неоднородных главных условий, когда разность двух любых функций из $\mathscr{H}_E^m$ принадлежит V_0, а разность двух любых функций из S^h принадлежит S_0^h, где S_0^h — подпространство в V_0. (Доказательство см. в разд. 4.4.) Поэтому оценка энергии деформации разности $u - u^h$ составляет вопрос чистой теории аппроксимации: оценить расстояние между u и S^h. Основные гипотезы таковы: во-первых, *степень пространства S^h равна $k - 1$* (S^h содержит в каждом элементе полный полином этой степени, подчиненный лишь у границы главным условиям) и, во-вторых, *базис его однороден при $h \to 0$*. Последнее в действительности представляет собой геометрическое условие на элементарные области: если диаметр области e_i равен h_i, то e_i содержит шар радиуса не менее τh_i, где τ строго больше нуля. Это ограничение не допускает произвольно малые углы в треугольных элементах. В четырехугольниках также не допускаются углы, близкие к π.

При этих условиях расстояние между u и S^h равно

$$a(u-u^h,\ u-u^h)\leqslant C^2h^{2(k-m)}|u|_k^2.$$

Поэтому скорость сходимости энергии деформации равна $h^{2(k-m)}$.

Вид этой оценки — типичный результат численного анализа. Отметим три факта. Показатель степени у h найти проще всего, так как он зависит лишь от степени полиномов. Он указывает скорость сходимости по мере измельчения сетки, этот эффект наблюдается при численном решении. Константа C зависит от конструкции элемента и его узловых параметров. Для правильных геометрических фигур можно найти хорошее асимптотическое значение C как ошибку в аппроксимирующих полиномах степени k (разд. 3.2). Третий множитель $|u|_k$ отражает свойства самой задачи, т. е. степень гладкости ее решения, и потому его легко оценить точно. Эта норма есть среднеквадратичное значение k-й производной от u и потому — в соответствии с теорией уравнений в частных производных — связана непосредственно с производными порядка $k-2m$ от функции f.

Заметим, что *сходимость имеет место тогда и только тогда, когда* $k>m$; другими словами, условие постоянной деформации таково, что *элементы должны воспроизводить точно любое решение, являющееся полиномом степени* m. Это требование для обеспечения сходимости постепенно появилось в технической литературе, оно возникло отчасти интуитивно, а отчасти из-за вычислительных неудач, связанных с нарушением этого правила (особенно заметных при изгибе пластины в бигармоническом случае $m=2$, когда пространство S^h не содержит член xy). Мы дадим строгое доказательство (насколько нам известно, первое) необходимости этого условия для сходимости в случае равномерной сетки. Такая теорема естественно соответствует абстрактной теории метода конечных элементов, допускающей наиболее общие пробные функции на равномерных разбиениях.

Мы сформулировали условие постоянной деформации так, как если бы оно было также необходимо и для неравномерных сеток, но это не так. По крайней мере не совсем так. Область Ω можно отобразить в другую область Ω' фиксированным гладким обратимым преобразованием T, не сохраняющим полиномов, а тогда элементы, удовлетворяющие на Ω условию постоянной деформации, не удовлетворяют ему в новых переменных. Тем не менее сходимость на Ω влечет сходимость соответствующей задачи на Ω', так что с помощью T можно переходить от одной области к другой. Влияние якобиана на ошибки для изопараметрических элементов разбирается в разд. 3.3.

Сходимость энергии деформации есть по существу сходимость производных порядка m от функции u^h к соответствующим производным от u. Эта производная, следовательно, особая, поскольку в методе Ритца минимизируется энергия. Для производной порядка s, где s может быть больше или меньше m, сходимость не может быть быстрее, чем $O(h^{k-1})$, т. е. чем порядок наилучшего приближения u в пространстве S^h степени $k-1$. Такая скорость сходимости обычно достигается решением u^h метода Ритца. Используя прием Нитше, дающий скорость $O(h^2)$ для перемещения в случае одномерных линейных элементов в разд. 1.6, покажем, что

$$\|u-u^h\|_s = O\left(h^{k-s}+h^{2(k-m)}\right). \tag{4}$$

Первый показатель почти всегда меньше, и скорость сходимости определяется теорией приближений. (Для $s=-1$ левая часть представляет собой осредненную по элементу ошибку в перемещении, и мы видим, что она может быть на один порядок лучше (h^{k+1}), чем сама ошибка в перемещении.) Тем не менее известны случаи, когда член $h^{2(k-m)}$ играет главную роль: если вообразить применение кубических сплайнов к задаче шестого порядка или, что реальнее, если для задачи изгиба пластины (уравнение четвертого порядка) взять только квадратичные функции на элементах, то скорость может быть ограничена порядком $2(k-m)=2$ даже для перемещений.

При условии, что в каждой точке решение u имеет k производных, скорость сходимости в отдельных точках ожидается такой же. (При наличии особенностей степень дифференцируемости и, следовательно, скорость сходимости совершенно различны в поточечном и среднеквадратичном смыслах. Мы не приводим детального доказательства оптимальных оценок ошибок в точках.) *В специальных точках ошибка действительно может сходиться быстрее, чем в среднем.* Например, для задачи $-u''=f$ узлы линейных элементов специфичны: $u^h=u_I$ и решение Ритца в этих узлах точное. Это вообще справедливо, если элементы служат решениями однородного дифференциального уравнения [X1, T5]. Для уравнения теплопроводности Томе отметил особую скорость сходимости в узлах сплайнов, Дуглас и Дюпон расширили этот принцип на свои методы коллокации.

Скорость сходимости сохраняется для неоднородных краевых условий (разд. 4.4), если только краевые данные интерполируются (или приближаются) полиномами по крайней мере той же степени $k-1$.

Для области Ω неправильной формы может появиться новый тип ошибки аппроксимации. Обычно бывает необходимо аппроксимировать границу Γ кусочно полиномиальной границей

Γ^h. В простейшем случае Γ^h кусочно линейна; Ω заменяется многоугольником Ω^h. Такой многоугольник можно разрезать на треугольники и применять далее метод конечных элементов без учета полосы $\Omega - \Omega^h$ между исходной границей Γ и многоугольником. Следовательно, мы как будто вдвигаем исходную дифференциальную задачу в Ω^h. В разд. 4.4 исследуется влияние этого *изменения области*. Кратко это влияние таково: ошибка m-й производной у границы равна $O(h)$, но быстро убывает внутри области. Это приграничный эффект, средняя ошибка равна $O(h^{3/2})$. Так как энергия деформации зависит от квадрата этой производной, то ошибка в энергии, обусловленная вычислением на Ω^h (с естественными или главными краевыми условиями), равна $O(h^3)$. Эта оценка применима к Ω, как и к Ω^h, если решение метода конечных элементов доопределить естественным образом, распространяя каждый полином вплоть до Γ. В противном случае вся энергия в полосе будет потеряна, а это составляет $O(h^2)$ — пропорционально объему полосы.

Заметим, что ошибка h^3, вызванная изменением области, преобладает, когда степень конечных элементов, используемых внутри области, превышает m. Если взяты полиномы минимальной степени m, необходимой для сходимости, то эффект от изменения области поглотится (по крайней мере внутри Ω^h) ошибкой $h^{2(k-m)} = h^2$, возникающей из обычной теории приближений.

Если граница Γ приближается кусочно полиномиальными функциями степени выше l, то соответственно уменьшается и ошибка от изменения области. Ошибка в m-й производной (деформация) у границы равна $O(h^l)$, а в общей энергии деформации в Ω^h равна $O(h^{2l+1})$. В случае главного условия $u = 0$ это означает, что оно точно выполняется для пробных полиномиальных функций на приближенной границе. Митчелл нашел изящную конструкцию кубических элементов, равных нулю на границе, составленной из гипербол, с ошибкой h^5 в энергии деформации и h^3 в перемещении.

Для главного условия на криволинейной границе (например, $u = 0$) можно также взять любые стандартные элементы и потребовать, чтобы они интерполировали условия в граничных узлах. В этом случае пробные функции не будут удовлетворять главному условию на всей границе, каждая пробная функция может обращаться в нуль вдоль некоторой кривой, близкой к Γ, но для разных функций эти нулевые кривые будут различны. В результате теория Ритца неприменима: пробные функции не принадлежат $\mathscr{H}_E^m$ ни на точной области Ω, ни на приближенной Ω^h. Кроме того, функция u^h, минимизирующая $I(v)$, не будет ближайшей пробной функцией к u. Тем не менее можно оценить ошибку, принимая во внимание, что каждая

функция v^h не равна нулю, но мала на границе. Наилучшая оценка ошибки в энергии деформации (удивляющая своей малостью) обычно порядка h^3. Грубо говоря, если мы работаем с кубическими функциями на треугольниках и интерполируем главные краевые условия, то наихудшая функция будет нарушать условия с величиной $O(h^{3/2})$ между узлами; в разд. 4.4 мы выведем из этого утверждения, что ошибка в энергии равна $O(h^3)$.

Существует еще одна альтернатива, несомненно, наиболее распространенная. Криволинейные элементы с помощью преобразования координат можно выпрямить. Такое преобразование может оказаться даже необходимым для достижения непрерывности между четырехугольниками с *уже* прямыми сторонами, если только это не прямоугольники. Это преобразование координат представляет собой центральный прием в технике метода конечных элементов.

Теоретически можно выпрямить почти любую граничную кривую, но практически это, конечно, неосуществимо. Кусочно полиномиальные функции являются наилучшими границами элементарных областей по тем же причинам, по каким они наилучшим образом приближают перемещения: с ними удобно работать на ЭВМ. В самом деле, выбор координат можно описать тем же классом полиномов, из которого берутся пробные функции; это метод *изопараметрических преобразований.* Идея эта превосходна. Выбор координат приводит к тем же трудностям, что и для пробных функций: преобразование должно быть непрерывным при пересечении границ элементов, так что элементы, соседние на исходной плоскости x, y, остаются соседними на плоскости ξ, η. Если преобразование $x(\xi, \eta)$, $y(\xi, \eta)$ построено стандартным образом из узловых параметров и мы убеждены в непрерывности по ξ и η (как для стандартных прямоугольных или треугольных элементов), то изопараметрические преобразования приведут к успеху даже для элементов, границы которых — полиномы степени $k-1$ по x и y. Этот прием ставит новые вопросы теории приближений, так как полиномы по ξ и η не будут более полиномами по x и y. Тем не менее изопараметрические преобразования не понижают порядка точности; если преобразования равномерно гладкие (разд. 3.3), то полная степень h^{k-s} в s-й производной достигается. В этом смысле *изопараметрический прием представляется наилучшим* для уравнений второго порядка и криволинейных границ. С главным краевым условием $u = g$ можно работать просто и эффективно без потерь в основном порядке точности.

Мы рассматривали пока вклад в $u - u^h$ только от перечисленных ошибок, считая, что приближение Ритца вычисляется

точно. На практике, однако, есть еще ошибки в численном интегрировании (получаем $\tilde{u}^h$ вместо u^h) и при решении заключительной линейной системы (получаем $\bar{u}^h$ вместо $\tilde{u}^h$). Надо знать величину ошибки интегрирования, чтобы выбрать квадратурную формулу, не требующую больших затрат времени и не нарушающую точность. В примере в гл. 1 упоминались две основные возможности: (1) неоднородные данные и любые существенные коэффициенты, которые изменяются вместе с x, можно заменить интерполирующими полиномами, и тогда интегралы для полученной задачи вычисляются точно; (2) интегрирование по всем элементарным областям можно выполнить с самого начала по стандартной квадратурной формуле, скажем по квадратурам Гаусса. Для простой задачи типа $-(pu')' + qu = f$ допускается некоторая свобода выбора. Напомним один теоретический результат: и интерполяция полиномами степени $k-1$, и квадратура Гаусса по $k-1$ точкам дают ошибку в деформациях порядка h^k. Для более сложной задачи нужно применять метод (2), и в действительности численное интегрирование стало одним из главных моментов метода конечных элементов. Оно дает решение $\tilde{u}^h$, сходящееся к u при условии определенной точности численных квадратурных формул: *производные порядка m у всех пробных функций должны интегрироваться точно.* Для каждой дополнительной степени точности ошибка $\tilde{u}^h - u^h$, возникающая из численных квадратурных формул, улучшается на некоторую степень величины h. Доказательство основано на тождестве, приведенном в разд. 4.3. Таким же способом определяется повышенная точность для изопараметрических элементов.

Наконец, ошибки округления. Их характер совершенно отличен от характера других ошибок — они пропорциональны *отрицательным* степеням величины h. При убывании h существует зона пересечения, в которой порядки ошибок округления и аппроксимации совпадают и до которой округление незначительно и неинтересно, а после нее чрезвычайно важно. Округление слабо зависит от степени полиномов и от размерности. Ключевой множитель h^{-2m} устанавливается шагом сетки и порядком самого уравнения. Для уравнения второго порядка шаги сетки, типичные для практических задач, еще не входят в зону пересечения, но для уравнений четвертого порядка это не так: вычислительная точность может потребовать проведение операций с двойной мантиссой. В гл. 5 мы обсудим и априорные оценки числа обусловленности, и апостериорные оценки округления, действительно совершаемого в данной задаче.

Это главные ошибки, анализируемые в линейных стационарных задачах $Lu = f$. В каждом случае анализ основан на основном вариационном уравнении $a(u, v) - (f, v) = 0$. Нет сомнения,

что по сравнению с конечными разностями этот анализ приводит к более последовательной и удовлетворительной математической теории. Частично эту технику можно распространить на *нелинейные уравнения*. В настоящее время выдающуюся математическую проблему представляет выделение классов нелинейных задач, физически важных и математически доступных, и изучение зависимости приближения от пространств пробных функций, областей и коэффициентов. Значительный успех в этом направлении достигнут Лионсом (см. [10]).

Уже выделены два класса как простейшие обобщения линейных эллиптических уравнений. Мы не знаем, насколько исчерпывающи их приложения в технике и физике, но появились они весьма естественно. Один из них — класс *строго монотонных операторов*, удовлетворяющих неравенству

$$\int_\Omega [M(u) - M(v)](u - v)\,dx \geqslant \sigma \| u - v \|_m^2.$$

Другой (близкий к нему) класс содержит *потенциальные операторы*, для которых M — градиент некоторого неквадратичного выпуклого функционала $I(v)$. Для знакомства с теорией рекомендуем книгу Вайнберга [4].

К этим операторам непосредственно применяется метод Галёркина: отыскивается такая функция $u^h = \sum Q_j \varphi_j$ из подпространства S^h, что

$$(M(u^h), \varphi_k) = 0 \quad \text{для всех} \quad \varphi_k.$$

Число (нелинейных) уравнений равно числу неизвестных коэффициентов, т. е. размерности пространства пробных функций S^h. Для двух описанных выше классов можно доказать существование такого решения u^h и его сходимость к u при условии, что на оператор наложены подходящие требования непрерывности. В самом деле, схема одного из возможных доказательств существования решения u такова: доказывается существование u^h в конечномерном пространстве и дается априорная оценка, устанавливающая, что все u^h принадлежат некоторому компактному множеству; тогда последовательность u^h должна иметь предельную точку при $h \to 0$; этой точкой и будет u. Сиарле, Шульц и Варга [C6] показали, что оценки ошибки для линейных и нелинейных монотонных задач отличаются незначительно.

Метод конечных элементов широко применяется в нелинейных задачах, например для упруго-пластичных и термовязкопластичных материалов. В дополнение к книге Одена [17] появилась и быстро разрастается техническая литература; укажем лишь ранние обзорные статьи Мартина [M5] и Маркала

[M4]. Здесь полезно повторить возможные формулировки метода конечных элементов, отличающиеся приемами работы с геометрическими нелинейностями, возникающими из-за сильного изгиба, и с существенными нелинейностями, без особого привлечения математики. Очевидно, что проблема сходимости хорошо поставлена, чрезвычайно интересна и созрела для решения. Надеемся, что в конечном счете эта математическая теория станет достаточно законченной и ей можно будет посвятить целую книгу (это кто-нибудь сделает!).

Приведем одно предостережение о нелинейных уравнениях, оно относится к задаче нелинейной упругости. Возьмем простую модель минимизации функционала вида

$$I(v)=\int [p(v,\ v_x)\, v_x^2 - 2fv]\, dx.$$

Так как он не квадратичен по v, равенство нулю первой вариации не приводит к линейному уравнению для u. Поэтому обычно проводятся итерационные процессы, простейшим из которых является метод *последовательной подстановки*: вычисляется нелинейный коэффициент для n-го приближения u_n и u_{n+1} определяется как решение линейной задачи.

Наше предостережение: если такой итерационный процесс проводится в вариационной задаче, так что u_{n+1} минимизирует интеграл

$$\int \left[p\left(u_n,\ \frac{du_n}{dx}\right) v_x^2 - 2fv\right] dx, \tag{d}$$

то *итерации сходятся к неверному ответу*. Читатель легко проверит, что предел u^* такого итерационного процесса удовлетворяет равенству

$$-\frac{d}{dx}\left(p(u^*,\ u_x^*)\frac{du^*}{dx}\right)=f,$$

которое не является уравнением виртуальной работы для $I(v)$. (Достаточно положить $p = v_x$.) Ошибка произошла из-за того, что последовательные подстановки осуществлялись до взятия первой вариации. Если сначала установить нелинейное уравнение для минимизирующей функции u, а *затем* решать это уравнение итерациями, то предел получился бы правильным.

Изложим одну частную нелинейную задачу, описывающую деформацию упруго-пластичного материала, которая иллюстрирует как возможности, так и трудности в доказательстве нелинейной сходимости. Она будет полезной для выявления некоторых деталей. Для простоты обозначений рассмотрим одномерную модель с напряжением du/dx; те же рассуждения применимы к системе напряжений ε_{ij} для двумерной и трехмерной

задачи упругости. Деформация представляет собой нелинейную функцию внешних сил, и ее нельзя определить лишь из окончательной нагрузки. Она зависит от истории задачи, т. е. от хронологического порядка, в котором силы прилагались к области. Это вводит искусственный параметр «времени», и в заданный момент t скорость изменения деформации $\dot{u}$ минимизирует функционал

$$I(\dot{v}) = \int [p(\dot{v}_x)^2 - 2f\dot{v}]\, dx.$$

Определяющей величиной служит модуль упругости $p(x)$. Если этот коэффициент не зависит от напряжений в материале, то в искусственном времени нет необходимости, окончательную деформацию $u(T)$ можно определить простой минимизацией (как и всюду в этой книге) с помощью окончательной нагрузки $f(T)$. В случае нелинейного закона напряжения-деформации коэффициент p в момент t зависит от напряжений, и не только от их значений в данный момент: *он зависит от всей истории напряжений.* Такая зависимость от способа действия физически вводится несколькими способами. Например, как только предел упругости превышен, нагрузка, следующая за равной разгрузкой, оставляет изменение сети напряжений в материале. Это явление действительно создает в каждый момент времени нелинейную задачу для $\dot{v}$, поскольку модуль упругости зависит тогда не только от прошлой истории, но и от текущего темпа изменений. На коэффициент p влияет *знак* произведения $v_x\dot{v}_x$, и он принимает разные значения для нагрузки и разгрузки. Для простоты мы будем избегать эту дополнительную трудность и предполагать отсутствие разгрузки. Несмотря на то, что напряжение — однозначная функция деформации, мы не требуем, чтобы она вычислялась из известных деформаций $u_x(\tau)$ в каждый момент времени $\tau \leqslant t$.

В аппроксимации Ритца $\dot{u}^h$ — это функция из подпространства пробных функций S^h, минимизирующая функционал

$$I(\dot{v}^h) = \int [p^h(\dot{v}^h_x)^2 - 2f\dot{v}^h]\, dx.$$

В заданный момент t коэффициент $p^h(x)$ не принимает то же значение, что и истинный модуль p. Вместо этого он зависит от истории аппроксимаций Ритца $u^h_x(\tau)$, $\tau \leqslant t$. Для доказательства сходимости надо предположить, что если эти приближенные деформации близки к истинным, то коэффициент p^h в данный момент близок к истинному значению p:

$$\max_x |p(x) - p^h(x)| \leqslant C \int_0^t \|\dot{u}_x(\tau) - \dot{u}^h_x(\tau)\|\, d\tau.$$

Для доказательства сходимости надо оценить, насколько функция $\dot{u}^h$, минимизирующая $I(\dot{v}^h)$, и ее производная по x, т. е. скорость изменения деформации, зависят от коэффициента p в вариационном принципе. Наш план состоит в расщеплении ошибки в момент времени t на две части:

$$\dot{u}_x(t) - \dot{u}_x^h(t) = \left(\dot{u}_x(t) - \dot{w}_x^h(t)\right) + \left(\dot{w}_x^h(t) - \dot{u}_x^h(t)\right),$$

где $\dot{w}^h$ — функция, минимизирующая $I(\dot{v}^h)$ на всем пространстве пробных функций S^h, если в момент t брался *истинный* коэффициент упругости p. Другими словами, первая часть ошибки вызвана аппроксимацией, вторая — изменением коэффициента упругости. Для первой части можно привлечь обычную теорию приближений: задача линейна в каждый момент времени, а ошибка в первых производных для пространства степени $k-1$ оценивается неравенством

$$\|\dot{u}_x(t) - w_x^h(t)\| \leqslant C' h^{k-1}.$$

Относительно второй части ошибки см. разд. 4.3. Там утверждается в качестве следствия, что эффект от изменения коэффициентов оценивается неравенством

$$\|\dot{w}_x^h(t) - \dot{u}_x^h(t)\| \leqslant C'' \max| p(x) - p^h(x) |.$$

Сравним два последних неравенства и запишем $\dot{e} = \dot{u}_x - \dot{u}_x^h$:

$$\|\dot{e}(t)\| \leqslant C' h^{k-1} + CC'' \int_0^t \|\dot{e}(\tau)\| d\tau.$$

Это в точности ситуация, к которой применимы доводы, приводящие к *лемме Гронуэлла*: разделим неравенство на его правую часть, умножим на CC'' и поинтегрируем по t:

$$\ln\left(C' h^{k-1} + CC'' \int_0^t \|\dot{e}(\tau)\| d\tau\right) - \ln(C' h^{k-1}) \leqslant CC'' t.$$

Потенцирование обеих частей и подстановка в предыдущее неравенство дают

$$\|\dot{e}(t)\| \leqslant C' h^{k-1} \exp(CC'' t).$$

Наконец, проинтегрируем по t:

$$\|e(t)\| \leqslant \int \|\dot{e}(t)\| dt \leqslant C''' h^{k-1}.$$

Итак, скорости сходимости по h для задач нелинейной пластичности и линейной упругости одинаковы.

Заметим, что вычисление аппроксимации Ритца предполагалось *непрерывным по времени*; до сих пор лишь дискретизация заменяла все допустимое пространство его подпространством S^h. Это соответствует изложению задачи Коши в гл. 7, где ошибки метода Ритца отделены от ошибок метода конечных разностей (или другого метода) по временному направлению. Для нелинейной задачи большие дискуссии вызвал наилучший «метод приращений», но мы полагаем, что все основные возможности сходятся в одном. Они просто вносят новую ошибку, пропорциональную степени приращения Δt в случае разностного уравнения.

Тем не менее в приведенном доказательстве есть одна техническая трудность, игнорировать которую не позволяет нам наша совесть. Это вопрос выбора нормы: если выбрана среднеквадратичная норма, то поточечные оценки для $p - p^h$ не верны. С другой стороны, для возможности использования максимальной нормы требуется новое изучение оценки h^{k-1} для $\dot{u}_x - \dot{w}_x^h$. Эта оценка следует из теории среднеквадратичного приближения, и, по-видимому, проще всего вывести ее, а не оценивать поточечно ошибку метода Ритца в статических линейных задачах. Другую возможность дает идея, предложенная Стренгом (последующие приложения см. в [B8]); она позволяет в случае гладкого решения переходить от одной нормы к другой. Такой прием часто бывает незаменим в нелинейных задачах, когда оценки ошибок носят глобальный характер, а неустойчивость может возникнуть локально. Третья возможность состоит в улучшении следствия в разд. 4.3, а именно в установлении зависимости от среднеквадратичной нормы возмущения $p - p^h$. Мы уверены в правильности основного доказательства и в том, что сочетание эксперимента и теории скоро приведет к более полному пониманию нелинейных ошибок.

В этот краткий обзор теории необходимо включить также *задачи на собственные значения и задачи с начальными условиями.* Метод конечных элементов успешно применяется непосредственно к обеим задачам. Для самосопряженных задач на собственные значения классический прием — вычисление оценок сверху при минимизации отношения Рэлея на подпространстве; он приводит к дискретной задаче на собственные значения $KQ = \lambda MQ$, где K и M — уже встречавшиеся матрицы жесткости и массы. В гл. 6 излагается эта дискретная формулировка и оцениваются ошибки в собственных векторах и функциях, зависящие от теории приближений: они возникают из-за замены исходного допустимого пространства $\mathscr{H}_E^m$ на S^h. Результаты описываются просто: $\lambda - \lambda^h$ есть величина порядка $h^{2(k-m)}$, а при $k \leqslant 2m$ порядок ошибок в s-х производных собственных функ-

ций, разрешаемый теорией приближений, не должен превосходить h^{k-s}.

Для задачи с начальными условиями $u_t + Lu = f$ положение столь же благоприятное. Решение методом конечных элементов имеет вид $u^h(t, x) = \sum Q_j(t)\varphi_j(x)$; временна́я переменная остается непрерывной, тогда как зависимость по x дискретизируется в терминах обычных кусочно полиномиальных базисных функций φ_j. Коэффициенты $Q_j(t)$ определяются из системы N обыкновенных дифференциальных уравнений, выражающих метод Галёркина: невязка $u_t^h + Lu^h - f$ не обращается тождественно в нуль, если истинное решение u не лежит в пространстве пробных функций S^h, но ее компоненты в S^h равны нулю. Таким образом, исходное уравнение удовлетворяется «на подпространстве».

В практических задачах время тоже должно быть дискретизировано, что предполагает применение метода конечных разностей. Например, схема Кранка — Николсона симметрична относительно $t_{n+1/2}$ при вычислении $u^h(t_{n+1})$ через $u^h(t_n)$ и потому имеет точность порядка Δt^2. Таким образом, окончательно вычисленное приближение содержит эту ошибку, как и ошибку метода Галёркина, вызванную дискретизацией по x. Последнюю из них мы проанализируем подробно и покажем, что при $k \geqslant 2m$ ее оптимальный порядок для s-й производной тоже равен h^{k-s}. Этот результат применяется к уравнениям *параболического* типа, например к уравнению теплопроводности; L — эллиптический оператор того же типа, что и в стационарных задачах. В случае *гиперболических* уравнений, не содержащих диссипативных членов, возможности метода конечных элементов несколько меньше; трудности в сравнении с явными разностными методами могут оказаться слишком большими. Тем не менее даже в этом случае достигнуты значительные результаты: исследование границ можно проводить почти автоматически; в гл. 7 включен набросок теории метода конечных элементов для гиперболического случая.

За недостатком места изучение изменения области, численного интегрирования и округления ограничено стационарным уравнением $Lu = f$. Результаты для задачи с начальными условиями и задачи на собственные значения очень похожи; для квадратурных ошибок эти обобщения теории осуществлены Фиксом (Симпозиум по методу конечных элементов в Балтиморе).

В последней главе излагаются результаты обширной серии численных экспериментов. Наиболее интересный из них касается задачи с сильными особенностями, порождаемыми трещиной в материале. Эта классическая задача механики разрыва по вычислению коэффициента интенсивности напряжения в

истоке трещины; вокруг этой точки напряжение изменяется подобно $r^{-1/2}$. Итак, возникает целый ряд вопросов:

1. Дадут ли наши оценки ошибок наблюдаемую скорость сходимости — и поточечной, и в среднем квадратичном?
2. Особенность понижает гладкость решения u и, следовательно, скорость сходимости; достаточна ли эта уменьшенная скорость в случае гладкого решения, когда особенности заведомо искажают расчет?
3. Можно ли выбором сетки или введением специальных пробных функций в особенностях восстановить нормальную скорость сходимости метода?

Ответ на каждый вопрос утвердителен и численные результаты очень убедительны. Как для стационарной задачи, так и для задачи на собственные значения замечательным средством против особенностей, вносимых острыми углами в области, служит введение пробных функций, правильно отражающих свойства особенностей.

Задача непосредственного касания материалов мало отличается от предыдущих. Производные от u имеют скачок в месте соприкосновения, и мы настоятельно рекомендуем следующее простое решение: ослабить требование непрерывности, налагаемое на производные от пробных функций, чтобы функция u^h могла повторить особенность в u. Мы не верим, что при нормальных обстоятельствах условие скачка (или любое другое естественное краевое условие) следует налагать.

Наконец, мы решили теоретически обосновать сравнительно новый технический прием — алгоритм Петерса — Уилкинсона для матричной задачи на собственные значения $Kx = \lambda Mx$. Конечно, решение линейной системы $KQ = F$ представляет собой еще более фундаментальную проблему — это тема для больших усовершенствований в упорядочении неизвестных или в выборе градиентной процедуры, но она сравнительно хорошо изучена. Задача на собственные значения сложнее и без эффективного алгоритма число неизвестных будет недостаточно — оно будет меньше числа, необходимого для выявления физических свойств задачи. Поэтому в гл. 6 мы излагаем идею Петерса — Уилкинсона (как и некоторые более укоренившиеся алгоритмы), а в гл. 8 применяем ее к численным экспериментам.

2.3. МЕТОД ГАЛЁРКИНА, КОЛЛОКАЦИЯ И СМЕШАННЫЙ МЕТОД

Описанный метод Ритца применяется только к задачам классического вариационного типа, в которых минимизируется выпуклый функционал. Соответствующее дифференциальное уравнение Эйлера самосопряжено и эллиптично. Однако хорошо

известно, что уравнения самого общего типа тоже можно записать в слабой форме, допускающей обобщение от метода Ритца к методу Галёркина. Такое применение к задаче с начальными условиями описано в гл. 7. Здесь мы обсудим два типа стационарных задач: первый, в котором производные нечетного порядка портят самосопряженность эллиптического уравнения, и второй, в котором соответствующий функционал не положительно определен — тогда проблема состоит в отыскании *стационарной точки, а не минимума* функционала $I(v)$. Ситуация возникает из принципа Хеллингера — Рейсснера в теории упругости и в соответствующем *смешанном методе* для конечных элементов и приводит к некоторым трудным математическим вопросам при доказательстве сходимости.

Сначала остановимся кратко на слабых формах дифференциального уравнения $Lu = f$. Их несколько, но все они обладают основным свойством: уравнение умножается на тестовую функцию $v(x)$ и интегрируется по области Ω; в результате получается уравнение

$$(Lu,\ v) = (f,\ v).$$

Оно справедливо для каждой функции v из некоторого тестового пространства V. Все зависит от выбора V. Если V содержит все δ-функции, то уравнение $Lu = f$ удовлетворяется в наиболее классическом (можно сказать, старомодном) смысле, т. е. в каждой точке. Дискретная форма этого тестового пространства приводит к *методу коллокации*, обсуждаемому ниже. Другой крайний случай — пространство V содержит лишь бесконечно гладкие функции, равные нулю в приграничной полосе. Формальное интегрирование по частям переносит все производные с u на v и приводит к уравнению

$$(u,\ L^*v) = (f,\ v),$$

где L^* — формально сопряженный к L оператор. В этой слабейшей форме u удовлетворяет уравнению только «в смысле распределений». Ясно, что изучается дискретная форма.

Между этими крайностями заключено очень много возможностей. Если $V = \mathscr{H}^0(\Omega)$, то говорят, что уравнение выполняется в сильном смысле, а иногда — в смысле L^2. Более общий выбор $V = \mathscr{H}^s(\Omega)$ позволяет перенести s производных с u на v. Если порядок оператора L равен $2m$, то решение u, вероятно, надо искать в $\mathscr{H}^{2m-s}$. Случай $s = m$ крайне важен, он описывается уравнением

$$a(u,\ v) = (f,\ v).$$

Здесь всюду должны играть свою роль краевые условия. В сильной форме ($s = 0$) на u налагается полное множество краевых

условий; решение должно принадлежать $\mathscr{H}_B^{2m}$. При увеличении s для определения пространства решений понадобятся лишь $2m-s$ производных от u, так что смысл имеют лишь краевые условия порядка, меньшего $2m-s$. Для $s=m$ они являются главными краевыми условиями[1]) и u принадлежит $\mathscr{H}_E^m$. В то же время число условий, налагаемых на v, возрастает. Они определяются производными порядка, меньшего s, которые входят в формулу Грина. Для $s=m$ функция v также принадлежит $\mathscr{H}_E^m$. Таким образом, в случае метода Ритца существует симметрия между пространством пробных функций и тестовым пространством.

Метод Галёркина представляет собой очевидную дискретизацию слабой формы. Вообще говоря, он содержит два семейства функций — подпространство S^h из пространства решений и подпространство V^h из тестового пространства V. Тогда решение Галёркина u^h — это такой элемент из S^h, что

$$(Lu^h,\ v^h)=(f,\ v^h) \quad \text{для всех} \quad v^h \in V^h. \tag{5}$$

В левой части надо s раз проинтегрировать по частям, как и для непрерывной задачи. Если размерности пространств S^h и V^h одинаковы (равны N), уравнение Галёркина обычным образом переходит в операторную форму: если $\varphi_1, \ldots, \varphi_N$ — базис пространства S^h, а $\psi_1, \ldots, \psi_N$ — базис пространства V^h, то решение $u^h=\Sigma Q_j\varphi_j$ удовлетворяет равенству

$$\left(\sum Q_j L\varphi_j,\ \psi_k\right)=(f,\ \psi_k), \quad k=1, \ldots, N.$$

В матричном виде это система $GQ=F$, $G_{kj}=(L\varphi_j, \psi_k)$, $F_k=(f, \psi_k)$. В методе Ритца $S^h=V^h\subset\mathscr{H}_E^m$, $\varphi_j=\psi_j$, $(L\varphi_j, \varphi_k)=a(\varphi_j, \psi_k)$ и G — матрица жесткости K.

---

[1]) С математической точки зрения главное условие $Bv=g$ определяется ограничеиным линейным оператором B на пространстве $\mathscr{H}^m$ всех функций с конечной энергией деформации. Функции, удовлетворяющие условию $Bv=0$, образуют замкнутое подпространство. Естественные условия — это те, что не налагаются и не могут налагаться на каждую функцию v, но благодаря специальному виду $I(v)$ они удовлетворяются минимизирующей функцией u. Обычный критерий для главных условий состоит в том, что оператор B должеи содержать лишь производные порядка, меньшего m, но это условие ни необходимо, ни достаточно. В двумерной задаче, например при $m=1$, нельзя требовать, чтобы функция v равнялась нулю в заданной отдельной точке P. Значение функции в P не будет ограниченным функционалом (оно может быть произвольно большим, как $\ln\ln r$, в то время как функция имеет единичную энергию деформации) и функции, удовлетворяющие условию $v(P)=0$, не образуют замкнутого подпространства. На самом деле они произвольно близки по энергии деформации к пробным функциям, *не* равным нулю в P, и минимизация по ним даст тот же результат, что и минимизация на всем пространстве $\mathscr{H}^1$.

Предположим, что S^h — пространство конечных элементов степени $k-1$, а V^h — пространство конечных элементов степени $l-1$. Тогда можно ожидать, что скорость сходимости s-х производных в методе Галёркина будет

$$\|u-u^h\|_s = O(h^{k-1}+h^{k+l-2m}). \tag{6}$$

Как и в (4), первый показатель у h отражает наилучший порядок аппроксимации, возможный в S^h, а на второй показатель влияют аппроксимация в тестовом пространстве и порядок $2m$ дифференциального уравнения. Теоретически возможно добиться некоторой экономии, выбирая $l<k$, скажем $k=4$ (кубические сплайны) и $l=2$ (линейные тестовые функции) в задаче второго порядка ([Б29]). Скорость сходимости та же, что и при $l=4$, а ширина ленты матрицы G уменьшена. Однако матрица G больше не симметрична даже для самосопряженной задачи; нам это кажется сомнительным.

Родственная возможность — «приближенно рассчитать» [1]) некоторые члены в дискретизации и тем самым отклониться от уравнений Ритца $KQ=F$ и $KQ=\lambda MQ$, используя подпространства меньшей степени для членов меньшего порядка в уравнении. Раньше это применялось в задачах на собственные значения, чтобы заменить *плотную матрицу массы* M диагональной *приближенной матрицей массы*. В новейших алгоритмах для задачи на собственные значения (гл. 6) не столь важно, чтобы матрица массы была диагональной; по вопросу анализа ошибок процедуры «приближенного расчета» см. [Т9].

В методе коллокации базис тестового пространства V^h состоит из δ-функций: $\psi_j(x)=\delta(x-x_j)$. Поэтому в уравнении Галёркина (5) требуется, чтобы дифференциальное уравнение выполнялось в каждом узле: $Lu(x_j)=f(x_j)$. Относительно границ ошибок (6) метод действует обычно так, как если бы было $l=0$, и скорость сходимости равна h^{k-2m}. Тем не менее существуют специальные точки коллокации, повышающие порядок сходимости и делающие метод крайне интересным [Д9, Б31]. Ширина ленты у G меньше, чем у K, и нет необходимости вычислять скалярные произведения и матрицы элементов; для более сложных нелинейных задач эти преимущества могут хорошо скомпенсировать тяжелую аппроксимацию и гладкость, требуемую от S^h.

Задачи, не являющиеся положительно определенными и симметричными

Мы хотим подробнее изучить два случая, когда пространство пробных функций S^h совпадает с тестовым пространством V^h

[1]) В оригинале lump (брать без разбора, смешивать в кучу). — *Прим. ред.*

(надо m раз проинтегрировать по частям, так что слабая форма — это по существу уравнение $a(u, v) = (f, v)$ метода Ритца) но задача отличается от классической задачи минимизации положительно определенного функционала. Первое отличие — уравнение несамосопряженное, как в примере с постоянными коэффициентами

$$Lu = -pu'' + \tau u' + qu = f.$$

Сопряженный оператор L^* имеет противоположный знак у производной нечетного порядка:

$$\int (-pu'' + \tau u' + qu)\, v = \int u\,(-pv'' - \tau v' + qv),$$

что приводит к энергетическому скалярному произведению

$$a(u, v) = \int pu'v' + \tau u'v + quv,$$

которое несимметрично: $a(u, v) \neq a(v, u)$. В самом деле, новый член *кососопряжен*; он соответствует *мнимой части* оператора L, вещественная часть которого $-pu'' + qu$ положительна, как и раньше. Это становится особенно ясно, если скалярное произведение расширить на комплексные функции:

$$a(u, v) = \int pu'\bar{v}' + \tau u'\bar{v} + qu\bar{v}.$$

Вещественная часть от $a(u, u)$ равна $\int p|u'|^2 + q|u|^2$, а новый член чисто мнимый.

Мы хотим показать, что можно строго установить, что скорость сходимости производной равна h^{k-1} (как и раньше), но для больших τ метод конечных элементов может оказаться несостоятельным. Доводы довольно общие. Доказательство сходимости, когда вещественная (самосопряженная) часть эллиптична, охватывает метод Ритца как частный случай, а метод Галёркина может оказаться неудовлетворительным, если член с нечетной производной (мнимая, или кососопряженная часть) очень велик.

Теорема 2.1. *Предположим, что* $|a(u, v)| \leqslant K\|u\|_m\|v\|_m$ *и вещественная часть задачи эллиптична:* $\operatorname{Re} a(u, u) \geqslant \sigma\|u\|_m^2$ *для* $u \in \mathscr{H}_E^m$. *Пусть* u^h *— функция из* $S^h \subset \mathscr{H}_E^m$, *удовлетворяющая уравнению Ритца — Галёркина* $a(u^h, v^h) = (f, v^h)$ *для всех* $v^h \in S^h$. *Тогда порядок сходимости по энергии (что эквивалентно порядку сходимости m-х производных) равен наилучшему возможному порядку аппроксимации, который можно достичь в* S^h:

$$\|u - u^h\|_m \leqslant \frac{K}{\sigma} \min_{S^h} \|u - v^h\|_m. \qquad (7)$$

В методе конечных элементов для пространства степени $k-1$ *этот порядок равен* h^{k-m}.

Доказательство. Так как $a(u, v) = (f, v)$ для всех v из $\mathscr{H}_E^m$, то вычитание дает $a(u-u^h, v^h) = 0$ для всех v^h. Следовательно,

$$\sigma \|u-u^h\|_m^2 \leqslant \operatorname{Re} a(u-u^h, u-u^h) =$$

$$= \operatorname{Re} a(u-u^h, u) = \operatorname{Re} a(u-u^h, u-v^h) \leqslant K \|u-u^h\|_m \|u-v^h\|_m.$$

Доказательство завершается делением неравенства на множитель $\|u-u^h\|_m$. Теперь можно применить метод Нитше и установить обычную скорость сходимости (4) для перемещений.

Несмотря на такую сходимость, метод Галёркина на практике может оказаться плохим. Предположим, что S^h — обычное подпространство кусочно линейных функций. Тогда в нашем примере уравнения Галёркина для $Q_j = v^h(x_j)$ будут просто разностными уравнениями

$$p\frac{-Q_{j+1}+2Q_j-Q_{j-1}}{h^2}+\tau\frac{Q_{j+1}-Q_{j-1}}{2h}+q\frac{Q_{j+1}+4Q_j+Q_{j-1}}{6} =$$

$$= h^{-1}\int f\varphi_j\,dx.$$

Заметим, что первая производная заменяется центральной разностью независимо от знака τ. Точное решение, однако, может сильно зависеть от знака τ: при $|\tau| \to \infty$ адвективный член $\tau u'$ доминирует над второй производной, и решение u есть решение типа пограничного слоя. На большей части интервала u по существу является решением *задачи с начальными условиями*, для которой центральная разность совершенно неприемлема, на дальнем конце для удовлетворения другого краевого условия появляются быстрые вариации и нужна особо мелкая сетка. Потребность в односторонних (против течения) разностях хорошо известна инженерам-химикам. Математически преобладание τ отражается в большом значении K/σ в оценке ошибки (7).

Второе отличие от классической формулировки Ритца возникает, когда функционал потенциальной энергии $I(v)$ не выпуклый (форма $a(u, u)$ не является положительно определенной) и задача состоит в отыскании не минимума, а стационарной точки. Это, естественно, встречается в *смешанном методе*, когда и перемещение, и его производные считаются *независимыми неизвестными*. Потенциальная энергия содержит произведения, которые могут быть и положительными, и отрицательными, это похоже на переход от x^2+y^2 к функции xy, имеющей вместо минимума седловую точку в начале координат.

Хороший пример дает уравнение $w^{(IV)} = f(x)$, описывающее изгиб нагруженной балки. Будем считать момент $M = w''$ новым неизвестным. Тогда исходное уравнение перейдет в $M'' = f$, и мы получим систему уравнений второго порядка

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} M \\ w \end{pmatrix}'' - \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} M \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ f \end{pmatrix}. \tag{8}$$

Это понижение порядка дает несколько преимуществ. От пробных функций в вариационной формулировке требуется лишь непрерывность между элементами, тогда как уравнению четвертого порядка соответствует потенциальная энергия $I(v) = \int (v'')^2 - 2fv$, конечная только тогда, когда наклон v' тоже непрерывен. Более того, число обусловленности матрицы жесткости совершенно изменяется при уменьшении порядка дифференциальных уравнений: четвертые разности заменяются вторыми, и число обусловленности переходит от $O(h^{-4})$ к $O(h^{-2})$. Может показаться чудом, что такое улучшение достигается чисто формальным введением производной M как новой неизвестной, но это улучшение вполне реальное.

Исследуем ошибки округления двумя способами. Рассмотрим свободно опертую балку с u и M, равными нулю на обоих концах. Тогда уравнения $M'' = f$ и $w'' = M$ можно решить отдельно, сначала для M, а затем для w, применяя для этого либо конечные разности, либо конечные элементы. Предположим, что приближенное решение задачи $M'' = f$ содержит ошибку округления ε_1, обычно порядка $h^{-2}2^{-t}$, для ЭВМ с длиной слова t. Тогда приближенное решение задачи $w'' = -f$ будет прежде всего содержать свою собственную ошибку округления ε_2 того же порядка и, кроме того, унаследованную ошибку ε_3. Последняя удовлетворяет равенству $\varepsilon_3'' = \varepsilon_1$, или, скорее, точно удовлетворяет используемой дискретизации этого уравнения, и потому порядок ошибки ε_3 также равен h^{-2}. *Ошибки округления не объединяются в* h^{-4}.

В качестве другой проверки вычислим число обусловленности дискретной системы. В случае конечных разностей матрица имеет вид

$$\begin{pmatrix} -1 & \delta^2 \\ \delta^2 & 0 \end{pmatrix}\begin{pmatrix} M \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ f \end{pmatrix}, \tag{9}$$

и собственные значения μ этой блочной матрицы связаны с собственными значениями $\lambda(\delta^2)$ равенством

$$\mu^2 + \mu = \lambda^2.$$

Мы знаем, что собственные значения λ оператора второй разности δ^2 располагаются от $O(1)$ до $O(h^{-2})$. Решая квадратное уравнение, получаем ту же область расположения собственных значений μ, и число обусловленности μ_{max}/μ_{min} сдвоенной системы действительно имеет порядок h^{-2}.

Мы хотим предложить объяснение этого чуда, основанное на нашем наблюдении, что обычное измерение числа обусловленности для этих матриц неестественно. В вычислительных целях мы будем рассматривать эти матрицы как преобразования евклидова пространства (дискретного $\mathscr{H}^0$) в себя и потому возьмем одну и ту же норму для невязки уравнения и для результирующей ошибки в решении. Это целиком противоположно тому, что делается в дифференциальной задаче, или тому, что происходит при оценке ошибки дискретизации: f измеряется в норме пространства $\mathscr{H}^0$, M и ее ошибка — в $\mathscr{H}^2$, w и ее ошибка — в $\mathscr{H}^4$. (В вариационной задаче соответственно $\mathscr{H}^{-2}$, $\mathscr{H}^0$ и $\mathscr{H}^2$.) В самом деле, оператор $L = d^2/dx^2$ с каким-либо обычным краевым условием вполне обусловлен как преобразование из $\mathscr{H}^2$ в $\mathscr{H}^0$. Ограниченность операторов L и L^{-1} была существенным моментом в разд. 1.2. Можно показать, что это верно и для разностного оператора δ^2, а также для любого приемлемого аналога в методе конечных элементов, если только эти естественные нормы сохраняются. Следовательно, должен быть алгоритм решения уравнения $KQ = F$, отражающий это свойство, и тогда чудо развеялось бы: ошибки в M и w соответствовали бы их положению.

До тех пор пока проводится стандартное исключение, будет разница в ошибках округления для задач четвертого и второго порядков.

Прежде чем продолжать далее, приведем сдвоенные дифференциальные уравнения смешанного метода к вариационной форме. Умножим первое уравнение в (8) на M, второе — на w и проинтегрируем по частям:

$$\int (w''M + M''w - M^2 - 2fw)\,dx = \\ = -\int (2w'M' + M^2 + 2fw)\,dx + (w'M + M'w)\Big|_0^\pi.$$

В случае свободно опертой балки w и M обращаются в нуль на каждом конце (допустимое пространство удовлетворяет полным условиям Дирихле) и проинтегрированный член исчезает. Для закрепленной балки условие $w = 0$ налагается на каждом конце, а *равенство нулю w' дает естественное краевое условие для M*. Если допустимое пространство V содержит все пары (M, w) с функцией M из пространства Неймана $\mathscr{H}^1$ и функцией

w из пространства Дирихле $\mathscr{H}_0^1$, то стационарная точка функционала

$$I(\vec{v}) = -\int (2w'M' + M^2 + 2fw)$$

служит в точности решением задачи о закрепленной балке. Грубо говоря, если на функцию M не наложено никаких условий на концах, а первая вариация от I равна нулю, то множитель w' при M должен равняться нулю. Подчеркнем, что функционал $I(\vec{v})$ не имеет определенного знака, как и для принципа упругости Рейсснера; задача состоит в отыскании стационарной точки, а не минимума для $I(\vec{v})$.

Отметим вычислительные следствия этой вариационной формулировки. Для кусочно линейных элементов и $Nh = \pi$ функции M^h и w^h представимы в виде

$$M^h = \sum_0^{N+1} z_j \varphi_j(x), \quad w^h = \sum_1^N q_j \varphi_j(x)$$

(так как M свободна на концах, то нужны две дополнительные базисные функции). Теперь неизвестных стало $2N + 2$, но *систему нельзя решить последовательно*, сначала для M^h, а затем для w^h. При $N = 2$ матрица коэффициентов такова:

$$\begin{pmatrix} -G & D \\ D^* & 0 \end{pmatrix},$$

где

$$G = \frac{h}{6}\begin{pmatrix} 2 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix}, \quad D = \frac{1}{h}\begin{pmatrix} -1 & 0 \\ 2 & -1 \\ -1 & 2 \\ 0 & -1 \end{pmatrix}.$$

G — матрица массы (или матрица Грама) функций φ_j, а D — прямоугольная матрица второй разности. Для больших N неизвестные можно упорядочить $z_0, q_1, z_1, q_2, \ldots$, и тогда получается ленточная матрица. Число обусловленности снова равно $O(h^{-2})$; G заменяется единичной матрицей; собственными значениями служат $\mu = 1$ (дважды) и корни уравнения $\mu^2 + \mu = \lambda$, где λ — собственное значение обычной матрицы четвертой разности D^*D. Ошибки в M^h и w^h будут порядка h^2, а в их производных — порядка h. Разумеется, вторая производная от w^h не будет равна M^h.

Для двумерных задач Хеллан и Герман разработали простые элементы, подходящие для смешанного метода. Вместо того чтобы описывать эти элементы, мы предпочитаем вернуться к главному с точки зрения математики моменту — к трудности

установления сходимости для неопределенного функционала. Проще всего объяснить задачу в терминах бесконечной симметричной системы линейных уравнений $Lu = b$. Предположим, что пространство S^h конечномерно, а P^h — симметричный проектор на S^h, т. е. $P^h v$ — компонента вектора v в S^h. *Тогда метод Галёркина (5) равносилен задаче отыскания приближенного решения $u^h \in S^h$, удовлетворяющего равенству*

$$P^h L P^h u^h = P^h b. \tag{10}$$

Оператор $P^h L P^h$ дискретной задачи, совпадающий с нашей матрицей жесткости K, положительно определен, если таков оператор L; это случай Ритца. На самом деле $P^h L P^h$ даже более положителен, чем L, так как наименьшее собственное значение возрастает при переводе задачи в подпространство. Это очевидно: если $v \in S^h$, то $(P^h L P^h v,\ v) = (L P^h v,\ P^h v) = (Lv,\ v)$. Пусть оператор L положительно определен на всем пространстве: $(Lv, v) \geqslant \sigma(v, v)$ для всех v. Тогда эта положительная определенность наследуется оператором $P^h L P^h$ на подпространстве S^h и без убывания σ.

Если оператор L симметричен, но неограничен, то таким же, по-видимому, будет и оператор Галёркина $P^h L P^h$. (Предполагается, что тестовое пространство V^h и пространство пробных функций S^h совпадают. В противном случае, если Q^h — проектор на V^h, то уравнением Галёркина будет $Q^h L P^h u^h = Q^h f$, а оператор $Q^h L P^h$ даже не симметричен.) Естественно ожидать, что u^h с увеличением размерности пространства S^h будет приближаться к u. Тем не менее эта сходимость *не автоматическая*; в подтверждение правильности гипотезы вернемся к основной теореме численного анализа: *согласованность и устойчивость влекут сходимость.*

Теорема 2.2. *Предположим, что метод Галёркина*

а) *согласован, т. е. $\|v - P^h v\| \to 0$ для каждой функции v, и*

б) *устойчив, т. е. дискретные операторы равномерно обратимы: $\|(P^h L P^h)^{-1}\| \leqslant C$.*

Тогда метод сходится: $\|u - u^h\| \to 0$.

Доказательство. Обозначим $(P^h L P^h)^{-1}$ через R. Так как $Lu = f$, то

$$P^h L P^h u + P^h L (u - P^h u) = P^h f,$$

или

$$u + R P^h L (u - P^h u) = R P^h f.$$

Вычитая $u^h = RP^h f$, находим, что

$$\|u - u^h\| = \|RP^h L(u - P^h u)\| \leqslant C\|L\|\|u - P^h u\| \to 0.$$

Заметим, что скорость сходимости зависит и от константы устойчивости C, и от аппроксимирующих свойств подпространства S^h, как и в теореме 2.1. (Полезен особый случай предложенной теоремы для $C = 1/\sigma$ и $\|L\| = K$. Более общие результаты получены Бабушкой [Б 5].) Можно расширить теорию и доказать также *необходимость* согласованности и устойчивости для сходимости, а существование решения u предполагать вовсе не обязательно. Браудер и Петришин показали, как вывести обратимость исходного оператора L.

Снова обращаем внимание на специальную роль положительной определенности: она делает устойчивость автоматической. Вот почему метод Ритца так надежен. В случае неограниченности оператора L предположим, что подпространство S^h образовано первыми N координатными направлениями; $P^h v$ задается первыми N компонентами вектора v. Тогда $P^h L P^h$ представляет собой N-й главный минор матрицы L (подматрица в ее верхнем левом углу), а устойчивость означает, что эти $(N \times N)$-подматрицы равномерно обратимы. Похоже, что их обратимость следует из обратимости всей матрицы L, но это неверно. Хорошим примером служит обратимая матрица

$$L = \begin{bmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} & & \\ & \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} & \\ & & \ddots \end{bmatrix};$$

для каждого нечетного числа N ведущий главный минор вырожден. Его последняя строка состоит из нулей, и для $f = (1, 1/2, 1/4, \ldots)$ функционал $(Lv, v) - 2(f, v)$ на подпространстве S^h не имеет седловой (стационарной) точки. (Мы признательны Маккарти за его помощь нам в этих вопросах.) Даже в случае 2×2 неопределенная квадратичная функция $2xy$ совершенно исчезает на подпространстве $x = 0$.

Заметим, что перенумерация неизвестных приводит к матрице $L = \begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix}$, очень близкой к сдвоенной системе в примере смешанного метода. Мы ожидали сходимость в этом примере, но она действительно подтверждается только при подходящем выборе подпространств S^h. *По-видимому, конструкция конечных элементов дает такой выбор.* Джонсон продолжил доказатель-

ство сходимости для двух наиболее важных смешанных элементов, в которых моменты соответственно постоянны и линейны в каждом треугольнике. Его доказательство устанавливает после исключения неизвестных перемещений и определения неизвестных моментов как функций, минимизирующих положительно определенное выражение (дополнительную энергию), особое свойство. Это свойство совпадает с известным условием Ритца: пробные моменты в дискретном случае содержатся в пространстве допустимых моментов (тех, которые достигают равновесия для предписанной нагрузки) полной непрерывной задачи. Поэтому, как и в методе Ритца, сходимость основана на теории приближений и ее можно доказать. Для других смешанных (и гибридных) элементов естественное доказательство проверяется из согласованности (или аппроксимируемости) и *устойчивости.* Тогда общая теорема Бабушки [Б5] дает сходимость. Бреззи доказал устойчивость для одного гибридного элемента, и его прием распространяется на общую теорию.

Для достижения численной устойчивости в смешанном методе, где матрица коэффициентов не ограничена, в ходе исключения Гаусса необходимо разрешить выбор главного элемента (перестановку строк и столбцов). Вычислительные результаты подтверждают предсказанное убывание числа обусловленности и ошибок округления, хотя Коннор и Уилл представили неудовлетворительные результаты для элементов высокой степени: смешанный метод обладает переменным успехом. Тем не менее понижение порядка — настолько ценное свойство, что стоит продолжать развивать эту идею.

2.4. СИСТЕМЫ УРАВНЕНИЙ; ЗАДАЧИ ОБ ОБОЛОЧКАХ; ВАРИАНТЫ МЕТОДА КОНЕЧНЫХ ЭЛЕМЕНТОВ

Можно возражать против того, что в нашей теории метода конечных элементов рассматриваются отдельные неизвестные, в то время как большинство приложений содержат системы r уравнений для вектора неизвестных $\vec{u} = (u_1, \ldots, u_r)$. К счастью, разница часто несущественна. Вариационный принцип для системы также состоит в минимизации квадратичного функционала $I(\vec{v})$, а оценки ошибок зависят, как и раньше, лишь от аппроксимирующих свойств пространств S^h.

Типичный пример дают двумерная и трехмерная задачи упругости. В каждой точке неизвестны перемещения по координатным направлениям, а решение $\vec{u}^h$ метода конечных элементов — вновь пробная функция, ближайшая к исходному решению в смысле энергии деформации, представляющей собой квадратичную функцию, содержащую все неизвестные. С помощью

теории приближений в следующей главе будет показано, что независимо от размерности и количества неизвестных, скорость сходимости связана со степенью конечных элементов и с аппроксимацией области.

Для оболочек возникают некоторые новые и более трудные задачи. Теория оболочек обычно строится как предельный случай трехмерной задачи упругости, когда область Ω становится очень тонкой в одном направлении (в направлении нормали к поверхности оболочки). В результате предельного перехода в функционал потенциальной энергии вводится вторая производная от поперечного перемещения w. Поэтому дифференциальные уравнения имеют четвертый порядок по w и второй порядок по перемещениям в плоскости. Эта диспропорция служит платой за снижение трехмерной задачи до двумерной.

Отметим сразу возрастающую популярность трехмерных элементов, для которых такая редукция *не проделывается*. Из предельной процедуры, управляющей поиском точного решения задач со специальными свойствами симметрии (как в теории оболочек), автоматически не следует, что тот же процесс упростит численное решение более общих задач. (Этот вопрос возникает для функций напряжений Эйри при изгибе пластины: чувствительны ли они с вычислительной точки зрения к понижению количества неизвестных и возрастанию порядка уравнений? Мы в этом сомневаемся.) Очевидно, что тонкая оболочка никогда не будет отражать типичную трехмерную задачу, так как всегда появятся трудности с областями, близкими к вырожденным. Экспериментально испытывался не только изопараметрический прием, но и специальный выбор узловых неизвестных и сокращенных формул интегрирования в направлении нормали. С теоретической точки зрения необходимо оценить эффект малого параметра толщины t (Фрид сделал это относительно численной устойчивости и числа обусловленности), но в общем аппроксимационный теоретический подход можно применить обычным образом.

Существует также целый ряд элементов для криволинейных оболочек в случае двух независимых переменных, к которым наша теория применима непосредственно — даже если они содержат производные разных порядков. Эти элементы построены на основе полностью согласованной теории оболочек: поверхность оболочки описывается параметрически набором трех уравнений $x_i = x_i(\theta_1, \theta_2)$, где θ_1 и θ_2 — независимые переменные в уравнениях оболочки, которые меняются по обычной плоской области Ω. Кривизна оболочки проявляется только через производные от x_i, входящие как переменные коэффициенты в дифференциальные уравнения. Почти обязательно для вычисления матриц жесткости элементов потребуется численное инте-

грирование, но по своей сути задача подобна любой другой задаче на плоскости. Купер и Линдберг построили согласованные элементы высокой точности (CURSHL [K12]), комбинируя редуцированные в $\mathscr{C}^1$ полиномы пятой степени и полные кубические функции класса $\mathscr{C}^0$ из разд. 1.9 для нормальных перемещений и перемещений в плоскости соответственно. Так как разность $k-m$ для всех компонент равна 3, то скорость сходимости в энергии деформации будет h^6.

Такая конструкция, точность которой подтверждена численными экспериментами, должна была бы прекратить поиск хороших элементов для оболочек, но такой конструкции нет. Для большинства приложений и программ элементы CURSHL очень сложны и не приняты широко. С одной стороны, есть много задач со специальными свойствами симметрии, в которых применимы цилиндрические или тонкие элементы оболочек [O7]. С другой стороны, можно аппроксимировать оболочку общего вида объединением плоских частей. Каждая из этих частей действительно становится простым элементом пластины, а деформации изгиба и растяжения объединяются только сборкой этих пластин. Один из вариантов этого подхода (неизбежно содержащий элементы, несогласованные с точки зрения чистой теории оболочек), по-видимому, считается наиболее простым и практически пригодным способом работы со сложными оболочками.

Мы бы очень хотели проанализировать сходимость (или ее отсутствие) этих комбинаций плоско-пластинчатых элементов. Предполагается, что в разумных условиях деформация оболочки многогранника (такого, как геодезический купол) приближает деформацию истинной криволинейной оболочки, но мы не в состоянии подтвердить эту догадку. Математические проблемы не изведаны и чрезвычайно интересны.

В одномерном случае, когда криволинейная дуга заменяется полиномиальной конструкцией, все выглядит проще. Энергия деформации для дуги является суммой энергий изгиба и растяжения, и ее можно нормализовать:

$$a(\vec{v}, \vec{v}) = C \int \left(\frac{dv_1}{ds} + \frac{v_2}{r}\right)^2 + \frac{t^2}{12}\left(\frac{1}{r}\frac{dv_1}{ds} - \frac{d^2v_2}{ds^2}\right)^2.$$

Здесь r — радиус кривизны, t — толщина, а условием допустимости служит принадлежность v_1 и v_2 классам $\mathscr{H}^1$ и $\mathscr{H}^2$ соответственно; первая производная dv_2/ds представляет собой вращение вектора нормали к дуге, и ей непозволительно иметь разрывный скачок. Пара $v_1 = u$ и $v_2 = w$, минимизирующая потенциальную энергию [$I(\vec{v}) = a(\vec{v}, \vec{v}) +$ линейные члены], представляет собой перемещения дуги по касательной и по нормали. За-

метим, что они определяются из дифференциальных уравнений (уравнений Эйлера для $I(\vec{v})$) разных порядков.

Для аппроксимирующей конструкции (собранной из отрезков прямых) радиус кривизны r становится бесконечным. Это разъединяет две компоненты вектора $\vec{v}$ в энергии деформации и, следовательно, приводит к простым и независимым дифференциальным уравнениям для u и w. Объединение происходит тем не менее для предотвращения разрыва конструкции в деформированном состоянии. Другими словами, каждая пробная функция $\vec{v} = (v_1, v_2)$ при подходе s слева и справа к s_0 должна передвинуть шарнир в одно и то же место и, таким образом, конструкция остается непрерывной. Это соответствует двум условиям, связывающим обе компоненты векторов $\vec{v}_-$ и $\vec{v}_+$ с шарнирным углом θ. (Условия главные, а не естественные.) Кроме того, есть главное условие непрерывности вращения dv_2/ds, сохраняющее угол в каждом шарнире. (Заметим, что функция v_2 не обязательно непрерывна и может содержать в шарнирах δ-функции, у которых производные и слева, и справа равны нулю.)

Эти условия вынуждают полиномы в методе конечных элементов сочленяться в узлах. Для v_2 типичны модифицированные эрмитовы кубические полиномы: непрерывность вращения остается неизменной, а функция может терпеть разрыв, связанный с разрывом v_1. Очевидно, что для задачи о дуге такие пробные функции неприемлемы, а так как энергия деформации тоже изменяется при отбрасывании r, то вопрос о сходимости остается открытым. Для случая дуги окружности и правильного многоугольника *отсутствие сходимости* было доказано Вальцем, Фултоном и Цирусом (Вторая Райт-Паттерсонская конференция). Уравнения метода конечных элементов оказались просто разностными, но согласованными с *неверным дифференциальным уравнением*. Главные члены были правильными (радиус кривизны проявился через угол θ в условии непрерывности рамки), но для отдельно избранного элемента появились также нежелательные члены нулевого порядка по h [1]). Это наводит на мысль о возможности условия сходимости, похожего на кусочное тестирование разд. 4.2.

Задачи для оболочек неизбежно сложнее. Даже для многогранников, построенных из плоских пластин, условия непрерывности трудно наложить на полиномиальные элементы. Непрерывность производных всегда труднее достичь для полиномов

---

[1]) Эти нежелательные члены сравнимы, однако, с другими, появляющимися в связи с уравнениями дуги в полной двумерной теории упругости. Таким образом, главные члены во всех теориях, содержащих приближение многоугольником, могут совпадать. То же самое можно сказать об оболочках.

степени, меньшей 5. Одну полезную, но математически туманную аппроксимацию дает *дискретная гипотеза Кирхгофа*: вращения (скорости изменения нормального перемещения в двух координатных направлениях на оболочке) выбираются независимыми переменными, а их истинное отношение к нормальному перемещению налагается только как ограничение в узлах, а не на всей поверхности. Аппроксимации такого типа имеют практическую ценность: слишком много разновидностей задач об оболочках и слишком велик выбор элементов, чтобы решить, что данный конкретный подход лучше других. Даже согласованные элементы, такие, как CURSHL, обладают недостатками из-за того, что движение твердого тела не воспроизводится точно. Довод таков: высокие сооружения, наклоняющиеся при ветре, могут допускать такие большие перемещения и как следствие такие большие ошибки конечных элементов, что затемняется внутренний изгиб, имеющий первостепенное значение. Мы склонны доверять точности и больше заботиться об удобствах элементов высокой степени.

Очевидно, задачи об оболочках поставляют методу конечных элементов очень мощный тест — из всех линейных задач, пожалуй, самый мощный. То же верно и для математического анализа. Но есть задачи, которые были действительно недоступны более ранней технике, и мы верим, что находимся на пути к их пониманию.

Кроме систем уравнений, есть еще одно большое упущение в описанной нами теории метода конечных элементов: мы рассматривали только *метод перемещений*, в котором перемещение задается в виде $\sum q_j \Phi_j$, а оптимальные коэффициенты Q_j определяются из принципа Ритца — Галёркина. Когда была установлена вычислительная эффективность этого метода, настал период совершенствования элементов и автоматизации ввода-вывода данных и требуется быстрое развитие машинной графики для обеспечения понятного вывода информации. Остались, однако, некоторые несовершенства, от которых можно избавиться только изменением самого математического метода. В этом разделе мы хотим описать некоторые из предложенных изменений, еще лежащих в рамках *метода взвешенных невязок*. Это означает, что приближенное решение по-прежнему будет комбинацией пробных функций, но способ выбора коэффициентов q_j среди возможных может быть другим, либо могут измениться сами неизвестные.

Наиболее важный вариант — *метод сил*, в котором в качестве неизвестных берутся напряжения — производные от u, а не сама функция u. Во многих задачах они играют первостепенную роль, и естественно приближать прямо их. Результат совершенно от-

личается от результата, полученного при использовании смешанного метода, в котором неизвестны как перемещения, так и напряжения, а функционал неограничен. Здесь минимизируется *функционал дополнительной энергии* и, не считая изменения допустимых условий (на пробные функции наложены новые ограничения между элементами и на границе), математическое содержание остается прежним. Порядок аппроксимации, достигаемый в подпространстве пробных функций, остается решающим.

Рассмотрим в качестве примера уравнение Лапласа

$$\Delta u = 0 \quad \text{в} \quad \Omega, \quad u = g \quad \text{на} \quad \Gamma,$$

в котором u определяется из условия минимума функционала $I(v) = \int\int v_x^2 + v_y^2$. Для принципа дополнительной энергии эти производные $v_x = \varepsilon_1$ и $v_y = \varepsilon_2$ взяты как основные зависимые переменные. От пары $(\varepsilon_1, \varepsilon_2)$ не требуется более быть градиентом некоторой функции v. Другими словами, тождество для смешанных производных $(\varepsilon_1)_y = (\varepsilon_2)_x$, или *условие совместимости*, более не налагается. (Отсюда следует, конечно, что, когда приближения ε_1^h и ε_2^h уже определены, не существует единственного способа интегрирования для отыскания соответствующей функции u^h. При точном применении принципа дополнительной энергии оптимальные функции ε_1 и ε_2 *будут* производными от истинного перемещения u, но при дискретной аппроксимации эта связь между градиентом и перемещением теряется.)

В примере с уравнением Лапласа, для того чтобы квадратичный функционал был конечным, нужно лишь, чтобы пробные функции для ε_1 и ε_2 принадлежали пространству $\mathcal{H}^0$. Однако есть еще и ограничение *равновесия*, налагаемое самим дифференциальным уравнением: $u_{xx} + u_{yy} = 0$ приводит к

$$(\varepsilon_1)_x + (\varepsilon_2)_y = 0. \tag{11}$$

В действительности это означает, что пара $(\varepsilon_2, -\varepsilon_1)$ будет градиентом некоторой функции w: $w_x = \varepsilon_2$ и $w_y = -\varepsilon_1$. Дополнительный процесс в этом частном случае можно интерпретировать следующим образом: вместо поиска гармонической функции u вычисляется соответствующая функция тока w. Функция w сопряжена с u (функция $u + iw$ аналитична).

Это делает два описанных принципа очень похожими внутри области Ω. На границе, однако, они заметно отличаются: условие Дирихле для u заменяется на условие Неймана для w. Чтобы убедиться в этом, напомним, что вдоль любой кривой функция u и ее функция тока связаны соотношением $w_n = -u_s$. (Это следует из уравнения Коши — Римана $w_x = -u_y$ для вер-

тикальных кривых, из уравнения $w_y = u_x$ для горизонтальных кривых и из их линейной комбинации в общем случае.) Поэтому на границе Г главное условие $u = g$ заменяется на $w_n = -g_s$. Как и в любой неоднородной задаче Неймана, это значит, что допустимое пространство не подвергается ограничениям у границы, но функционал I заменяется на

$$I' = \iint (w_x^2 + w_y^2)\, dx\, dy + \int g_s w\, ds.$$

Интегрируя по частям последний член, получаем $-\int g w_s ds$, так что функцию w можно исключить, заменяя ее производными $w_x = \varepsilon_2$ и $w_y = -\varepsilon_1$. Принцип дополнительной энергии утверждает, что пара $(\varepsilon_1, \varepsilon_2)$, минимизирующая I', будет градиентом перемещения u, минимизирующего I.

При аппроксимации конечными элементами пробные функции для w должны принадлежать $\mathscr{H}^1$, чтобы интеграл I' был конечным. Во что это выливается для переменных ε_1 и ε_2? Во-первых, так как пробные функции для w непрерывны на границах между элементами, то таковы же и их производные по направлению стороны. Поэтому тангенциальная компонента градиента $(\varepsilon_2, -\varepsilon_1)$, являющаяся нормальной компонентой вектора $(\varepsilon_1, \varepsilon_2)$, должна быть непрерывна. *Именно это ограничение вызывает наибольшие трудности при аппроксимации принципа дополнительной энергии.* Каждая из функций ε_1 и ε_2 может не быть непрерывной, но непрерывность нормальной компоненты вектора $(\varepsilon_1, \varepsilon_2)$ обязательна.

Важно вывести это ограничение прямо из *условия равновесия* (11), минуя доказательство с помощью функции тока, так как для более общих задач идея функции тока неуместна. Сначала мы должны придать некоторый смысл уравнению $(\varepsilon_1)_x + (\varepsilon_2)_y = 0$, когда ε_1 и ε_2 принадлежат лишь $\mathscr{H}^0$. У них может не быть производных, так что условие нужно понимать в слабом смысле. Умножим его на гладкую функцию z, равную нулю на Г, и проинтегрируем по частям. Надлежащее ограничение тогда дается слабой формой уравнения (11):

$$\iint \left(\varepsilon_1 \frac{\partial z}{\partial x} + \varepsilon_2 \frac{\partial z}{\partial y}\right) dx\, dy = 0 \quad \text{для всех таких } z.$$

Для пробного пространства кусочно полиномиальных функций к этой форме ограничения можно применить теорему Грина, беря каждый раз по одному элементу. После этого исходное условие (11) должно выполняться внутри каждого элемента, а на каждой границе между элементами должно происходить сокращение граничных интегралов, или натяжений, возникающих из-за двух соседних областей. Это как раз то сокращение,

которое требует непрерывность нормальной компоненты вектора $(\varepsilon_1, \varepsilon_2)$.

Публикации по методу дополнительной энергии, по-видимому, начались с Треффца [Т10], установившего, что метод дает нижнюю границу интеграла энергии. Затем Фридрихс [Ф18] открыл, что основополагающая идея, как и в примере для уравнения Лапласа, состоит в применении метода Ритца к сопряженной задаче и что краевые условия, главные для одной задачи, становятся естественными для сопряженной. Ортогональность допустимых пространств для двух задач была развита Синжем в гиперкруговом методе [19] после его фундаментальной статьи совместно с Прагером [П9]. Комбинация прямых и сопряженных принципов приводит как к нижним, так и к верхним границам для энергии деформации и для перемещения u. Абстрактное сообщение об этом дали Обэн и Бушар [О3], а Вейнбергер [В7] применил идею к некоторым модельным задачам. Программы были составлены Моутом и Янгом.

Метод конечных элементов первоначально применялся (де Вебек [В5]) к напряжениям, т. е. к дополнительному принципу. Сейчас литература по этой теме неисчерпаема [1]).

Условия, наложенные между элементами, всегда вызывают практические трудности, и это приводит к конструкциям *метода множителей* и *гибридного метода*. Например, Андерхегген [А9] предложил использовать для задачи четвертого порядка о пластине и обычного метода Ритца минимизации функционала потенциальной энергии $I(v)$ кубические полиномы, для которых наклон нормали обычно разрывался между элементами. Наложение ограничения на непрерывность наклона связывает с краем каждого элемента множитель Лагранжа и заменяет метод Ритца методом минимизации с ограничением. Требуемые изменения в программах очень просты. Однако матрица жесткости становится неопределенной и (из-за неизвестных на сторонах) вычислительное время для кубических функций оказывается сравнимым с обычным методом жесткости для редуцированных полиномов пятой степени, предложенных в разд. 1.9.

Вторая интересная модификация — гибридный метод, впервые предложенный Пианом и Тонгом [П4, П3]. Они остроумно справились с трудностями, возникающими между элементами, построив семейство аппроксимаций и для поля напряжений внутри каждого элемента, и для перемещений на границах элементов. Поля напряжений удовлетворяют дифференциальному

---

[1]) Теория метода сил исходит из тех же прииципов и фактически тех же теорем об аппроксимации, которые применяются в этой книге к методу перемещений. Недостаток места не позволяет нам параллельно развивать здесь всю эту теорию.

уравнению внутри каждого элемента (так что однородный случай $f = 0$ гораздо проще), а перемещения, которые задаются независимым множеством кусочно полиномиальных функций, непрерывны. Для каждого перемещения модели дополнительная энергия сначала минимизируется отдельно внутри каждого элемента. Это приводит к семейству перемещений v^h, определенных теперь не только на границах элементов, но и всей области Ω, для которой можно применить метод Ритца. Окончательная гибридная аппроксимация определяется минимизацией по этим v^h. Обычно энергия в этом методе лежит между нижней границей, предоставляемой чистым методом Ритца, и верхней границей сопряженного к нему метода и во многих случаях дает аппроксимацию, существенно лучшую любой другой.

Мы еще раз вернемся к основному методу Ритца и заново рассмотрим вопрос: можно ли изменить вариационный принцип так, чтобы не было необходимости в главных краевых условиях? Принцип дополнительной энергии дает один из возможных ответов, но есть и другие. В самом деле, сейчас известен стандартный прием работы с неудовлетворяемыми ограничениями: ввести в минимизируемое выражение *штрафную функцию*. (Это было главной темой замечательной лекции Куранта [К15]; метод конечных элементов пришел позднее!) Для $-\Delta u = f$ и $u = g$ на Γ функционал $I(v)$ заменяется функционалом

$$I^h(v) = \iint_\Omega (v_x^2 + v_y^2 - 2fv) + C_h \int_\Gamma (v - g)^2 \, ds.$$

Точный минимум на всем допустимом пространстве $\mathscr{H}^1$ без ограничений достигается на функции U^h, удовлетворяющей задаче

$$-\Delta U^h = f, \quad C_h^{-1} U_n^h + U^h = 0 \quad \text{на} \quad \Gamma.$$

Следовательно, если $C_h \to 0$, то эти решения U^h сходятся к решению u, равному нулю на границе Γ. Теперь в методе Ритца функционал I^h, содержащий штрафной член, минимизируется на всем пространстве S^h без каких-либо краевых ограничений. Предположим, что S^h — пространство кусочно полиномиальных функций, содержащее полные полиномы степени $k-1$. Кроме того, существует равновесие между ошибкой $u - U^h$ (из-за невыполненных краевых условий и штрафа) и ошибкой $U^h - u^h$, вызванной минимизацией на подпространстве. Этот баланс привел Бабушку [Б4] к определению оптимальной зависимости C_h от h: $C_h = ch^{1-k}$.

Бабушка дал также строгие оценки ошибок для родственного метода множителей Лагранжа, в котором вновь пробные функции не подвержены ограничениям на границе. Для уравне-

ния Пуассона этот метод отыскивает стационарную точку $(u(x, y), \lambda(s))$ неограниченного функционала

$$F(v, \Lambda) = \int_{\Omega} (v_x^2 + v_y^2 - 2fv)\, dx\, dy - 2 \int_{\Gamma} \Lambda (v - g)\, ds.$$

Множитель Лагранжа пробегает все допустимые функции, определенные на Г, а в истинной стационарной точке он связан с решением равенством $\lambda = \partial u/\partial n$ [1]). Ошибку в стационарной точке (u^h, λ^h) на подпространстве метода конечных элементов легко оценить [Б6].

Главные краевые условия можно также отбросить с помощью совсем другого подхода, имеющего длинную историю: *метода наименьших квадратов*. Вместо функционала $I(v) = (Lv, v) - 2(f, v)$ в этом методе минимизируется невязка $\|Lw - f\|^2$. Забывая на время о краевых условиях, представим минимизируемый функционал в виде

$$\|Lw - f\|^2 = (Lw, Lw) - 2(f, Lw) + (f, f) = (L^*Lw, w) - 2(L^*f, w) + (f, f).$$

Последний член не зависит от w, и потому *метод наименьших квадратов на самом деле минимизирует функционал* $I' = (L^*Lw, w) - 2(L^*f, w)$, *который является функционалом Ритца для задачи* $L^*Lu = L^*f$. Эта новая задача автоматически самосопряжена, но заметим, что порядок уравнения удвоился.

С краевыми условиями метод наименьших квадратов был должным образом проанализирован впервые в недавних работах Брамбла и Шатца [Б26, Б27]. Рассматривая в качестве примера задачу Дирихле $-\Delta u = f$ в Ω и $u = g$ на Г, они ввели функционал

$$I''(w) = \iint (\Delta w + f)^2\, dx\, dy + ch^{-3} \int_{\Gamma} (w - g)^2\, ds. \qquad (12)$$

Множитель h^{-3} не имеет отношения к степени k подпространства S^h, как было в методе штрафов. Множитель устанавливает естественный баланс между слагаемыми внутри области и на границе. Минимизация функционала I'' становится теперь задачей *одновременной аппроксимации* в Ω и на многообразии Г меньшей размерности.

---

[1]) Мы верим, что этот метод будет перспективным вариантом стандартного метода перемещений, потому что он направлен непосредственно на проблему вычисления решения (или, точнее, его нормальной производной) на границе. Часто нужна именно эта информация, а определение ее из приближенного решения внутри области численно неустойчиво и не вполне удовлетворительно.

Разумно ожидать, что пространство кусочно полиномиальных функций даст ту же степень аппроксимации на Γ, что и в Ω. Если Γ — прямая (возьмем простейший случай), то полный полином от $x_1, \ldots, x_n$ сводится на Γ к полному полиному от краевых переменных $s_1, \ldots, s_{n-1}$. Мы полагаем, что для криволинейной границы теория приближений настолько же эффективна: s-я производная от u отличается от своего интерполянта на величину $O(h^{k-s})$ на Γ. Однако если дано только, что в Ω можно достичь k-ю степень аппроксимации, то одновременная аппроксимация с коэффициентом h^{-3} будет гораздо более трудной задачей. Ее решение, найденное Брамблом и Шатцем, показывает, что баланс степеней в (12) абсолютно верен. При условии, что $k \geqslant 4m$, их оценка ошибки для решения u_{LS}^h методом наименьших квадратов оптимальна:

$$\| u - u_{LS}^h \|_0 \leqslant ch^k \| u \|_k. \tag{13}$$

Практические трудности в методе наименьших квадратов связаны с повышением порядка дифференциального уравнения с $2m$ до $4m$. Чтобы новый функционал был конечным, пробные функции должны принадлежать пространству $\mathscr{H}^{2m}$. Это означает, что условие допустимости представляет собой требование непрерывности всех производных вплоть до порядка $2m-1$, что трудно достижимо. Более того, ширина ленты матрицы K возрастает, а ее число обусловленности по существу возводится в квадрат, переходя от $O(h^{-2m})$ к $O(h^{-4m})$[1]). Поэтому численное решение неминуемо должно сходиться медленнее. Скорость сходимости в $\mathscr{H}^1$ будет обычно равна h^ρ, где ρ — меньшее из чисел $k-s$ и $2k-4m$.

Наконец, упомянем два технических приема, недавно изобретенных для задач с однородными дифференциальными уравнениями, скажем $\Delta u = 0$, и неоднородными краевыми условиями. Есть множество важных приложений, когда u и $\partial u/\partial n$ интересуют нас только на границе и вычисление решения всюду в Ω оказывается неэффективным.

Одна из возможностей состоит в отыскании семейства точных решений $\varphi_1, \ldots, \varphi_N$ для дифференциального уравнения и в выборе комбинации $\sum Q_j \varphi_j$, удовлетворяющей возможно ближе краевым условиям. Это означает минимизацию некоторых выражений для граничной ошибки. В методе наименьших квадратов это приводит к линейному уравнению для Q_j. Фокс, Генричи и Молер [Ф11] достигли больших успехов в минимизации ошибки на дискретном множестве граничных точек вместо применения принципа минимакса. Если φ_j — собственные функции диф-

[1]) Это возражение опровергнуто модификацией Брамбла и Нитше.

ференциального оператора, то возможны значительные упрощения в некоторых отношениях; снова серьезные трудности возникают около границы. Ясное и детальное обсуждение этого более классического круга идей можно найти в книгах Михлина [12—14].

Второй способ использования точной информации об однородном уравнении состоит во введении функции Грина и в преобразовании задачи в интегральное уравнение на границе Г. В некоторых первоначальных попытках приближенное решение на Г бралось как кусочно полиномиальная функция, а ее коэффициенты определялись коллокацией. Теоретического обоснования этой идеи, по-видимому, не существует, но его время наступит.

3 АППРОКСИМАЦИЯ

3.1. ПОТОЧЕЧНАЯ АППРОКСИМАЦИЯ

Этот раздел открывает обсуждение вопроса, который с математической точки зрения лежит в основе теории метода конечных элементов, а именно аппроксимация пространствами S^h. Начнем с поточечной аппроксимации, где легко установить специальную роль полиномов. Затем эту модель распространим на пространства $\mathcal{H}^s(\Omega)$, т. е. на аппроксимации в энергетических нормах, на которых основан метод конечных элементов Ритца — Галёркина.

Предположим сначала, что задана гладкая функция $u = u(x)$, определенная в каждой точке $x = (x_1, \ldots, x_n)$ n-мерной области Ω. Пусть S — пространство узловых конечных элементов, натянутое на функции $\varphi_1(x), \ldots, \varphi_N(x)$. Как и в разд. 2.1, это означает, что каждой функции φ_j соответствует такой узел z_j и такая производная D_j, что

$$D_i\varphi_j(z_i) = \delta_{ij}. \tag{1}$$

Предположим, наконец, что *пространство имеет степень* $k-1$, *т. е. каждый полином от* $x_1, \ldots, x_n$ *с полной степенью меньше* k *можно представить в виде комбинации базисных функций* φ_j *и, следовательно, он принадлежит* S. (Например, полная степень полинома x_1x_2 равна 2; его наличие требуется в пространстве степени $k-1=2$, но не в пространстве первой степени, даже несмотря на то, что он линеен и по x_1, и по x_2.) Если $P(x)$ — такой полином и ему соответствует линейная комбинация $P = \sum p_j\varphi_j$, то по свойству интерполяции (1) весовые коэффициенты — это как раз узловые значения полинома:

$$p_j = D_jP(z_j).$$

Узловые производные D_j должны быть порядка $\leqslant k$.

Степень пространства S обычно легко вычислить. Для линейной и билинейной аппроксимации степень равна 1 (другими словами, $k=2$), для кубической и бикубической $k=4$; для редуцированной аппроксимации пятой степени $k=5$ и т. д.

Исходя из этих предположений, попытаемся вывести порядок аппроксимации, достигаемый пространством S. Вспоминая, что

область Ω разбита на элементарные области e_1, e_2, ..., будем при измерении точности аппроксимации использовать расстояния

$$h_i = \operatorname{diam}(e_i), \quad h = \max h_i.$$

Рассматривая *последовательность* узловых подпространств S^h с параметром h, мы надеемся установить, что ошибка аппроксимации убывает, как степенная функция от h. Это потребует предположения однородности, которое можно выразить следующим образом: *базисные функции* φ_j^h *однородны порядка* q *при условии, что существуют такие постоянные* c_s, *что для всех* h, i *и* j

$$\max_{\substack{x \in e_i \\ |\alpha| = s}} \left| D^\alpha \varphi_j^h(x) \right| \leqslant c_s h_i^{|D_j| - s}. \tag{2}$$

Это условие налагается на все производные $D^\alpha = \partial^{|\alpha|}/\partial x_1^{\alpha_1} \ldots \partial x_n^{\alpha_n}$ *вплоть до порядка* q, *т. е. оно выполняется для всех* α, *для которых* $|\alpha| = \alpha_1 + \ldots + \alpha_n \leqslant q$. Напомним, что $|D_j|$ — порядок производной, интерполируемой базисной функцией φ_j^h в узле z_j^h; $|D_j| = 0$, если φ_j^h соответствует значению интерполируемой функции v в вершине, $|D_j| = 1$, если φ_j^h соответствует производной v_x, v_y или v_n и т. д.

Хороший пример в одномерном случае дают эрмитовы кубические полиномы ($k = 4$), определяемые на каждом интервале своими значениями и значениями производных на концах. Для единичного интервала две такие базисные функции ψ и ω изображены на рис. 1.8. На уменьшенном интервале $[0, h]$ эти функции принимают вид $\psi^h = \psi(x/h)$ и $\omega^h = h\omega(x/h)$ соответственно: дополнительный множитель h вводится в ω^h для того, чтобы наклон в начале координат оставался равным 1. Этим объясняется наличие в (2) члена $h^{|D_j|}$, введенного для совпадения размерностей обеих частей неравенства. Такие базисные функции однородны вплоть до $q = 2$, но третьи производные содержат δ-функцию в узлах, и однородность пропадает. Ситуация типична для конечных элементов: допускается, чтобы в (2) ступенчатые функции содержались в производных q-го порядка, но не выше. Поэтому q — параметр, соответствующий гладкости подпространства; S^h содержится в пространстве $\mathscr{C}^{q-1}$ функций с $q - 1$ непрерывными производными и, что важнее, в пространстве $\mathscr{H}^q$ функций с q производными, суммируемыми в среднем квадратичном. Соответствующим условием возможности применения метода Ритца к дифференциальному уравнению порядка $2m$ будет просто $q \geqslant m$.

Условие однородности становится особенно важным для размерности 2 и выше. Обычно его можно записать в виде геомет-

рических ограничений на элементарные области e_i. Сначала рассматривается стандартный элемент, скажем правильный треугольник T с вершинами $(0, 0)$, $(1, 0)$, $(0, 1)$. По отношению к T базисные функции и их производные вплоть до порядка q должны быть ограничены. Затем преобразованием координат треугольник T переводится в заданный треугольник e_i и исследуется якобиан этого преобразования. В результате получаем, что *базис однороден, если при* $h \to 0$ *все углы в триангуляции превосходят некоторую нижнюю границу* θ_0. В этом случае нетрудно найти такие постоянные c_s, что

$$c_s \leqslant \frac{\text{const}}{(\sin \theta_0)^s}.$$

Подчеркнем, что влияние геометрических свойств на аппроксимацию целиком охватывается этой оценкой. При разбиении на четырехугольники все углы к тому же должны быть строго меньше 180°, чтобы избежать вырождение в треугольники.

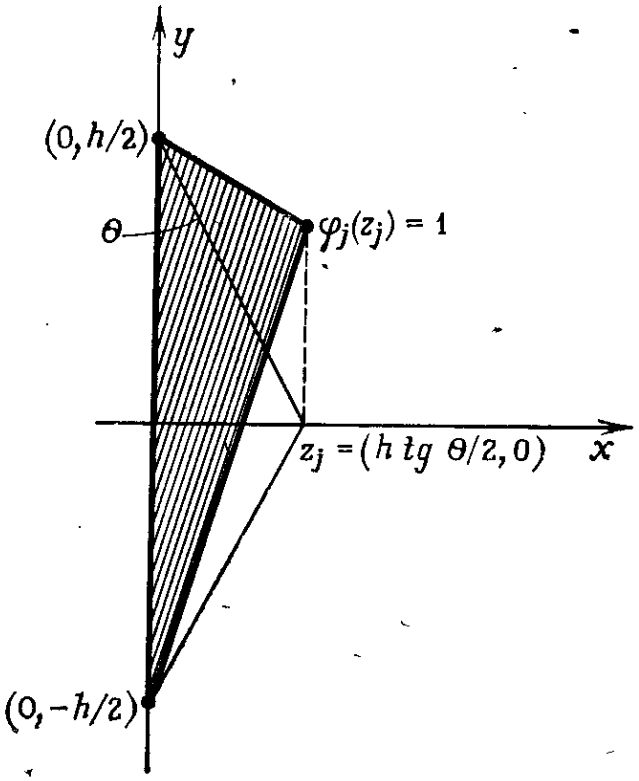


Рис. 3.1.
Функция линейного типа на узком треугольнике.

Отметим ситуацию, в которой эта оценка для c_s неверна. Она, безусловно, нарушается, если в элементарной области полиномы не определяются однозначно узловыми параметрами, т. е. если матрица H в разд. 1.10, связывающая коэффициенты полинома a_i с узловыми параметрами q_j, для какой-нибудь конфигурации элементарных областей будет неограниченной. Как только такой вырожденный случай выявлялся, его избегали в литературе. Опасность наиболее велика, когда область e_i первоначально ограничена кривой, и, чтобы применить одну из стандартных конструкций конечных элементов, ее отображают в многоугольник. Если якобиан этого преобразования обращается в нуль (см., например, разд.3.3), математические свойства конструкции теряются.

Легко понять геометрические условия для линейных функций на треугольниках. Пирамида φ_j^h равна 1 в j-й вершине и 0 в остальных. Между ними всегда $|\varphi_j^h(x)| \leqslant 1$, поэтому для нулевой производной однородность выполняется: $c_0 = 1$. Рассмотрим производную по x на рис. 3.1. Так как $\varphi_j(z_j) = 1$ и $\varphi_j(P) = 0$, на-

клон функции φ_j между этими точками равен

$$\frac{\partial}{\partial x}\varphi_j = \left(\frac{h}{2}\operatorname{tg}\theta\right)^{-1}.$$

Из сравнения этого соотношения с условием однородности (2) видно, что постоянная c_1 должна быть не меньше $2/\operatorname{tg}\theta$. Поэтому, если треугольник может вырождаться в «стрелку», базис не будет однороден.

На таких треугольниках линейная интерполяция может привести к существенным ошибкам. Возьмем, например, функцию $u(x, y) = y^2$. Ее интерполянт u_I равен нулю в z_j и $h^2/4$ в двух других узлах (и потому в P). Следовательно, для производной по x (равной нулю, так как функция $u = y^2$ не зависит от x) может быть серьезная ошибка в интерполяции:

$$\frac{\partial}{\partial x}(u - u_I) = \frac{h^2/4}{(h/2)\operatorname{tg}\theta} = \frac{h}{2\operatorname{tg}\theta}.$$

Эта ошибка в производной равна $O(h)$, только если угол θ ограничен снизу. Заметим, что $u - u_I = O(h^2)$ независимо от θ; трудности возникают от безразмерного множителя $h_{\max}/h_{\min}$, вносимого производной.

Вырожденных треугольников любого вида избегают также и по другой причине: они могут нарушить численную устойчивость метода Ритца, поскольку отражаются на числе обусловленности матрицы K. Поэтому было построено несколько алгоритмов такой триангуляции произвольной области Ω, чтобы все углы оставались более $\theta_0 \approx \pi/8$. Для более общих n-мерных элементов геометрические требования можно выразить в терминах вписанных шаров, а не углов: область e_i должна содержать шар радиуса не менее νh_i, где ν — фиксированная постоянная.

Теперь мы готовы к поточечной аппроксимации в узловом методе. Заданная функция u будет приближаться интерполянтом в пространстве S, другими словами, функцией u_I, у которой те же узловые параметры, что и у u:

$$u_I(x) = \sum D_j u(z_j)\,\varphi_j(x).$$

Предположим, что u имеет k производных в обычном поточечном смысле, и оценим s-е производные от $u - u_I$.

Теорема 3.1. *Пусть степень пространства S равна $k - 1$, а базис удовлетворяет условию однородности* (2). *Тогда для* $s \leqslant q$

$$\max_{\substack{x\in e_i \\ |\alpha|=s}} |D^\alpha u(x) - D^\alpha u_I(x)| \leqslant C_s h_i^{k-s} \max_{\substack{x\in e_i \\ |\beta|=k}} |D^\beta u(x)|. \qquad (3)$$

Доказательство. Выберем произвольную точку x_0 в e_i и разложим u в ряд Тейлора

$$u(x) = P(x) + R(x),$$

где P — полином степени $k-1$, а R — остаточный член. Нам нужна обычная оценка для остаточного члена R и его производных:

$$\max_{\substack{x \in e_i \\ |\alpha|=s}} |D^\alpha R(x)| \leqslant C h_i^{k-s} \max_{\substack{x \in e_i \\ |\beta|=k}} |D^\beta u(x)|. \tag{4}$$

Ее можно доказать, выражая R как интеграл по прямой от x_0 до x.

Интерполянт u_I разлагается по линейности в сумму двух интерполянтов

$$u_I(x) = P_I(x) + R_I(x).$$

Решающий момент здесь — совпадение P_I и P: любой полином степени $k-1$ точно воспроизводится его интерполянтом. Именно здесь полиномы играют особую роль. Другими словами, $P - P_I$ совпадает на e_i с пробной функцией, у которой все определяющие узловые параметры равны нулю; поэтому $P \equiv P_I$ на e_i. Таким образом, $u - u_I = R - R_I$, и осталось оценить лишь производные от R_I:

$$D^\alpha R_I(x) = \sum D_j R(z_j) \cdot D^\alpha \varphi_j(x).$$

В этой сумме не более d ненулевых членов, так как другие базисные функции на этом элементе равны нулю. Поэтому если (4) объединить с условием однородности (2), то

$$|D^\alpha R_I(x)| \leqslant dC h_i^{k-|D_j|} \max |D^\beta u| \cdot c_s h_i^{|D_j|-s} \leqslant$$
$$\leqslant C' h_i^{k-s} \max |D^\beta u|.$$

Следовательно, R_I оценивается так же, как R; отсюда немедленно получаем

$$|D^\alpha (u - u_I)| = |D^\alpha (R - R_I)| \leqslant (C + C') h_i^{k-s} \max |D^\beta u|.$$

Это и есть оценка (3). Теорема доказана [1]).

Заметим, что эта оценка полностью локальна; u_I имитирует свойства u в каждом элементе. Этого нельзя ждать от решения

---

[1]) Однородность базиса *не* является необходимым условием для выполнения теоремы об аппроксимации. Для билинейной и бикубической аппроксимаций на прямоугольниках выбор очень мелкой сетки в одном координатном направлении испортит однородность, но не порядок аппроксимации. Даже для треугольников можно указать более слабое условие (условие Синжа в [19]), а именно наибольший угол должен быть строго меньше π, т. е. *второй наименьший* угол в каждом треугольнике должен превышать некоторую величину θ_0.

методом конечных элементов u^h, так как оно получается минимизацией глобальной функции $I(v)$ — потенциальной энергии на всей области Ω. В самом деле, особенность функции u действительно распространяется с помощью u^h на всю область — иногда с очень медленным затуханием. Этот эффект анализируется в гл. 8.

Для абстрактного метода конечных элементов справедлива аналогичная теорема об аппроксимации. Мы потратим время на доказательство этой второй теоремы, несмотря на то что они частично совпадают в случае обычного узлового метода на равномерной сетке. Сплайны не охватываются предыдущей теоремой, потому что им нехватает локального интерполирующего базиса, однако их аппроксимирующие свойства чрезвычайно важны. В самом деле, специальная регулярность математической структуры допускает более изящный результат. *Аппроксимация на равномерной сетке зависит от наличия пробной функции ψ со следующим замечательным свойством: для любого полинома P степени, меньшей k,*

$$P(x) \equiv \sum_l P(l)\,\psi(x-l). \tag{5}$$

Такую суперфункцию ψ можно найти в пробном подпространстве S тогда и только тогда, когда его степень равна $k-1$. Функция ψ будет отличаться от нуля на малом участке элементов.

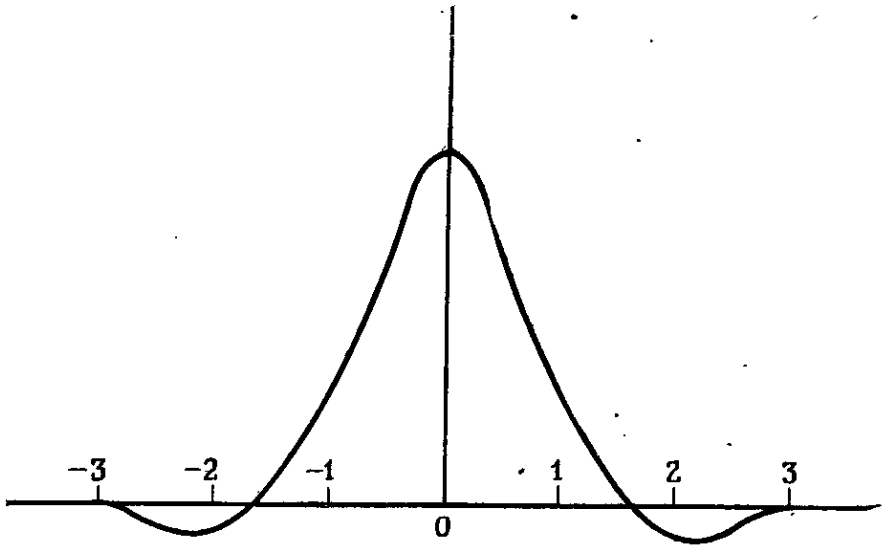
В простейшем одномерном примере пространство S^h состоит из непрерывных кусочно линейных функций. На каждый интервал сетки приходится одно неизвестное ($M=1$), а базис порождается функцией-крышкой Φ_1. Другими словами, базисными функциями служат $\varphi_l^h(x) = \Phi_1(x/h - l)$, а степень подпространства S^h равна 1. В этом примере в качестве функции ψ можно взять Φ_1, так как для любого полинома $P(x) = \alpha x + \beta$

$$P(x) = \sum P(l)\,\Phi_1(x-l).$$

Обе части равенства совпадают в узлах и линейны между ними, поэтому они должны совпадать всюду. То же самое справедливо в двумерном случае для линейных функций на треугольниках, где графиком функции $\Phi_1 = \psi$ является пирамида, и для билинейных функций на прямоугольниках, где график похож на пагоду. В этих случаях аппроксимирующая функция u_Q в приведенной ниже теореме совпадает с интерполянтом u_I.

Построение ψ для кубических функций не так тривиально. Для эрмитовых кубических функций в разд. 1.7 были описаны вместе с B-сплайном Φ_1 две порождающие функции. Ни одна из

них не удовлетворяет условию (5) на ψ вплоть до $k = 4$. Тем не менее существует кубический сплайн, удовлетворяющий этому условию и отличающийся от нуля на шести интервалах (рис. 3.2). Это комбинация B-сплайна и двух его соседей, а так как каждый сплайн автоматически является эрмитовой кубической функцией, то ψ в таком виде подходит также и для суперфункции в пространстве эрмитовых кубических функций. Обращаем внимание на то, что вовсе не обязательно знать, какова



$$\psi(x) = -(1/6)\varphi(x+1) + (4/3)\varphi(x) - (1/6)\varphi(x-1)$$

Рис. 3.2.

Суперфункция для кубических полиномов.

именно функция ψ: важно лишь, что где-то в пробном пространстве лежит функция, отвечающая за аппроксимирующие свойства пространства.

Если дано, что существует функция $\psi(x)$, равная нулю при $|x| \geqslant p$ и удовлетворяющая тождеству суперфункции, приблизим заданную функцию u функцией

$$u_Q(x) = \sum_l u(lh)\,\psi\left(\frac{x}{h} - l\right)$$

из S^h. Будем называть функцию u_Q *квазиинтерполянтом*, основанным на ψ. Он зависит, как и u_I, от локальных свойств функции u, но не совсем интерполирует ее в узлах. Преимущество функции u_Q состоит в том, что ее можно записать легко и явно, тогда как интерполяция сплайнами требует решения системы уравнений (несколько B-сплайнов отличны от нуля в каждом

узле). Так как для аппроксимации достаточно лишь *некоторой* функции из S^h, близкой к u, то мы свободны в выборе ради удобства.

Теорема 3.2. *Пусть степень пространства S^h абстрактного метода конечных элементов равна $k-1$, а ψ обладает ограниченными производными вплоть до порядка q. Тогда для любой производной D^α порядка $|\alpha| \leqslant q$*

$$\max_{\substack{0 \leqslant x_i < h \\ |\alpha|=s}} |D^\alpha u(x) - D^\alpha u_Q(x)| \leqslant c_s h^{k-s} \max_{\substack{|x| \leqslant \rho h \\ |\beta|=k}} |D^\beta u(x)|. \qquad (6)$$

Доказательство. Рассуждения почти такие же, как в теореме 3.1. Разложение Тейлора вблизи начала координат дает $u(x) = P(x) + R(x)$, где P — полином степени $k-1$, а R — остаточный член. Разбивая u_Q в сумму $P_Q + R_Q$, видим прежде всего, что P_Q совпадает с P:

$$P(x) = P_Q(x) = \sum_l P(lh)\,\psi\left(\frac{x}{h} - l\right).$$

Это есть не что иное, как тождество (5) с измененным (множителем h) масштабом на оси x; другими словами, тождество (5) применено к полиному $p(x) = P(xh)$, а затем x заменено на xh.

Остаточный член формулы Тейлора оценен в предыдущей теореме; остается оценить R_Q:

$$D^\alpha R_Q(x) = \sum R(lh)\, D^\alpha\left[\psi\left(\frac{x}{h} - l\right)\right].$$

Для x из куба $0 \leqslant x_i < h$ сетки

$$|D^\alpha R_Q(x)| \leqslant \sum C\,|lh|^k \max_{\substack{|x| \leqslant \rho h \\ |\beta|=k}} |D^\beta u(x)|\, h^{-|\alpha|} C',$$

где C' — граница для производных от ψ, множитель $h^{-|\alpha|}$ возникает при дифференцировании D^α, а остальное служит верхней гранью для $R(lh)$. Как и в теореме 3.1, существенна конечность суммы: поскольку $\psi \neq 0$ лишь для $|x| \leqslant \rho$, то действительно присутствует лишь конечное множество значений l. Итак, мы получили требуемую оценку при $|\alpha| = s$:

$$|D^\alpha(u - u_Q)| = |D^\alpha(R - R_Q)| \leqslant c_s h^{k-s} \max |D^\beta u|.$$

Тот же результат получается и для любого другого куба сетки, если разложить функцию u около одной из его вершин.

3.2. СРЕДНЕКВАДРАТИЧНОЕ ПРИБЛИЖЕНИЕ

Мы хотим доказать, что при тех же предположениях относительно подпространства S^h (его степень равна $k-1$, а базис однороден) для среднеквадратичной нормы возможна такая же степень аппроксимации. Именно вариационный принцип делает естественным и неизбежным работу с этими нормами пространств $\mathscr{H}^s$: энергия деформации есть не что иное, как интеграл от квадратов производных функции u. Поскольку u^h — ближайший к u элемент в этой энергетической норме, то мы будем изучать в основном эту ошибку.

Итак, вообще говоря, норма пространства $\mathscr{H}^m$ содержит все производные D^α порядка $|\alpha| \leqslant m$:

$$\| v \|^2_{m,\,\Omega} = \sum_{|\alpha| \leqslant m} \int_\Omega \left| \frac{\partial^{|\alpha|} v(x_1, \ldots, x_n)}{\partial x_1^{\alpha_1} \ldots \partial x_n^{\alpha_n}} \right|^2 dx.$$

Нам понадобится также *полунорма* $|v|_{m,\Omega}$, содержащая только производные порядка $|\alpha| = m$. Она называется полунормой, потому что обладает двумя свойствами нормы $|cv| = |c|\,|v|$ и $|v + w| \leqslant |v| + |w|$, но не является положительно определенной: для настоящей нормы $\| v \| = 0$, только если $v = 0$, а для этой полунормы $|P|_m = 0$, где P — любой полином степени, меньшей m. Очевидно, что $\| v \|^2_m = |v|^2_0 + |v|^2_1 + \ldots + |v|^2_m$.

В одном смысле распространение на эти новые нормы поточечных результатов совершенно тривиально. Используя интерполянт u_I из предыдущего раздела, возведем поточечную ошибку в квадрат и проинтегрируем: если $|\alpha| = s$, то

$$\int_\Omega |D^\alpha u - D^\alpha u_I|^2 dx \leqslant (\operatorname{mes}\Omega) \max |D^\alpha u - D^\alpha u_I|^2 \leqslant$$

$$\leqslant C_s^2 (\operatorname{mes}\Omega) h^{2(k-s)} \max_{\substack{x \in \Omega \\ |\beta| = k}} |D^\beta u(x)|^2.$$

Таким образом, если k-е производные функции u ограничены, ошибка будет того же порядка h^{k-s}, как и раньше. Это решает вопрос о скорости сходимости для гладких решений.

Такая оценка, однако, никогда полностью не соответствует требованиям. Она содержит две совершенно разные нормы, и чтобы получить среднеквадратичную оценку ошибки s-х производных, мы предположили равномерную ограниченность k-х производных. Мы вправе рассчитывать на более симметричную теорему, в которой употребляются нормы только одного вида. Более того, такое улучшение необходимо для задач с особенностями. На пластине с разрезом решение u возрастает, только как квадратный корень из расстояния до конца разреза. Функ-

ция $r^{1/2}$ даже недифференцируема; поточечная ошибка будет вести себя, как $h^{1/2}$ или, возможно, $h^{1/2} \ln h$. Однако решение должно принадлежать $\mathscr{H}^1$. В самом деле, u имеет примерно «полторы производной» в среднеквадратичном смысле: s-я производная ведет себя, как $r^{1/2-s}$, и для всех $s < 3/2$

$$\int | r^{(1/2)-s} |^2 r \, dr \, d\theta < \infty.$$

Короче, мы хотели бы доказать, что *порядок среднеквадратичной ошибки равен примерно* $h^{3/2}$, а такой результат можно вывести только из теоремы, в которой сделаны предположения о *среднеквадратичной дифференцируемости* функции u.

Теорема 3.3. *Пусть степень подпространства* S^h *равна* $k-1$, *а его базис однороден порядка* q. *Предположим, что порядок всех производных* D_j, *связанных с узловыми параметрами, меньше* $k - n/2$. *Тогда для любой функции* $u(x_1, \ldots, x_n)$, *обладающей* k *суммируемыми в квадрате производными, и для любой производной* D^α *порядка* $s \leqslant q$

$$\int_{e_i} | D^\alpha u(x) - D^\alpha u_I(x) |^2 \, dx \leqslant C_s^2 h_j^{2(k-s)} | u |^2_{k, e_i}. \tag{7}$$

Так как интеграл по Ω *равен сумме интегралов по* e_i, *то*

$$| u - u_I |_{s, \Omega} \leqslant C_s h^{k-s} | u |_{k, \Omega}. \tag{8}$$

Замечание 1. Предположение $|D_j| < k - n/2$ необходимо для того, чтобы определить интерполянт u_I. Лемма Соболева (упоминаемая в разд. 1.8) гарантирует, что для функции $u(x_1, \ldots, x_n)$, обладающей k суммируемыми в квадрате производными, производная $D_j u$ корректно определена в любой точке z_j:

$$| D_j u(z_j) | \leqslant c \| u \|_k. \tag{9}$$

К счастью, предположение $|D_j| < k - n/2$ выполняется для всех практических конечных элементов, и теорема непосредственно приводит к оценке скорости сходимости метода конечных элементов.

Для того чтобы показать более сложный случай, обратимся к аппроксимации кусочно постоянными функциями ($k = 1$) на плоскости ($n = 2$). Тогда, например, для функции $\ln \ln r$ из разд. 1.8 лемма Соболева не гарантирует, что интерполянт u_I имеет смысл: если узел попадет в начало координат, что будет означать $\ln \ln 0$? Тем не менее даже в этом случае подпространство S^h *обязательно* содержит функцию, дающую правильный порядок h аппроксимации метода наименьших квадратов; функ-

цию u можно подходящим образом сгладить, прежде чем ее интерполировать. Такая же конструкция возможна и в общем случае [С8]: *если степень подпространства S^h равна $k-1$, а его базис однороден, то для любой размерности n*

$$|u-\bar{u}_I|_s \leqslant C_s h^{k-s}|u|_k,$$

где $\bar{u}$ — сглаженная функция u. Таким образом, каждое пространство метода конечных элементов S^h содержит функцию (обычно u_I, а если необходимо, то $\bar{u}_I$), аппроксимирующую u с ожидаемым порядком. Оценки для u_I доказываются намного проще и достаточны для всех практических целей — или были бы такими, если расширить их на дробные производные, как потребовалось в примере пластины с разрезом. Это расширение на самом деле непосредственно выводится из теоремы, если воспользоваться теорией «интерполяционных пространств», которую мы опускаем, не желая вдаваться в подробности.

Замечание 2. Для доказательства даже более простой теоремы нужна одна вспомогательная лемма. Разложение Тейлора, на котором основана поточечная аппроксимация, здесь применить нельзя: функция u может иметь достаточно производных для определения ее интерполянта, но не для разложения Тейлора с остаточным членом h^k. Следовательно, нам необходимо привлечь функциональный анализ для построения полинома, настолько же близкого к функции u в среднем квадратичном, насколько были близки к u главные члены ряда Тейлора в поточечном смысле. Основной результат (см. работы [15] и [Б21]) таков: *для каждого элемента найдется такой полином P_{k-1}, что остаточный член $R = u - P_{k-1}$ удовлетворяет неравенству*

$$|R|_{s,\,e_i} \leqslant c h_i^{k-s}|u|_{k,\,e_i}, \quad s \leqslant k, \tag{10a}$$

а в каждом узле z_j

$$|D_j R(z_j)| \leqslant c' h_i^{k-|D_j|-n/2}|u|_{k,\,e_i}. \tag{10b}$$

На области диаметра $h_i = 1$ неравенство (10a) представляет собой стандартную лемму [15]; (10b) следует из (10a) и неравенства Соболева (9). Далее, указанная степень h_i появляется при изменении масштаба независимых переменных так, чтобы сжать область до e_i [1]). Константы c и c' зависят от наименьшего угла в e_i. Проще всего предположить ограниченность этого угла снизу, что естественно при однородном базисе.

---

[1]) h_i^k возникает при изменении масштаба k-х производных, а величину $h_i^{n/2}$ дает квадратный корень из объема области e_i.

Доказательство теоремы. Рассуждения те же, что и в поточечном случае — благодаря этой лемме. Для каждой элементарной области e_i запишем $u = P_{k-1} + R$, где R удовлетворяет неравенствам (10). Интерполянт u_I в силу линейности равен сумме двух интерполянтов:

$$u_I = (P_{k-1})_I + R_I = P_{k-1} + R_I,$$

так как полином P_{k-1} интерполируется точно. Следовательно, $u - u_I = R - R_I$. Рассмотрим

$$R_I(x) = \sum (D_j R)(z_j)\,\varphi_j(x).$$

Не равно нулю лишь конечное число d этих членов. Учитывая однородность базиса и оценку (10b), находим, что для любой производной порядка $|\alpha| = s$

$$\left|D^\alpha R_I(x)\right| \leqslant dc' h_i^{k-|D_j|-n/2} |u|_{k,\,e_i} \cdot c_s h_i^{|D_j|-s}.$$

Теперь возведем в квадрат, проинтегрируем по e_i и извлечем квадратный корень:

$$\left(\int_{e_i} |D^\alpha R_I(x)|^2\,dx\right)^{1/2} \leqslant c'' h_i^{k-s} |u|_{k,\,e_i}.$$

Это и есть оценка (10a) для R; правая часть совпадает с правой частью неравенства (7). Тем самым доказательство закончено. Технический прием, состоящий в применении неравенств (10) с учетом особой роли полиномов, известен под названием *леммы Брамбла — Гилберта.*

Такая же теорема с аналогичным доказательством справедлива для абстрактного метода конечных элементов на равномерной сетке (в частности, для сплайнов); снова берем u_Q вместо u_I [C7].

Теория *неравенств* в частных производных приводит к вопросу *об одностороннем приближении.* Для стандартных линейных элементов мы уже установили, что при заданной функции $u \geqslant 0$ оценки теорем 3.1—3.3 остаются справедливыми, если потребовать $0 \leqslant v^h \leqslant u$. (Интерполянт $v^h = u_I$, конечно, бесполезен, так как он не обязательно лежит ниже u.) Заметим, что Дюво и Лионс сумели сформулировать в виде вариационных неравенств несколько важных физических задач (в том числе задач упруго-пластичности), в дифференциальной формулировке приводящих к чрезвычайно неудобным элиптико-гиперболическим системам с неизвестной свободной границей раздела. Моско и Стренг и независимо от них Фальк подтвердили обычную ошибку h^2 в энергии для линейной аппроксимации *задачи Сен-Венана о кручении,* типичной для класса вариационных неравенств, так называемых *задач с ограничениями.*

Если элементы содержат несколько полиномиальных членов более высокой степени k, то возникает дополнительный вопрос: все ли производные этого порядка нужны в правой части неравенства (7)? Билинейные элементы, например, воспроизводят член кручения xy, и потому кажется излишним включать в оценку ошибки $ch^{2-s}|u|_2$ смешанную производную u_{xy}. Этот вопрос решен Брамблом и Гилбертом [Б22]: действительно, достаточно включить в оценку ошибки лишь u_{xx} и u_{yy}. Отсюда будут вытекать ценные следствия для теории прямоугольных изопараметрических элементов в следующем разделе.

Может случиться, что неравенство (7) выполняется внутри каждого элемента для производных всех порядков $s \leqslant k$. Ограничение $s \leqslant q$ в (8) возникает, когда u_I имеет лишь $q-1$ производных на границах между элементами.

Было бы полезно узнать что-нибудь о постоянных C_s в теореме. Они прямо указывают на свойства отдельного элемента: если для одного элемента они больше, чем для другого той же степени, то первый элемент сравнительно неточен; или «жесток». Для кусочно линейной аппроксимации на прямой с равномерно расположенными узлами $x_j = jh$ эти оптимальные постоянные можно вычислить. Две функции представляют особый интерес: первый тригонометрический полином $f(x) = \sin \pi x/h$, равный нулю в каждой точке сетки, и первый алгебраический полином $g(x) = x^2$, не совпадающий с его линейным интерполянтом. Для функции f и интерполянт, и наилучшее линейное приближение тождественно равны нулю. Поэтому сама функция f будет ошибкой, и легко подсчитать, что

$$|f|_0 = \frac{1}{\pi^2} h^2 |f|_2, \quad |f|_1 = \frac{1}{\pi} h |f|_2.$$

Из доказательства теоремы 1.3 в разд. 1.6 следует, что эти постоянные оптимальны, так как для любой функции u и ее интерполянта u_I

$$|u - u_I|_0 \leqslant \frac{1}{\pi^2} h^2 |u|_2, \quad |u - u_I|_1 \leqslant \frac{1}{\pi} h |u|_2.$$

Следовательно, для линейной аппроксимации $C_0 = 1/\pi^2$ и $C_1 = 1/\pi$.

Какая роль отводится функции $g(x) = x^2$? Для каждой *фиксированной* функции u она дает постоянную, *асимптотически правильную при* $h \to 0$. В предыдущем случае мы зафиксировали h и нашли наихудшую функцию $\sin \pi x/h$. Здесь же мы фиксируем u и ищем пределы

$$c_0 = \lim_{h \to 0} \frac{\min |u - v^h|_0}{h^2 |u|_2} \quad \text{и} \quad c_1 = \lim_{h \to 0} \frac{\min |u - v^h|_1}{h |u|_2}. \tag{11}$$

Минимум берется по всем $v^h \in S^h$, т. е. по всем кусочно линейным функциям. Для каждого значения h эти отношения ограничены величинами $1/\pi^2$ и $1/\pi$ соответственно и, значит, пределы не могут превышать эти постоянные. Можно ожидать, что они будут меньше, так как фиксированная функция u не может походить сразу на все осциллирующие функции $\sin \pi x/h$. Эти новые постоянные c_0 и c_1 (если они существуют) кажутся более естественными при оценке улучшения, ожидаемого в практической задаче при измельчении сетки, поскольку здесь фиксировано решение, а h изменяется.

Сначала рассмотрим специфическую функцию $u = g = x^2$. На интервале $[-1, 1]$ наилучшие линейные приближения для перемещений и наклонов минимизируют соответственно интегралы

$$\int_{-1}^{1} |x^2 - a_1 - a_2 x|^2 \, dx \quad \text{и} \quad \int_{-1}^{1} |2x - a_2|^2 \, dx.$$

Эти интегралы равны

$$\frac{2}{5} - \frac{4a_1}{3} + 2a_1^2 + \frac{2a_2^2}{3} \quad \text{и} \quad \frac{8}{3} + 2a_2^2.$$

В обоих случаях $a_2 = 0$, так как наилучшее приближение четной функции на симметричном интервале $[-1, 1]$ тоже четно. Оптимальное значение a_1 есть $1/3$, и отношения, образующие c_0 и c_1 (на интервале длины $h = 2$), принимают вид

$$\frac{\left|x^2 - \frac{1}{3}\right|_0}{2^2 |x^2|_2} = \frac{1}{12\sqrt{5}}, \quad \frac{\left|x^2 - \frac{1}{3}\right|_1}{2 |x^2|_2} = \frac{1}{2\sqrt{3}}.$$

Последнее не слишком отличается от $1/\pi$; оно меньше $1/\pi$, как и следовало ожидать.

Заметим, что $x^2 - 1/3$ представляет собой второй полином Лежандра; это ошибка в методе наименьших квадратов при аппроксимации квадратичной функции $g(x) = x^2$ линейными. Она принимает одинаковые значения $2/3$ на обоих концах интервала. Это позволяет легко увязать ее с наилучшим приближением для x^2 на соседнем интервале $1 \leqslant x \leqslant 3$. На нем функция ошибки оптимальной аппроксимации имеет вид $(x-2)^2 - 1/3$, т. е. тот же полином Лежандра, но смещенный вдоль оси x на две единицы. Эта схема сохраняется: на каждом интервале $[2n-1, 2n+1]$ оптимальная функция ошибки $x^2 - v^h$ есть $(x-2n)^2 - 1/3$ и функция ошибки периодична с периодом $h = 2$ (рис. 3.3). Постоянные $1/12\sqrt{5}$ и $1/2\sqrt{3}$ одинаковы на любой совокупности этих интервалов. *Более того, размерность этих постоянных правильна и они не изменяются, если сначала изменить масштаб*

независимых переменных так, чтобы перейти к произвольному исходному интервалу $[-h/2, h/2]$, *а затем сдвинуть начало координат в произвольную точку*. Ось x на рис. 3.3 масштабирована в отношении $h/2:1$, а функция ошибки на оси y — в отношении $(h/2)^2:1$. Легко проверить, что отношения инвариантны. Причина такой специфичности функции x^2 заключается в том, что переход к $(x-x_0)^2$ изменяет ее на линейное выражение $2xx_0 - x_0^2$, которое можно выразить через пробные функции. Заметим,

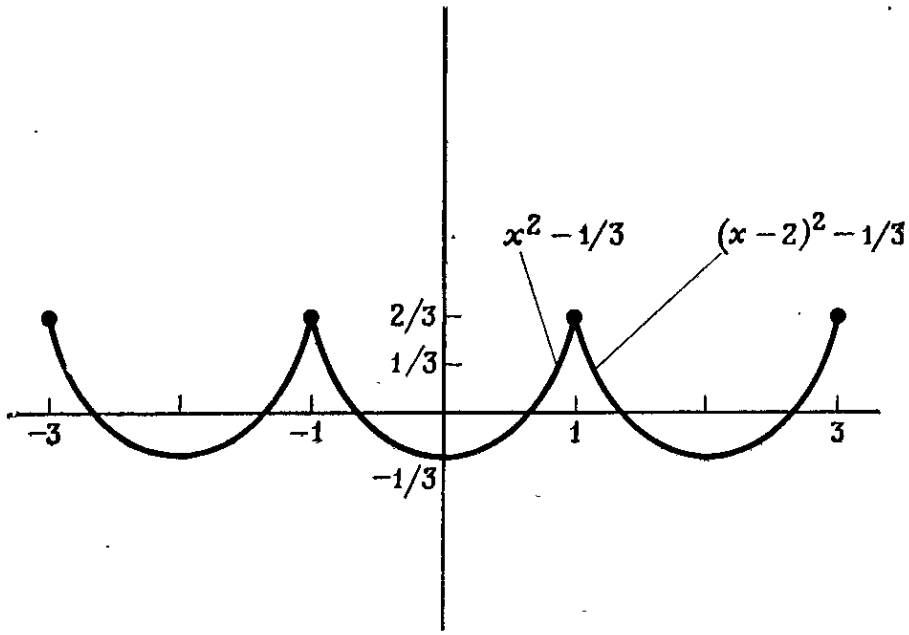


Рис. 3.3.
Ошибка линейной аппроксимации функции x^2.

что *наилучшее приближение никоим образом не будет линейным интерполянтом*: интерполянт дает правильный показатель степени у h, но слишком большую постоянную. В процедуре Ритца достигается наилучшая постоянная, потому что минимизация осуществляется по всему пространству S^h.

Замечательно, что асимптотическая постоянная не зависит от u.

Теорема 3.4. *Для произвольной функции* $u(x)$ *предельные значения отношений при* $h \to 0$ *те же, что и для специфической функции* x^2: $c_0 = 1/12\sqrt{5}$ *и* $c_1 = 1/2\sqrt{3}$.

Из этой теоремы вытекает, что при $h \to 0$ каждая функция ведет себя, как кусочно квадратичная, т. е. функция ошибки после линейной аппроксимации локально подобна функции, изображенной на рис. 3.3. Другими словами, *чем пристальней вы смотрите на функцию, тем больше она напоминает вам полином*. Это лежит в основе разложений в ряд Тейлора. Поэтому постоянная c_0 асимптотически правильна при аппроксимации мето-

дом наименьших квадратов не только для какой-то особой функции u, но для любого ее выбора. Аналогично постоянная c_1 асимптотически правильна для ошибки $u - u^h$ метода конечных элементов в примере гл. 1 для уравнения второго порядка. Вообще c_1 будет решающей постоянной *даже в ошибке перемещения* для $u - u^h$, поскольку функция u^h выбрана так, чтобы минимизировать энергию деформации в ошибке. Прием Нитше в теореме 1.5 указывает, что c_1^2 правильнее c_0 отражает ошибку перемещения.

Теорему 3.4 докажем сначала для гладкой функции u. На каждом интервале длины h со средней точкой x_0

$$u(x) = u(x_0) + (x - x_0)\,u'(x_0) + \frac{(x - x_0)^2}{2}\,u''(x_0) + O(h^3). \quad (12)$$

Наилучшим (в смысле наименьших квадратов) линейным приближением для трех первых членов будет

$$l(x) = u(x_0) + (x - x_0)\,u'(x_0) + \frac{1}{3}\left(\frac{h}{2}\right)^2 \frac{u''(x_0)}{2}.$$

Используя такие линейные функции на каждом подынтервале, мы действительно возвращаемся к квадратичному случаю. Учитывая ошибку $O(h^3)$, возникающую из остаточного члена в разложении Тейлора, видим, что отношения на каждом интервале все еще удовлетворяют равенствам

$$|u - l|_0 = \frac{h^2 |u|_2}{12\sqrt{5}}(1 + O(h)), \quad |u - l|_1 = \frac{h |u|_2}{2\sqrt{3}}(1 + O(h)).$$

Возведем в квадрат и просуммируем по всем интервалам, тогда для кусочно линейной функции L, образованной этими кусками l,

$$\frac{|u - L|_0}{h^2 |u|_2} = \frac{1 + O(h)}{12\sqrt{5}}. \quad (13)$$

Остается одна трудность: *L разрывна в узлах*. Линейная аппроксимация l зависит от разложения Тейлора на своем собственном интервале, и нельзя ожидать, что соседние функции соединятся с ней. Расхождение, однако, имеет порядок лишь

$$\frac{1}{3}\left(\frac{h}{2}\right)^2 \left(\frac{u''(x_0)}{2} - \frac{u''(x_0 + h)}{2}\right) = O(h^3),$$

так что изменив каждый кусок l на $O(h^3)$, получим непрерывную функцию $\tilde{L}$ (в норме $|\ |_1$, изменение дает $O(h^2)$). После такого изменения и равенство (13), и его аналог в норме $|\ |_1$ продолжают выполняться для $\tilde{L}$. Поэтому при $h \to 0$ отношения приближаются к постоянным $1/12\sqrt{5}$ и $1/2\sqrt{3}$. Ясно, что ни один другой выбор v^h в (11) не дает меньшей постоянной, так как

эти постоянные уже были верны для функции L, образованной из оптимальных l на каждом подынтервале. Таким образом, теорема доказана для случая, когда функция u достаточно гладка, чтобы допустить разложение Тейлора (12).

Техника расширения на все функции $u \in \mathscr{H}^2$ стандартна. Зададим линейный оператор P_s^h из $\mathscr{H}^2$ в $\mathscr{H}^s$ условием «$h^{s-2}P_s^h u$ — компонента функции u, ортогональная к S^h». Два свойства этого оператора уже доказаны:

1. $|P_s^h u| \leqslant C_s |u|_2$ (теорема 3.3 для $k=2$).
2. $|P_s^h u|_s \to c_s |u|_2$ для гладкой функции u (см. предыдущий абзац).

Второе свойство для всех u следует из обычного доказательства полноты, не интересного для инженеров и скучного для математиков. Теорема 3.4 доказана.

Обращаем внимание (и снова будем это делать), что нули функции $x^2 - 1/3$ являются специфическими точками. Поскольку это нули полинома Лежандра, они участвуют в гауссовых квадратурах: на интервале $[jh, (j+1)h]$ они переходят в $(j + 1/2 \pm \pm 1/\sqrt{3})h$. Для целей метода конечных элементов они специфичны еще и по другой причине: в этих точках наилучшее приближение для квадратичной функции равно нулю, а u^h абсолютно точна. (Известно, что в методе коллокации это так; см. разд. 2.3.) Будем называть их *точками перемещения*. Есть также *точки напряжения*, открытые Барлоу, которые еще важнее. Это точки, где производные от функции ошибки равны нулю (точка $x = 0$ в нашем простом примере); в разд. 3.4 мы покажем, что *ошибки напряжений в этих точках меньше на добавочную степень* h.

Предыдущая теорема распространяется на любой конечный элемент на n-мерной равномерной сетке и даже на любой пример абстрактного метода конечных элементов. Для n переменных существует несколько производных D^β порядка $k = |\beta|$, возможно, связанных с разными постоянными в ошибках аппроксимации. В самом деле, если x^β оказывается в S^h, то соответствующая постоянная равна нулю. Локально можно считать функцию u разложенной в ряд Тейлора вплоть до члена степени k. Члены степени $k-1$ точно воспроизводятся пробным подпространством, и аппроксимация асимптотически зависит лишь от производных $D^\beta u$ порядка k. Это обобщение теоремы 3.4 можно сформулировать, употребляя матрицы K_s вместо числовых постоянных c_s.

Теорема 3.5. *Если степень пространства S^h на равномерной сетке равна $k-1$, то существуют такие неотрицательно опреде-*

ленные матрицы K_s, что для любой функции $u \in \mathcal{H}^k$

$$h^{2(s-k)} \min_{S^h} |u - v^h|_s^2 \to \sum_{|\alpha|=|\beta|=k} K_s^{\alpha\beta} \int (D^\alpha u)(D^\beta u)\, dx. \quad (14)$$

Диагональные элементы $K_s^{\beta\beta}$ можно определить аппроксимацией одночленов $u = x^\beta = x_1^{\beta_1} \dots x_n^{\beta_n}$, для которых $D^\beta u$ — постоянная, а другие производные $D^\alpha u$ порядка k равны нулю.

Удобно возвести в квадрат выражения, фигурирующие в теореме 3.4, т. е. $K_0 = c_0^2 = 1/720$ и $K_1 = c_1^2 = 1/12$. Для двумерной линейной аппроксимации матрицы K_s будут третьего порядка соответственно трем производным $\partial^2/\partial x^2$, $\partial^2/\partial x \partial y$, $\partial^2/\partial y^2$ порядка $k = 2$. Для пространств сплайнов произвольной степени k постоянные $K_s^{\alpha\beta}$ вычислены [С7] в терминах чисел Бернулли. Мы

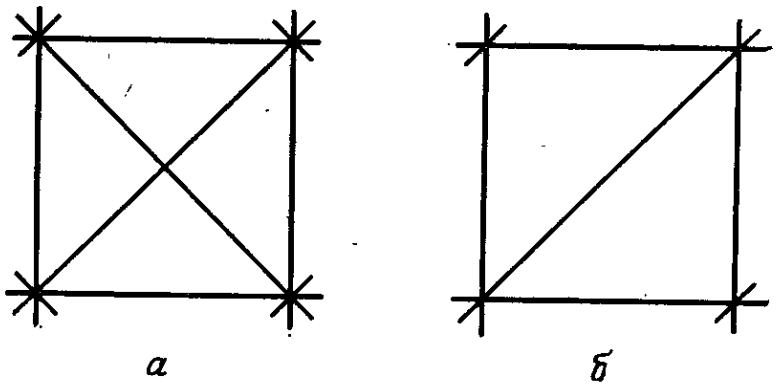


Рис. 3.4.
Две возможные триангуляции.

думаем, что для пространств эрмитовых функций они будут такими же. Известны и минимальные постоянные C_s для сплайнов, если h фиксировано, а функция u меняется (Бабушка). Здесь снова крайний случай представляют синусы с длиной волны $2h$, их наилучшие приближения — тождественные нули, а постоянные равны $C_s = \pi^{s-k}$.

С помощью теоремы 3.5 можно количественно сравнить два различных полиномиальных элемента или два одинаковых элемента с разными геометрическими формами. Рассмотрим две регулярные триангуляции плоскости: одна содержит диагонали в обоих направлениях, другая — только в одном (рис. 3.4). Комбинационно они совершенно различны. В триангуляции *а* одни узлы связаны с четырьмя соседними, а другие — с восемью. В *б* каждый узел имеет шесть соседних. Рассмотрим пространство Куранта S^h непрерывных кусочно линейных функций на этих треугольниках. Так как в триангуляции *а* вдвое больше узлов, чем в *б*, то размерность пространства S_A^h (соответствующего триангуляции *а*) вдвое больше размерности пространства

S_B^h (соответствующего триангуляции *б*). Более того, пространство S_A^h содержит S_B^h и потому по крайней мере такое же хорошее в смысле аппроксимации. Возникает вопрос: *вдвое ли оно лучше* (компенсация за удвоенное количество параметров)?

В двумерном случае три одночлена: x^2, xy и y^2. Предположим, что для каждого из них мы нашли функцию $u^h \in S^h$, минимизирующую $|u - u^h|_1$. Эта функция u^h и будет решением метода конечных элементов уравнения Пуассона, когда точное решение u есть квадратичная функция.

Начнем с функции $u = x^2$ на квадрате со стороной $h = 2$, симметричном относительно начала координат. Из соображений симметрии значения оптимальной функции u^h в четырех вершинах одинаковы и равны, скажем, α в случае *а* и β в случае *б*. В центре $u^h = \beta$ в случае *б*, а в случае *а* u^h может принимать другое значение, скажем γ. Это означает, что в *б* функция u^h постоянна на этом квадрате, причем равна $\beta = 1/3$, как и ранее в одномерном случае. Ошибка на квадрате сетки есть

$$|u - u^h|_1^2 = \iint (x^2 - \beta)_x^2 + (x^2 - \beta)_y^2 = \iint (2x)^2 \, dx \, dy = \frac{16}{3}.$$

Она равна $h^2 |u|_2^2/12$ и, как в одномерном случае, $K_1^{11} = 1/12$. В конфигурации *а* минимизация проводится по γ, так что постоянная должна быть меньше, и окончательно получаем

$$|u - u^h|_1^2 = \frac{h^2 |u|_2^2}{18}. \qquad (15)$$

Так можно сравнить две конфигурации. Размерность в случае *а* вдвое больше, чем в случае *б*; другими словами, в *а* эффективнее работать с квадратами со стороной $h/\sqrt{2}$ вместо h в случае *б*. При такой замене постоянная в (15) становится равной $1/9$ и *конфигурация б лучше, чем а в отношении* 12 : 9. Коэффициент в ошибке при x^2 для эквивалентного множества свободных параметров будет меньше в случае *b* в $\sqrt{3/4}$ раз. По симметрии это справедливо и для коэффициента при y^2. Для члена кручения xy обе конфигурации оказываются одинаково эффективны, и нечего выбирать.

Эти подсчеты подтверждаются численными экспериментами, описываемыми в технической литературе, которая отдает предпочтение конфигурации *б*. Для элементов более высокого порядка на ЭВМ смогли вычислить постоянные $K_{\alpha\beta}$, решая методом конечных элементов задачу, истинным решением которой было $u = x^\beta$. Одновременно вычисляются главные члены в ошибке усечения ряда Тейлора для конечно-разностной схемы, возникающей на равномерной сетке (см. разд. 1.3 и 3.4).

Из теоремы 3.5 вытекает также важный теоретический результат.

Следствие. *Для достижения аппроксимации порядка h^{k-s} для s-х производных пробное пространство S^h на равномерной сетке должно быть по крайней мере степени $k-1$. Поэтому метод конечных элементов в случае дифференциального уравнения порядка $2m$ сходится, только если $k > m$. Это и есть условие постоянной деформации, состоящее в том, что все полиномы степени m должны принадлежать S^h.*

Это следствие является обратным к теореме 3.2 на равномерной сетке. Для его доказательства предположим, что степень пространства S^h равна лишь $l-1\,(l<k)$. Пусть x^α имеет степень l и не принадлежит S^h. Тогда по теореме 3.4

$$h^{s-l}\min|x^\alpha - v^h|_s \to \text{const} \neq 0.$$

Следовательно, порядок аппроксимации для x^α равен только $l-s$, а не $k-s$, и следствие доказано. Ясно, что эта теорема,

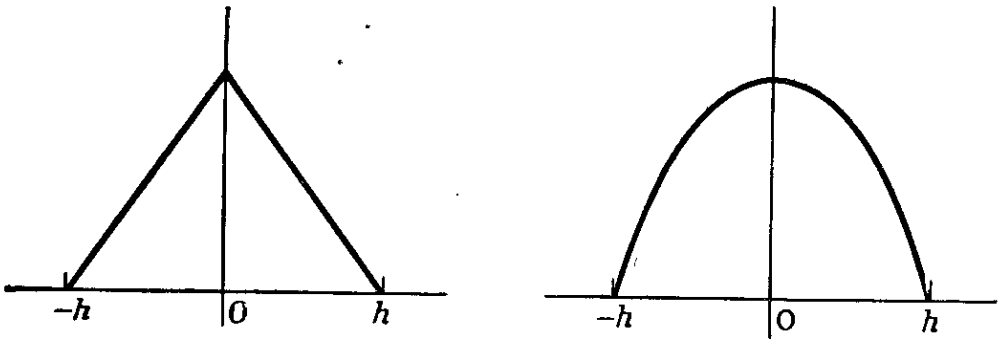


Рис. 3.5.
Функции неполиномиального типа: *а* — функция-крышка, *б* — функция Cos.

дающая точное условие сходимости, намного сильнее, чем ее следствие. Первая применяется для всех u, в то время как для последнего необходимо исследовать аппроксимацию полиномов наименьшей степени, не принадлежащих пробному пространству.

Вывод довольно интересен: все зависит от присутствия полиномов. Это значит, что *кусочно полиномиальные функции являются наилучшими пробными функциями не только из-за их удобства, но также из-за их аппроксимирующих свойств.* С самого начала метод конечных элементов работал с подпространствами оптимального типа.

Результат следствия можно доказать непосредственно [С12]; учитывая его важность, докажем его в простейшем случае. Мы покажем, что для аппроксимации порядка h пробному пространству должна принадлежать постоянная функция 1. Допустим, например, что функция-крышка заменена косинусом (рис. 3.5).

Пробное пространство содержит все комбинации

$$v^h = \sum q_j \operatorname{Cos}[\pi(x - jh)/2h];$$

запись Cos означает, что вне центральной дуги $|\theta| \leqslant \pi/2$ косинус продолжен нулем. В пределе при $h \to 0$ могла бы проявиться какая-нибудь разница по сравнению с кусочно линейной аппроксимацией.

Предположим, однако, что мы пытаемся аппроксимировать функцию $u \equiv 1$ на единичном интервале. В линейном случае эта функция принадлежит пробному пространству, а в нашем случае — нет, и функция ошибки E^h периодична с периодом h. Забудем о краевых условиях или лучше предположим, что они периодические, поскольку в любом случае они оказывают второстепенное влияние на аппроксимацию внутри интервала. Таким образом, ошибка равна

$$\min_{S^h} \| 1 - v^h \|_0^2 = \int_0^1 (E^h(x))^2 \, dx.$$

Этот интеграл охватывает $1/h$ периодов и на каждом из них равен $K_0 h$, где постоянная K_0 не зависит от h. Если изменить h, то функция ошибки (как на рис. 3.3) просто будет в новом масштабе в соответствии с новым периодом. Следовательно, ошибка аппроксимации равна постоянной K_0 (как и предсказано в теореме 3.5) и *не убывает с* h. Пространство Косинусов не имеет никаких аппроксимирующих свойств.

Конечно, это неблагоприятное заключение не относится к обыкновенным косинусам, которые числятся среди наиболее ценных пробных функций в методе Ритца. В некотором смысле они имеют бесконечную точность, $k = \infty$, поскольку пользуются любой дополнительной степенью гладкости аппроксимируемого решения u. Так как каждый косинус не равен нулю на всем интервале, этот случай не охватывается теорией метода конечных элементов, и условие обязательной для успешной аппроксимации принадлежности полиномов пробному пространству больше не имеет силы.

Закончим этот раздел несколькими историческими замечаниями о случае равномерной сетки, другими словами, о теории аппроксимации в абстрактном методе конечных элементов. Естественно подойти к проблеме, используя преобразование Фурье. Среднеквадратичные нормы $\int\int |D^\alpha u|^2 \, dx$ можно превратить с помощью формулы Парсеваля в $\int\int |\xi^\alpha \hat{u}|^2 \, d\xi$, а условие, что функция φ порождает полиномы, дает нули в ее преобразовании Фурье. Функция-крышка, например, порождает все линейные полиномы, ее преобразованием будет $\hat{\varphi}(\xi) = (\sin(\xi/2)/\xi/2)^2$

с нулями порядка $k = 2$ во всех точках $\xi = \pm 2\pi, \pm 4\pi, \ldots$. Для B-сплайнов произвольной степени $k-1$ показатель в $\hat{\varphi}$ становится просто k; это результат свертки *функции-ящика* с самой собой k раз. Связь между полиномами от x степени $k-1$ и нулями порядка k в точках $\xi = 2n\pi$ была обнаружена Шёнбергом в его первой статье по сплайнам [Ш2] и три раза переоткрывалась в литературе по методу конечных элементов [Г4], [Б3], [Ф9].

В статьях [С7, С12] и в книге Обэна [16] основательно изучается анализ Фурье абстрактного метода конечных элементов в n-мерном случае. Сформулированное выше следствие о том, что пространство S^h на равномерной сетке должно иметь степень $k-1$ для достижения аппроксимации порядка h^k, было сначала доказано методом Фурье, причем вместе с существованием суперфункции ψ, обсуждаемой в разд. 3.1. С большим сожалением мы признаем, что анализ Фурье не сможем изложить подробно; для нас реально лишь выбрать результаты, единственные в своем роде для равномерной сетки и в то же время важные для общей теории: асимптотическая теорема 3.5 и ее следствие, конечно-разностный аспект системы $KQ = F$, описанный в разд. 3.4, и обсуждение числа обусловленности в гл. 5.

Теперь вернемся к главному результату этого раздела: на элементарной области e_i разность между функцией u и ее интерполянтом u_I удовлетворяет неравенству

$$|u - u_I|_{m, e_i} \leqslant C_m h_i^{k-m} |u|_{k, e_i}. \tag{16}$$

Что происходит, если в решении u есть особенность, препятствующая его принадлежности пространству $\mathscr{H}^k$? На равномерной сетке порядок сходимости определенно будет понижен. Если u обладает только r производными, суммируемыми в квадрате, то ошибка в энергии будет убывать, как $h^{2(r-m)}$, а не как $h^{2(k-m)}$. И поточечная ошибка в напряжениях будет явно хуже. Вопрос заключается в следующем: можно ли достичь какого-нибудь улучшения «градуировкой» сетки, т. е. изменением шага h для измельчения сетки около особенности?

Когда неудобно вводить специальные сингулярные пробные функции, существует один полезный кустарный способ, предоставляемый формулой (16): *градуировка должна быть такой, чтобы величины* $h_i^{k-m} |u|_{k, e_i}$ *были примерно равны на двух соседних элементах.* В одномерном случае для особенности x^α в начале координат это означает, что функция $h^{k-m+1/2} x^{\alpha-k}$ должна быть примерно постоянной; чем больше x, тем больше может быть $h = \Delta x$. Оказывается, что это правило имеет замечательное следствие: *подходящей градуировкой сетки можно достичь одинакового порядка точности для сингулярного и для регулярного решений* u. Другими словами, пусть для неравномерной n-мер-

ной сетки с N элементами средний размер сетки $\bar{h}$ вычислен по формуле $N\bar{h}^n = \operatorname{mes}\Omega$. Тогда правильная градуировка может дать $|u - u_I|_{m,\Omega} = O(\bar{h}^{k-m})$ даже для сингулярной функции u с нормой $|u|_{k,\Omega} = \infty$.

3.3. КРИВОЛИНЕЙНЫЕ ЭЛЕМЕНТЫ И ИЗОПАРАМЕТРИЧЕСКИЕ ПРЕОБРАЗОВАНИЯ

Основная идея проста. Предположим, что мы собираемся использовать обычный полиномиальный элемент, например один из тех, что определены на прямоугольниках или треугольниках в разд. 1.8. Предположим также, что области, на которые разбивается Ω, неподходящей формы, т. е. могут иметь одну и более криволинейных сторон или быть непрямоугольными четырехугольниками. Выбирая новую систему координат ξ, η, можно привести элементы к правильной форме. Матрицы жесткости элементов тогда вычисляются интегрированием в новых переменных на треугольниках или прямоугольниках, а минимизация приводит к решению $u^h(\xi, \eta)$ метода конечных элементов, которое можно преобразовать обратно в переменные x и y [1]).

Подчеркнем несколько важных моментов. Во-первых, так как типичный интеграл по двумерному элементу преобразуется по формуле

$$\iint_{e_l} p(x, y)(v_x)^2\,dx\,dy \to$$

$$\to \iint_{E_l} p(x(\xi, \eta), y(\xi, \eta))(v_\xi \xi_x + v_\eta \eta_x)^2 J(\xi, \eta)\,d\xi\,d\eta, \qquad (17)$$

то преобразование координат и его производные должны вычисляться легко. Далее, преобразование координат не должно чрезмерно искажать элемент, иначе якобиан $J = x_\xi y_\eta - x_\eta y_\xi$ может обратиться в нуль внутри области интегрирования; это может произойти удивительно легко. Чрезмерное искажение также разрушит точность, заложенную в полиномиальный элемент. Полиномы в новых переменных не соответствуют полиномам в старых переменных, и для сохранения теории аппроксимации требуется, чтобы преобразование координат было равномерно гладким. Наконец, для того чтобы согласованные элементы в переменных ξ, η были согласованными в переменных x, y, должно выполняться *глобальное* условие непрерывности для преобразования координат: если энергия содержит m-е производные, то преобразование координат должно быть класса $\mathscr{C}^{m-1}$ между элементами. Пока

---

[1]) Изопараметрическая техника так же важна для трехмерного случая. Ее проще продемонстрировать на примерах в плоскости, но теоретически никакой разницы иет.

мы обсудим только случай $m = 1$, возникающий из дифференциальных уравнений второго порядка, где отображение должно быть непрерывным между элементами: точка, общая для e_i и e_j, не должна при $e_i \to E_i$, $e_j \to E_j$ распадаться на две отдельные точки.

Как удовлетворить эти условия, особенно требование легкой вычислимости? *Изопараметрическая техника состоит в выборе кусочно полиномиальных функций для определения преобразования координат* $x(\xi, \eta)$ *и* $y(\xi, \eta)$. Строго говоря, термин *изопараметрический* означает, что для преобразования координат выбираются *такие же полиномиальные элементы*, как и для самих пробных функций; термин *субпараметрический* означает, что

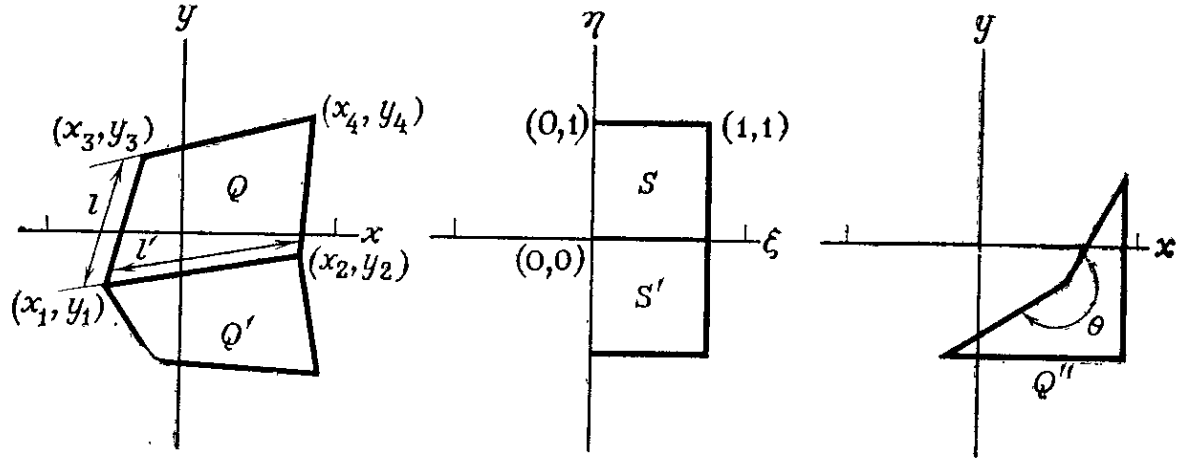


Рис. 3.6.
Изопараметрические отображения в четырехугольники.

берется подмножество полиномов меньшей степени. В любом случае мы требуем непрерывности между элементами и невырожденности матрицы Якоби.

Основной изопараметрический пример состоит в билинейном преобразовании квадрата в четырехугольник (рис. 3.6). Преобразование координат из S в Q осуществляется по формулам

$$\begin{aligned} x(\xi, \eta) &= x_1 + (x_2 - x_1)\,\xi + (x_3 - x_1)\,\eta + (x_4 - x_3 - x_2 + x_1)\,\xi\eta, \\ y(\xi, \eta) &= y_1 + (y_2 - y_1)\,\xi + (y_3 - y_1)\,\eta + (y_4 - y_3 - y_2 + y_1)\,\xi\eta; \end{aligned} \tag{18}$$

каждая сторона квадрата S линейно переходит в соответствующую сторону четырехугольника Q. Например, если $\eta = 0$, а ξ изменяется от 0 до 1, то точка (x, y) движется линейно от одного угла (x_1, y_1) к другому (x_2, y_2). Для этой границы положение других вершин (x_3, y_3) и (x_4, y_4) абсолютно безразлично. Это гарантирует согласованность в переменных x и y для любого согласованного элемента в переменных ξ и η (скажем, билинейного или бикубического эрмитова элемента, для которого преобразование изопараметрическое или субпараметрическое соответственно).

Надо проверить, что преобразование (18) обратимо, другими словами, что каждая точка (x, y) в Q соответствует одной и только одной паре (ξ, η) в S. Разрешая уравнения (18) относительно ξ и η через x и y, мы получили бы сложные квадратные корни, и это ничего бы нам не дало. Поэтому, так как уже проверено соответствие границ для S и Q, мы только покажем, что внутри S якобиан нигде не обращается в нуль:

$$J = x_\xi y_\eta - x_\eta y_\xi = \begin{vmatrix} x_2 - x_1 + A\eta & x_3 - x_1 + A\xi \\ y_2 - y_1 + B\eta & y_3 - y_1 + B\xi \end{vmatrix},$$

где $A = x_4 - x_3 - x_2 + x_1$, а $B = y_4 - y_3 - y_2 + y_1$. Якобиан в действительности *линеен*, а не билинеен, поскольку коэффициент при $\xi\eta$ в этом (2×2)-определителе равен нулю: $AB - AB = 0$. Следовательно, *если J имеет один и тот же знак во всех четырех углах квадрата S, он не может равняться нулю внутри S.* В угле $\xi = 0$, $\eta = 0$ якобиан равен

$$J(0, 0) = (x_2 - x_1)(y_3 - y_1) - (y_2 - y_1)(x_3 - x_1),$$

что в свою очередь равно $ll' \sin\theta$; стороны l, l' и угол θ между ними изображены на рис. 3.6. Поэтому $J > 0$ в этом угле, если внутренний угол θ меньше π. Это справедливо и для любого другого угла. Следовательно, якобиан J *всюду отличен от нуля тогда и только тогда, когда четырехугольник Q выпуклый*, т. е. все его углы должны быть меньше π. В противном случае, как для Q'' на рис. 3.6, J изменит знак где-нибудь внутри S. Тогда преобразование координат будет незаконным.

Заметим, что, хотя полиномы от x и y не переходят, вообще говоря, в полиномы от ξ и η, *линейные полиномы* 1, x, y *представляют исключение.* Само преобразование координат выражает x и y как билинейные функции от ξ и η, и, конечно, постоянная функция остается постоянной. Если эти три полинома принадлежат пробному пространству, то выполнение условия сходимости гарантировано, т. е. все решения $u = \alpha + \beta x + \gamma y$ с постоянной деформацией точно воспроизводятся в S^h. *Это всегда верно для изопараметрических преобразований* и сходимость обеспечена. Субпараметрический случай еще лучше: если пробное пространство содержит все биквадратичные или бикубические функции от ξ и η, то оно содержит и все биквадратичные или бикубические функции от x и y, а S^h имеет степень 2 или 3 соответственно. Поэтому в предположении, что углы четырехугольника Q заключены строго между 0 и π, аппроксимация в полной мере возможна, а ошибка в деформациях равна $O(h^{k-1})$, как и должно быть.

Для треугольников наиболее важен пример треугольника с одной криволинейной стороной, лежащей на границе Γ

(рис. 3.7). Простейшая из возможных кривых — кривая второй степени, и естественный выбор элементов — квадратичные функции. На треугольнике в плоскости ξ, η пробная функция

$$v^h = a_1 + a_2\xi + a_3\eta + a_4\xi^2 + a_5\xi\eta + a_6\eta^2$$

определяется своими значениями в шести узлах фигуры: в трех вершинах и в серединах трех сторон.

Представим себе два отображения исходного криволинейного треугольника в плоскости x, y. Сначала простое линейное преобразование нормализует криволинейный треугольник, переводя две прямые стороны в координатные оси на плоскости x', y'.

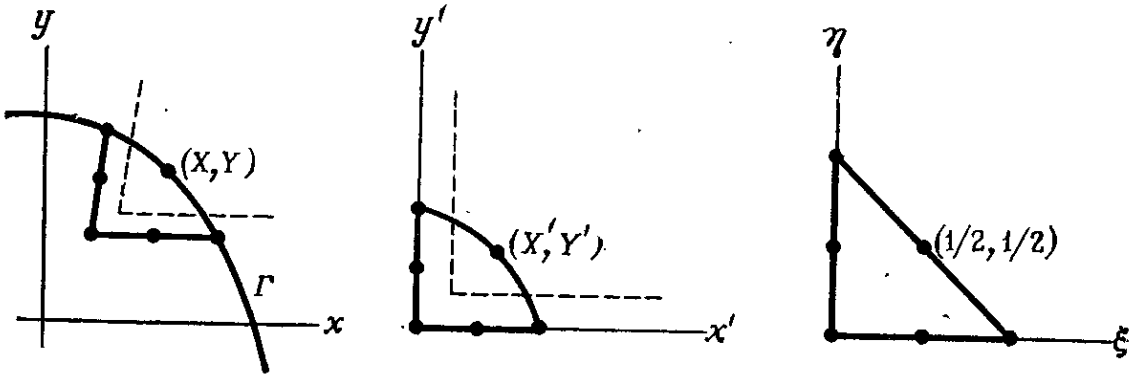


Рис. 3.7.
Отображения криволиниейного треугольника.

Якобиан этого преобразования постоянен, так что этот шаг предпринимается для удобства. А вот следующий шаг важен: надо связать координаты x', y' с ξ, η так, чтобы заданная граничная точка (X', Y') переходила в среднюю точку $(1/2, 1/2)$. На практике (X, Y) и, следовательно, (X', Y') часто выбираются на середине дуги, но это необязательно. Отображение билинейно:

$$\begin{aligned} x' &= \xi + (4X' - 2)\,\xi\eta, \\ y' &= \eta + (4Y' - 2)\,\xi\eta. \end{aligned} \tag{19}$$

Легко проверить, что прямые стороны сохраняются и непрерывность при переходе в соседние элементы обеспечивается. Якобиан имеет вид

$$J = \begin{vmatrix} 1 + (4X' - 2)\,\eta & (4X' - 2)\,\xi \\ (4Y' - 2)\,\eta & 1 + (4Y' - 2)\,\xi \end{vmatrix} =$$
$$= 1 + (4X' - 2)\,\eta + (4Y' - 2)\,\xi.$$

Снова он линеен; он равен 1 при $\xi = 0$, $\eta = 0$ и всюду в треугольнике отличен от нуля тогда и только тогда, когда он поло-

жителен в двух других вершинах. Это условие, впервые высказанное нам Митчеллом, есть просто

$$4X' - 2 > -1, \quad \text{или} \quad X' > 1/4, \quad \text{в точке} \quad (0, 1),$$
$$4Y' - 2 > -1, \quad \text{или} \quad Y' > 1/4, \quad \text{в точке} \quad (1, 0).$$

Следовательно, (X', Y') может лежать где-нибудь в квадранте, образованном штриховыми линиями, и тогда (X, Y) — в отмеченном секторе (рис. 3.7). Заметим, что даже для исходного треугольника с прямыми сторонами точка (X, Y) должна лежать в средней части ее стороны, иначе сдвиг ее внутрь может привести к обращению якобиана в нуль. (Конечно, в этом случае нет причин ее сдвигать: на треугольнике с прямыми сторонами можно было бы взять квадратичные элементы в переменных x, y даже с произвольно расположенными средними узлами. Отображение в плоскость ξ, η действительно предназначается для случая, когда надо выпрямить криволинейные стороны.)

В этом примере криволинейная сторона была параболой. В общем изопараметрическом случае как с треугольниками, так и с прямоугольниками отображения $x(\xi, \eta)$, $y(\xi, \eta)$ задаются тем же типом полиномиальных элементов, что и для перемещений, а все стороны могут быть полиномами степени $k - 1$. Ограничения те же, что и на сами элементы, т. е. когда неизвестные содержат несколько производных в узле, это означает, что соответствующие производные граничных кривых должны быть непрерывны в узлах. Случай Лагранжа поэтому будет простейшим для изопараметрических преобразований, так как неизвестны только значения функции, а единственное ограничение — непрерывность между элементами, необходимая в любом случае. В самом деле, все особенно просто, если, как в сирендиповом прямоугольном элементе на рис. 3.8, нет внутренних узлов. Отображение между границами тогда полностью определяет преобразование координат, которое в противном случае *очень чувствительно к передвижению внутренних узлов.*

Подчеркнем, что вся изопараметрическая техника основана на применении *численного интегрирования* (в переменных ξ, η) для вычисления элементов матриц K и F. Из выбора переменных в интеграле (17) по элементарной области видно, что даже для изотропного материала ($p = \text{const}$) математический эквивалент переменных свойств материала выражается функциями ξ_x, η_y и $J(\xi, \eta)$. Вообще говоря, две первые функции рациональны, а последняя — полином, гладкость которого зависит от искажения элементарной области. В разд. 4.3 мы установим влияние ошибок численного интегрирования на окончательный результат и требуемый порядок точности.

Здесь мы рассматриваем вопрос аппроксимации: насколько можно приблизить изопараметрическими элементами истинное решение $u(x, y)$? Ответ должен зависеть от величины производных в преобразовании координат F:

$$F(\xi, \eta) = (x(\xi, \eta), y(\xi, \eta)).$$

Пусть $\hat{u}(\xi, \eta)$ обозначает решение $u(x, y)$, преобразованное к новым координатам. Тогда если степень пространства S^h равна

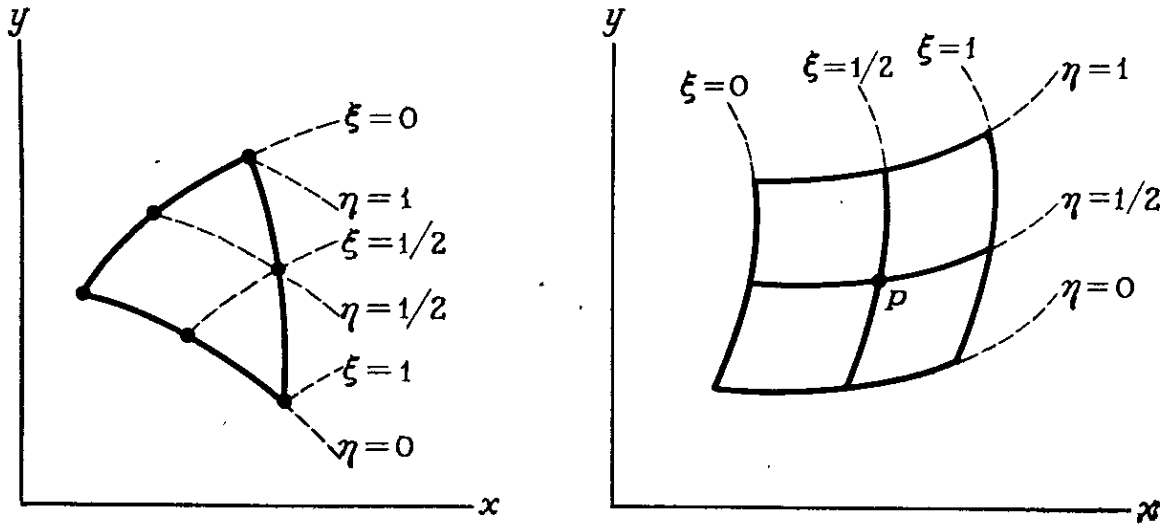


Рис. 3.8.
Обычные изопараметрические элементы.

$k-1$ в переменных ξ и η, то интерполянт $\hat{u}_I$ удовлетворяет неравенству

$$\sum_{|\alpha| \leqslant s} \int_{E_i} |D^\alpha \hat{u} - D^\alpha \hat{u}_I|^2 \, d\xi \, d\eta \leqslant C^2 h_i^{2(k-s)} \sum_{|\beta| \leqslant k} \int_{E_i} |D^\beta \hat{u}|^2 \, d\xi \, d\eta. \quad (20)$$

При замене на переменные x, y каждая производная от $\hat{u}$ переходит в сумму:

$$\frac{\partial}{\partial \xi} \hat{u} \to u_x x_\xi + u_y y_\xi,$$

$$\frac{\partial^2}{\partial \xi^2} \hat{u} \to u_x x_{\xi\xi} + u_y y_{\xi\xi} + u_{xx} x_\xi^2 + u_{yy} y_\xi^2.$$

Вообще производная порядка $|\beta| \leqslant k$ должна быть ограничена:

$$|D^\beta \hat{u}(\xi, \eta)| \leqslant \|F\|_k \sum_{|\gamma|=1}^{k} |D^\gamma u(x, y)|,$$

где постоянная $\|F\|_k$ вычисляется из степеней производных x_ξ, y_ξ, x_η, ... внутри элементарных областей вплоть до порядка k. Сиарле и Равьяр [C5], рассуждения которых мы излагаем далее,

подробно расписывают это выражение. Обратная величина якобиана J также входит в преобразование интеграла:

$$\int_{E_i} |D^\beta \hat{u}|^2 \, d\xi \, d\eta \leqslant \frac{\|F\|_k^2}{\min\limits_{E_i} |J(\xi, \eta)|} \int_{e_i} \sum_1^k |D^\nu u|^2 \, dx \, dy.$$

Неравенство справедливо для всех $|\beta| \leqslant k$, так что

$$\|\hat{u}\|_{k, E_i}^2 \leqslant \frac{\|F\|_k^2}{\min |J|} \|u\|_{k, e_i}^2. \tag{21}$$

Это дает верхнюю границу для правой части в (20).

Для левой части нужна нижняя граница; поэтому обратим рассуждения, которые привели нас к (21):

$$\|u - u_I\|_{s, e_i}^2 \leqslant \|F^{-1}\|_s^2 \max_{E_i} |J| \, \|\hat{u} - \hat{u}_I\|_{s, E_i}^2. \tag{22}$$

На этот раз множитель $\|F^{-1}\|_s$ зависит от степеней производных вплоть до порядка s *обратного* отображения F^{-1}, переводящего переменные x, y в ξ, η. Если якобиан J не равен нулю, как мы предположили, то производные от F^{-1} можно выразить через производные от F. Для первых производных справедливо тождество матриц Якоби

$$\begin{pmatrix} \xi_x & \xi_y \\ \eta_x & \eta_y \end{pmatrix} = \begin{pmatrix} x_\xi & x_\eta \\ y_\xi & y_\eta \end{pmatrix}^{-1}.$$

Для производных более высокого порядка это тождество дифференцируют, и основной вопрос снова состоит в оценке производных от F.

Подытожим результаты. Подстановка (21) и (22) превращает исходное неравенство $\|\hat{u} - \hat{u}_I\|_s \leqslant C h_i^{k-s} \|\hat{u}\|_k$ на элементарной области E_i в аналогичное неравенство на e_i:

$$\|u - u_I\|_{s, e_i} \leqslant C' h_i^{k-s} \|u\|_{k, e_i}, \tag{23}$$

где

$$C' = C \left(\frac{\max |J|}{\min |J|} \right)^{1/2} \|F\|_k \|F^{-1}\|_s.$$

Это основной результат. Порядок аппроксимации для изопараметрических функций тот же, что и для обычных полиномов, *при условии, что постоянная C' остается ограниченной.* Заметим, что h_i — диаметр элементарной области на плоскости ξ, η, даже хотя аппроксимационное неравенство (23) выполняется для пе-

ременных x, y. Можно, однако, предположить, что диаметры одинаковы в этих двух системах координат. (Изменение масштаба для ξ, η не меняет неравенство, как и должно быть: если h_i переходит в αh_i, то нормы для F и F^{-1} умножаются на α^{-k} и α^s соответственно.) Эта нормализация удобна, поскольку для равных диаметров задача изопараметрической аппроксимации сводится как раз к оценке функции F и ее производных и к вопросу о необращении в нуль якобиана J, равномерно при $h \to 0$.

Для четырехугольников существует важная модификация. Она необходима даже для преобразования, определяемого равенствами

$$x(\xi, \eta) = \xi, \qquad y(\xi, \eta) = \eta - \frac{\xi\eta}{2h}.$$

Это преобразование переводит три точки $(0, 0)$, $(h, 0)$ и $(0, h)$ в них самих, а четвертый угол квадрата переходит в новую точку $(h, h/2)$. Другими словами, это совершенно типичное отображение, преобразующее квадрат на плоскости ξ, η в четырехугольник, сравнимый с ним формой и размерами. Тем не менее смешанная производная $y_{\xi\eta}$, входящая в норму $\|F\|_k$ в аппроксимационном неравенстве, имеет порядок $1/h$. Это нарушит порядок аппроксимации. Сиарле и Равьяру удалось, однако, воспользоваться присутствием члена кручения $\xi\eta$ в пробном пространстве: билинейная интерполяция воспроизводит его точно (в дополнение к линейным членам $a_1 + a_2\xi + a_3\eta$ в случае треугольника). Результат этих авторов для четырехугольников состоит в том, что смешанные производные $x_{\xi\eta}$ и $y_{\xi\eta}$ не появляются в множителе $\|F\|_k$ и можно достичь ожидаемого порядка аппроксимации. Аналогичное заключение верно для биквадратичных и бикубических функций.

Следствия неравенства (23) сформулируем в виде теоремы.

Теорема 3.6. *Предположим, что пробное пространство на плоскости ξ, η имеет степень $k-1$, а преобразования F элементарных областей в плоскость x, y равномерны при $h \to 0$. Тогда если $u \in \mathcal{H}^k(\Omega)$, то интерполянт $u_I \in S^h$ удовлетворяет неравенствам*

$$\left(\int_\Omega |u - u_I|^2 \, dx\, dy\right)^{1/2} \leqslant C' h^k \|u\|_k,$$

$$\left(\int_\Omega |\operatorname{grad}(u - u_I)|^2 \, dx\, dy\right)^{1/2} \leqslant C' h^{k-1} \|u\|_k.$$

Постоянная C' та же, что в (23). Следовательно, скорость сходимости решения u^h метода конечных элементов для дифферен-

циального уравнения второго порядка равна $h^{2(k-1)}$ для энергии деформации.

Неравенства теоремы относятся только к функциям и их первым производным, т. е. к случаям $s = 0$ и $s = 1$ в (23). Аппроксимация порядка h^{k-s} выполняется также для производных высших порядков внутри каждой элементарной области, но когда элементы объединены, интерполянт u_I не более чем непрерывен и принадлежит лишь $\mathscr{H}^1$ на всей области Ω.

Требуемая в теореме равномерность — условие довольно суровое и заключается по существу в следующем:

1. Якобианы должны быть строго больше нуля, так что все углы должны быть строго между 0 и π.
2. Стороны элементов на плоскости x, y должны быть полиномами с равномерно ограниченными производными, т. е. кривизны и производные высших порядков вдоль сторон должны оставаться ограниченными. В частности, стороны могут отклоняться от прямых лишь на величину $O(h^2)$. Тогда для удачно выбранного отображения F будет $\|F\|_k \leqslant \text{const}$.

Вычисления Фрида [Ф13] демонстрируют необходимость условия 2. Он увеличивал кривизну сторон, пока их отклонение от исходного квадрата не стало того же порядка, что и размеры самого квадрата. На единичном квадрате это еще означает ограниченность кривизны, но изменение масштаба до размера h делает кривизну (т. е. вторую производную) величиной порядка $1/h$. Численные результаты были соответственно бедны. Элементы, рассматриваемые в практических задачах, занимают промежуточное положение между этими чрезмерно искаженными элементами и элементами, близкими к прямолинейным, требуемым по предположению равномерности в теореме 3.6.

С другой (счастливой) стороны, предположим, что криволинейный элемент расположен на границе области Ω, т. е. одна его криволинейная сторона принадлежит истинной границе Γ. (Обычно полиномиальная аппроксимация Γ и состоит в интерполировании истинной границы в указанных узлах.) Тогда *для гладкой границы Γ условие равномерности на этой стороне выполняется автоматически.* Интерполированная полиномами сторона будет отклоняться лишь на $O(h^2)$ от прямой, и кривизны будут оценены через кривизну границы Γ. Следовательно, *изопараметрическая техника для уравнения второго порядка позволяет работать с главными краевыми условиями без потерь в точности и простоте по сравнению с естественными краевыми условиями.* То же относится к внутренней границе.

Улучшение, которого можно достичь этим приемом, огромно по сравнению с аппроксимацией границы многоугольником.

Зламал [37] в экспериментах с кубическими треугольными элементами, например, обнаружил разницу в порядке величины перемещений, а еще больше в деформациях. Без сомнения, эта техника успешно применяется к задачам второго порядка [1]).

Для уравнений четвертого порядка (например, задачи о пластине и об оболочке) ситуация гораздо менее удачна. По теории преобразование координат обязательно должно быть класса $\mathscr{C}^1$: его первые производные должны быть непрерывны между элементами, иначе пробные функции будут несогласованными. (Сходимость для несогласованных элементов все еще возможна, как

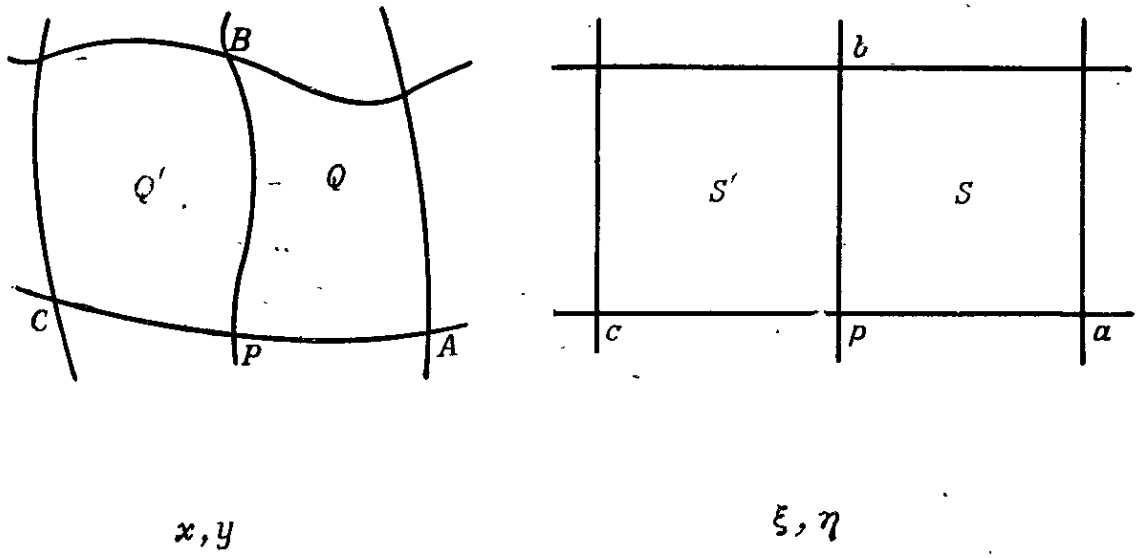


Рис. 3.9.
Ограничения на отображение класса $\mathscr{C}^1$.

мы докажем в следующей главе. Однако для элементов в задаче о пластине даже эта надежда рушится, потому что они не могут выдержать *кусочное тестирование*.) Преобразование координат класса $\mathscr{C}^1$ теоретически возможно, но требование дополнительной непрерывности чрезвычайно обременительно (рис. 3.9). Как только в точке известны два направления (в точке P касательные к PA и PB устанавливаются, когда S отображается в Q), все другие направления полностью определяются. (Для любой функции $f(x, y)$ класса $\mathscr{C}^1$ все производные по направлению можно вычислить из градиента, т. е. из производных f_x и f_y в двух направлениях.) На рис. 3.9 это значит, что касательная к PC определена. В самом деле, кривая CPA должна иметь непрерывную касательную в точке P, так как cpa — отрезок

[1]) Гордон и Холл недавно предложили отображать в квадрат сразу всю область Ω, а не поэлементно. Для этого они используют *составные функции*, вариант обычных конечных элементов. Если сама область Ω не слишком отличается от квадрата (отображение круговой области создаст искусственные особенности в углах), это даст выигрыш во времени по сравнению с изопараметрическим методом на каждом элементе.

прямой. Отсюда следует, что *общую четырехугольную или треугольную сетку (даже прямолинейную) нельзя перевести в равномерную сетку преобразованием координат класса* $\mathscr{C}^1$. Отображение на рис. 3.9 можно осуществить бикубическими эрмитовыми элементами, например допускающими кубические криволинейные стороны. Но касательные к сторонам, а в случае эрмитовых полиномов еще и смешанные производные $x_{\xi\eta}$ и $y_{\xi\eta}$ должны быть непрерывны в вершинах. Это суровое и почти неприемлемое теоретическое ограничение на изопараметрическую технику для задач четвертого порядка.

По-видимому, это знак того, что не стоит работать с уравнениями четвертого порядка вместо систем второго порядка. Аналитически исключение неизвестных может иметь огромное значение, но не численно. Конструкция полиномов класса $\mathscr{C}^1$ на элементах с прямыми границами уже трудна, а на криволинейных элементарных областях (особенно, когда мы приступаем к численному интегрированию) — бесконечно хуже.

3.4. ОЦЕНКИ ОШИБОК

В этом разделе мы применим предыдущие теоремы об аппроксимации для достижения главной цели всей нашей теории: нахождение оценки ошибки $u - u^h$ метода конечных элементов. Функция u служит решением n-мерной эллиптической краевой задачи порядка $2m$, а u^h — ее приближением Ритца, вычисленным в пространстве метода конечных элементов S^h. На равномерной сетке уравнения метода конечных элементов $KQ = F$ становятся системой разностных уравнений, и мы находим одновременно порядок точности этих разностных уравнений.

Основной вопрос заключается в том, насколько хорошо подпространства S^h приближают все допустимое пространство $\mathscr{H}^m$. В энергетической норме ничего другого и не происходит: u^h близко к u насколько возможно, а ошибка в энергии должна быть оптимального порядка $h^{2(k-m)}$:

$$a(u - u^h, u - u^h) \leqslant C^2 h^{2(k-m)} \| u \|_k^2. \tag{24}$$

Это простейшая и тем не менее основная оценка ошибки. Так как выражение $a(v, v)$ для энергии деформации положительно определено (другими словами, задача эллиптична: $a(v, v) \geqslant \geqslant \sigma \| v \|_m^2$), то неравенство (24) равносильно неравенству

$$\| u - u^h \|_m \leqslant c' h^{k-m} \| u \|_k. \tag{24'}$$

Для перемещения, или вообще для s-x производных, сразу же возникает вопрос: имеет ли $u - u^h$ снова оптимальный порядок h^{k-s}? Тогда бы все определялось степенью $k - 1$ конечных

элементов, но это не совсем правильно. Если зафиксировать S^h и увеличить порядок задачи $2m$, то в итоге пробные функции более не будут допустимыми, и метод Ритца потерпит неудачу. Поэтому порядок точности должен как-то зависеть от m, а не только от k и s.

Правильный порядок можно вычислить с помощью изящного вариационного доказательства, принадлежащего Обэну и Нитше. Доказательство известно как *прием Нитше*; оно было расширено Шульцем и теперь представляет собой стандартный подход к ошибкам в перемещениях. В разд. 1.6 была установлена ошибка h^2 для линейных элементов; общий случай сложнее, но результаты (25)—(26) очень просты. Верхний предел $2(k-m)$ скорости сходимости получен первым автором [C12].

Теорема 3.7. *Предположим, что степень пространства метода конечных элементов S^h равна $k-1$, а энергия деформации имеет гладкие коэффициенты и удовлетворяет условию эллиптичности $\sigma\|u\|_m^2 \leqslant a(v, v) \leqslant K\|v\|_m^2$. Тогда приближение метода конечных элементов u^h отличается от истинного решения и на величину*

$$\|u-u^h\|_s \leqslant Ch^{k-s}\|u\|_k, \qquad s \geqslant 2m-k, \tag{25}$$

$$\|u-u^h\|_s \leqslant Ch^{2(k-m)}\|u\|_k, \qquad s \leqslant 2m-k. \tag{26}$$

Эти показатели степени h оптимальны, так что порядок точности никогда не превышает $2(k-m)$ в любой норме; почти во всех реальных случаях порядок равен $k-s$.

Доказательство. Прием Нитше состоит во введении *вспомогательной задачи* $Lw = g$, вариационная форма которой (уравнение виртуальной работы) есть

$$a(w, v) = (g, v) \quad \text{для всех} \quad v \in \mathscr{H}_E^m. \tag{27}$$

Из теории уравнений в частных производных известно, что существует единственное решение w с производными на $2m$ порядков больше, чем у правой части f: $\|w\|_{2m-s} \leqslant c\|g\|_{-s}$.

Возьмем $v = u - u^h$ в (27):

$$|(g, u-u^h)| = |a(w, u-u^h)| = |a(w-v^h, u-u^h)| \leqslant$$
$$\leqslant K\|w-v^h\|_m\|u-u^h\|_m. \tag{28}$$

Неравенство справедливо для любой функции $v^h \in S^h$, так как $a(v^h, u-u^h) = 0$ по основной теореме Ритца 1.1. Теперь предположим, что v^h — ближайшее приближение к w в норме пространства $\mathscr{H}^m$. (Или, что по существу то же самое, пусть v^h — решение вспомогательной задачи (27) методом конечных элементов и потому наилучшее приближение к w по энергии. Заме-

тим, что в доказательстве используется аппроксимация только по энергии и никогда — прямая аппроксимация в норме пространства $\mathscr{H}^s$.) В соответствии с теоремой об аппроксимации 3.3

$$\| w - v^h \|_m \leqslant \begin{cases} ch^{m-s} \| w \|_{2m-s}, & \text{если} \quad k \geqslant 2m - s, \\ ch^{k-m} \| w \|_k, & \text{если} \quad k \leqslant 2m - s. \end{cases} \tag{29}$$

В первом случае число k сводится к $2m - s$ до применения теоремы об аппроксимации; если подпространство полно для степени $k - 1$, то оно, несомненно, полно для любой меньшей степени.

Подставляя (24′) и (29) в (28), получаем

$$|(g, u - u^h)| \leqslant Kc \left\{ \begin{matrix} h^{m-s} \\ h^{k-m} \end{matrix} \right\} \| w \|_{2m-s} c' h^{k-m} \| u \|_k \leqslant$$

$$\leqslant C \left\{ \begin{matrix} h^{k-s} \\ h^{2(k-m)} \end{matrix} \right\} \| g \|_{-s} \| u \|_k.$$

По двойственности в определении отрицательных норм (формула (58) в гл. 1)

$$\| u - u^h \|_s = \max_g \frac{|(g, u - u^h)|}{\| g \|_{-s}} \leqslant C \left\{ \begin{matrix} h^{k-s} \\ h^{2(k-m)} \end{matrix} \right\} \| u \|_k.$$

Доказательство закончено. Оно наиболее прямолинейно для перемещений, $s = 0$, поскольку в этом случае правая часть g во вспомогательной задаче есть в точности $u - u^h$; такой выбор был сделан в разд. 1.6.

Аналогичный результат верен для неоднородных главных краевых условий [C8]. Кроме того, оценки ошибок распространены на задачи *вынужденных колебаний*, в которых основное условие эллиптичности $a(v, v) \geqslant \sigma \| v \|_m^2$ нарушено добавлением нового члена нулевого порядка. Действительно, исходное дифференциальное уравнение $Lu = f$ заменяется на $Lu - \alpha u = f$: если α попадает между двумя собственными значениями оператора L, то оператор $L - \alpha$ уже положительно определен, но уравнение все же имеет решение. Шульц [Ш5] доказал, что скорость сходимости не изменяется.

Замечания. Случай $s = m$, соответствующий ошибке в энергии, всегда дает правильную степень h^{k-m}. Ошибка в производных более высоких порядков для $s > m$ не устанавливается в теореме в том виде, как она сформулирована, поскольку w не принадлежит $\mathscr{H}^m$ и (29) не имеет смысла. Тем не менее показатель $k - s$ будет правильным внутри каждого элемента, если только подпространство S^h удовлетворяет *обратной гипотезе*:

каждое дифференцирование пробной функции v^h увеличивает ее максимум самое большее в c/h раз. Эта гипотеза означает почти то же, что и условие однородности (2), за исключением того, что там был множитель c/h_i; поэтому обратная гипотеза выполняется, если все элементарные области сравнимы по размеру. Иначе в оценке ошибок для производных порядка $s > m$ появится множитель $(h_{\max}/h_{\min})^{s-m}$.

Теорема и ее доказательство переносятся без изменений на случай $s < 0$. Удивительно, но *скорость сходимости в отрицательных нормах имеет не только академический интерес.* Причина в том, что ошибка в норме $\| \ \|_{-1}$ служит границей ошибки, усредненной по области:

$$\| u - u^h \|_{-1} = \max \frac{\left| \int_\Omega (u - u^h)\, v\, dx \right|}{\| v \|_1} \geqslant \frac{\left| \int_\Omega (u - u^h)\, dx \right|}{(\operatorname{mes} \Omega)^{1/2}},$$

если положить $v \equiv 1$. Поэтому в обычном случае $k > 2m$ из теоремы 3.7 вытекает такое следствие для $s = -1$: *усредненная ошибка намного меньше обычной ошибки перемещения в точке.* Точнее,

$$\int | u - u^h |^2 dx \sim h^{2k}, \qquad \text{но} \qquad \left| \int (u - u^h)\, dx \right|^2 \sim h^{2(k+1)}.$$

Это должно означать, что *знак ошибки быстро меняется.* В действительности это происходит внутри каждого элемента, и практическая задача — отыскать хотя бы приблизительно «специальные точки», где меняется знак. Вблизи таких точек точность перемещения u^h будет особая.

(Если представить себе, что уравнение $-u'' = 1$ решается с помощью линейных элементов, это даст отличный пример: u^h совпадает с интерполянтом u_I, так что точность в узлах будет особой [1]). Однако исследуя тщательнее, получим, что усреднен-

---

[1]) Явление «сверхсходимости» в узлах недавно прояснилось с помощью изящного рассуждения Дюпона и Дугласа. Пусть $G_0(x)$ — фундаментальное решение, соответствующее точке x_0, т. е. G_0 — реакция на точечную нагрузку $f_0 = \delta(x - x_0)$. Тогда

$$| u(x_0) - u^h(x_0) | \leqslant C \| u - u^h \|_m \| G_0 - v^h \|_m \quad \text{для всех} \quad v^h \in S^h.$$

Доказательство. $u(x_0) - u^h(x_0) = (u - u^h, f_0) = a(u - u^h, G_0) = a(u - u^h, G_0 - v^h)$. Наиболее интересен случай, когда x_0 — узел, так как аппроксимация решения G_0, вероятно, наиболее удачна. В одномерном случае для уравнения $-u'' = f$ решение G_0 линейно изменяется с переменой угла наклона в x_0 и его можно точно воспроизвести функцией v^h. Это подтверждает неограниченную точность в этом специальном случае. Как правило, член $\| G_0 - v^h \|$ будет добавлять к степени h^{k-m}, возникающей из $\| u - u^h \|_m$, некоторую конечную степень h. В настоящее время тщательно изучается вопрос поточечной сходимости в целом и, в частности, эти увеличения степени h (сверхсходимость) в специальных точках.

ная ошибка есть величина того же порядка h^2, что и ошибка в обычной точке. В самом деле, линейный интерполянт никогда не проходит выше истинного (квадратичного) решения, так что ошибка целиком односторонняя. Объясняется это тем, что условие $k > 2m$ для появления ошибки h^{k+1} не выполнено: $k = 2m = 2$. Узлы составляют исключение потому, что пробные функции служат решениями однородного дифференциального уравнения $-u'' = 0$ [X1, T5]. Это вовсе не пример быстрых перемен в знаке.)

Напряжения находятся в аналогичном положении. Для задач второго порядка их ошибки равны h^{k-1} в обычных точках и h^k в среднем. Следовательно, эти ошибки также должны менять знак и должны существовать особые *точки напряжения*. Их наличие заметил в реальных расчетах Барлоу; для квадратичных функций на треугольниках *особыми оказываются середины сторон*. Точность в этих узлах лучше, чем в вершинах, где даже после усреднения результатов по соседним элементам приближения напряжений неудовлетворительны. Так как середины сторон служат также узлами для квадратичных элементов, то ситуация чрезвычайно благоприятна. Она может испортиться лишь ошибками от изменения области, которые необязательно меняют знак.

Мы думаем, что точки напряжений можно обнаружить следующим образом. Главный член в ошибке определяется задачей аппроксимации пробных функций из S^h полиномами P_k степени k в энергетическом смысле. На равномерной сетке эту задачу можно решить точно. Точки напряжений определяются тем свойством, что истинные напряжения (производные от P_k) совпадают с их приближениями (производными от полиномов низшей степени). В одномерном случае для элементов первой степени равенство выполняется в серединах интервалов, т. е. там, где наклон квадратичной функции равняется наклону ее линейного интерполянта. (Или, что эквивалентно, там, где функция ошибки на рис. 3.3 имеет горизонтальную касательную.) По соображениям симметрии середины должны были бы быть особыми точками также для элементов высшей степени. (Особые точки для перемещений расположены иначе, а в простейшем случае это нули второго полинома Лежандра. Они оказались чувствительнее к краевым условиям, чем точки напряжения.) Для двумерного случая результаты могут зависеть от выбора полиномов P_k и особые точки для одной компоненты напряжения необязательно будут такими для других. Середины сторон, вероятно, окажутся особыми для производных вдоль сторон, но не для напряжений в направлении нормали. Пока это объект исследования, но в конце концов эти специальные точки будут изучены и поняты полностью.

Для равномерных сеток возникают три дополнительные проблемы:

1. Интерпретировать $KQ = F$ как систему конечно-разностных уравнений с соответствующими локальными ошибками усечения.

2. Доказать оптимальность показателей степени в теореме 3.7.

3. Показать, что для гладких решений те же скорости сходимости не только в среднем, но в каждой отдельной точке.

Мы не собираемся обсуждать технические подробности, в частности, проблемы 3. Грубо говоря, раз поведение ошибок отсечения установлено, центральный вопрос — устойчивость разностного оператора K. Это свойство трудно установить в максимальной норме, но для упомянутой ранее обратной гипотезы оно почти наверняка выполняется. Мы проверили его для модельных задач с постоянными коэффициентами, ошибка в каждой точке области оказалась правильного порядка. Более точные результаты получили Нитше, Брамбл и Сиарле с Равьяром, но общая задача не решена.

Для проблемы 1 предположим сначала, что на каждом квадрате сетки только один узел (и одно связанное с ним неизвестное); это случай линейных элементов на правильных треугольниках, билинейных элементов на квадратах и сплайнов. Тогда $KQ = F$ будет выглядеть точно как общепринятое разностное уравнение. Этот факт привел к бесчисленным дискуссиям о связи между конечными элементами и конечными разностями. Ясно, что не все разностные уравнения можно получить подходящим выбором элемента: матрица K должна быть симметричной и положительно определенной, но даже при этих ограничениях соответствующий элемент может отсутствовать. С другой стороны, достаточно терпеливый читатель может пожелать рассматривать все уравнения метода конечных элементов (даже на неравномерной сетке с многими узловыми неизвестными) как конечно-разностные уравнения. Мы приветствуем это намерение. Вообще система $KQ = F$ дает новый тип объединенных разностных уравнений, который в принципе можно было изобрести без вариационного принципа в качестве посредника. Исторически, конечно, это почти никогда не случалось. Метод конечных элементов систематически приводит к специальному классу уравнений (*пересечению всевозможных разностных уравнений со всевозможными уравнениями Ритца — Галёркина*), удивительно удачному при вычислениях.

Если коэффициенты исходного дифференциального уравнения $Lu = f$ постоянны, то порядок локальной ошибки отсечения можно проверить следующим образом. Пусть $f(x)$ — чисто показательная функция $e^{i\xi x}$, так что u можно найти точно и подставить

в разностное уравнение. Ошибка отсечения будет иметь вид $e^{i\xi x}E(h, \xi)$. Например, в задаче $-u'' = f$ решением будет $u = e^{i\xi x}/\xi^2$, и ошибка отсечения для $h^{-2}(-u_{j+1} + 2u_j - u_{j-1}) = f_j$ равна

$$\left(\frac{-e^{i\xi h} + 2 - e^{-i\xi h}}{h^2\xi^2} - 1\right)e^{i\xi x} = \left(\frac{2 - h^2\xi^2 - 2\cos h\xi}{h^2\xi^2}\right)e^{i\xi x} = Ee^{i\xi x}.$$

Вообще коэффициент E можно выразить через преобразования Фурье пробных функций. (Здесь мы попали прямо в суть абстрактного метода конечных элементов: эта техника на неравномерной сетке была бы невозможна.) При разложении E в ряд по степеням h старший показатель равен как раз порядку точности разностного уравнения. Мы вычислили этот показатель и убедились, что он равен меньшему из чисел k и $2(k - m)$, т. е. скорость сходимости, указанная теоремой 3.7, правильна.

Предположим, наконец, что найдется M неизвестных, связанных с каждым квадратом сетки, так что уравнение метода конечных элементов $KQ = F$ становится объединенной системой M разностных уравнений. Неизвестными могут быть значения функции u^h в различных узлах или значения функции и производных в кратном узле. Это не вносит сложностей, если ошибка оценивается из вариационных соображений (теорема 3.7); результат зависит только от порядка аппроксимации, достигаемого подпространством S^h, и любой дополнительный факт о подпространстве к делу не относится. Тем не менее при $M > 1$ аспект разностного уравнения становится намного тоньше.

Короче говоря, проблема состоит в том, что не все ошибки отсечения в разностных уравнениях имеют ожидаемый порядок. Поэтому не так просто оценить эти ошибки, а затем, применяя устойчивость для обращения матрицы K, превратить их в оценки ошибки $u - u^h$. Дело в том, что u^h задается специальной комбинацией пробных функций и, если другие комбинации почти не вносят вклад в задачу аппроксимации, их вклад в u^h также оказывается малым. Напомним, что в абстрактном методе функции $\Phi_1, \ldots, \Phi_m$ порождают аппроксимацию порядка k тогда и только тогда, когда можно построить из них отдельную функцию ψ, обладающую свойством (5), требуемым в теореме 3.2, т. е. функцию, которая *сама* подходит для аппроксимации. Можно считать пространство S^h порожденным этой суперфункцией ψ и $M - 1$ более или менее бесполезными функциями. Образуя соответствующую комбинацию разностных уравнений $KQ = F$, перепишем нашу систему метода конечных элементов в виде совокупности разностных уравнений специальной формы: одно уравнение системы — точный аналог исходного дифференциального уравнения, остальные $M - 1$ уравнений (связанные с функ-

циями в S^h, бесполезными для аппроксимации) нам совершенно безразличны. (Для больших M конструкция становится очень сложной и в качестве инструмента возможен лишь анализ Фурье; см. [C7] и [C12].) Окончательный вывод: метод Ритца приписывает почти весь вес одному разностному уравнению, соответствующему функции ψ, и порядок точности этого уравнения совпадает с показателем $\min(k, 2(k-m))$ в теореме 3.7. Это и есть порядок точности (для перемещения) метода конечных элементов.

4 НАРУШЕНИЯ ВАРИАЦИОННОГО ПРИНЦИПА

4.1. НАРУШЕНИЯ ЗАКОНОВ РЭЛЕЯ—РИТЦА

Одно из основных правил в теории Ритца состоит в том, что пробные функции в вариационном принципе должны быть допустимы. В наших обозначениях каждая функция v^h должна принадлежать пространству $\mathscr{H}_E^m$, а u^h — минимизировать $I(v^h)$. Сформулировать это правило просто, но оно нарушается повседневно и по важным причинам. В самом деле, это правило включает три условия и все они представляют вычислительные трудности — возможно преодолимые, но серьезные:

1. Пробные функции должны обладать m производными, суммируемыми с квадратом и потому быть класса $\mathscr{C}^{m-1}$ на границах элементарных областей.
2. Главные краевые условия должны быть выполнены.
3. Функционал $I(v^h) = q^T K q - 2q^T F$ должен вычисляться точно.

Наша цель — исследовать последствия нарушения этих условий.

Первое нарушается несогласованными элементами; в следующем разделе мы покажем, что в этом случае *сходимость может быть и может не быть*. Совершенно необязательно (и даже не всегда вероятно), что дискретная задача совместима с непрерывной. Наоборот, к пробным функциям применяется *кусочное тестирование*, которое определяет, согласованно или нет они воспроизводят состояния постоянной деформации. Если да, то процесс сходится внутри каждой элементарной области.

Главные краевые условия выполняются по крайней мере вдоль границы, приближенной изопараметрическими или субпараметрическими элементами. Однако есть много обстоятельств, при которых эти элементы непригодны: либо они слишком сложны для программирования вручную, либо сама задача слишком сложна — например, полная система четвертого порядка для уравнения оболочек. В таких случаях правило можно частично удовлетворить следующим образом: главные краевые условия могут налагаться *в граничных узлах*. Между узлами полиномиальные пробные функции не могут совпасть с общей

границей и потому главные условия можно только *интерполировать*. В разд. 4.4 мы покажем, что сходимость все же есть.

Условие 3 нарушается чаще остальных, поскольку неудобно вычислять точно функционал $I(v)$, но очень легко вычислить его приближенно. Приближение осуществляется двумя путями: интегралы по каждой элементарной области вычисляются численным интегрированием или же сама область интегрирования Ω заменяется совокупностью простых элементарных фигур. В обоих случаях $I(v)$ заменяется новым функционалом $I_*(v)$. Поэтому математическая проблема состоит в определении зависимости минимизирующей функции u^h от самого функционала: если деформировать выпуклый параболоид, то насколько передвинется минимум?

Каждое из этих трех возможных нарушений требует тщательного анализа, если мы хотим оправдать фактическое использование метода конечных элементов. Прежде чем начать изучение последствий, полезно обратить внимание на две идеи, возникающие снова и снова на протяжении всей главы; они относятся к анализу всех трех проблем. Во-первых, всегда рассматривается равенство нулю первой вариации, или, в физических терминах, уравнение виртуальной работы. Даже когда правила нарушены, минимизирующая функция u_*^h все же удовлетворяет равенству

$$a_*(u_*^h, v^h) = (f, v^h)_* \quad \text{для всех} \quad v^h \in S^h.$$

В случае метода Ритца это сопоставимо с

$$a(u, v) = (f, v) \quad \text{для всех} \quad v \in \mathscr{H}_E^m.$$

Если $S^h \subset \mathscr{H}_E^m$ и $I = I_*$ (так что звездочку можно убрать), то одно уравнение можно вычесть из другого, и, как в теореме 1.1,

$$a(u^h - u, v^h) = 0 \quad \text{для всех} \quad v^h \in S^h.$$

В нашем случае это выражение отлично от нуля и вся проблема сводится к тому, чтобы показать, что это отличие мало.

Вторая идея специфична для метода конечных элементов и состоит в использовании специальных свойств полиномов. Мы уже отмечали, как к полиномиальным решениям применяется кусочное тестирование. Ситуация аналогична численному интегрированию, где точность зависит от степени полиномов, интегрируемых точно. Отметим еще одно свойство, полезное для анализа изменений области: полиномы не могут значительно меняться в полосе между заданной областью Ω и ее аппроксимацией Ω^h.

Подчеркнем, что анализ не зависит от оценок возмущений в матрице жесткости при переходе от K к $\tilde{K}$ или K^*. Это измене-

ние может быть очень большим (уравнение может оказаться совершенно другим), несмотря на то, что мы все еще работаем с полиномами степени m. На языке функционального анализа это значит, что множество таких полиномов плотно в допустимом пространстве при $h \to 0$, так что их поведение играет решающую роль.

4.2. НЕСОГЛАСОВАННЫЕ ЭЛЕМЕНТЫ И КУСОЧНОЕ ТЕСТИРОВАНИЕ

Некоторые из часто используемых элементов не согласованы — их производные порядка $m-1$ разрывны на границах элементарных областей — и тем не менее они довольно хорошо работают. Точнее, иногда они работают хорошо, а иногда — нет. Чаще всего так рискуют в задачах четвертого порядка, где элементы должны принадлежать $\mathscr{C}^1$. Подбор наклонов между элементами может оказаться трудным для нормальных перемещений w пластины при изгибе, и он чрезвычайно труден для оболочек. Поэтому технический прием состоит в обычном вычислении энергий внутри каждого отдельного элемента, а затем сложении результатов. Вследствие этого истинный функционал $I(v)$ заменяется суммой интегралов элементов.

$$I_*(v) = \sum_e [a_e(v, v) - 2(f, v)_e] = a_*(v, v) - 2(f, v)_*.$$

Функционал I отличается от I_* тем, что в I_* игнорируются особенности на границах элементов, в то время как для несогласованных элементов $I(v) = \infty$.

Приближение по Ритцу u_*^h является (возможно, несогласованной) функцией из пробного пространства, минимизирующей $I_*(v^h)$. Такое свойство u_*^h выражается, как обычно, равенством нулю первой вариации от I_*:

$$a_*(u_*^h, v^h) = (f, v^h)_* \quad \text{для всех} \quad v^h \in S^h. \tag{1}$$

Это приближенное уравнение виртуальной работы. Оно совпадает с обычным уравнением, но только опять интегралы вычисляются поэлементно, а затем суммируются, причем разрывы между элементами не учитываются.

Наша цель — найти условия, при которых такая аппроксимация u_*^h методом конечных элементов сходится к u вопреки ее незаконной конструкции. Этот вопрос оставался неясным, пока Айронс (см. [Б7]) не выдвинул простую, но блестящую идею, известную как *кусочное тестирование*. Предположим, что произвольная группа элементов находится в состоянии постоянной деформации: $u(x, y) = P_m(x, y)$, где P_m — полином степени m. Тогда, поскольку этот полином принадлежит S^h (даже на несо-

гласованные элементы наложено условие постоянной деформации: степень $k-1$ подпространства должна быть не менее m), истинное решение Ритца u^h тождественно совпадает с P_m. (На границе рассматриваемой группы налагаемые условия выбираются согласованными с постоянной деформацией, т. е. для перемещений требуется, чтобы $u^h = Pm$ на границе группы элементов.) Тогда *тестирование состоит в том, чтобы выяснить, совпадает ли решение u_*^h метода конечных элементов с полиномом P_m*, несмотря на несовпадение I и I_* в результате игнорирования границ между элементами. Можно предположить (что мы и сделаем), что в задаче все коэффициенты постоянны, так как их изменения на элементе вносят вклад в u_*^h лишь $O(h)$.

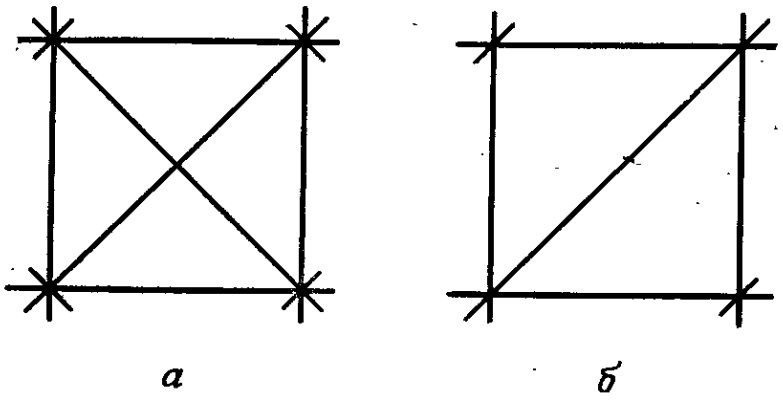


Рис. 4.1.
Успешное и неудачное кусочные тестированиия.

В [Б7] приведен знаменитый пример, в котором элементы формы *б* (рис. 4.1) выдерживают это тестирование, а для элементов формы *а* несовпадение составляет 1,5%. (Такова ошибка для крупномасштабной энергии деформации, а поточечная ошибка напряжений достигает 25%.) Здесь рассматривается кубическая функция с параметрами v, v_x и v_y в вершинах. Десятое неизвестное исключается с помощью ограничения на коэффициенты. Авторы неодобрительно относятся к приравниванию коэффициентов при x^2y и xy^2, потому что это ограничение на правильном треугольнике может нарушаться: все девять узловых параметров функции $xy(1-x-y)$ на стандартном треугольнике равны нулю — ограничение выполнено, а полином не равен тождественно нулю. Поэтому авторы выбрали условие, инвариантное относительно вращения и никогда не вырождающееся.

Для неправильных треугольников много шансов, что они не выдержат кусочного тестирования. Тем не менее может быть приемлемой ошибка в несколько процентов, особенно когда она совершается в направлении, противоположном истинной ошибке метода Ритца. Как объяснялось в разд. 1.10, последняя всегда

возникает из-за слишком жесткой конструкции: $a(u^h, u^h) \leqslant a(u, u)$; требования к несогласованному решению u_*^h слабее, а его перемещения, как и деформации, часто превышают оценку.

Один подход к теории заключается в том, чтобы сразу рассматривать дискретную систему $KQ = F$ как конечно-разностное уравнение, забывая, что оно не так выведено. Если это уравнение согласовано и устойчиво, то решение u_*^h должно сходиться к u. Для равномерной сетки это составляет как раз суть кусочного тестирования: *оно представляет собой контроль согласования*[1]). Обычно несогласованные элементы имеют только на одну производную меньше, так что $v^h \in \mathscr{C}^{m-2}$, и согласование в дифференциальном уравнении гарантируется для членов меньшего порядка. Тогда кусочное тестирование в случае *б* (рис. 4.1) подтверждает согласование для разделенной разности самого высокого порядка, в то время как в случае *а* доказано, что уравнение $KQ = F$ будет аналогом *неверного дифференциального уравнения* с новым главным членом. Эту связь между согласованием и кусочным тестированием можно установить методом Фурье. Мы опускаем детали, потому что этот метод ограничивается на деле равномерной сеткой и требует много подготовительной работы.

Второй подход к теории несогласованных элементов — вариационный и потому более общий. Мы утверждаем, что вариационный смысл успешного кусочного тестирования таков: *для каждого полинома* P_m *и каждой* (*несогласованной*) *базисной функции* φ_j

$$a_*(P_m, \varphi_j) = a(P_m, \varphi_j). \tag{2}$$

Равенство (2) справедливо тогда и только тогда, когда кусочное тестирование выдержано. Заметим, что правая часть корректно определена. Главные члены представляют собой произведение m-х производных от P_m, являющихся постоянными, на m-е производные от несогласованных функций φ_j, являющихся δ-функциями, когда направление производной нормально к границам элементов. Поэтому вклады границы конечны: они вычисляются из скачка в нормальной производной порядка $m-1$, как мы покажем на примере. Эти вклады *необязательно нулевые*: кусочное тестирование может оказаться неудачным и сходимости не будет.

Для того чтобы доказать эквивалентность равенства (2) кусочному тестированию, рассмотрим участок, вне которого функция φ_j тождественно равна нулю. Предположим, что истин-

[1]) Напомним, что для разностных уравнений согласование проверяется исследованием нескольких первых членов ряда Тейлора, т. е. рассматриваются полиномы вплоть до определенной степени. Кусочное тестирование делает точно то же для конечных элементов.

ное решение u равно P_m, и заметим, что соответствующая нагрузка f (если она есть), несомненно, не содержит δ-функций на границах элементов:

$$(f, \varphi_j)_* = (f, \varphi_j).$$

Вводя-два уравнения виртуальной работы, получаем

$$a_*(u_*^h, \varphi_j) = a(P_m, \varphi_j).$$

Если кусочное тестирование выдержано, так что приближенное решение u_*^h совпало с P_m, то очевидно, что равенство (2) справедливо. Обратно, предположим, что (2) справедливо для всех φ_j. Тогда $a_*(u_*^h, \varphi_j) = a_*(P_m, \varphi_j)$ и обязательно $u_*^h = P_m$: кусочное тестирование выдержано.

Следует упомянуть об одной технической трудности в этом доказательстве. Обычно первая вариация не должна обращаться в нуль в направлении функции φ_j, лежащей вне $\mathscr{H}^m$. Однако для гладкого решения $u = P_m$ можно показать (на основе интегрирования по частям [C9]), что это обязательно происходит: $a(u, \varphi_j) = (f, \varphi_j)$, а приведенные выше рассуждения верны. Короче, равенство (2) как раз и проверяется кусочным тестированием.

Возможно, простейший пример, на котором можно проверить кусочное тестирование, доставляют прямоугольные элементы Вильсона [B10]. К обычным билинейным функциям (на квадрате $-1 \leqslant x, y \leqslant 1$ четыре базисные функции $(1 \pm x)(1 \pm y)/4$) он добавляет две новые: φ и ψ. Внутри квадрата $\varphi = 1 - x^2$ и $\psi = 1 - y^2$, а вне его они доопределены нулем, и потому *они разрывны на границе*. Так как они равны нулю вне элемента, то решение окончательной линейной системы допускает статическую конденсацию, а эффект от φ и ψ должен позволить улучшить представление внутри каждого элемента. Вследствие разрыва тем не менее понадобилось кусочное тестирование.

Для простоты предположим, что дифференциальное уравнение имеет вид $-\Delta u = f$. Энергетическое скалярное произведение равно $a(u, v) = \iint u_x v_x + u_y v_y$, и если $P = a + bx + cy$ — произвольный полином степени $m = 1$, то

$$a_*(P, \varphi) = \int_{-1}^{1} \int_{-1}^{1} b(-2x)\, dx\, dy = 0$$

С другой стороны, по формуле Грина

$$a(P, \varphi) = \iint (-\Delta P)\varphi - \int \varphi \frac{\partial P}{\partial n}\, ds.$$

Выберем достаточно большую область, чтобы внутри целиком содержался заданный квадрат; тогда функция φ обратится в нуль на границе и интеграл по прямой будет равен 0. Так как $-\Delta P = -P_{xx} - P_{yy} = 0$ для любого линейного полинома P, то $a(P, \varphi) = 0 = a_*(P, \varphi)$, и кусочное тестирование выдержано.

Кусочное тестирование, очевидно, представляет собой очень простое правило: истинное значение $a(P_m, \varphi_j)$ равно нулю, как мы только что видели, и потому требуется, *чтобы интеграл от каждой деформации* $D^m\varphi_j$ (*вычисленный несогласованным образом — без учета границы элементов*) *также был равен нулю.* Это можно проверить аналитически — кусочное тестирование осуществляется без ЭВМ. Величина этих интегралов, когда они не равны нулю, определяет степень несовместности несогласованных уравнений.

Другой способ достижения этого результата состоит в применении теоремы Грина на отдельном квадрате:

$$a(P, \varphi) - a_*(P, \varphi) = -\int \varphi \frac{\partial P}{\partial n}\, ds.$$

Так как $\varphi = 0$ на вертикальных сторонах $x = \pm 1$, остается один интеграл вдоль нижней стороны и другой — в обратном направлении по верхней стороне. Так как производная $\partial P/\partial n$ в одном случае равна $-c$, а в другом $+c$, то

$$\int \varphi \frac{\partial P}{\partial n}\, ds = \int_{-1}^{1} (1 - x^2)(-c)\, dx + \int_{1}^{-1} (1 - x^2)\, c\, (-dx) = 0.$$

Таким образом, опять тестирование выдержано. Тестирование *не* выдержано для четырехугольника произвольной формы, когда он построен изопараметрически: $\varphi = 1 - \xi^2$, $\psi = 1 - \eta^2$, x, y — билинейные функции от ξ и η, отображающие обычный квадрат в заданный четырехугольник. На самом деле кусочное тестирование могут не пройти даже четырехугольники разумной формы, так что элементы становятся непригодными. Для того чтобы прошли тестирование изопараметрические элементы, Тейлор изменил несогласованные функции φ и ψ, а затем модифицировал также численные квадратуры — зло *исправляется* злом.

Существует второй несогласованный элемент, в равной степени простой и полезный. Он составлен из кусочно линейных функций на треугольниках. Предпочтительнее выбирать узлы не в вершинах, что дает непрерывность на каждой границе между элементами, а помещать их *в середины сторон.* Поэтому непрерывность между элементами пропадает (за исключением этих середин) и пробное пространство получается большей размерности — грубо говоря, его размерность в три раза больше

размерности обычного пространства Куранта, так как она равна отношению числа сторон к числу вершин. Это большее пространство позволяет наложить еще условие $\operatorname{div} v^h = 0$ на стороне и все же сохранить достаточно степеней свободы для аппроксимации.

Для успешной работы с этим пространством необходимо пройти кусочное тестирование. Мы проведем тестирование, как и раньше, вычисляя интеграл $\int \varphi\, \partial P/\partial n\, ds$ вдоль каждой стороны. Фактически *мы интегрируем скачок в функции* φ, чтобы выяснить влияние прохождения по стороне в одном направлении и затем возвращения по другой стороне. Скачок в φ' будет линейной функцией, поскольку φ линейна в каждом из двух треугольников с общей стороной. В середине стороны функция φ непрерывна и скачок равен нулю. Наконец, $\partial P/\partial n$ — постоянная, так как P — линейный полином. Поскольку интеграл от линейной функции равен нулю, при условии что функция равна нулю в середине пути интегрирования, каждая пробная функция выдерживает кусочное тестирование.

Заметим, что от сетки даже не требовалась равномерность! Темам установил для этих элементов неравенство Пуанкаре: $\| v^h \|_0^2 \leqslant C a_* (v^h, v^h)$.

Мы собираемся доказать, что решения метода конечных элементов, основанные на несогласованных элементах Вильсона, сходятся к u. Скорость сходимости будет минимальной $-O(h^2)$ по энергии, хотя это может не дать правильного отражения ее точности при больших h. (Существенная черта метода конечных элементов — успех на грубой сетке; даже элементы, не являющиеся сходящимися и не выдерживающие кусочное тестирование, для реальных h могут дать удовлетворительные результаты.) Если элемент к тому же выдерживает тестирование для полиномов более высокой степени P_n, то скорость сходимости по энергии должна была бы возрасти до $h^{2(n-m+1)}$, но это не так.

Наш план — начать с общей оценки ошибки, пригодной для любого несогласованного элемента, и тем самым выделить величину, которая играет решающую роль в определении ошибки. Это величина Δ, определяемая формулой

$$\Delta = \max_{v^h \in S^h} \frac{| a_*(u, v^h) - a(u, v^h) |}{| v^h |_*}, \quad \text{где} \quad | v^h |_* = [a_*(v^h, v^h)]^{1/2}.$$

Затем для отдельного элемента, изучаемого выше, оценим Δ и выведем скорость сходимости.

Граница ошибки, на которой все основано, такова:

$$| u - u_*^h |_* \leqslant \Delta + \min_{w^h \in S^h} | u - w^h |_*. \tag{3}$$

Это неравенство отражает, как и должно, порядок аппроксимации (последний член) и влияние несогласованности: $a = a_*$ и $\Delta = 0$ для согласованных элементов. Для доказательства (3) сравним уравнения виртуальной работы: предположим, что функция f гладкая, тогда

$$a_*(u_*^h, v^h) = (f, v^h)_* = (f, v^h) \quad \text{и} \quad a(u, v^h) = (f, v^h).$$

Отсюда следует, что для всех $v^h \in S^h$

$$a_*(u - u_*^h, v^h) = a_*(u, v^h) - a(u, v^h). \tag{4}$$

Здесь мы сразу получаем ценный результат: левая часть по неравенству Шварца ограничена величиной $|u - u_*^h|_* |v^h|_*$. Разделим на $|v^h|_*$ и возьмем максимум по S^h; имеем *нижнюю границу ошибки* (мы признательны Р. Скотту за его помощь)

$$|u - u_*^h|_* \geqslant \Delta. \tag{5}$$

Это показывает, что оценка (3), если она доказана, совершенно реальна и влечет за собой, что любой сходящийся элемент на равномерной сетке должен выдержать кусочное тестирование. Только если тестирование выдержано, $\Delta \to 0$.

Для того чтобы закончить доказательство верхней оценки (3), выберем функцию w ближайшей к u в S^h. По неравенству треугольника

$$|u - u_*^h|_* \leqslant |u - w|_* + |w - u_*^h|_* = \min_{w^h} |u - w^h|_* + |w - u_*^h|_*.$$

Таким образом, неравенство (3) будет доказано, если установить равенство последнего члена величине Δ. Этот член есть максимум по всем v^h отношения $R = |a_*(w - u_*^h, v^h)| / |v^h|_*$. А так как w — ближайшая к u функция в S^h (т. е. проекция элемента u на S^h), то она удовлетворяет равенству $a_*(w, v^h) = a_*(u, v^h)$ для всех $v^h \in S^h$. Подставляя это равенство в числитель в R и учитывая тождество (4), видим, что максимум отношения R совпадает с Δ. Это доказывает верхнюю оценку (3).

Покажем, что для прямоугольных элементов Вильсона, выдерживающих кусочное тестирование, $\Delta = O(h)$; это дает правильную скорость сходимости для $|u - u_*^h|_*$. Для получения этой оценки запишем пробную функцию в виде

$$v^h = \sum a_i \varphi_i + \sum b_i \psi_i + \sum c_i \omega_i,$$

где φ_i и ψ_i — новые несогласованные базисные функции, сдвинутые в i-й элемент. Тогда $\varphi_i = 1 - ((x - x_i)/2h_i)^2$, где x_i — координата центра i-го квадрата. Функции ω_i образуют обычный базис для согласованного билинейного элемента; мы назвали их функциями «пагоды». Так как они согласованы, то они не

влияют на Δ. В самом деле, на обычном квадрате числитель в Δ равен

$$(a_* - a)(u, v^h) = (a_* - a)(u, a_i\varphi_i + b_i\psi_i) = \\ = (a_* - a)(u - P_1, a_i\varphi_i + b_i\psi_i).$$

Здесь использовано кусочное тестирование или скорее эквивалентное тождество (2); введение линейного полинома P_1 не оказывает влияния. Запасшись небольшим терпением, можно оценить этот вклад в Δ на i-м квадрате:

$$C\|u - P_1\|_{1, e_i}\|a_i\varphi_i + b_i\psi_i\|_1 \leqslant C'h\|u\|_{2, e_i}\|a_i\varphi_i + b_i\psi_i\|_1.$$

Мы выбрали P_1, что позволяется теорией аппроксимации, для того чтобы производная от $u - P_1$ была порядка $O(h)$ (напомним, что в этом примере $m = 1$). Теперь просуммируем вклады по всем квадратам и применим неравенство Шварца:

$$|a_*(u, v^h) - a(u, v^h)| \leqslant C'h\sum\|u\|_{2, e_i}\|a_i\varphi_i + b_i\psi_i\|_1 \leqslant \\ \leqslant C''h\left(\sum\|u\|^2_{2, e_i}\right)^{1/2}\left(\sum\|v^h\|^2_{1, e_i}\right)^{1/2} = C''h\|u\|_2|v^h|_*.$$

Деля на $|v^h|_*$, приходим к правому неравенству

$$\Delta \leqslant C''h\|u\|_2.$$

Теперь из основной оценки (3) следует, что несогласованные элементы Вильсона дают ошибку $\|u - u^h_*\|_* = O(h)$ в норме энергии деформации. Мы уверены, что эта скорость сходимости верна, но отметим два момента. (1) Из эксперимента ясно, что постоянный множитель перед h намного меньше, чем он был бы без дополнительных несогласованных пробных функций. (2) Не обязательно энергия деформации в u^h_* меньше, чем в u. Фактически сходимость сверху скорее правило, нежели исключение, как для энергии деформации, так и для перемещений. Доказательство сходимости после удачного кусочного тестирования обсуждается далее в указателе обозначений. Айронс и Раззак в Трудах Балтиморского симпозиума [6] описали много других элементов, выдерживающих тестирование, в том числе

(1) прямоугольный элемент с 12 степенями свободы, открытый Эйри, Адини, Клафом и Мелошем (в углах узловыми параметрами служат v, v_x и v_y, а образующей функцией является сумма полной кубической функции и x^3y, xy^3),

(2) элемент кручения постоянной кривизны, предложенный Морли, который можно рассматривать как согласованный для принципа дополнительной энергии,

(3) новый элемент типа элементов Олмана — Пиана.

Обширные вычисления с элементом (1) для задачи о пластине на собственные значения (элемент непрерывен, но несогласован для уравнений четвертого порядка) приведены в [Л3]. Они описывают определенную скорость сходимости порядка h^2 для ошибок в собственных значениях, соответствующих ошибкам в энергии для статических задач. Это согласуется с нашим предсказанием. Заметим, что за несогласованность пришлось заплатить скоростью сходимости: для этого элемента $k = 4$ и теория аппроксимации должна была бы допустить скорость $O(h^{2(k-m)}) = O(h^4)$.

Оказывается, что элемент (3) выдерживает тестирование даже на неравномерной сетке, что довольно замечательно. В самом деле, равенство граничных интегралов нулю, требуемое в тестировании, было бы, вообще говоря, довольно счастливым случаем даже на равномерной сетке, и нужно ожидать, что в будущем для достижения необходимой инженерам точности будут использоваться не только несогласованные, но и *несходящиеся* элементы.

Приведенный список содержит также ряд численно интегрируемых элементов, но мы предпочитаем *не рассматривать численное интегрирование как производящее несогласованные элементы.* Влияние такого интегрирования на самом деле состоит в замене истинного функционала $I(v)$ новым, но разность между ними не содержит граничных интегралов. Поэтому численное интегрирование и ошибку, которую оно вносит в аппроксимацию u^h метода конечных элементов, мы изучим отдельно.

4.3. ЧИСЛЕННОЕ ИНТЕГРИРОВАНИЕ

Численное интегрирование становится все более важной частью метода конечных элементов. На ранних стадиях метода одно из основных его преимуществ заключалось как раз в обратном, а именно интегрирование полиномов на треугольниках и прямоугольниках было основано на точных формулах. В настоящее время, по-видимому, особая простота полиномов более не играет существенной роли и рациональные функции, и функции даже еще более общего вида так же удобны. Фактически же нет ничего более неверного: залог успеха численного интегрирования в методе конечных элементов — присутствие полиномов.

Основной вопрос таков: какая степень точности интерполяционной формулы требуется для сходимости? Необязательно, чтобы *каждый появляющийся полином интегрировался точно.* Подынтегральное выражение в энергии $a(v, v)$ содержит *квадраты* полиномов, и формула, точная этой степени, может обойтись слишком дорого. Очень важно должным образом прокон-

тролировать ту часть времени работы ЭВМ, которая тратится на численное интегрирование.

Математически мы вновь сталкиваемся с изменением функционала $I(v)$; это влияние вычислительных квадратурных формул. Предположим, что истинный функционал равен

$$I(v) = a(v, v) - 2b(f, v) = \int\int [p(x, y)(v_x^2 + v_y^2) - 2fv]\, dx\, dy.$$

(Отметим новое обозначение $b(f, v)$ для линейного члена.) Тогда

$$I^*(v) = a^*(v, v) - 2b^*(f, v) = \\ = \sum w_i [p(\xi_i)(v_x^2(\xi_i) + v_y^2(\xi_i)) - 2f(\xi_i)\, v(\xi_i)].$$

Каждая элементарная область дает определенное количество квадратурных узлов $\xi_i = (x_i, y_i)$ с весами w_i, зависящими от размеров и формы элементарной области и от выбранного правила численного интегрирования. Квадратурная формула называется *точной степени q*, если интеграл от любого полинома P_q правильно вычисляется суммой $\sum w_i P_q(\xi_i)$.

Предположим, что I^* минимизируется по всем пробным функциям v^h. Тогда минимизирующая функция $\tilde{u}^h = \sum \tilde{Q}_j \varphi_j$ определяется приближенной системой метода конечных элементов $\tilde{K}\tilde{Q} = \tilde{F}$, в которой матрица жесткости и вектор нагрузки найдены численным, а не точным интегрированием. *Это и есть система* (не считая ошибок округления), *на самом деле решаемая с помощью ЭВМ.* Наша цель — оценить разность $u^h - \tilde{u}^h$, и мы повторим здесь основную мысль: нет необходимости в близости энергий I и I^* для того, чтобы разность $u^h - \tilde{u}^h$ была мала.

Сначала обсудим основную теорему, а потом докажем ее и приведем примеры. *Главное условие на квадратурную формулу совпадает с кусочным тестированием для несогласованных элементов: $\tilde{u}^h$ сходится к u^h в энергии деформации* (т. е. $\|u^h - \tilde{u}^h\|_m \to 0$) *тогда и только тогда, когда для всех полиномов степени m и всех пробных функций*

$$a^*(P_m, v^h) = a(P_m, v^h) + O(h). \qquad (6)$$

Дополнительное условие положительной определенности требует, чтобы приближенная энергия деформации a^* была эллиптична на подпространствах S^h, т. е. $a^*(v^h, v^h) \geqslant \theta \|v^h\|_m^2$. Член $O(h)$ входит в условие только потому, что, если свойства материала внутри элемента изменяются, это не выражается точно в квадратурах. Это второстепенный эффект.

Проверка (6) в действительности применяется только к главным членам в энергии деформации, содержащим m-е произ-

водные. Так как эти производные от P_m постоянны, то скалярное произведение $a(P_m, v^h)$ содержит только интегралы от m-х производных от v^h и условие сходимости сводится к следующему: *m-е производные от каждой пробной функции должны интегрироваться точно.* Если пробные функции — полиномы степени $k-1$, то это условие означает, что квадратурная формула должна быть верной по крайней мере до степени $k-m-1$. Элементы на четырехугольниках намного капризнее: есть члены в пробных функциях, ничего не добавляющие к степени аппроксимации, но производные их должны тем не менее интегрироваться точно. Например, для билинейных элементов в задаче второго порядка у члена кручения xy производная линейна, поэтому квадратурная формула должна быть точной и для этих членов, а не только для постоянных деформаций, возникающих из линейных членов $a+bx+cy$.

Практически равенство (6) очень часто выполняется для всех линейных полиномов P_n некоторой степени n, большей m. В этом случае точность выше минимальной: ошибка в деформациях имеет порядок h^{n-m+1}. Каждая дополнительная степень точности в квадратурной схеме вносит дополнительную степень h в оценку ошибки. Другими словами, если степень пробных функций равна $k-1$, а квадратура точна степени q, то порядок ошибки равен $q-k+m+2$. Если пробные функции содержат какой-нибудь член степени выше $k-1$, что всегда бывает для элементов на прямоугольниках, то порядок ошибки портится. В любом случае правильное тестирование выражается формулой $a^*(P_n, v^h) = a(P_n, v^h)$: точно должен интегрироваться полный полином степени $n-m$, умноженный на пробные деформации $D^m v^h$ [1]).

Мы не будем пытаться построить новые квадратурные формулы, но удивительно, что даже на треугольниках и прямоугольниках эта классическая проблема полностью не решена. Айронс [А5] показал, что еще можно сделать в этом направлении, добившись заданной точности q с гораздо меньшим количеством точек, чем требуется при суперпозиции обычных гауссовых квадратур. Русская школа Соболева, Люстерника и др. провела глубокое изучение «кубатурных формул» на регулярных областях и получила несколько замечательных формул: читателя

[1]) Интересно подсчитать точность квадратурной формулы, требуемую для того, чтобы разность $u^h - \tilde{u}^h$ была того же порядка h^{k-m} в деформациях, что и осиовная ошибка аппроксимации. Если этот показатель совпадает с $n-m+1$, то $n=k-1$, и точно должны интегрироваться m-е производные от всех полиномов степени $k-1$, умноженные на m-е производные от всех пробных полиномов v^h. Если v^h сами содержат все полиномы степени $k-1$, как часто бывает на треугольниках, а существенные коэффициенты в энергии деформации постоянны, то главные члены в энергии — *квадраты m-х производных* — для сохранения точности в целом должны вычисляться точно.

может заинтересовать 14-точечная формула на кубе со стороной $\sqrt{2}$, порядок точности которой равен 5:

$$\frac{2\sqrt{5}}{361}\left(\frac{121}{8}\sum_{1}^{8} f(\xi_i) + 40\sum_{1}^{6} f(\eta_i)\right),$$

где ξ_i — вершины куба, расположенного подобно исходному, со стороной $\sqrt{38/33}$, а η_i — вершины правильного октаэдра, лежащие на описанной сфере радиуса $\sqrt{19/60}$! Мы повторяем, однако, что численное интегрирование для конечных элементов должно принимать во внимание члены высшего порядка, часто встречающиеся в пробных функциях на прямоугольниках, даже если они не вносят вклада в теорию аппроксимации. На треугольниках элемент обычно представляет собой полный полином степени $k-1$ либо близок к нему, и тогда весь вопрос сводится к правильному интегрированию полиномов возможно большей степени. Таблица 4.1 взята у Купера [K13], добавившего несколько новых квадратурных формул на треугольниках к формулам, приведенным Зенкевичем [7]. Формулы симметричны относительно пространственных координат, поэтому если встречается квадратурный узел $\xi_i = (\zeta_1, \zeta_2, \zeta_3)$, то *обязательно встретятся и все его перестановки.* Если все ζ_i различны, то таких узлов в квадратуре 6; если два значения ζ_i совпадают, то таких узлов три, если в формуле используется центральная точка (1/3, 1/3, 1/3), то лишь один раз.

Изопараметрический метод не может существовать без численного интегрирования, поскольку подынтегральное выражение — рациональная функция от новых координат ξ и η. Поначалу кажется невозможным, чтобы даже численное интегрирование было успешным, так как для рациональных функций оно никогда не бывает точным. Элементы $a^*(\varphi_j, \varphi_k)$ матрицы $\tilde{K}$ будут совершенно отличны от элементов $K_{jk} = a(\varphi_j, \varphi_k)$, и доказательство на основе теории возмущений невозможно. Тем не менее мы вычисляем $a(P_m, v^h) - a^*(P_m, v^h)$ и применяем тестирование. Решающий момент состоит в том, что тестирование включает только одну пробную функцию, а не обе φ_j и φ_k одновременно, и это нас спасает.

Типичное преобразование приведено в разд. 3.3 для $m = 1$; производная $\partial P_m/\partial x$ равна постоянной c и

$$\iint_{e_i} p(x, y)\frac{\partial P_m}{\partial x}\frac{\partial v^h}{\partial x}\,dx\,dy \to$$

$$\to c\iint_{E_i} p(x(\xi, \eta), y(\xi, \eta))\left(\frac{\partial v^h}{\partial \xi}\frac{\partial \xi}{\partial x} + \frac{\partial v^h}{\partial \eta}\frac{\partial \eta}{\partial x}\right) J\,d\xi\,d\eta.$$

Таблица 4.1

w_i	ζ_1	ζ_2	ζ_3	Кратность
	3-точечная формула	*Степень точности* 2		
0,33333 33333 33333	0,66666 66666 66667	0,16666 66666 66667	0,16666 66666 66667	3
	3-точечная формула	*Степень точности* 2		
0,33333 33333 33333	0,50000 00000 00000	0,50000 00000 00000	0,00000 00000 00000	3
	4-точечная формула	*Степень точности* 3		
−0,56250 00000 00000	0,33333 33333 33333	0,33333 33333 33333	0,33333 33333 33333	1
0,52083 33333 33333	0,60000 00000 00000	0,20000 00000 00000	0,20000 00000 00000	3
	6-точечная формула	*Степень точности* 3		
0,16666 66666 66667	0,65902 76223 74092	0,23193 33685 53031	0,10903 90090 72877	6
	6-точечная формула	*Степень точности* 4		
0,10995 17436 55322	0,81684 75729 80459	0,09157 62135 09771	0,09157 62135 09771	3
0,22338 15896 78011	0,10810 30181 68070	0,44594 84909 15965	0,44594 84909 15965	3
	7-точечная формула	*Степень точности* 4		
0,37500 00000 00000	0,33333 33333 33333	0,33333 33333 33333	0,33333 33333 33333	1
0,10416 66666 66667	0,73671 24989 68435	0,23793 23664 72434	0,02535 51345 51932	6
	7-точечная формула	*Степень точности* 5		
0,22503300003300000	0,33333 33333 33333	0,33333 33333 33333	0,33333 33333 33333	1
0,12593 91805 44827	0,79742 69853 53087	0,10128 65073 23456	0,10128 65073 23456	3
0,13239 41527 88506	0,47014 20641 05115	0,47014 20641 05115	0,05971 58717 89770	3
	9-точечная формула	*Степень точности* 5		
0,20595 05047 60887	0,12494 95032 33232	0,43752 52483 83384	0,43752 52483 83384	3
0,06369 14142 86223	0,79711 26518 60071	0,16540 99273 89841	0,03747 74207 50088	6
	12-точечная формула	*Степень точности* 6		
0,05084 49063 70207	0,87382 19710 16996	0,06308 90144 91502	0,06308 90144 91502	3
0,11678 62757 26379	0,50142 65096 58179	0,24928 67451 70910	0,24928 67451 70911	3
0,08285 10756 18374	0,63650 24991 21399	0,31035 24510 33785	0,05314 50498 44816	6
	13-точечная формула	*Степень точности* 7		
−0,14957 00444 67670	0,33333 33333 33333	0,33333 33333 43333	0,33333 33333 33333	1
0,17561 52574 33204	0,47930 80678 41923	0,26034 59660 79038	0,26034 59660 79038	3
0,05334 72356 08839	0,86973 97941 95568	0,06513 01029 02216	0,06513 01029 02216	3
0,07711 37608 90257	0,63844 41885 69809	0,31286 54960 04875	0,4869 03154 253160	6

Заметим, что матрица преобразования имеет вид

$$\begin{pmatrix} \xi_x & \xi_y \\ \eta_x & \eta_y \end{pmatrix} = \begin{pmatrix} x_\xi & x_\eta \\ y_\xi & y_\eta \end{pmatrix}^{-1} = \frac{1}{J}\begin{pmatrix} y_\eta & -x_\eta \\ -y_\xi & x_\xi \end{pmatrix}.$$

Очевидно, что $\xi_x = y_\eta/J$ и $\eta_x = -y_\xi/J$. С учетом этих подстановок интеграл равен

$$c \iint\limits_{E_i} p(\xi, \eta)\,(v_\xi^h y_\eta - v_\eta^h y_\xi)\,d\xi\,d\eta. \tag{7}$$

Рациональные функции исчезли и сходимость будет, если этот интеграл вычисляется правильно. (Для уравнений четвертого порядка рациональные функции *не* исчезают и численное интегрирование нельзя оправдать, если преобразование координат не удовлетворяет условию гладкости $\|F\|_k \leqslant C$ (стр. 193). В этом случае элементы искажены слабо и J на вид не хуже, чем переменный коэффициент p; ошибка интегрирования *того же порядка, что и для постоянных коэффициентов без изопараметрических преобразований.*)

В случае билинейных функций на четырехугольниках все производные v_ξ^h, y_η, ... линейны и, по-видимому, квадратурная формула должна быть верна для квадратичных полиномов. Однако выходит так, что члены второго порядка в конкретной комбинации $K = v_\xi^h y_\eta - v_\eta^h y_\xi$ уничтожаются, так что для сходимости вполне достаточно точности первой степени. Практически квадратурная формула будет точнее (вероятно, она должна быть точнее, чтобы форма a^* была положительно определена), и выигрыш от этого — скорость сходимости больше минимальной. Предположим, что условия разд. 3.3 выполнены: якобиан строго больше нуля, а коэффициенты преобразования координат $x(\xi, \eta)$ и $y(\xi, \eta)$ ограничены. Тогда, как и на плоскости x, y, каждая дополнительная степень полиномов от ξ и η увеличивает порядок аппроксимации на 1. Поэтому мы надеемся на улучшение ошибки интегрирования при тех же затратах; поскольку квадратурная формула первого порядка в билинейном случае достаточна для сходимости, ошибка в деформациях должна быть величиной $O(h^q)$, если квадратурная формула точна степени q.

Еще одно замечание по изопараметрическому методу: поскольку преобразования координат $x(\xi, \eta)$ и $y(\xi, \eta)$ имеют тот же вид, что и функция $v^h(\xi, \eta)$, комбинация K имеет тот же вид, что и якобиан J. Следовательно, наше правило о точной интегрируемости K совпадает с правилом [7], основанным на интуитивном предположении Айронса о том, что объем каждого элемента (интеграл от J) должен точно вычисляться по квадратурным формулам. В трехмерном случае это предположение

требует более высокий порядок точности: K и J содержат произведения трех, а не двух производных. Для субпараметрических элементов эти два правила отличаются: матрица жесткости зависит от K, а матрица массы и вектор нагрузки — от J. Повторяем, что эти правила применимы к сильно искаженным изопараметрическим элементам; для малых искажений якобиан J гладок и при проверке сходимости может не приниматься в расчет.

Теперь приступим к теории, целиком основанной на следующем простом тождестве.

Лемма 4.1. *Предположим, что* u^h *и* $\tilde{u}^h$ *минимизируют функционалы* $I(v^h)$ *и* $I^*(v^h)$ *соответственно, так что уравнения виртуальной работы (уравнения Эйлера* $\delta I = \delta I^* = 0$*) принимают вид*

$$a(u^h, v^h) = b(f, v^h) \quad \text{и} \quad a^*(\tilde{u}^h, v^h) = b^*(f, v^h) \quad \text{для всех} \quad v^h \in S^h.$$

Тогда

$$a^*(u^h - \tilde{u}^h, u^h - \tilde{u}^h) = (a^* - a)(u^h, u^h - \tilde{u}^h) - (b^* - b)(f, u^h - \tilde{u}^h). \quad (8)$$

Доказательство. Левая часть тождества равна

$$a^*(u^h, u^h - \tilde{u}^h) - a^*(\tilde{u}^h, u^h - \tilde{u}^h) =$$
$$= (a^* - a)(u^h, u^h - \tilde{u}^h) + a(u^h, u^h - \tilde{u}^h) - a^*(\tilde{u}^h, u^h - \tilde{u}^h).$$

При $v^h = u^h - \tilde{u}^h$ в уравнениях виртуальной работы последние два члена дают $(b - b^*)(f, u^h - \tilde{u}^h)$, и доказательство закончено. Можно записать похожее тождество, заменив a, b и u^h на a^*, b^* и $\tilde{u}^h$, но оно не так полезно. Заметим также, что члены в a и b фактически уничтожаются в (8), но важно сохранить их и работать с разностями $a^* - a$ и $b^* - b$.

Из этого тождества непосредственно вытекает наша главная теорема.

Теорема 4.1. *Предположим, что приближенная энергия деформации положительно определена:* $a^*(v^h, v^h) \geqslant \theta \| v^h \|_m^2$, *и*

$$|(a^* - a)(u^h, v^h)| + |(b^* - b)(f, v^h)| \leqslant Ch^p \| v^h \|_m. \quad (9)$$

Тогда ошибка приближенного интегрирования в энергии деформации равна

$$\| u^h - \tilde{u}^h \|_m \leqslant \theta^{-1} Ch^p. \quad (10)$$

Доказательство. Левая часть тождества (8) ограничена снизу величиной $\theta \| u^h - \tilde{u}^h \|_m^2$, а правая часть ограничена

сверху согласно неравенству (9), где $v^h = u^h - \tilde{u}^h$. Сокращая на общий множитель, получаем (10).

Лемма и теорема не ограничиваются только численными квадратурами. Они также применимы *к изменению коэффициентов исходного дифференциального уравнения*. Другими словами, они описывают способ зависимости решения как непрерывной, так и дискретной задач от коэффициентов и неоднородного члена. Рассмотрим одномерную задачу $-(pu')' + qu = f$ и предположим, что p, q и f заменены на $\tilde{p}$, $\tilde{q}$ и $\tilde{f}$. Тогда тождество, примененное ко всему пространству $\mathscr{H}_E^1$ вместо подпространства S^h, превращается в

$$\int \tilde{p}\,(u' - \tilde{u}')^2 + \tilde{q}\,(u - \tilde{u})^2 =$$

$$= \int (\tilde{p} - p)\,u'\,(u' - \tilde{u}') + (\tilde{q} - q)\,u\,(u - \tilde{u}) - (\tilde{f} - f)\,(u - \tilde{u}). \qquad (11)$$

Левая часть, равная a^*, положительно определена, если $\tilde{p} > 0$ и $\tilde{q} \geqslant 0$. В правой части каждый член меньше произведения $\|u - \tilde{u}\|_1$ на постоянный множитель возмущения. Это дает простую оценку, не самую точную из возможных, результирующего возмущения в решении.

Следствие. *Предположим, что коэффициенты и неоднородный член возмущены менее, чем на* ε:

$$\max_x (|p - \tilde{p}|, |q - \tilde{q}|, |f - \tilde{f}|) < \varepsilon.$$

Тогда решение также возмущается на $O(\varepsilon)$:

$$\|u - \tilde{u}\|_1 < C\frac{\varepsilon}{\theta}. \qquad (12)$$

В терминах конечных элементов неравенство (12) допускает следующую интерпретацию. Предположим, что p, q и f заменены их интерполянтами в пространстве метода конечных элементов. Это возмущение — величина порядка h^k. Если результирующую задачу решить точно методом конечных элементов (в интегралах по элементарным областям появятся произведения *трех* полиномов), то на основании следствия $\|u^h - \tilde{u}^h\|_1 = O(h^k)$. Таким образом, интерполяция представляет собой возможную альтернативу численного интегрирования и ей уделяется львиная доля внимания в литературе по численному анализу; Дуглас и Дюпон [Д10] успешно исследовали даже нелинейные параболические задачи. В технических расчетах, однако, всегда предпочиталось непосредственное численное интегрирование; для изопараметрических элементов или элементов оболочек по существу вы-

бирать не приходится. Грубый подсчет числа операций наводит на мысль, что и в других задачах непосредственное применение квадратур эффективнее и потому мы намерены остановиться подробнее на этой технике.

Ситуация наиболее ясна, как обычно, на простейших примерах. Поэтому начнем с уравнения $-u'' = f$ и применим численное интегрирование к $I(v) = \int (v')^2 - 2fv$. Сначала рассмотрим требование положительной определенности:

$$a^*(v^h, v^h) = \sum w_i (v_x^h(\xi_i))^2 \geqslant \theta \int (v_x^h)^2 \, dx. \tag{13}$$

Интервал разбивается на элементы, другими словами, на подынтервалы, и на каждом элементе применяется обычная квадратурная формула. Если v^h — полином степени $k-1$, а квадратурные весовые коэффициенты w_i положительны, то требование положительной определенности сводится к следующему: *на подынтервалах должно быть по крайней мере* $k-1$ *точек интегрирования* ξ_i. В противном случае на каждом подынтервале найдется ненулевой полином степени $k-2$, равный нулю в каждой точке ξ_i. Объединяя эти полиномы и интегрируя полученную функцию, мы строим пробную функцию v^h, для которой не выполняется требование (13):

$$\int (v_x^h)^2 \, dx > 0, \quad \text{но} \quad v_x^h(\xi_i) = 0.$$

Если есть отрицательные веса, то для положительной определенности понадобится еще больше точек интегрирования. Мы не рассчитываем на популярность таких квадратурных формул; они к тому же чувствительны к ошибкам округления.

В двумерном случае для билинейных пробных функций на прямоугольниках условие положительной определенности не выполняется, если взять для квадратуры *правило средней точки*:

$$\int_{-h/2}^{h/2} \int_{-h/2}^{h/2} g(x, y) \, dx \, dy \sim h^2 g(0, 0).$$

Это одноточечная формула Гаусса на квадрате и она точна для полиномов $g = 1, x, y, xy$. Тем не менее *она не определена*: для пробной функции $v^h = xy$ численное интегрирование выражения $(v_x^h)^2 + (v_y^h)^2$ дает нуль. Если положить $v^h = +1$ или -1 в шахматном порядке на всем множестве узлов в Ω, то в результате получим высокую частоту осцилляции (простое кручение с наименьшей длиной волны $2h$, допускаемой сеткой), численная энергия которой равна нулю. Это находит отражение в дискретном приближении лапласиана, возникающего из правила средней

точки; типичная строка матрицы жесткости K дает

$$4\tilde{u}_{j,k} - \tilde{u}_{j+1,k+1} - \tilde{u}_{j-1,k+1} - \tilde{u}_{j+1,k-1} - \tilde{u}_{j-1,k-1}.$$

Это 5-точечная схема, повернутая на 45°, она, по-видимому, сильно неустойчива. На самом деле наше осциллирующее кручение дает решение однородного уравнения $\tilde{K}\tilde{Q} = 0$ [1]).

Аналогичная одноточечная формула с узлами в центре тяжести треугольников для линейных пробных функций *не* будет неопределенной. Деформации постоянны и не могут обратиться в нуль в центре, не будучи тождественно равны нулю. Фактически лапласиан порождает обычную 5-точечную схему.

Основная проверка определенности состоит в обнаружении пробных функций, которые при численном интегрировании теряют всю свою энергию деформации. Практически это выясняется из ранга матрицы жесткости элемента: если единственное нулевое собственное значение появляется от перемещений твердого тела, то квадратурная формула правильна. Если еще есть нулевые собственные значения, то квадратурная формула может все же быть приемлемой: надо проверить, можно ли собрать полиномы, грешащие на отдельных элементах, в пробную функцию v^h, обладающую слишком малой энергией на всей области (как в случае кручения, описанного выше). Например, четырехточечная формула Гаусса (2×2) не удовлетворяет нашему условию устойчивости для биквадратичных функций с девятью параметрами. Для гауссовых узлов $(\pm\xi, \pm\xi)$ на квадрате с центром в начале координат функция $(x^2 - \xi^2)(y^2 - \xi^2)$ имеет нулевую энергию деформации; этот шаблон можно передвигать и тогда трудности будут на всей области. (Матрица K на самом деле может не быть вырожденной, если эта схема не отвечает краевым условиям (скажем, $v^h = 0$) задачи. В этом случае можно рискнуть и испытать такую четырехточечную формулу интегрирования, даже если $\tilde{K}$ намного ближе к вырождению, чем позволено теорией.)

Вопрос об определенности становится довольно тонким для важных элементов с восемью параметрами, полученных из биквадратичных элементов исключением члена x^2y^2 и узла в центре каждого квадрата сетки. Так как пробные функции более не содержат функцию $(x^2 - \xi^2)(y - \xi^2)$, то четырех узлов Гаусса, по-видимому, достаточно для устойчивой аппроксимации уравнения Лапласа конечными элементами. Тейлор показал, однако, что для плоской задачи упругости с *двумя* зависимыми

[1]) В недавней работе Жиро проанализирована дискретная система, возникающая из этой квадратурной формулы. В некоторых ситуациях она оказывается удивительно полезной, хотя нарушает нашу гипотезу положительной определенности.

переменными ситуация иная: комбинация $u = x(y^2 - \xi^2)$, $v = -y(x^2 - \xi^2)$ лежит в пробном пространстве, а ее энергия деформации равна нулю для квадратурной формулы 2×2. Но Тейлор доказал также, что *эту схему нельзя продолжить в соседний элемент*. Ранг матрицы отдельного элемента слишком мал, но глобальная матрица жесткости абсолютно невырожденна и численное интегрирование дает правильные результаты.

Итак, хватит об определенности, это вопрос достаточного для интегрирования количества узлов. Теперь займемся точностью, определяемой полиномами, для которых квадратурная формула точна. Есть два пути развития теории. Один состоит в непосредственном вычислении показателя p, фигурирующего в уравнении (9) теоремы 4.1. Мы изложим этот подход в двух следующих абзацах, выписывая явные оценки (14)—(15) для ошибок численного интегрирования. Затем в заключительных абзацах этого раздела дадим более простое и четкое доказательство, непосредственно приводящее к связи между точностью квадратурной формулы и порядком результирующей ошибки $u^h - \tilde{u}^h$.

Первый подход. Для квадратурной формулы, точной степени q, ошибка численного интегрирования $\int g(x)\,dx$ ограничена величиной $Ch^{q+1}\int |g^{(q+1)}(x)|\,dx$. Это утверждение — точная копия теоремы 3.3 об аппроксимации, и доказывается оно таким же образом. Применяя его к неравенству (9) из теоремы 4.1, получаем

$$|(a^* - a)(u^h, u^h - \tilde{u}^h)| \leqslant Ch^{q+1} \sum_i \int_{e_i} \left| \left(\frac{d}{dx}\right)^{q+1} [p(x)\, u_x^h (u_x^h - \tilde{u}_x^h) + q(x)\, u^h (u^h - \tilde{u}^h)] \right| dx, \quad (14)$$

$$(b^* - b)(f, u^h - \tilde{u}^h)| \leqslant Ch^{q+1} \sum_i \int_{e_i} \left| \left(\frac{d}{dx}\right)^{q+1} [f(u^h - \tilde{u}^h)] \right| dx. \quad (15)$$

Предположим, что правая часть f гладкая, а также гладок переменный коэффициент $p(x)$. Тогда u — тоже гладкая функция, и таково же ее приближение u^h методом конечных элементов. Следовательно, единственными неконтролируемыми членами в правой части будут функция $u^h - \tilde{u}^h$ и ее производная. Каждое дифференцирование этих пробных функций может добавлять множитель h^{-1} [1]). Казалось бы, $q+1$ дифференцирований могут сократить множитель h^{q+1} и разрушить доказательство сходимости. *Это и есть то место, где существенно, что пробные*

[1]) Точнее, $|v^h|_{s+1} \leqslant Ch^{-1}|v^h|_s$ для всех s.

функции — полиномы. Так как степень полинома $u_x^h - \tilde{u}_x^h$ равна $k-2$, то это максимально возможное число множителей h^{-1}; дальнейшее дифференцирование аннулирует полином. Следовательно, порядок выражения в (14) равен $h^{(q+1)-(k-2)} \| u^h - \tilde{u}^h \|_1$. То же верно и для выражения в (15): первое дифференцирование дает $u_x^h - \tilde{u}_x^h$, и мы приходим к тому же рассуждению. (Мы еще вернемся к этому моменту и детально покажем, почему $b - b^*$ имеет тот же порядок, что и $a - a^*$, если применить одинаковые квадратурные формулы.) Таким образом, $p = q - k + 3$ в теореме 4.1, так что в результате численного интегрирования

$$\| u^h - \tilde{u}^h \|_1 \leqslant Ch^{q-k+3};$$

оценка совпадает с показателем $q - k + m + 2$, полученным выше. Для N-точечной гауссовой квадратурной формулы степень точности равна $q = 2N - 1$, а результирующая ошибка есть $O(h^{2N-k+2})$. Поэтому использование $k-1$ узлов в квадратурной формуле для одномерного случая очень удачно: их достаточно, чтобы удовлетворить требованию определенности, и они приводят к ошибке порядка h^k. Этот порядок даже ниже, чем у ошибки аппроксимации в деформациях.

В двумерном и трехмерном случаях принцип тот же. Для определенных полиномов степени $q+1$, скажем $x^\alpha y^\beta$, квадратурная формула не будет точна. В ошибке $a^* - a$ появляются соответствующие члены вида

$$Ch^{q+1} \sum_i \iint \left| \left(\frac{\partial}{\partial x}\right)^\alpha \left(\frac{\partial}{\partial y}\right)^\beta [D^m u^h \cdot D^m v^h] \right|, \qquad v^h = u^h - \tilde{u}^h.$$

Деформация $D^m u^h$ гладка, но каждое дифференцирование функции $D^m v^h$ может внести множитель h^{-1}. Условие сходимости состоит в том, что должно быть не более q этих множителей. Поэтому $q+1$ дифференцирований $(\partial/\partial x)^\alpha (\partial/\partial y)^\beta$ должны аннулировать $D^m v^h$ для любой пробной функции v^h. Другими словами, функция $D^m v^h$ не должна содержать членов $x^\alpha y^\beta$, для которых квадратурная формула неточна. Это как раз и есть данный ранее критерий сходимости (6), а именно m-е производные каждой пробной функции должны интегрироваться точно.

Второй подход. Предположим, что с помощью квадратурной формулы точно вычисляется интеграл от любого полинома степени $n - m$, умноженного на m-ю производную $D^m v^h$ любой пробной функции. Мы хотим показать, следуя [C9], что $p = n - m + 1$ в теореме 4.1; тогда ошибка численного интегрирования в деформациях будет того же порядка h^p. Для этого рассмотрим две величины $a - a^*$ и $b - b^*$ из неравенства (9).

Типичный член в $(a-a^*)(u^h, v^h)$ с коэффициентом $c(x, y)$, выражающим свойства материала, выглядит так:

$$\iint c(x, y) D^m u^h D^m v^h \, dx\, dy - \sum w_i c(\xi_i) D^m u^h(\xi_i) D^m v^h(\xi_i). \quad (16a)$$

Существенный момент, вытекающий из нашего условия $a(P_n, v^h) = a^*(P_n, v^h)$ на точность квадратурной формулы, заключается в том, что этот член можно переписать в виде

$$\iint (cD^m u^h - P_{n-m}) D^m v^h \, dx\, dy - \sum w_i (cD^m u^h - P_{n-m})(\xi_i) D^m v^h(\xi_i). \quad (16b)$$

Так как численное интегрирование производится поэлементно, то можно остановиться на определенном элементе и выбрать P_{n-m} возможно ближе к $cD^m u^h$. В соответствии с нашей теорией среднеквадратичной аппроксимации разность между ними будет порядка h^{n-m+1} (при условии достаточной гладкости $cD^m u^h$, а это будет, если гладки данные первоначальной задачи). Суммируя результаты по отдельным элементам, видим, что ошибка (16a) имеет правильный порядок h^{n-m+1}.

Небольшие изменения этого рассуждения приводят к тем же оценкам в энергии для любых производных порядка меньше m, а также для члена $b - b^*$. Мы опишем все шаги по оценке ошибки в последнем выражении, предполагая для простоты, что m-е производные пробных функций v^h состоят из всех полиномов некоторой степени t. Тогда наше условие точности квадратурных формул сводится к правильному интегрированию всех полиномов степени $n - m + t$. Следовательно,

$$(b - b^*)(f, v^h) = \iint f v^h \, dx\, dy - \sum w_i (f v^h)(\xi_i) =$$
$$= \iint (f v^h - P_{n-m+t}) \, dx\, dy - \sum w_i (f v^h - P_{n-m-t})(\xi_i).$$

Для правильно выбранных полиномов P в каждом элементе порядки этих величин равны произведению $h^{n-m+t+1}$ на интеграл от абсолютного значения производных вплоть до того же порядка от fv^h. Но функцию v^h можно дифференцировать только $t + m$ раз, после чего она (будучи полиномом) исчезает. Следовательно, предполагая, что f имеет $n - m + t + 1$ производных, получаем, что порядок ошибки $b - b^*$ равен

$$h^{n-m+t+1} \| v^h \|_{t+m} \leqslant h^{n-m+1} \| v^h \|_m.$$

И, наконец, удаление t дифференцирований из v^h оплачивается появлением множителя h^{-t}, как в примечании на стр. 223.

Таким образом, и $a - a^*$, и $b - b^*$ имеют правильный порядок h^{n-m+1}, а по теореме 4.1 таков же порядок и у ошибки в деформациях. Это основной результат настоящего раздела: *если* $a(P_n, v^h) = a^*(P_n, v^h)$, *то* $\|u^h - \tilde{u}^h\|_m = O(h^{n-m+1})$. Сиарле и Равьяр смогли показать [6], что даже при применении изопараметрического метода ошибка в перемещении обычно оказывается меньше ошибки в деформациях. (Их доказательство состоит в модификации приема Нитше.) Таким образом, порядок ошибки перемещения равен h^{n+1}, и теория численного интегрирования дает удовлетворительный результат: для сходимости необходимо, чтобы $n = m$, а условия $n = k - 1$ достаточно для сведения ошибок численного интегрирования к уровню ошибок аппроксимации полиномиальными пробными функциями.

4.4. АППРОКСИМАЦИЯ ОБЛАСТИ И КРАЕВЫХ УСЛОВИЙ

Одновременно с приближением допустимых функций из $\mathscr{H}_E^m$ кусочно полиномиальными в методе конечных элементов производятся и другие приближения, совершенно отличные от первых. Прежде всего можно менять саму область: Ω заменяется на близкий многоугольник Ω^h или, в изопараметрическом методе, на область с кусочно полиномиальной границей. Любое другое приближение произвольной области вызвало бы большие трудности. Далее, сами краевые условия служат объектом аппроксимации. Если в задаче указано, что $u = g(x, y)$ на Γ или $u_n + \alpha u = b(x, y)$, то эти функции g и b почти неизбежно интерполируются в узлах на границе Γ (или на ее приближении). Мы хотим оценить ошибку этих приближений.

В настоящем разделе подробно обсуждаются четыре проблемы:

1. Изменение области для однородного условия Дирихле $u = 0$ на Γ.
2. Изменение области для однородного условия Неймана $u_n = 0$ на Γ.
3. Аппроксимация неоднородного условия Дирихле $u = g(x, y)$ на Γ.
4. Аппроксимация произвольного неоднородного условия Неймана $u_n + \alpha(x, y)u = b(x, y)$ на Γ.

В каждом случае рассматривается двумерное уравнение второго порядка, скажем уравнение Пуассона $-\Delta u = f$. Большей частью оно берется для удобства и простоты описания: для нескольких неизвестных и трехмерного пространства изменения незначительны. Для чистой задачи Дирихле или Неймана высокого порядка, например для пластины с закрепленными или

свободными краями, порядок ошибки в энергии деформации такой же, как мы получим ниже.

Если комбинируются главные и естественные краевые условия, ситуация совершенно другая. Например, в задаче о *свободно опертой пластине* решается бигармоническое уравнение $\Delta^2 u = f$ для внутренней нагрузки, а коэффициент Пуассона ν входит в естественное краевое условие:

$$u = 0 \quad \text{и} \quad \nu \Delta u + (1 - \nu) u_{nn} = 0 \quad \text{на} \quad \Gamma. \tag{17}$$

В этой физически важной задаче, математически корректно поставленной, *нельзя гарантировать, что решение на многоугольнике* Ω^h *близко к решению и на* Ω, *если многоугольник* Ω^h *близок к* Ω. Это относится как к точному решению U^h на Ω^h, так и к аппроксимации u^h методом конечных элементов. Требуется нечто большее, чем простая сходимость границ.

Трудность легко увидеть, когда Ω^h — многоугольник. На каждой его стороне условие $U^h = 0$ сразу заставляет производную по касательной U^h_{tt} обратиться в нуль. Следовательно, второе краевое условие в (17) эквивалентно на прямолинейной стороне условию $\Delta U^h = 0$ и зависимость от ν исчезает. Бабушка [Б1] предложил ввести новое неизвестное $V^h = \Delta U^h$; это дает два уравнения второго порядка

$$\Delta U^h = V^h \quad \text{и} \quad \Delta V^h = f \quad \text{в} \quad \Omega^h, \quad U^h = V^h = 0 \quad \text{на} \quad \Gamma^h.$$

Для такой системы второго порядка сходимость *гарантируется*, а U^h, V^h приближают решения задачи

$$\Delta U = V \quad \text{и} \quad \Delta V = f \quad \text{в} \quad \Omega, \quad U = V = 0 \quad \text{на} \quad \Gamma.$$

Это как раз задача о закрепленной пластине с $\nu = 1$. Таким образом, *предельная функция не зависит от коэффициента Пуассона, входящего в краевые условия.* Сходимость есть, но почти всегда к неверному решению. Соответствующие трудности для расчетов методом конечных элементов представлены в [Р1] и обсуждаются в [Б10]. С другой стороны, мы предчувствуем успех изопараметрического метода, если аппроксимация границы Γ по крайней мере кусочно квадратична; в этом случае кривизна границы сходится. Если же предположить, что главное условие $u = 0$ заменяется в граничных узлах условием $u^h = \partial u^h / \partial t = 0$, использовать пространство Z_3 (см. разд. 1.9) и взять производную $\partial / \partial t$ вдоль истинной границы Γ, то сходимость можно ожидать даже на многоугольнике. В таком изложении, однако, требуемой теории не существует.

Вернемся к четырем проблемам, перечисленным выше. В каждом случае теория довольно сложна, но выводы делаются непосредственно.

1. Пусть Ω заменяется вписанным многоугольником Ω^h, а пробные функции приравниваются нулю на прямых сторонах границы Γ^h. Представим себе, что они доопределены нулем с внешней стороны границы Γ^h. Тогда эти функции допустимы для вариационной задачи: они равны нулю на истинной границе Γ, а пробное пространство S^h — настоящее подпространство в $\mathscr{H}_0^1(\Omega)$. Следовательно, основная теорема 1.1 метода Ритца гарантирует, что u^h минимизирует ошибку в энергии деформации:

$$\iint_{\Omega} (u-u^h)_x^2 + (u-u^h)_y^2 = \min_{S^h} \iint_{\Omega} (u-v^h)_x^2 + (u-v^h)_y^2. \qquad (18)$$

Так как каждая функция v^h равна нулю вне Ω^h, то интеграл на полосе $\Omega - \Omega^h$ фиксирован: это интеграл от $u_x^2 + u_y^2$. Поэтому u^h минимизирует не только на Ω, но и на Ω^h:

$$\iint_{\Omega^h} (u-u^h)_x^2 + (u-u^h)_y^2 = \min_{S^h} \iint_{\Omega^h} (u-v^h)_x^2 + (u-v^h)_y^2. \qquad (19)$$

Весь вопрос теперь просто в оценке выражений в (18) и (19). До сих пор мы выбирали функцию v^h равной интерполянту u_I.

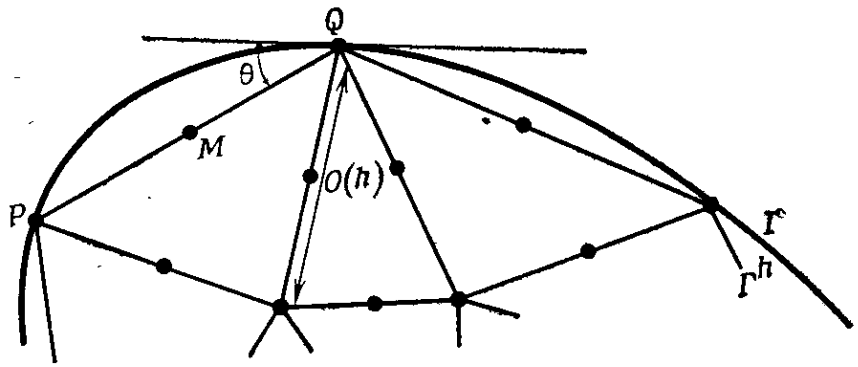


Рис. 4.2.
Аппроксимация границы многоугольниками.

Этот выбор все еще хорош, если S^h — пространство Куранта кусочно линейных функций с узлами в вершинах треугольников. Так как u_I интерполирует $u = 0$ в граничных узлах, то $u_I = 0$ вдоль всей границы многоугольника и $u_I \in S^h$. Стандартная теорема 3.3 об аппроксимации дает величину h^2 ошибки в энергии. Несомненно, что приближение около границы очень плохое.

Предположим, что используется более сложный элемент, например кусочно квадратичный (рис. 4.2). Как обычно, узлы расположены в вершинах и в серединах сторон. Каждая пробная функция будет равна нулю вдоль Γ^h при условии, что она равна нулю во всех граничных узлах, но *это исключает функцию* u_I *из пробного пространства* S^h. Истинное решение u равно

нулю в вершинах P и Q, но не в середине M, так что вдоль границы интерполянт отличен от нуля. Так как M находится на расстоянии $O(h^2)$ от истинной границы Γ, то по теореме о среднем значении

$$|u(M)| \leqslant Ch^2 \max_{\Omega} |\operatorname{grad} u|. \tag{20}$$

В (19) больше нельзя выбирать $v^h = u_I$. Вместо этого удобно в качестве v^h взять кусочно квадратичную функцию u_I^*, интерполирующую *и только во внутренних узлах* и равную нулю на Γ^h. При таком выборе ошибка в энергии деформации на Ω^h равна

$$\iint_{\Omega^h} (u-u_I^*)_x^2 + (u-u_I^*)_y^2 = |u-u_I+u_I-u_I^*|_1^2 \leqslant$$

$$\leqslant (|u-u_I|_1 + |u_I-u_I^*|_1)^2 \leqslant$$

$$\leqslant 2|u-u_I|_1^2 + 2|u_I-u_I^*|_1^2$$

(здесь мы применили неравенство треугольника). Член $|u-u_I|_1^2$ уже известен: по теореме 3.3 он равен $O(h^4)$. Новый член $u_I - u_I^*$ представляет собой кусочно квадратичную функцию, равную нулю во всех внутренних треугольниках: единственные узлы, где $u_I \neq u_I^*$, лежат в серединах M_i сторон на границе Γ^h; в этих точках $u_I^* = 0$, $u_I = u = O(h^2)$. Таких сторон $O(1/h)$, так что

$$|u_I - u_I^*|_1^2 = \sum \iint_{e_i} |u(M_i)|^2 (\varphi_x^2 + \varphi_y^2)\, dx\, dy \leqslant$$

$$\leqslant O\left(\frac{1}{h}\right) c^2 h^4 \max |\operatorname{grad} u|^2 \iint (\varphi_x^2 + \varphi_y^2)\, dx\, dy.$$

Здесь φ — квадратичная функция, равная 1 в среднем узле M_i и нулю в 5 других узлах треугольника. Ее первые производные имеют порядок $1/h$, а площадь треугольника — порядок h^2; окончательно получаем

$$|u_I - u_I^*|_1^2 = O(h^3).$$

Теорема 4.2. *Ошибка в энергии на Ω^h при аппроксимации области многоугольником удовлетворяет неравенству*

$$|u-u^h|_1^2 \leqslant |u-u_I^*|_1^2 = O(h^3). \tag{21}$$

Ошибка в значении функции есть $u - u^h = O(h^2)$.

Последняя оценка (которую мы не будем доказывать) вытекает из непрерывной задачи на приближенной области: $-\Delta U^h = f$ на Ω, $U^h = 0$ на Γ^h. Так как $\Delta(u - U^h) = -f + f = 0$, то по принципу максимума $u - U^h$ достигает максимума на Γ^h.

Но на этой границе $U^h = 0$, а $u = O(h^2)$, так что порядок ошибки $u - U^h$ везде равен h^2. В теореме утверждается, что тот же результат верен для $u - u^h$ [Б19].

Отсюда следует, что если судить лишь по показателю степени h и приближать Ω многоугольником, то при вычислении деформации нет смысла выходить за пределы квадратичных функций формы. Более того, для достижения наилучшего из возможных порядков h^2 ошибки в перемещениях достаточно даже линейных полиномов. Возможно, это тот самый случай, когда показатель степени h недостаточен для отражения истинной точности: действительная ошибка в вычислениях с линейными элементами может оказаться чрезмерной.

Подобная оценка верна и для трехмерной задачи Дирихле более высокого порядка. Самое необычное, что показатель должен быть нечетным; средняя ошибка в производных от $u - u^h$ имеет *дробный* порядок $h^{3/2}$. Сама оценка тем не менее верна, и в действительности точное решение U^h уравнения на многоугольнике (которое ближе к u, чем u^h, потому что минимизирует функционал на пространстве $\mathscr{H}_0^1(\Omega^h)$, содержащем S^h) тоже имеет ошибку порядка $h^{3/2}$ [Б9]. При аппроксимации границы прямоугольниками, что намного грубее, порядок становится равным $h^{1/2}$; вычислительные результаты совершенно неудовлетворительны. В этом и заключается причина использования треугольных элементов.

Объясняется дробный показатель 3/2 так: имеется *пограничный слой* толщиной в два элемента, внутри которого ошибки в производных равны $O(h)$. Легко проверить, что угол θ на рис. 4.2 есть величина того же порядка, так что истинная производная от u вдоль хорды равна $O(h)$, а не 0. Этот слой в ошибку в энергии вносит $O(h^3)$. За пределами пограничного слоя ошибка совершенно другая: оптимальный порядок h^2 для ошибки в перемещениях, т. е. $u - u^h = O(h^2)$, и она настолько гладка, что ее первые производные также порядка h^2. Ошибка ведет себя внутри области намного лучше, чем у границы, демонстрируя тем самым свойства гладкости, присущие всем эллиптическим задачам. В аппроксимации методом конечных элементов $u - u^h$ вдоль каждой хорды PMQ изменяется от нуля до $O(h^2)$ и обратно; это особое поведение быстро стушевывается в направлении нормали к границе, что хорошо наблюдается в расчетах и проверяется непосредственно. Это вариант *принципа Сен-Венана*, относящегося больше к геометрии, чем, что более привычно, к краевым данным. Закон одинаков: *вдали от границы важно поведение в среднем, а не локальные коротковолновые осцилляции.*

Ситуация не отличается от отображения единичного круга в круг с волновой границей. Предположим, что последняя описывается функцией $R(\theta) = 1 + h^2 \cos \theta/h$. Тогда при конформном отображении между этими «кругами» точка (r, θ) переходит в $(r + h^2 r^{1/h} \cos \theta/h, \theta)$. Вся энергия заключена внутри пограничного слоя $r \geqslant 1 - Ch$; для малых r амплитуда члена «волновых» возмущений экспоненциально мала:

$$h^2 r^{1/h} \leqslant h^2 (1 - Ch)^{1/h} \sim h^2 e^{-C}.$$

Производные от отображения равны $O(h)$ около границы. Внутри они фактически равны нулю из-за множителя $r^{1/h}$. Заметим, что площади обоих кругов одинаковы; если бы один из них был вписан в другой, то появился бы дополнительный член rh^2, производные от которого не исчезают, но всюду в области имеют меньший порядок h^2. В терминах принципа Сен-Венана усреднение по локальным осцилляциям отлично от нуля и распространяется далее. Более того, если вместо волн $\cos \theta/h$ у круга были бы зубцы $|\cos \theta h|$, то конформное отображение обладало бы слабыми особенностями в местах стыка. Однако при аппроксимации методом конечных элементов эти особенности смазываются и средняя ошибка в производных имеет порядок h у границы и h^2 внутри.

Известно несколько путей достижения оптимальной скорости сходимости для квадратичных функций. Уже упоминались преобразование координат изопараметрическими элементами и использование $x; y$ элементов Митчелла с кусочно гиперболическими (!) функциями в качестве границы. Еще одна возможность — вычислить поправки к аппроксимации u^h многоугольниками [Б23] или модифицировать исходный функционал $I(v)$ [Н3]. Во всех этих способах несущественно, что Ω^h лежит внутри Ω; фактически ошибка от изменения области частично гасится, если Γ^h систематически проходит то внутри, то снаружи Г. Снова по принципу Сен-Венана главный член ошибки зависит от площади $\Omega - \Omega^h$ (с алгебраическим знаком) и предоставляет следующую возможность: усреднить приближенное решение для вписанного и описанного многоугольников. Все эти предложения, за исключением изопараметрического метода, в практических задачах в основном не проверены.

2. Рассмотрим уравнение $-\Delta u + qu = f$ с естественным краевым условием $u_n = 0$. Здесь не возникает вопрос о наложении условий на неверную границу: пробные функции на границе не подчинены никаким ограничениям. Однако область подвергается изменению, если неудобно проводить интегрирование по истинной области Ω. Если потенциальная энергия вычисляется по приближенной области Ω^h, то это значит, что вводится

новое определение потенциальной энергии:

$$I^h(v) = a^h(v, v) - 2(f, v)^h = \iint_{\Omega^h} (v_x^2 + v_y^2 + qv^2 - 2fv)\, dx\, dy.$$

Если предположить, что u^h минимизирует истинный функционал I по S^h, а $\tilde{u}^h$ минимизирует I^h, то задача сведется к оценке $e^h = u^h - \tilde{u}^h$.

С математической точки зрения это по существу тот же вопрос, что возникает для квадратур Гаусса: интегралы вычисляются не точно. Поэтому мы применяем вариант тождества леммы 4.1; непосредственно из равенства нулю первых вариаций для u, u^h и $\tilde{u}^h$ вытекает, что

$$E^2 = a^h(e^h, e^h) = a^h(u^h - u, e^h) + (a^h - a)(u, e^h) + (f, e^h) - (f, e^h)^h.$$

Первое слагаемое в правой части по неравенству Шварца не превышает

$$[a^h(u^h - u, u^h - u) - a^h(e^h, e^h)]^{1/2} \leqslant C_1 h^{k-1} E.$$

Остальные дают интеграл по полосе $\Omega - \Omega^h$:

$$B = \iint_{\Omega - \Omega^h} (-u_x e_x^h - u_y e_y^h - que^h + fe^h)\, dx\, dy.$$

Предполагая, что u_x, u_y, q и f ограничены, и снова применяя неравенство Шварца, получаем

$$|B| \leqslant C_2 A^{1/2} \left[\iint_{\Omega - \Omega^h} (e_x^h)^2 + (e_y^h)^2 + (e^h)^2 \right]^{1/2},$$

где A — площадь полосы $\Omega - \Omega^h$. Осталось оценить последний интеграл с помощью E, т. е. связать размер элемента e^h в полосе с его размером внутри области. Для произвольных функций это было бы невозможно. Однако функция $u^h - \tilde{u}^h$ никоим образом не произвольна: в каждом треугольнике она совпадает с некоторым полиномом, и эти полиномы не могут вдруг взорваться — оценка на Ω^h влечет за собой оценку на Ω.

Обозначим через T типичный криволинейный треугольник у границы, а через T^h — вписанный в него прямолинейный треугольник. Их разность $T - T^h$ служит одним из кусков полосы $\Omega - \Omega^h$. Пробные функции представляют собой полиномы на каждом треугольнике T и нам надо оценить ошибку $e^h = u^h - \tilde{u}^h$, вызванную интегрированием только по T^h. Эта ошибка на каждом треугольнике T сама будет полиномом $P(x, y)$, и для нее справедлива лемма, принадлежащая Бергеру.

Лемма 2.2. *Пусть* $\rho =$ площадь $(T - T^h)$/площадь (T^h). *Тогда существует такая постоянная* c, *зависящая только от степени полинома* $P(x, y)$, *что*

$$\iint\limits_{T-T^h} P_x^2 + P_y^2 + P^2 \leqslant c\rho \iint\limits_{T^h} P_x^2 + P_y^2 + P^2.$$

Суммирование по всем граничным треугольникам и увеличение правой части включением в рассмотрение внутренних узлов приводят неравенство леммы к неравенству

$$\iint\limits_{\Omega-\Omega^h} (e_x^h)^2 + (e_y^h)^2 + (e^h)^2 \leqslant c\rho E^2.$$

Подставляя это в оценку для $|B|$, получаем

$$E^2 \leqslant [C_1 h^{k-1} + C_2 (Ac\rho)^{1/2}] E.$$

Другими словами, ошибка в энергии деформации, возникающая при интегрировании по Ω^h вместо Ω, удовлетворяет неравенству

$$E^2 \leqslant C_3 (h^{2(k-1)} + A\rho).$$

Первое слагаемое отражает энергию деформации $u - u^h$ и не добавляет ничего нового. Интерес представляет второе, чисто геометрическое слагаемое. Предположим, например, что Ω^h — многоугольник. Тогда площадь A равна $O(h^2)$, а отношение ρ для площадей соседних треугольников равно $O(h)$. Таким образом, *ошибка энергии деформации при замене области многоугольником имеет один и тот же порядок* h^3 *как для естественного, так и для главного краевого условия.* Мы думаем, что снова есть пограничный слой. Для изопараметрических, а не субпараметрических элементов $A = O(h^k)$ и $\rho = O(h^{k-1})$; их произведение h^{2k-1} поглощается обычной ошибкой аппроксимации h^{2k-2} в энергии деформации. Поэтому изопараметрические элементы должны быть удачны на криволинейных областях независимо от того, какие наложены краевые условия — главные или естественные.

3. Следующая проблема связана с неоднородным главным краевым условием $u = g(x, y)$ на Γ. Это условие выполняется для каждого элемента v истинного допустимого пространства $\mathscr{H}_E^1$. Поэтому две любые допустимые функции отличаются на функцию $v_0 \in \mathscr{H}_0^1$, значения которой на границе равны нулю.

Предположим, что ситуация та же для пробных функций $v^h \in S^h$: все пробные функции принимают одинаковые значения на границе Γ (необязательно $v^h = g$; нельзя ожидать от полиномов слишком многого, пусть, например, $v^h = g^h$). Тогда две лю-

бые пробные функции отличаются на функцию v_0^h, равную нулю на Г. Эти функции v_0^h образуют пространство S_0^h, являющееся подпространством истинного однородного пространства $\mathscr{H}_0^1$.

Теорема 4.3. *Пусть u минимизирует $I(v)$ на $\mathscr{H}_E^m$, а u^h минимизирует $I(v)$ на S^h. Тогда равенство нулю первых вариаций выражается формулами*

$$a(u, v_0) = (f, v_0) \quad \text{для всех} \quad v_0 \in \mathscr{H}_0^1, \tag{22}$$

$$a(u^h, v_0^h) = (f, v_0^h) \quad \text{для всех} \quad v_0^h \in S_0^h. \tag{23}$$

Как и в теореме 1.1, *u^h обладает дополнительным минимизирующим свойством*

$$a(u - u^h, u - u^h) = \min_{v^h \in S^h} a(u - v^h, u - v^h). \tag{24}$$

Доказательство такое же, как для теоремы 1.1. Так как u и u^h минимизируют функционал I, любые возмущения εv_0 и εv_0^h увеличивают его:

$$I(u) \leqslant I(u + \varepsilon v_0) \quad \text{и} \quad I(u^h) \leqslant I(u^h + \varepsilon v_0^h).$$

Раскрывая формулу, видим, что коэффициенты при ε должны равняться нулю, а это и есть уравнения виртуальной работы (22) и (23).

Для доказательства того, что u^h минимизирует ошибку в энергии деформации, запишем

$$\begin{aligned} a(u - v^h, u - v^h) &= a(u - u^h + u^h - v^h, u - u^h + u^h - v^h) = \\ &= a(u - u^h, u - u^h) + 2a(u - u^h, u^h - v^h) + \\ &\quad + a(u^h - v^h, u^h - v^h). \end{aligned}$$

Равенство нулю среднего слагаемого получаем автоматически, вычитая (23) из (22) и полагая $v_0^h = u^h - v^h$. Последнее слагаемое будет положительным, если $v^h \neq u^h$. Таким образом, в этой точке достигается минимум, и теорема доказана.

Следствие. *Пусть Ω — многоугольник, а главное условие $u = g(x, y)$ на Г интерполируется в методе конечных элементов: во всех граничных узлах пробные функции удовлетворяют равенству $v^h(z_j) = g(z_j)$ (или, более общо, $D_j v^h(z_j) = D_j g(z_j)$). Тогда*

$$a(u - u^h, u - u^h) \leqslant a(u - u_I, u - u_I). \tag{25}$$

Поэтому оценка ошибки в методе конечных элементов сводится к обычной оценке аппроксимации $u - u_I$.

Доказательство. Надо проверить следующие условия:

а) Интерполянт u_I должен принадлежать S^h, так что в (24) возможен выбор $v^h = u_I$. Поскольку u_I удовлетворяет краевым условиям, налагаемым в следствии (т. е. $u_I(z_j) = u(z_j) = g(z_j)$ в граничных узлах), это требование выполнено.

б) Все пробные функции v^h должны принимать одинаковые значения на границе Г, так что их разности принадлежат $\mathscr{H}_0^1$. Другими словами, пробные функции v^h должны определяться на Г своими значениями в граничных узлах. Так как Г состоит из плоских сторон и мы предполагаем, что используется согласованный элемент, то условие действительно будет выполнено.

Мы колебались, давать ли обычную оценку $h^{2(k-1)}$ для (25), поскольку такая оценка требует, чтобы решение было гладким ($u \in \mathscr{H}^k$). В углах многоугольных областей почти автоматически появляются особенности в производных от u. Для уравнений второго порядка решение u обычно не принадлежит $\mathscr{H}^{1+\pi/\alpha}$, где α — наибольший внутренний угол. Первые производные будут, действительно, аппроксимироваться до порядка $h^{\pi/\alpha}$, а энергия деформации — до порядка $h^{2\pi/\alpha}$. В гл. 8 мы рассмотрим, как сохранить обычный порядок $h^{2(k-1)}$, измельчая сетку в углах или (еще лучше) вводя специальные пробные функции с правильными особенностями.

Предположим, что область Ω, не являющаяся многоугольником, заменяется многоугольником Ω^h либо непосредственно в плоскости x, y, либо в плоскости ξ, η после изопараметрического преобразования. Если вычисления произведены на Ω^h, то опять граничные узлы полностью определяют все краевые значения и можно применить следствие: окончательная ошибка поглощается суммой ошибки в $u - u_I$ на приближенной области и *изученной ранее ошибки, вызванной изменением области.*

Остается еще возможность работать на криволинейной области Ω в заданных координатах x, y и интерполировать $v^h = g$ в граничных узлах. Вероятно, интегрирование следует проводить численно, особенно на криволинейных элементах у границы, хотя эксперименты в [K14] обрабатывались точными алгебраическими операциями. В этом методе краевые значения меняются от одной пробной функции к другой, кроме, разумеется, значений в самих узлах. Разности между пробными функциями малы, но на Г отличны от нуля. Поэтому законы Ритца нарушаются и возникают теоретические вопросы:

а) Как велики могут стать эти разности на границе?

б) Отражается ли на ошибке $u - u^h$ это наихудшее из возможных поведение на Г, или u^h все еще оптимальное в S^h приближение к u?

Вопросы относятся даже и к однородным краевым условиям $u = g = 0$ на Г. В этом случае пробные функции равны нулю в граничных узлах и в а) спрашивается, насколько велики они могут быть на остальной части границы Г. Удивительно, но ответ не зависит от их степени. Максимальная величина v^h на Г одинакова и для линейной функции v^h, равной нулю только в вершинах граничных треугольников, и для квадратичных или кубических функций, обращающихся в нуль в одном или двух дополнительных узлах на каждом куске границы Г. Она одинакова даже для пространства кубических функций Z_3 в граничных вершинах, обращающихся в нуль вместе со своими истинными производными по касательным. Ответ таков: *существует особая пробная функция V^h с единичной энергией, среднее значение которой на* Г *имеет порядок* $h^{3/2}$:

$$\iint_{\Omega} (V_x^h)^2 + (V_y^h)^2 = 1, \quad ch^3 \leqslant \int_{\Gamma} |V^h|^2 ds \leqslant Ch^3.$$

Для ответа на вопрос б), т. е. для нахождения оценки для $u - u^h$, необходимо расширить классическую теорию Ритца. Верно, что первые вариации все еще обращаются в нуль для минимизирующих функций u и u^h:

$$a(u, v) = (f, v) \quad \text{для} \quad v \in \mathscr{H}_0^1, \quad a(u^h, v^h) = (f, v^h) \quad \text{для} \quad v^h \in S^h.$$

Однако S^h даже при $g = 0$ не будет подпространством в $\mathscr{H}_0^1$; в неоднородном случае $g \neq 0$ и S_0^h не содержится в $\mathscr{H}_0^1$. Тем не менее еще можно привлечь формулу Грина: так как $-\Delta u = f$, то

$$a(u, v^h) = \iint_{\Omega} u_x v_x^h + u_y v_y^h = \iint_{\Omega} f v^h + \int_{\Gamma} u_n v^h \, ds.$$

При вычитании $a(u^h, v^h) = (f, v^h)$ остаются граничные слагаемые

$$a(u - u^h, v^h) = \int_{\Gamma} u_n v^h \, ds \quad \text{для всех} \quad v^h \in S^h. \tag{26}$$

Это и есть тот член, который контролирует ошибку, вызванную нарушением главных краевых условий.

Мы хотим показать, что ответ на вопрос б) самый худший из возможных: порядок ошибки будет определяться пробной функцией V^h, наибольшей на Г. В самом деле, для $v^h = V^h$ в (26) это легко показать. Так как V^h имеет единичную энергию, то левая часть, согласно неравенству Шварца, ограничена квадратным корнем из $a(u - u^h, u - u^h)$. В правой части V^h принимает среднее значение $h^{3/2}$ и, если нет специальных упрощений

(см. далее!), следует ожидать, что интеграл от $u_n V^h$ будет того же порядка. Поэтому

$$a(u-u^h,\ u-u^h)^{1/2} \sim h^{3/2}.$$

Это значит, что интерполирование краевых условий и интегрирование по Ω даже для элемента высокого порядка не лучше, чем замена Ω многоугольником Ω^h. (По крайней мере показатель у h не лучше, а оценок привлекаемых констант у нас нет.) Одно объяснение таково: каждый полином, интерполирующий $u=0$ в граничных узлах, будет равен нулю вдоль кривой, проходящей близко от истинной границы, но две кривые могут отличаться на $O(h^2)$ — то же расстояние, что и для многоугольника Ω^h.

В [Б9] показано, что $h^{3/2}$ служит также *верхней* границей для ошибки интерполяции краевых условий. Из преобразований, описанных в этой статье, всплывает второй, еще более удивительный факт: даже если точность ограничена существованием нежелательной функции V^h порядка $h^{1/2}$ на Γ, *ошибка истинного перемещения* $u-u^h$ *на самом деле имеет порядок* h^3 *на* Γ. Другими словами, краевые значения решения Ритца выглядят обманчиво хорошими. Наибольшей ошибкой должны обладать нормальные производные у границы.

Теперь исследуем возможность упрощений в интеграле $\int u_n v^h\,ds$ в правой части равенства (26). Если краевые значения v^h осциллируют около нуля, то даже, хотя их средний модуль $|v^h|$ может быть порядка $h^{3/2}$, сам интеграл будет стремиться к величине меньшего порядка. *Эти осцилляции происходят для некоторых элементов, но не для всех.* Например, они бывают у квадратичных функций, когда условие $u=0$ (или $u=g$) интерполируется в граничных вершинах и в точках границы Γ, расположенных посередине между ними. Главный член в v^h представляет собой кубическую функцию $s(s-h/2)(s-h)$ от длины дуги на Γ, равную нулю в граничных узлах, ее среднее значение равно нулю. Другими словами, v^h осциллирует около нуля. В соответствии с этим Скотт показал, что ошибка в энергии, вызванная интерполяцией $u=g$, улучшается до $O(h^{5/2})$. (Это не так для кубических функций [Б9], если только узловые точки не расположить специальным образом на Γ, чтобы среднее значение v^h сделать нулевым.) Так как обычная ошибка аппроксимации для квадратичных функций равна $O(h^2)$ в энергии, граничная ошибка теперь оказывается совершенно приемлемой. Мы не знаем, пригодится ли в будущем этот подход в изопараметрическом методе для граничных треугольников, а именно интегрирование непосредственно по криволинейным элементам в плоскости x, y.

4. Последняя проблема в нашем списке возникает из неоднородного естественного условия

$$u_n + \alpha(x, y)\, u = b(x, y) \quad \text{на} \quad \Gamma, \quad \alpha \geqslant 0.$$

Для уравнения Пуассона это приводит к появлению граничных интегралов в потенциальной энергии:

$$I(v) = \iint_{\Omega} v_x^2 + v_y^2 + \int_{\Gamma} \alpha v^2 - 2 \iint_{\Omega} fv - 2 \int_{\Gamma} bv.$$

Строго говоря, для этих граничных интегралов надо исследовать три вопроса: полиномиальную аппроксимацию на Г, численное интегрирование и изменение области. Два последних вопроса возникают потому, что для криволинейной границы Г вычислить точно граничные интегралы почти невозможно. Некоторые численные процедуры можно приспособить, а когда пробные функции определены на Ω^h вместо Ω, интегралы переносятся на Γ^h. Мы хотим опустить подробный анализ этих ошибок на Γ^h, так как убеждены в том, что они не больше тех, которые уже изучены на Ω^h. Очень хотелось бы иметь строгое доказательство нашего убеждения.

Новый вопрос связан с полиномиальной аппроксимацией на Г. Он возникает обычным образом: метод минимизирует ошибку в энергии:

$$a(u - u^h, u - u^h) = \min \iint_{\Omega} (u - v^h)_x^2 + (u - v^h)_y^2 + \int_{\Gamma} \alpha (u - v^h)^2 \, ds.$$

Для пространства метода конечных элементов степени $k - 1$ и выбора $v^h = v_I$ интеграл по Ω имеет порядок $h^{2(k-1)}$. К счастью, интеграл по Г имеет даже более высокий порядок, *скорость сходимости не уменьшается от присутствия граничных интегралов.* Это очевидно для границы Г, образованной прямыми: сужение пробных функций на Г дает полный полином степени $k - 1$ от граничной переменной s, а интеграл по Г имеет порядок h^{2k}. Нитше получил такой же результат для криволинейной границы [6].

Для изопараметрических элементов граница в плоскости ξ, η прямолинейна и все граничные интегралы вычисляются непосредственно численным интегрированием. В действительности основное заключение теории для краевой задачи второго порядка, по-видимому, таково: *изопараметрический метод устанавливает локальное преобразование координат в направлении нормали и по касательной, точнее и удобнее того, которое достигалось конечно-разностным методом.*

5 УСТОЙЧИВОСТЬ

5.1. НЕЗАВИСИМОСТЬ БАЗИСА

В определенном смысле здесь не должно быть проблемы устойчивости. Решение эллиптической вариационной задачи зависит непрерывным образом от исходных данных; если нагрузка f и все принятые на границе перемещения и силы малы, то энергия деформации в u мала. Иными словами, задача корректно поставлена. Кроме того, *независимо от выбора подпространства* S^h энергия деформации в приближении Ритца u^h автоматически ограничена энергией деформации в u; метод Ритца проектирует u на S^h, и это может лишь уменьшить энергию (следствие из теоремы 1.1). Поэтому приближенные задачи *равномерно корректно поставлены* и всегда должна быть возможность построения «устойчивого» алгоритма для численного расчета u^h.

Трудность заключается в том, что для достижения полной численной устойчивости при стремлении шага сетки к нулю требуемый алгоритм для решения возникающей системы уравнений может оказаться просто слишком сложным. Для обычной пятиточечной разностной схемы можно систематически использовать все ограничения на матрицу коэффициентов, вытекающие из согласованности с оператором Лапласа: суммы вдоль каждой строки матрицы, а также первые моменты равны нулю на всех стадиях метода исключения Гаусса. Однако соответствующие ограничения для нерегулярных конечных элементов будет чрезвычайно трудно использовать. Поэтому мы будем исследовать обычный алгоритм исключения, допуская увеличение ошибки округления при $h \to 0$, однако имея в виду, что численная устойчивость должна быть по возможности понятной; излишние численные неустойчивости приниматься во внимание не будут.

Известно, что ключ к устойчивости лежит в *равномерной линейной независимости элементов базиса* φ_j. Даже если u^h совершенно не зависит от выбора базиса, округление, которое входит в вычисленное значение $\bar{u}^h$, зависит от него. (С. Г. Михлин ввел понятие *сильной минимальности* базиса, и советские авторы часто исследуют с этой точки зрения численные методы, а также и устойчивость относительно изменений коэффициентов в дифференциальном уравнении.) Обычная процедура определения линейной независимости базиса состоит в рассмотрении *матрицы*

Грама (в физических терминах матрицы массы), элементы которой есть скалярные произведения элементов базиса:

$$M_{jk} = (\varphi_j, \varphi_k) = \int_\Omega \varphi_j(x)\,\varphi_k(x)\,dx_1 \ldots dx_n.$$

Так как метод Ритца всегда оперирует с энергией $a(v, v)$, присущей задаче, то во многих случаях мы будем предпочитать работать с матрицей жесткости: $K_{jk} = a(\varphi_j, \varphi_k)$. Обе матрицы эрмитовы и положительно определены.

В качестве первой меры независимости базиса мы предлагаем *число обусловленности*

$$\varkappa(M) = \frac{\lambda_{\max}(M)}{\lambda_{\min}(M)}.$$

Если бы базис был ортонормальным, то матрица M была бы единичной и $\varkappa = 1$. Это не так для конечных элементов, но важно то, что на равномерной сетке *базисные функции метода конечных элементов равномерно линейно независимы*: $\varkappa(M) \leqslant \leqslant \text{const}$. Другими словами, все собственные значения матрицы M одного и того же порядка. Как заметил Шульц, при аппроксимации по методу наименьших квадратов (который есть не что иное, как метод Ритца, примененный к дифференциальному уравнению нулевого поряка $u = f$) кусочные полиномы значительно более устойчивы, чем последовательность $1, x, y, x^2, \ldots$ обычных полиномов. Число обусловленности матрицы массы, соответствующей этой последовательности и являющейся матрицей Гильберта (1.6), растет по экспоненциальному закону.

Существуют приложения, в которых в качестве более реальной меры независимости можно взять *оптимальное число обусловленности*

$$c(M) = \min_D \varkappa(DMD).$$

Здесь D — любая положительная диагональная матрица, соответствующая *изменению масштаба* элементов базиса; $c = 1$, если исходный базис лишь ортогонален, но не ортонормален. При работе с нерегулярными элементами одни пробные функции могут быть намного меньше других, и изменение масштаба может привести к значительному уменьшению числа обусловленности. Мы смотрим на изменение масштаба так: *если число обусловленности для M или K улучшается при изменении масштаба, это позволяет предположить, что численные трудности, которые тем самым устраняются, вероятно, никогда не распространяются на всю задачу.* Изменение масштаба устраняет *локальные* трудности; если конкретная функция φ_j плоха из-за масштаба, так что, например, диагональный элемент K_{jj} слишком мал, то ошибки округления разрушат доверие к вычисленному значению весового

коэффициента Q_j. Измеряемый эффект зависит от выбора нормы; так как функция φ_j мала, такая ошибка округления не будет означать большой ошибки в энергии, но в узле z_j аппроксимация может быть сравнительно плохой.

Обычное правило при масштабировании разреженных положительно определенных матриц заключается в том, что *все диагональные элементы должны быть одинаковыми*: в случае конечных элементов, в первую очередь регулируемом матрицей жесткости, это означает, что энергии деформации K_{jj} в элементах базиса равны. Это правило дает диагональную масштабирующую матрицу D, которая почти оптимальна [Ш3]. Вопрос масштабирования возникает даже в случае равномерной сетки для конечных элементов эрмитова типа, когда среди неизвестных Q_j появляются как значения функции, так и производные. Возможно, естественная процедура состоит в том, чтобы неизвестные сохраняли правильные размерности за счет вращения $\theta = hv_j'$, а не $\theta = v_j'$.

Фрид [Ф16] заметил, что в одних задачах можно с успехом изменять масштаб, а в других — нельзя. Возьмем, например, двухточечную краевую задачу $-u'' = f$ с кусочно линейными элементами. Если все подынтервалы имеют длину h, за исключением первого, у которого она равна h/c, то матрицы жесткости при естественном краевом условии $u'(0) = 0$ или главном условии $u(0) = 0$ пропорциональны соответственно матрицам

$$K_{\text{ест}} = \begin{pmatrix} c & -c & & \\ -c & 1+c & -1 & \\ & -1 & 2 & \cdot \\ & & \cdot & \cdot \end{pmatrix} \quad \text{или} \quad K_{\text{глав}} = \begin{pmatrix} 1+c & -1 & \\ -1 & 2 & \cdot \\ & \cdot & \cdot \end{pmatrix}.$$

Наибольшее собственное значение растет с ростом c, тогда как наименьшее имеет обычный порядок N^{-2}, где N — порядок матрицы. Поэтому обусловленность ухудшается при $c \to \infty$.

Масштабирование приводит к тому, что все диагональные элементы становятся равными 2, а нижний конец матриц не изменяется:

$$DK_{\text{ест}}D = \begin{pmatrix} 2 & -2\sqrt{c/1+c} & -\sqrt{2/1+c} & \\ -2\sqrt{c/1+c} & 2 & 2 & -1 \\ & -\sqrt{2/1+c} & -1 & \cdot \end{pmatrix},$$

$$DK_{\text{глав}}D = \begin{pmatrix} 2 & -\sqrt{2/1+c} & \\ -\sqrt{2/1+c} & 2 & -1 \\ & -1 & \cdot \end{pmatrix}.$$

Наибольшие собственные значения теперь ограничены; трудный вопрос всегда — величина $\lambda_{\min}$. В рассматриваемом случае вторая матрица не доставляет хлопот: c очень слабо влияет на значение $\lambda_{\min}$, которое снова оказывается порядка N^{-2}. Однако первая матрица начинается с блока порядка 2, который при больших c почти вырожден, и наименьшее собственное значение всей матрицы обязательно меньше собственных значений этого блока. Таким образом, $\lambda_{\min} \to 0$ при $c \to \infty$. Итак, изменение масштаба успешно при главном краевом условии, но не при естественном.

Это численный аналог хорошо изученной физической ситуации: жесткая система, соединенная с землей посредством мягкой пружины, чрезвычайно неустойчива, в то время как неподвижное соединение (главное условие) дает устойчивую систему. Следует подчеркнуть, что в то время, как плохую обусловленность, связанную с числами, можно предвидеть, а вырождение элементов можно избежать, возникающую *физическую плохую обусловленность* нельзя обойти, разве лишь заменяя метод жесткостей на метод сил с напряжениями в качестве неизвестных. Эта ситуация встречается при резком изменении жесткости среды или когда коэффициент Пуассона приближается к пределу несжимаемости $\nu = {}^1/_2$ [Ф17]. Для оболочек трудности связаны с большой жесткостью в направлении толщины или с очень тонкими оболочками. Грубо говоря, пространственные гармоники могут включать в себя отношение $(r/th)^2$ [20], в то время как гармоники изгиба выявляют четвертый порядок задачи и округление пропорционально h^{-4}.

Чтобы закончить это введение, мы должны разъяснить связь между числом обусловленности матрицы и ее чувствительностью к возмущениям. Крайний случай возникает, когда вектор нагрузки F в заданной линейной системе $KQ = F$ совпадает с единичным собственным вектором матрицы K, в частности с собственным вектором $V_{\max}$, соответствующим $\lambda_{\max}$. Тогда решением системы будет $Q = V_{\max}/\lambda_{\max}$. Предположим, что вектор нагрузки слегка возмущен другим собственным вектором $V_{\min}$, соответствующим $\lambda_{\min}$, так что $\tilde{F} = V_{\max} + \varepsilon V_{\min}$. Тогда решение принимает вид $\tilde{Q} = V_{\max}/\lambda_{\max} + \varepsilon V_{\min}/\lambda_{\min}$. Поэтому относительное изменение в решении равно

$$\frac{|Q - \tilde{Q}|}{|Q|} \sim \varepsilon \frac{\lambda_{\max}}{\lambda_{\min}} = \varkappa\varepsilon.$$

Таким образом, возмущение в F порядка ε дает возмущение в Q порядка $\varkappa\varepsilon$.

Легко заметить, что этот случай крайний и что всегда

$$\frac{|\delta Q|}{|Q|} \leqslant \varkappa \frac{|\delta F|}{|F|}.$$

Доказательство. $|F| = |KQ| \leqslant \lambda_{\max} |Q|$ и $|\delta F| = |K\delta Q| \geqslant \geqslant \lambda_{\min} |\delta Q|$.

Из этого простого неравенства и подобного ему неравенства с $|\delta K|/|K|$ в правой части (см. [20]) вытекают далеко идущие следствия. Если число обусловленности $\varkappa$ порядка 10^{-s}, то *в решении уравнения $KQ = F$ можно потерять до s значащих цифр.* Если это близко к числу точных знаков в вычислениях на ЭВМ, то для достижения нужной точности вычислений может потребоваться двойная точность расчетов на ЭВМ.

Было высказано возражение, что в конкретной задаче число обусловленности может не иметь ничего общего с вектором нагрузки. Постоянная $\varkappa$ может, а на самом деле должна быть по крайней мере отчасти пессимистична. Айронс [A2] предложил несколько разных чисел, которые в процессе исключения формировались в машине и которые автоматически учитывали масштабирование и значение *f данной* задачи. Мы допускаем, что для немедленного принятия решений (окончание особых вычислений или переход к двойной точности) вычисляемые величины такого рода будут наилучшими. Однако наша цель здесь состоит в отыскании некоторой априорной меры чувствительности, и число обусловленности для этого вполне удовлетворительно. Правило, к которому оно приводит, а именно что ошибка округления будет увеличиваться пропорционально h^{-2m}, в обычных вычислениях определенно выполняется. Мы подчеркиваем, что *здесь нет зависимости от общего числа элементов в области, а есть зависимость от числа элементов на каждой стороне.* Другими словами, нет существенной зависимости от числа пространственных переменных.

5.2. ЧИСЛО ОБУСЛОВЛЕННОСТИ

Наша цель — оценить отношение $\varkappa = \lambda_N(K)/\lambda_1(K)$ наибольшего к наименьшему собственному значению матрицы жесткости.

Теорема 5.1. *Для каждой вариационной задачи и каждого выбора конечного элемента существует такая постоянная c, что*

$$\varkappa \leqslant c h_{\min}^{-2m}. \tag{1}$$

Постоянная обратно пропорциональна наименьшему собственному значению λ_1 заданной непрерывной задачи и увеличивается, если геометрия элементов становится вырожденной.

Дадим два корректных, но довольно неформальных доказательства. Первое применяется лишь к специальному случаю равномерной сетки и иллюстрирует, как анализ Фурье (или Тёплица) позволяет вполне точно вычислить собственные значения

и число обусловленности. Второе применяется к конечным элементам произвольной формы.

Начнем с предположения, что сетка равномерна и коэффициенты данной задачи $Lu = f$ постоянны. Матрицы жесткости элементов k_i, а также матрицы массы m_i все будут одинаковы. Это означает, что по существу K и M — *тёплицевы матрицы*. Элементы тёплицевой матрицы постоянны вдоль каждой диагонали. K_{ij} зависит только от разности $j - i$ номеров столбца и строки. (В непрерывном случае это соответствует интегральному оператору $\int K(s-t)f(t)\,dt$, т. е. *свертке*, ядро которого зависит лишь от $s - t$.) Хотя это свойство может теряться на границах, особенно с естественными краевыми условиями, мы хотим показать, как все же полезны и легки вычисления с тёплицевыми матрицами.

Предположим, сравниваются две матрицы жесткости, обе возникающие из уравнения $-u'' = f$ при кусочно линейных элементах. Первая имеет порядок N и подчиняется условиям $u(0) = u'(\pi) = 0$; это основной пример на протяжении всей книги. Вторая соответствует случаю полного отсутствия границ, т. е. когда интервал $[0, \pi]$ расширяется до $(-\infty, \infty)$ в результате добавления еще и еще матриц элементов:

$$K = \frac{1}{h}\begin{bmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 1 \end{bmatrix}, \quad K_\infty = \frac{1}{h}\begin{pmatrix} \ddots & \ddots & \ddots & \\ -1 & 2 & -1 & \\ & -1 & 2 & -1 \\ & & \ddots & \ddots \end{pmatrix}.$$

Так как K образуется из K_∞ (исключением всех элементов вне $[0, \pi]$ и наложением главного условия $Q_0 = 0$), то крайние собственные значения матрицы K ограничены соответствующими собственными значениями матрицы K_∞:

$$\lambda_{\min}(K_\infty) \leqslant \lambda_{\min}(K) \leqslant \lambda_{\max}(K) \leqslant \lambda_{\max}(K_\infty).$$

То же справедливо и для матриц массы, так что линейную независимость базиса можно проверить прежде всего на бесконечном интервале.

С тёплицевой матрицей K_∞, которую можно описать как матрицу дискретной свертки, работать просто. Точно так же, как собственные функции любого дифференциального уравнения с постоянными коэффициентами, ее собственные векторы являются чистыми экспонентами. Возьмем вектор с компонентами $v_j = e^{ij\theta}$ и применим к нему матрицу:

$$K_\infty v_j = \frac{1}{h}(-e^{-i\theta} + 2 - e^{-i\theta})\,v_j.$$

Поэтому собственное значение равно

$$\lambda(\theta)=\frac{1}{h}(-e^{-i\theta}+2-e^{i\theta})=\frac{2(1-\cos\theta)}{h}.$$

Так как это число расположено между 0 и $4/h$, то

$$0\leqslant\lambda_{\min}(K)\leqslant\lambda_{\max}(K)\leqslant\frac{4}{h}.$$

В этом случае результат совпадает с утверждением теоремы Гершгорина (1.4): каждое собственное значение лежит в круге с центром в точке $K_{ii}=2/h$, радиус которого равен сумме $\sum_{j\neq i}|K_{ij}|=2/h$. В общем случае результат, полученный по теореме Гершгорина, совсем не так хорош: даже для рассматриваемых ниже билинейных элементов с внедиагональными элементами одного и того же знака он неточен.

Так же исследуются матрицы массы с элементами $h/6$, $4h/6$, $h/6$ вдоль каждой строки. Собственными значениями в случае бесконечного интервала будут

$$\mu(\theta)=\frac{h}{6}e^{-i\theta}+\frac{4h}{6}+\frac{h}{6}e^{i\theta}=\frac{2h}{3}+\frac{h}{3}\cos\theta.$$

Так как $\cos\theta$ изменяется от -1 до 1, то собственные значения матрицы M_∞ заключены между $h/3$ и h. Отметим способ, который в применении к матрицам массы дает лучший результат: получается не только правильная верхняя граница для наибольшего собственного значения, но и хорошая *нижняя* граница для $\lambda_{\min}(M)$. Число обусловленности матрицы M очень точно задается соотношением

$$\varkappa(M)\leqslant\varkappa(M_\infty)=\frac{h}{h/3}=3.$$

Так как эта оценка не зависит от h, то кусочно линейные функции-крышки линейно независимы равномерно по $h\to 0$. Нулевая нижняя граница для $\lambda_{\min}(K)$ возникла из-за наличия у K_∞ нулевого собственного значения: постоянная функция удовлетворяет уравнению $-u''=0$ на всей прямой и соответствует дискретному собственному вектору (...111...), которого нет у конечной матрицы K только из-за главного краевого условия. Поэтому в конечном счете для достижения строгих оценок для $\varkappa(K)$ приходится привлекать матрицы массы.

Для того чтобы лучше понять структуру матрицы K_∞, рассмотрим еще два примера.

1. Билинейные элементы на квадратной сетке для $-\Delta u=f$. Здесь в каждом узле одно неизвестное и типичная строка матрицы K (умноженная на $3h$) состоит из 8 на главной диагонали

и -1 на восьми соседних диагоналях. Уравнение $KQ = F$ имеет вид

$$8Q_{jk} - \sum Q_{j'k'} = 3hF_{jk},$$

где (j', k') означают восемь ближайших к (j, k) точек на квадратной сетке. Собственные векторы v матрицы K_∞ снова являются чистыми экспонентами, но теперь у них уже две частоты: компоненты вектора v равны $v_{jk} = e^{i(j\theta+k\varphi)}$. Вычислим собственные значения:

$$3hK_\infty v_{jk} = (8 - e^{i\theta} - e^{-i\theta} - e^{i\varphi} - e^{-i\varphi} - e^{i(\theta+\varphi)} - e^{-i(\theta+\varphi)} - e^{i(\theta-\varphi)} - e^{-i(\theta-\varphi)})\, v_{jk}.$$

Крайние собственные значения определяются из условий максимума и минимума выражения в скобках. Так как оно линейно относительно $\cos\theta$ и $\cos\varphi$, то экстремум должен быть, когда $\cos\theta$ и $\cos\varphi$ равны ± 1, так что

$$\lambda_{\min}(K_\infty) = 0, \quad \lambda_{\max}(K_\infty) = \frac{12}{3h} \text{ (при } \theta = \pi,\ \varphi = 0).$$

Теорема Гершгорина дала бы $\lambda_{\max} \leqslant 16/3h$.

2. Кубические элементы в одномерном случае для $u^{IV} = f$. Здесь в каждом узле уже два неизвестных и $KQ = F$ — система двух разностных уравнений. Соответственно K и K_∞ — *блочные тёплицевы матрицы*: вдоль главной диагонали расположен блок A размера 2×2, с одной стороны от него находится блок B того же размера, а с другой — транспонированный блок B^T. В соответствии с результатами разд. 1.7

$$K_\infty = \begin{pmatrix} \ddots & \ddots & & \\ B^T & A & B & \\ & B^T & A & B \\ & & \ddots & \ddots \end{pmatrix},$$

$$A = \frac{1}{h^3}\begin{pmatrix} 24 & 0 \\ 0 & 8h^2 \end{pmatrix}, \quad B = \frac{1}{h^3}\begin{pmatrix} -12 & 6h \\ -6h & 2h^2 \end{pmatrix}.$$

По аналогии с предыдущими примерами собственные значения можно вычислить из блока $\lambda(\theta) = B^T e^{-i\theta} + A + Be^{i\theta}$. Это (2×2)-матрица; максимум ее собственных значений на всем отрезке $-\pi \leqslant \theta \leqslant \pi$ совпадает с $\lambda_{\max}(K_\infty)$. Крайний случай снова должен быть при $\theta = 0$ или $\theta = \pi$, поскольку отношение Рэлея $x^T\lambda(\theta)x/x^Tx$ линейно по $\cos\theta$. Подсчитываем

$$\lambda(0) = \frac{1}{h^3}\begin{pmatrix} 0 & 0 \\ 0 & 12h^2 \end{pmatrix}, \quad \lambda(\pi) = \frac{1}{h^3}\begin{pmatrix} 48 & 0 \\ 0 & 4h^2 \end{pmatrix}.$$

Итак, $\lambda_{\max}(K_\infty) = \max(12/h, 48/h^3)$. На этот результат может повлиять изменение относительного масштаба двух типов базисных функций, соответствующих перемещению и наклону.

Крайние собственные значения любого блока тёплицевой матрицы K_∞ можно найти тем же способом. Если порядок блоков равен M (число неизвестных на квадрате сетки), то и $\lambda(\theta)$ будет матрицей того же порядка, в n-мерной задаче $\theta = (\theta_1, \ldots, \theta_n)$. Так как в методе (узловых) конечных элементов каждый квадрат связан только с ближайшими к нему, то матрица λ будет линейной по $\cos\theta_1, \ldots, \cos\theta_n$. Поэтому $\lambda_{\max}(K_\infty)$ можно вычислить, пробуя все возможные комбинации ± 1 для этих косинусов и определяя наибольшие собственные значения из 2^n получающихся в результате матриц λ. Итак, работая только с матрицами порядка M (меньшего, чем у матриц элементов; для некоторых элементов значение M приведено в таблицах в разд. 1.8), можно вычислить точно число обусловленности чистой тёплицевой матрицы. Для матриц массы (также обозначаемых через M!) получаем $\varkappa(M) \leqslant \varkappa(M_\infty)$, причем постоянная $\varkappa(M_\infty)$ зависит только от элемента. *Конечные элементы равномерно независимы.*

Для матриц жесткости есть хорошая верхняя граница значения $\lambda_{\max}$, а нижней хорошей — нет. Здесь нам придется оставить точные вычисления и вернуться к неравенствам. Согласно принципу Рэлея, наименьшее собственное значение матрицы K равно

$$\lambda_1(K) = \min \frac{x^T K x}{x^T x} \geqslant \min \frac{x^T K x}{x^T M x} \min \frac{x^T M x}{x^T x} = \lambda_1\left(M^{-1}K\right)\lambda_1(M). \qquad (2)$$

С появлением $M^{-1}K$ легче использовать вариационные доказательства. Задача на собственные значения $KQ = \lambda MQ$ есть в точности аналог непрерывной задачи $Lu = \lambda u$, решаемой методом конечных элементов, и в гл. 6 мы показываем, что *главное собственное значение* $\lambda_1(M^{-1}K)$ *дискретной задачи всегда не меньше главного собственного значения* $\lambda_1(L)$ *непрерывной задачи.* Это и есть нижняя граница, которая не зависит от h.

В примере $Lu = -u''$ с краевыми условиями $u(0) = u'(\pi) = 0$ наименьшее собственное значение равно $\lambda_1(L) = 1/4$. Так как мы уже доказали, что $\lambda_1(M) \geqslant \lambda_1(M_\infty) \geqslant h/3$, то отсюда следует, что $\lambda_1(K) \geqslant h/12$. Наименьшее собственное значение матрицы K равно $\sin^2\left(\frac{h}{4}\right)\Big/\left(\frac{h}{4}\right)^2$, т. е. примерно $h/4$. В оценке (2) теряется множитель 3, по существу число обусловленности матрицы M, так как истинный минимум отношения x^TKx/x^Tx достигается на основной гармонике x, компоненты которой все одного знака. Это соответствует более $\lambda_{\max}(M)$, чем значению $\lambda_{\min}(M)$, возникающему в (2).

Однако существенно то, что показатель степени h в числе обусловленности определяется правильно:

$$\varkappa(K)=\frac{\lambda_{\max}(K)}{\lambda_{\min}(K)}\leqslant\frac{4/h}{h/12}=\frac{48}{h^2}.$$

Для уравнения порядка $2m$ на регулярной области тот же подход к тёплицевой матрице (вместе с (2)) дает число обусловленности порядка h^{-2m}. Постоянная здесь, как и должно быть, включает в себя $\lambda_1(L)$; если физическая задача плохо обусловлена, то это должно отражаться на ее аналоге в методе конечных элементов.

Второе доказательство теоремы 5.1. Осталось лишь распространить оценку h^{-2m} для числа обусловленности на случай нерегулярных элементов. Тёплицева структура пропадает, и все доказательство должно быть основано на матрицах отдельных элементов. Мы будем следовать Фриду [Ф13]. Чтобы найти верхнюю границу для $\lambda_{\max}(K)$, вспомним, что глобальная матрица жесткости составлена из матриц элементов k_i:

$$x^TKx=\sum_i x_i^Tk_ix_i. \tag{3}$$

Вектор x и матрица K имеют порядок N, вектор x_i и матрица k_i — порядок d_i, т. е. число степеней свободы внутри i-го элемента. Действительно, x_i получается из x, если удалить все, кроме d_i соответствующих компонент; это можно представить себе как умножение вектора x на *матрицу инциденций*, составленную соответствующим образом из 0 и 1.

Если Λ — максимальное из всех собственных значений матриц жесткости элементов, так что, согласно принципу Рэлея, $x_i^Tk_ix_i\leqslant\Lambda x_i^Tx_i$, то из (3) вытекает, что

$$x^TKx\leqslant\Lambda\sum_i x_i^Tx_i.$$

Предположим теперь, что q — максимальное число элементов, «встречающихся» в любом узле. Тогда ни одна из компонент вектора x не входит более чем в q векторов x_i, так что $\sum x_i^Tx_i\leqslant qx^Tx$. Поэтому

$$x^TKx\leqslant\Lambda qx^Tx \quad\text{и}\quad \lambda_{\max}(K)\leqslant\Lambda q.$$

Чтобы найти нижнюю границу для $\lambda_{\min}(M)$, рассуждаем точно так же, как раньше:

$$x^TMx=\sum x_i^Tm_ix_i\geqslant\theta\sum x_i^Tx_i\geqslant\theta rx^Tx,$$

где θ — наименьшее из собственных чисел матриц массы элементов, а r — минимальное число элементов, «встречающихся»

в узле. (Часто $r = 1$ или $r = 2$ из-за элементов на границе; конечно, $r = 1$, если существуют узлы, внутренние по отношению к элементам.) Снова, согласно принципу Рэлея,

$$\lambda_1(M) = \min \frac{x^T M x}{x^T x} \geqslant \theta r.$$

Поэтому в силу (2) число обусловленности матрицы жесткости меньше, чем

$$\varkappa(K) \leqslant \frac{\lambda_{\max}(K)}{\lambda_1(L)\,\lambda_1(M)} \leqslant \frac{\Lambda q}{\lambda_1(L)\,\theta r}. \tag{4}$$

Если геометрия элементов не вырождается, так что базис однороден в смысле гл. 3, то непосредственное вычисление дает ожидаемое значение $\varkappa(K) = O(h^{-2m})$. Если же вырождение происходит, например треугольники становятся очень тонкими или четырехугольники приближаются к треугольникам, оно отражается на параметрах Λ и θ. Так как это собственные значения матриц малого размера, то влияние вырождения можно строго оценить. Фрид [Ф13] вычислил эту зависимость от геометрии для нескольких примеров (оценка в теореме 5.1 иногда пессимистична) и дал также нижнюю границу для числа обусловленности.

Он получил также красивую оценку в одномерном случае для неравномерных сеток, пользуясь неравенством $a(u^h, u^h) \leqslant a(u, u)$; дискретная структура всегда жестче непрерывной. Предположим, что в j-й узел помещена точечная нагрузка $f = \delta(x - z_j)$. Тогда, как отмечалось в разд. 1.10, у вектора нагрузки F только одна ненулевая компонента

$$a(u^h, u^h) = Q^T K Q = F^T K^{-1} F = (K^{-1})_{jj}.$$

Таким образом, величина $(K^{-1})_{jj}$ меньше истинной энергии $a(u, u)$ и ограничена *независимо от распределения узлов.* (Мы предполагаем, что рассматривается задача $-(pu')' + qu = f$ и неизвестны значения функции.) Так как матрица K^{-1} положительно определена и тем самым $(K^{-1})_{ij}^2 \leqslant (K^{-1})_{ii}(K^{-1})_{jj}$, то ее наибольший элемент должен быть на главной диагонали. Поэтому сумма абсолютных значений элементов каждой ее строки ограничена величиной cN, где N — порядок матрицы. Можно взять N^{-1} в качестве среднего значения шага сетки $\bar{h}$. Так как h_i фигурирует в знаменателе в i-й матрице элемента, то соответствующая сумма вдоль каждой строки матрицы K ограничена величиной $c/h_{\min}$. Следовательно, число обусловленности даже в более сильном смысле сумм абсолютных значений по строке (что соответствует поточечным, а не среднеквадратичным оценкам) меньше, чем $C/\bar{h}h_{\min}$. Для равномерной сетки опять будет C/h^2.

Важный вывод: *ошибка округления не зависит сильно от степени полиномиального элемента.* Главным образом она зависит от h, от порядка рассматриваемой задачи и от основного собственного значения непрерывной задачи. Поэтому при наличии ошибок округления достичь численную точность можно, увеличивая степень пробных функций. Число обусловленности для кубических элементов лишь немного хуже, чем для линейных, так что ошибки округления в этих случаях для заданного значения h сравнимы. Однако ошибка дискретизации для кубических элементов на порядок меньше. Поэтому в переходный момент, когда округления не позволяют получать более точные результаты за счет уменьшения h, с помощью кубического элемента этого можно достичь. Особенно это относится к вычислению напряжений, где дифференцирование (или взятие разностей) перемещений вводит в численную ошибку дополнительный множитель h^{-1}. Даже в задачах второго порядка ошибка округления становится значительной и наилучший выход из положения — увеличить степень пробных функций.

6 ЗАДАЧИ НА СОБСТВЕННЫЕ ЗНАЧЕНИЯ

6.1. ВАРИАЦИОННАЯ ФОРМУЛИРОВКА И ПРИНЦИП МИНИМАКСА

Задачи на собственные значения, которые мы будем записывать в виде $Lu = \lambda u$ или, более общо, $Lu = \lambda Bu$, очень часто встречаются в приложениях. Назовем здесь лишь задачи о продольном изгибе стержней и выпучивании оболочек, колебании упругих тел и о многогрупповой диффузии в ядерных реакторах. К счастью, как и для стационарных уравнений $Lu = f$, для этих задач также полезна идея Рэлея — Ритца. В самом деле, эта идея исходит из описания Рэлея основной частоты как наименьшего значения *отношения Рэлея*. Поэтому шаг, который был предпринят в последние 15 лет, вполне естествен и неизбежен: применить новые идеи метода конечных элементов к этой давно установленной вариационной форме задачи на собственные значения.

С практической точки зрения это означает, что кусочно полиномиальные функции можно подставить непосредственно в отношение Рэлея в качестве пробных функций. Вычисление этого отношения становится как раз той задачей, которая уже обсуждалась и для выполнения которой к настоящему времени создано множество мощных вычислительных машин. Эта задача представляет собой вычисление матриц жесткости и массы K и M. Следующий шаг, однако, приводит к другой, более трудной вычислительной задаче линейной алгебры: вместо решения линейной системы $KQ = F$ надо решить дискретную задачу на собственные значения $KQ = \lambda MQ$. К счастью, сейчас известно, как можно использовать свойства этих двух матриц: симметричность, разреженность, положительную определенность матрицы M, для ускорения численного алгоритма. В разд. 6.4 мы рассмотрим несколько эффективных численных методов.

С теоретической точки зрения, основные этапы в построении оценок погрешностей снова зависят от аппроксимирующих свойств подпространства метода конечных элементов S^h. Следовательно, здесь непосредственно можно применить теоремы аппроксимации из гл. 3. Математически новый необходимый шаг заключается в том, чтобы исходя из этих теорем аппрокси-

мации вывести удовлетворительные оценки ошибок в собственных значениях и собственных функциях. Этот шаг и является нашей основной целью.

Прежде чем приступить к изложению общей теории, исследуем несколько специальных примеров, чтобы проиллюстрировать поведение аппроксимаций методом конечных элементов. Затем применим теорию аппроксимации к задаче на собственные значения и найдем границы для ошибок $\lambda_l - \lambda_l^h$ и $u_l - u_l^h$ в l-м собственном значении и в l-й собственной функции соответственно, достаточно точные для того, чтобы правильно отразить зависимость их от l и h.

Начнем с задачи на собственные значения вида

$$Lu = -\frac{d}{dx}\left(p(x)\frac{du}{dx}\right) + q(x)u = \lambda u, \quad 0 < x < \pi. \tag{1}$$

Оператор L этой задачи изучался в гл. 1, его простота заключается в его одномерности. Продемонстрируем еще раз различие между естественными и главными краевыми условиями, задав по одному из них:

$$u(0) = 0, \quad u'(\pi) = 0. \tag{2}$$

Известно, что такая задача Штурма — Лиувилля имеет бесконечную последовательность вещественных собственных значений

$$\lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_j \leqslant \ldots \to \infty$$

и соответствующую полную систему ортонормальных собственных функций.

$$(u_j, u_k) = \int_0^\pi u_j(x)\,u_k(x)\,dx = \delta_{jk}. \tag{3}$$

Отсюда немедленно вытекает, что эти собственные функции ортогональны также и в смысле *энергетического скалярного произведения*:

$$(Lu_j, u_k) = (\lambda_j, u_j, u_k) = \lambda_j \delta_{jk}, \tag{4}$$

или, после интегрирования по частям,

$$a(u_j, u_k) = \int_0^\pi (pu_j'u_k' + qu_ju_k)\,dx = \begin{cases} \lambda_j, & j = k, \\ 0, & j \neq k. \end{cases} \tag{5}$$

В случае постоянных p и q, который будет служить модельной задачей, графиками собственных функций служат синусоиды

$$u_j(x) = \sqrt{\frac{\pi}{2}}\sin\left(j - \frac{1}{2}\right)x, \quad \lambda_j = p\left(j - \frac{1}{2}\right)^2 + q.$$

Первый шаг в методе Рэлея — Ритца состоит в том, чтобы переписать $Lu = \lambda u$ как вариационную задачу. Есть две возможности, соответствующие минимизации по Ритцу и слабой форме записи по Галёркину стационарных уравнений $Lu = f$. Обе приводят к одному и тому же результату. Первая заключается в том, чтобы ввести *отношение Рэлея* в виде

$$R(v) = \frac{a(v, v)}{(v, v)} = \frac{\int p(v')^2 + qv^2}{\int v^2}. \tag{6}$$

Мы утверждаем, что стационарные (или критические) точки функционала $R(v)$, т. е. точки, где градиент от R обращается в нуль, являются в точности собственными функциями задачи. Чтобы понять это, допустим, что пробная функция v заменена соответствующим разложением в ряд $\sum \alpha_j u_j$ по собственным функциям $\alpha_j = (v, u_j)$. Тогда, используя условия ортогональности (3) и (5), получаем

$$R(v) = \frac{a\left(\sum \alpha_j u_j, \sum \alpha_k u_k\right)}{\left(\sum \alpha_j u_j, \sum \alpha_k u_k\right)} = \frac{\sum \lambda_j \alpha_j^2}{\sum \alpha_j^2}. \tag{7}$$

В стационарной точке функционала $R(v)$ производная по каждой переменной α_k должна равняться нулю:

$$\frac{\partial R}{\partial \alpha_k} = \frac{2\alpha_k \sum \alpha_j^2 (\lambda_j - \lambda_k)}{\left(\sum \alpha_j^2\right)^2} = 0.$$

Если v — одна из собственных функций u_j, то это условие, несомненно, выполняется: $\alpha_k = 0$ для всех k, кроме $k = j$, а j-е слагаемое равно нулю, поскольку $\lambda_j - \lambda_k = 0$. Других стационарных точек, отличных от собственных функций, нет. (Для кратного собственного значения $\lambda_j = \lambda_{j+1}$ все комбинации $v = \alpha_j u_j + \alpha_{j+1} u_{j+1}$ будут собственными функциями и существует целая плоскость стационарных точек.) Значение $R(v)$ в стационарной точке $v = u_j$ легко вычисляется: оно точно равно собственному значению λ_j:

$$R(u_j) = \frac{a(u_j, u_j)}{(u_j, u_j)} = \lambda_j. \tag{8}$$

Полезно попытаться представить себе график $R(v)$. Числитель $a(v, v)$, как и для $I(v)$ в гл. 1, соответствует выпуклой поверхности «рюмки для яйца», единственное отличие в том, что линейный член $-2(f, v)$ теперь отсутствует, так что дно рюмки лежит в начале координат. Влияние знаменателя (v, v) можно исследовать двумя способами. Так как он делает отношение однородным, т. е. $R(\alpha v) = R(v)$, то можно рассмотреть лишь еди-

ничные векторы v. Другими словами, можно положить знаменатель равным 1 и рассматривать сечение рюмки круговым цилиндром $(v, v) = 1$. Минимум на этом сечении соответствует основной частоте λ_1.

Второй способ — положить числитель равным 1. Это означает, что рюмка разрезается горизонтальной плоскостью, лежащей на 1 выше основания. Поперечное сечение будет эллипсом $a(v, v) = 1$, или точнее, бесконечномерным эллипсоидом $\sum \lambda_j \alpha_j^2 = 1$. Его главная ось расположена в направлении первой собственной функции u_1, так как именно по этой оси эллипсоид наиболее вытянут. Другими словами, при фиксированном числителе, равном 1, отношение Рэлея минимально, когда знаменатель максимален. Если затем рассмотреть эллипс, перпендикулярный к этой главной оси, т. е. фиксировать $\alpha_1 = 0$ и получить пространство на единицу меньшей размерности, то длина главной оси *этого* эллипса будет равна $\sqrt{1/\lambda_2}$. (И вообще длина главной оси любого другого поперечного сечения исходного эллипсоида будет меньше длины главной оси эллипсоида $\sum \alpha_j^2 \lambda_j = 1$, но больше $\sqrt{1/\lambda_2}$, что соответствует принципу минимакса, описанному ниже в (13): наименьшее собственное значение λ_1' при любом дополнительном условии удовлетворяет соотношению $\lambda_1 \leqslant \lambda_1' \leqslant \lambda_2$.)

Оказывается, что для четырехмерного случая поперечное сечение рюмки для яйца есть оболочка всего яйца.

До сих пор мы не уточняли, какие функции v допустимы в вариационной характеризации (т. е. в отношении Рэлея) собственных значений и собственных функций. Это в точности тот же вопрос, что возникал в стационарной задаче в разд. 1.3. Там функционал $I(v)$ в конце концов минимизировался по всем функциям из $\mathscr{H}^1$, удовлетворяющим главным краевым условиям, т. е. на пространстве $\mathscr{H}_E^1$. Это пространство естественным образом появилось в процессе пополнения, который заключается в присоединении к множеству гладких функций, удовлетворяющих *всем* краевым условиям, любой функции v, равной пределу функций v_N из этого множества в смысле $a(v - v_N, v - v_N) \to 0$. Так как этот процесс не изменяет поверхность $R(v)$, то мы здесь будем выполнять ту же процедуру пополнения; все стационарные точки сохраняются. (И если бы функция $p(x)$ была разрывной, этот процесс заполнения дыр на поверхности $R(v)$ в действительности заполнял бы множество собственных функций: так как они не гладкие, то они первоначально исключались.)

В результате допустимым пространством, как и ранее, будет $\mathscr{H}_E^1$. Это означает, что для дискретной аппроксимации можно

использовать все те же подпространства метода конечных элементов с единственным отличием, что теперь мы ищем стационарные точки функционала $R(v)$, а не минимум для $I(v)$.

Ранее упоминалось, что есть второй способ переформулировки задачи на собственные значения. Он состоит в том, чтобы записать уравнение $Lu = \lambda u$ в его *слабой форме*, или *форме Галёркина*. Умножим $Lu = \lambda u$ на функцию v и проинтегрируем по частям:

$$\int_0^\pi (pu'v' + quv)\,dx = \lambda \int_0^\pi uv\,dx. \tag{9}$$

Теперь задача на собственные значения заключается в том, чтобы найти скаляр λ и функцию $u \in \mathscr{H}_E^1$, при которых (9) выполняется для всех $v \in \mathscr{H}_E^1$. Краевые условия для u получаются правильными, так как уравнение (9) фактически эквивалентно уравнению

$$\int Lu(x)\, v(x)\,dx - pu'v\Big|_0^\pi = \lambda \int uv\,dx.$$

При $x = 0$ проинтегрированный член $pu'v$ автоматически обращается в нуль. Из равенства, содержащего остальные члены, в силу произвольности выбора v в $\mathscr{H}_E^1$ вытекает, что естественное условие $u' = 0$ при $x = \pi$ выполняется и на всем интервале функция u подчиняется дифференциальному уравнению $Lu = \lambda u$.

Уравнение (9) представляет собой частный случай обычной слабой формы задачи на собственные значения: *найти скаляр* λ *и функцию* u *в допустимом пространстве* V, *при которых*

$$a(u, v) = \lambda(u, v) \text{ для всех } v \text{ в } V. \tag{10}$$

Это соотношение похоже на условие $a(u, v) = (f, v)$ обращения в нуль первой вариации в стационарной задаче минимизации $I(v)$. Действительно, это уравнение означает *обращение в нуль первой вариации от* R *в стационарной точке* u:

$$R(u + \varepsilon v) = \frac{a(u + \varepsilon v,\ u + \varepsilon v)}{(u + \varepsilon v,\ u + \varepsilon v)} = \frac{a(u, u) + 2\varepsilon a(u, v) + \ldots}{(u, u) + 2\varepsilon(u, v) + \ldots} =$$

$$= R(u) + 2\varepsilon \frac{a(u, v)(u, u) - a(u, u)(u, v)}{(u, u)^2} + \ldots =$$

$$= R(u) + 2\varepsilon \frac{a(u, v) - \lambda(u, v)}{(u, u)} + O(\varepsilon^2).$$

Таким образом, формулировка задачи в слабой форме и постановка ее как задачи отыскания стационарных точек эквивалентны.

Для случая λ_1 и u_1, т. е. для основной частоты и соответствующего ей собственного колебания, стационарная точка фактически является минимумом:

$$\lambda_1 = \min_{v \text{ в } \mathscr{H}_E^1} R(v). \tag{11}$$

Это становится очевидным, если v представить в виде $\sum \alpha_j u_j$, так как тогда $R(v) = \sum \alpha_j^2 \lambda_j / \sum \alpha_j^2 \geqslant \lambda_1$.

Полезно также описать собственные функции более высокого порядка исходя из минимизации, поскольку анализировать сходимость к минимуму намного проще, чем к стационарной точке. Одна возможность в этом направлении — заставить функцию v быть ортогональной к первым $l-1$ собственным функциям: $\alpha_1 = (v, u_1) = 0, \ldots, \alpha_{l-1} = (v, u_{l-1}) = 0$. При этих условиях минимум отношения Рэлея равен λ_l:

$$\lambda_l = \min_{v \perp E_{l-1}} R(v). \tag{12}$$

Здесь E_{l-1} — пространство, натянутое на собственные функции $u_1, \ldots, u_{l-1}$.

Есть другая формула для λ_l, не требующая знания $u_1, \ldots, u_{l-1}$. Ее открыли Пуанкаре, Курант и Фишер. В дальнейшем нашем изложении она играет фундаментальную роль.

Принцип минимакса: *если $R(v)$ максимизируется на l-мерном подпространстве S_l, то λ_l — наименьшее (при всевозможных наборах l-мерных подпространств $S_l \subset \mathscr{H}_E^1$) из наибольших значений $R(v)$, т. е.*

$$\lambda_l = \min_{S_l} \max_{v \text{ в } S_l} R(v). \tag{13}$$

Если $S_l = E_l$, то максимум $R(v)$ в точности-равен λ_l.

Для доказательства формулы (13) надо показать, что для любого выбора подпространства S_l

$$\max_{v \text{ в } S_l} R(v) \geqslant \lambda_l. \tag{14}$$

Рассуждение основано на выборе функции $v^* \in S_l$, которая должна быть ортогональна к E_{l-1}, т. е. удовлетворять $l-1$ уравнениям $(v, u_i) = 0$, $1 \leqslant i < l$. Так как мы налагаем лишь $l-1$ однородных условий в пространстве с l параметрами, то такая функция v^* существует. Поскольку $v^* \perp E_{l-1}$, из (12) вытекает, что $\lambda_l \leqslant R(v^*)$. Другими словами, (14) выполняется и формула (13) справедлива.

Из этой формулы сразу можно вывести грубую оценку для собственного значения λ_l, сравнивая рассматриваемую задачу

с задачей с постоянными коэффициентами. Ясно, что для любой функции v

$$\int p_{\max}(v')^2 + q_{\max}v^2 \geqslant \int p(x)(v')^2 + q(x)v^2 \geqslant \int p_{\min}(v')^2 + q_{\min}v^2.$$

Разделим на $\int v^2$. Легко видеть, что в центре стоит отношение Рэлея для задачи с переменными коэффициентами и оно заключено между отношениями Рэлея для задач, собственные значения которых известны в явном виде. По принципу минимакса каждое собственное значение λ_l рассматриваемой задачи должно лежать между двумя известными собственными значениями:

$$p_{\max}\left(l-\frac{1}{2}\right)^2 + q_{\max} \geqslant \lambda_l \geqslant p_{\min}\left(l-\frac{1}{2}\right)^2 + q_{\min}. \qquad (15)$$

В частности, λ_l при $l \to \infty$ имеет порядок l^2.

Наконец, мы подошли к главной цели этого раздела — установить *принцип Рэлея — Ритца для приближенного вычисления собственных значений.* Можно начать либо с постановки задачи в слабой форме $a(u, v) = \lambda(u, v)$, либо с описания собственных значений как критических (стационарных) точек отношения $R(v) = a(v, v)/(v, v)$. В любом случае идея состоит в том, чтобы работать в пределах конечномерного подпространства S^h из всего допустимого пространства $\mathscr{H}_E^1$. В этом подпространстве мы ищем такие λ^h и u^h, что

$$a(u^h, v^h) = \lambda^h(u^h, v^h) \text{ для всех } v^h \text{ в } S^h. \qquad (16)$$

Другими словами,

$$\int p(x)(u^h)'(v^h)' + q(x)u^hv^h = \lambda^h \int u^hv^h.$$

Второй подход: *приближенные собственные векторы являются критическими точками для $R(v^h)$ на S^h.*

Чтобы убедиться, что оба метода приводят к одним и тем же аппроксимациям, выберем в S^h базис $\varphi_1, \ldots, \varphi_N$. Тогда любую функцию $v^h \in S^h$ можно представить в виде

$$v^h = \sum q_j\varphi_j, \qquad (17)$$

где q_j — обобщенные координаты (узловые параметры функции v^h, если S^h — пространство метода конечных элементов). Под-

ставим в отношение Рэлея:

$$R(v^h) = \frac{a(v^h, v^h)}{(v^h, v^h)} = \frac{\sum\sum q_j q_k \int (p(x)\varphi_j'\varphi_k' + q(x)\varphi_j\varphi_k)\,dx}{\sum\sum q_j q_k \int \varphi_j\varphi_k\,dx}. \quad (18)$$

Интегралы в числителе и знаменателе уже известны — это элементы матрицы жесткости K^h и матрицы массы M^h. Таким образом, отношение Рэлея выражается через вектор $q = (q_1, \ldots, q_N)$ формулой

$$R(v^h) = \frac{q^T K^h q}{q^T M^h q}. \quad (19)$$

Критические точки этого дискретного отношения служат решениями матричной задачи на собственные значения

$$K^h Q^h = \lambda^h M^h Q^h. \quad (20)$$

Это и есть та задача, которую надо решить. Можно ожидать, что собственные значения λ_l^h будут приближать λ_l (по крайней мере для малых значений l), а собственные векторы Q_l^h приведут к соответствующим приближенным собственным функциям

$$u_l^h = \sum_1^N (Q_l^h)_j \varphi_j. \quad (21)$$

Таким образом, компоненты дискретных собственных векторов матричной задачи $KQ = \lambda MQ$ дают значения собственных функций в узлах в методе конечных элементов.

Слабая форма задачи на собственные значения приводит непосредственно к тому же результату. Пусть в уравнении (16) $v^h = \varphi_k$, тогда

$$a\left(\sum Q_j^h\varphi_j,\ \varphi_k\right) = \lambda^h\left(\sum Q_j^h\varphi_j,\ \varphi_k\right),$$

а это просто k-я строка матричного уравнения $K^hQ^h = \lambda^h M^h Q^h$.

Если базисные функции φ_j ортонормальны, то матрица массы M^h будет единичной и дискретная задача состоит в отыскании собственных значений матрицы K^h. Однако условие ортогональности для φ_j несовместимо с более важным свойством конечных элементов, а именно с тем, что функция φ_j должна равняться нулю на всех элементах, не содержащих узел z_j. Поэтому мы должны либо принять $M^h = I$, либо нарушить идею Рэлея, допустив приближенный расчет масс. Мы предпочитаем первое, поскольку сейчас появляются численные алгоритмы решения общей задачи на собственные значения $KQ = \lambda MQ$, сравнимые по эффективности с алгоритмами для задачи $KQ =$

$= \lambda Q$. Подчеркнем, что матрица массы во всех случаях симметрична и положительно определена; это матрица Грама для линейно независимых векторов $\varphi_1, \ldots, \varphi_N$.

Во многих приложениях наиболее важна основная частота λ_1, и мы особенно надеемся, что λ_1^h обеспечит хорошую аппроксимацию для λ_1. Заметим, что так как λ_1^h — наименьшее значение $R(v)$ на подпространстве S^h, а λ_1 — минимум на всем допустимом пространстве $\mathscr{H}_E^1$, то *всегда* $\lambda_1^h \geqslant \lambda_1$. Естественно ожидать, что если истинную собственную функцию u_1 можно хорошо аппроксимировать в подпространстве S^h, то λ_1^h будет автоматически близко к λ_1; это будет основной результат теории.

Принцип минимакса с одинаковым успехом применяется к дискретной задаче (с тем же доказательством), так что приближенные собственные значения можно охарактеризовать формулой

$$\lambda_l^h = \min_{S_l} \max_{v^h \in S_l} R(v^h). \tag{22}$$

Здесь S_l *пробегает последовательность всевозможных* l*-мерных подпространств пространства* S^h. Конечно, определение имеет смысл только при $l \leqslant N$, так как если N — размерность пространства S^h, то существует всего лишь N приближенных собственных значений. Сравнивая принципы минимакса (22) и (13), сразу видим, что *каждое собственное значение* λ_l *оценивается сверху*:

$$\lambda_l^h \geqslant \lambda_l \text{ для всех } l. \tag{23}$$

Каждое пространство S_l, рассматриваемое при минимизации в (22), также допускается и в (13), так что значение λ_l (минимум в (13)) по крайней мере так же мало, как и λ_l^h.

Принцип минимакса без изменений распространяется на случай $Lu = \lambda Bu$, где B — положительно определенный оператор; отношением Рэлея здесь будет $R(v) = (Lv, v)/(Bv, v)$. Матрица массы дискретной задачи принимает вид $M_{jk} = (B\varphi_j, \varphi_k)$.

6.2. НЕСКОЛЬКО ПРОСТЫХ ПРИМЕРОВ

В этом разделе мы рассмотрим несколько специальных примеров из тех, что встречаются в общей теории аппроксимации собственных значений. В основном займемся задачей с постоянными коэффициентами

$$-p\frac{d^2u}{dx^2} + qu = \lambda u, \quad 0 < x < \pi,$$

с краевыми условиями $u = 0$, $u'(\pi) = 0$,

В качестве первого пробного пространства возьмем пространство кусочно линейных функций с равномерно расположенными узлами $x_j = jh$. Если базис образован обычными функциями-крышками φ_j, то основные матрицы таковы:

$$M^h = \frac{h}{6}\begin{bmatrix} 4 & 1 & & & \\ 1 & 4 & 1 & & \\ & \cdot & \cdot & \cdot & \\ & & 1 & 4 & 1 \\ & & & 1 & 2 \end{bmatrix}, \quad K_1^h = \frac{1}{h}\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \cdot & \cdot & \cdot & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{bmatrix}$$

Матрицу жесткости можно записать в виде $K^h = pK_1^h + qM^h$. Оптимальные весовые коэффициенты для собственной функции

$$u^h(x) = \sum_{j=1}^{N} Q_j \varphi_j(x),$$

отыскиваемой по принципу Рэлея — Ритца, совпадают со значениями u^h в узлах, и система $K^hQ^h = \lambda^h M^h Q^h$ представляет собой не что иное, как разностное уравнение

$$\frac{p}{h}\left(2Q_j^h - Q_{j+1}^h - Q_{j-1}^h\right) + \frac{qh}{6}\left(4Q_j^h + Q_{j+1}^h + Q_{j-1}^h\right) = \\ = \frac{\lambda^h h}{6}\left(4Q_j^h + Q_{j+1}^h + Q_{j-1}^h\right), \quad 1 \leqslant j \leqslant N. \quad (24)$$

Чтобы это уравнение было справедливо в граничной точке $j = N$, положим $Q_{N+1}^h = Q_{N-1}^h$. Вспомним, что условие Дирихле $u(0) = 0$ дает $Q_0 = 0$.

Система (24) без краевых уравнений имеет тёплицеву структуру в том смысле, что (i, j)-е элементы матриц жесткости и массы зависят лишь от разности $i - j$. Поэтому разумно ожидать, что собственные векторы будут тригонометрическими и на самом деле компоненты l-го собственного вектора равны

$$(Q_l^h)_j = \sqrt{\frac{\pi}{2}} \sin\left(\left(l - \frac{1}{2}\right) jh\right).$$

Для того чтобы выписать выражение для соответствующего собственного значения λ_l^h, обозначим

$$k_h(l) = 2h^{-2}\left(1 - \cos\left(l - \frac{1}{2}\right)h\right), \quad m_h(l) = \frac{2 + \cos\left(l - \frac{1}{2}\right)h}{3}.$$

Тогда

$$\lambda_l^h = \frac{pk_h(l) + qm_h(l)}{m_h(l)}. \quad (25)$$

В этой частной задаче собственные функции u_l совпадают в узлах со своими приближениями u_l^h:

$$u_l^h(jh) = u_l(jh) = \sqrt{\frac{\pi}{2}} \sin\left(\left(l - \frac{1}{2}\right) jh\right). \tag{26}$$

Отсюда следует, что функция u_l^h равна интерполянту для u_l и, согласно теоремам об аппроксимации,

$$\|u_l - u_l^h\|_s \leqslant Ch^{2-s} \|u_l\|_2 = O\left(l^2 h^{2-s}\right). \tag{27}$$

Хотя точность в узлах собственных функций u_l^h в методе Ритца специфична, оценка (27) характерна и для общего случая. *Ошибки в собственных функциях $u_l - u_l^h$ в методе Ритца того же порядка, что и ошибки приближенного решения стационарных задач $Lu = f$.*

Вернемся к приближенным собственным значениям λ_l^h (формула (25)). Разложим в ряд введенные выше функции $k_h(l)$ и $m_h(l)$:

$$k_h(l) = \left(l - \frac{1}{2}\right)^2 - \frac{h^2}{12}\left(l - \frac{1}{2}\right)^4 + O\left(l^6 h^4\right),$$

$$m_h(l) = 1 - \frac{h^2}{6}\left(l - \frac{1}{2}\right)^2 + O\left(l^4 h^4\right).$$

Вспоминая, что $\lambda_l = p(l - 1/2)^2 + q$, запишем ошибку в собственном значении в виде

$$\lambda_l^h = \lambda_l + h^2 \frac{p}{12}\left(l - \frac{1}{2}\right)^4 + O\left(l^6 h^4\right) = \lambda_l + O\left(h^2 l^4\right). \tag{28}$$

Заметим, что точность собственных значений та же, что и у *энергии* в собственных функциях. Это утверждение справедливо и в общем случае:

$$\lambda_l^h - \lambda_l \leqslant C \|u_l^h - u_l\|_m^2$$

для задачи $2m$-го порядка. Причина в том, что около критической точки график отношения Рэлея $R(v)$ как функции от v плоский. Поэтому умеренно точные пробные функции дают очень точные приближения к собственным значениям.

Собственные значения фактически не зависят от постоянной q, так как добавление к оператору члена qu просто *сдвигает спектр* на постоянную величину $\lambda_l = p(l - 1/2)^2$ при $q = 0$ и $\lambda_l = p(l - 1/2)^2 + q$ в общем случае. Решающий момент здесь — тот же эффект в аппроксимации λ_l^h по Ритцу — Галёркину. Удобно использовать инвариантность ошибки $\lambda_l^h - \lambda_l$ относительно q и гарантировать на протяжении всей главы, что $\lambda_1 > 0$, добавляя к оператору, если необходимо, достаточно большую постоянную.

При *приближенном расчете* матрица массы M^h заменяется диагональной, в рассматриваемом простом случае — единичной матрицей I. Оказывается, в нашем примере собственные функции от этого не меняются. Напротив, собственные значения заменяются на

$$\tilde{\lambda}_l^h = pk_h(l) + q = \lambda_l - \frac{ph^2}{12}\left(l - \frac{1}{2}\right)^4 + O(l^6h^4). \qquad (29)$$

Таким образом, $\tilde{\lambda}_l^h$ — *нижняя граница* для λ_l и ее точность того же порядка $O(h^2)$, что и для λ_l^h.

Приближенный расчет M^h допускает ясную физическую интерпретацию на языке жесткости системы $K^hQ^h = \lambda^hM^hQ^h$. С этой точки зрения замена матрицы массы M^h единичной матрицей I делает систему «мягче» и, значит, уменьшает величины приближенных собственных значений. Так как аппроксимации по Рэлею — Ритцу λ_l^h всегда служат верхними границами для λ_l, т. е. $\lambda_l^h \geqslant \lambda_l$, то можно надеяться, что уменьшение величины λ_l^h будет увеличивать точность аппроксимации. С другой стороны, такое нарушение правил Ритца может, на наш взгляд, сделать систему слишком мягкой и тем самым неблагоприятно отразиться на точности результатов. В рассматриваемой задаче ущерб был небольшим, но на более типичном примере в конце этого раздела мы покажем, что возможна значительно более серьезная потеря точности.

Исследуем пример кусочно полиномиальной аппроксимации более высокой степени, а именно возьмем кубическое эрмитово пространство S^h на равномерной сетке. Здесь базисные функции соответствуют значениям функции, а ω_j — значениям ее производной, так что [1])

$$v^h(x) = \sum_{j=1}^{N} u_j\psi_j(x) + \sum_{j=0}^{N-1} u'_j\omega_j(x).$$

Матричная задача на собственные значения для отыскания оптимальных весовых коэффициентов при $p = 1$, $q = 0$ принимает вид

$$\frac{6}{5h}(2u_j - u_{j+1} - u_{j-1}) + \frac{1}{10}(u'_{j+1} - u'_{j-1}) =$$
$$= \lambda^h\left[\frac{h}{70}(52u_j + 9u_{j+1} + 9u_{j-1}) - \frac{13h^2}{420}(u'_{j+1} - u'_{j-1})\right], \qquad (30)$$
$$\frac{h}{30}(8u'_j - u'_{j+1} - u'_{j-1}) - \frac{1}{10}(u_{j+1} - u_{j-1}) =$$
$$= \lambda^h\left[\frac{h^3}{420}(8u'_j - 3u'_{j+1} - 3u'_{j-1}) + \frac{13h^2}{420}(u_{j+1} - u_{j-1})\right]. \qquad (31)$$

[1]) Ради простоты мы будем требовать, чтобы все функции $v^h \in S^h$ удовлетворяли как главному краевому условию в точке $x = 0$, так и естественному условию при $x = \pi$.

Строго говоря, (30) и (31) выполняются только для $1 \leqslant j \leqslant N-1$, но эти уравнения справедливы и для граничных точек, если положить

$$u_0 = u_{-1} + u_1 = u'_1 - u'_{-1} = u'_N = u'_{N+1} + u'_{N-1} = u_{N+1} - u_{N-1} = 0.$$

(Заметим, что уравнение (30) сводится к равенству $0 = 0$ при $j = 0$, а (31) — при $j = N$.)

Оказывается, что как и в кусочно линейном случае, приближенные собственные функции этой простой задачи совпадают в узлах с тригонометрическими полиномами:

$$u_j = \sqrt{\frac{\pi}{2}} \sin\left(\left(l - \frac{1}{2}\right) jh\right),$$
$$u'_j = \sqrt{\frac{\pi}{2}} \left(l - \frac{1}{2}\right) \alpha_l \cos\left(\left(l - \frac{1}{2}\right) jh\right). \qquad (32)$$

Действительно, подставив (32) в (30)—(31) и обозначив $\nu_l = l - 1/2$, получим задачу на собственные значения размера 2×2:

$$\begin{bmatrix} \frac{12}{5h}(1 - \cos \nu_l h) & -\frac{\sin \nu_l h}{5} \\ -\frac{\sin \nu_l h}{5} & \frac{h}{15}(4 - \cos \nu_l h) \end{bmatrix} \begin{bmatrix} 1 \\ \nu_l \alpha_l \end{bmatrix} =$$
$$= \lambda_l^h \begin{bmatrix} \frac{h}{70}(52 + 18 \cos \nu_l h) & \frac{13}{210} h^2 \sin \nu_l h \\ \frac{13}{210} h^2 \sin \nu_l h & h^3\left(\frac{2}{105} - \frac{\cos \nu_l h}{70}\right) \end{bmatrix} \begin{bmatrix} 1 \\ \nu_l \alpha_l \end{bmatrix}, \qquad (33)$$

Заметим, что (33) дает для каждого целого числа $l < N$ два собственных значения $\lambda^h_{l,0}$ и $\lambda^h_{l,1}$. Так как система (30)—(31) имеет порядок $2N - 2$, то они обязательно будут собственными значениями задачи в методе конечных элементов. Непосредственным вычислением находим

$$\lambda^h_{l,0} = \lambda_l + O\left(l^8 h^6\right). \qquad (34)$$

Порядок ошибки для s-й производной от соответствующей собственной функции равен $l^4 h^{4-s}$.

Порядок остальных собственных значений $\lambda^h_{l,1}$ есть $O(h^{-2})$, и *они не приближают никакое собственное значение дифференциального уравнения.* На первый взгляд это кажется серьезным препятствием к применению кубических эрмитовых элементов для вычисления собственных значений, поскольку *по крайней мере половина собственных значений матричной задачи совершенно бесполезна при аппроксимации.* Однако это явление до-

вольно типично для аппроксимаций методом конечных элементов. Более пристальное рассмотрение даже кусочно линейных приближений показывает, что для каждого числа h найдется такое целое число l_h, что годятся лишь первые l_h собственных значений; более того, $hl_h \to 0$ при $h \to 0$. Это имеет важное значение при выборе метода решения задачи $KQ = \lambda MQ$: *нужно вычислять только главные собственные значения.*

Закончим этот раздел замечанием о том, что «приближенный расчет» матрицы M^h может привести к серьезным потерям точности [T9]. Например, типичный подход — заменить в (30) выражения

$$\frac{1}{70}(52u_j + 9u_{j+1} + 9u_{j-1}) \quad \text{и} \quad \frac{1}{10}(u'_{j+1} - u'_{j-1})$$

на u_j и 0 соответственно и в (31)

$$\frac{1}{420}(8u'_j - 3u'_{j+1} - 3u'_{j-1}), \quad \frac{13}{420}(u_{j+1} - u_{j-1})$$

на $u'_j/210$ и 0. Это приведет к задаче $\tilde{K}\tilde{Q} = \lambda Q$, более простой, поскольку M заменится единичной матрицей. Однако прямое вычисление показывает, что ошибка собственных значений порядка $O(h^6)$ увеличится тогда до $O(h^2)$.

6.3. ОШИБКИ В СОБСТВЕННЫХ ЗНАЧЕНИЯХ И СОБСТВЕННЫХ ФУНКЦИЯХ

В этом разделе мы изложим общую теорию аппроксимации по Рэлею — Ритцу в применении к эллиптическим задачам на собственные значения $Lu = \lambda u$. Постановка задачи хорошо известна: интегрирование по частям преобразует (Lv, v) в симметричную форму $a(v, v)$, определенную для всех v из пространства допустимых функций $\mathscr{H}_E^m$. Собственными функциями будут точки u_l, в которых отношение Рэлея $R(v) = a(v, v)/(v, v)$ стационарно, а соответствующими собственными значениями будут $\lambda_l = R(u_l)$. Собственные функции ортогональны, и в силу симметричности оператора L собственные значения вещественны.

На подпространстве S^h отношение Рэлея принимает вид

$$R(v^h) = \frac{a(v^h, v^h)}{(v^h, v^h)} = \frac{q^T K q}{q^T M q}$$

и стационарные точки Q дают приближение для u_l^h и λ_l^h. Эти точки определяются матричной задачей на собственные значения $KQ = \lambda^h MQ$, и любые два собственных вектора Q_i и Q_l удовлетворяют обычным соотношениям ортогональности:

$$Q_i^T M Q_l = \delta_{il}, \quad Q_i^T K Q_l = \lambda_l^h \delta_{il}. \tag{35}$$

В терминах собственных функций $u_i^h = \sum (Q_i)_j \varphi_j$ и $u_l^h = \sum (Q_l)_j \varphi_j$ это означает, что

$$(u_i^h, u_l^h) = \delta_{il}, \quad a(u_i^h, u_l^h) = \lambda_l^h \delta_{il}. \tag{36}$$

Таким образом, приближения отражают основные свойства точных решений. Для них также справедлив принцип минимакса:

$$\lambda_l^h = \min_{S_l \subset S^h} \max_{v^h \in S_l} R(v^h), \tag{37}$$

где S_l — любое l-мерное подпространство.

Рассмотрим сначала оценки для собственных значений. Пусть P — проектор Рэлея — Ритца, определенный следующим образом: если функция u принадлежит $\mathscr{H}_E^m$, то Pu — составляющая в подпространстве S^h (относительно энергетического скалярного произведения):

$$a(u - Pu, v^h) = 0 \quad \text{для всех} \quad v^h \in S^h. \tag{38}$$

Это означает, что в энергетической норме $a(v, v)$ функция Pu — ближайшая в S^h к заданной функции u. Другими словами, если бы u было решением стационарной задачи $Lu = f$, то Pu было бы в точности его аппроксимацией u^h по методу Ритца. Вместе с нашим предыдущим результатом по аппроксимации (теорема 3.7) это гарантирует, что

$$\| u - Pu \|_s \leqslant C [h^{k-s} + h^{2(k-m)}] \| u \|_k. \tag{39}$$

Мы будем оценивать $\lambda_l^h - \lambda_l$ так: обозначим через E_l подпространство, натянутое на точные собственные функции $u_1, \ldots, u_l$, и возьмем $S_l = PE_l$ в качестве подпространства в S^h, используемого в принципе минимакса. Таким образом, S_l натягивается на пробные функции $Pu_1, \ldots, Pu_l$. Они не совпадают тождественно с приближенными собственными функциями $u_1^h, \ldots, u_l^h$, но в доказательстве важно лишь, что они близки к последним.

Лемма 6.1. *Пусть e_l — множество единичных векторов в E_l и*

$$\sigma_l^h = \max_{u \in e_l} |2(u, u - Pu) - (u - Pu, u - Pu)|. \tag{40}$$

Тогда при условии $\sigma_l^h < 1$ приближенные собственные значения ограничены сверху:

$$\lambda_l^h \leqslant \frac{\lambda_l}{1 - \sigma_l^h}. \tag{41}$$

Доказательство. Чтобы применить принцип минимакса, надо быть уверенным, что подпространство $S_l = PE_l$ l-мерно. Ясно, что E_l само l-мерно, так что вопрос сводится к следующему: может ли равенство $Pu^* = 0$ выполняться для ненулевого вектора $u^* \in E_l$? Нормализуем вектор u^* так, чтобы он был единичным, т. е. принадлежал e_l. Тогда равенство $Pu^* = 0$ означает, что

$$\sigma_l^h \geqslant |2(u^*, u^* - Pu^*) - (u^* - Pu^*, u - Pu^*)| = |(u^*, u^*)| = 1.$$

Так как это противоречит условию $\sigma_l^h < 1$, то S_l должно быть l-мерным.

Теперь из принципа минимакса (37) получаем

$$\lambda_l^h \leqslant \max_{v^h \in S_l} R(v^h) = \max_{u \in e_l} \frac{a(Pu, Pu)}{(Pu, Pu)}.$$

Поскольку P — проектор в энергетической норме [1]), числитель здесь ограничен сверху: $a(Pu, Pu) \leqslant a(u, u)$. Знаменатель ограничен снизу:

$$(Pu, Pu) = (u, u) - 2(u, u - Pu) + (u - Pu, u - Pu) \geqslant 1 - \sigma_l^h.$$

Поэтому

$$\lambda_l^h \leqslant \max_{u \in e_l} \frac{a(u, u)}{1 - \sigma_l^h} = \frac{\lambda_l}{1 - \sigma_l^h},$$

и лемма доказана.

Теперь задача — оценить σ_l^h; для этого нам понадобится одно тождество.

Лемма 6.2. *Если* $u = \Sigma_1^l c_i u_i$ *принадлежит* e_l, *то*

$$(u, u - Pu) = \sum c_i \lambda_i^{-1} a(u_i - Pu_i, u - Pu). \qquad (42)$$

Доказательство. Так как u_i — истинная собственная функция, то в силу (10)

$$(u_i, u - Pu) = \lambda_i^{-1} a(u_i, u - Pu).$$

Кроме того, если в определении (38) проектора P положить $v^h = Pu_i$, то $a(Pu_i, u - Pu) = 0$. Вычитая, получаем

$$(u_i, u - Pu) = \lambda_i^{-1} a(u_i - Pu_i, u - Pu).$$

---

[1]) Из теоремы 1.1 следует (и легко вывести непосредственно), что

$$a(v, v) = a(Pv, Pv) - 2a(v - Pv, Pv) + a(v - Pv, v - Pv).$$

Последнее слагаемое неотрицательно, а предпоследнее по определению (38) проектора P равно нулю.

Умножив последнее равенство на c_i и просуммировав по i, придем к (42).

Теперь найдем основную оценку ошибок в собственных значениях.

Теорема 6.1. *Если S^h — пространство метода конечных элементов степени $k-1$, то существует такая постоянная δ, что для малых h приближенные собственные значения ограничены:*

$$\lambda_l \leqslant \lambda_l^h \leqslant \lambda_l + 2\delta h^{2(k-m)}\lambda_l^{k/m}. \tag{43}$$

Это согласуется с явными оценками (28) *и* (34) *для линейных и кубических элементов в одномерном случае.*

Доказательство. Мы хотим оценить σ_l^h. Так как $|a(v, w)| \leqslant K\|v\|_m\|w\|_m$, то для первого слагаемого в σ_l^h имеем

$$2|(u, u-Pu)| = 2\left|\sum_1^l c_i\lambda_i^{-1}a(u_i - Pu_i, u-Pu)\right| \leqslant$$
$$\leqslant 2K\left\|(I-P)\sum c_i\lambda_i^{-1}u_i\right\|_m\|(I-P)u\|_m.$$

Применим теорему аппроксимации:

$$2|(u, u-Pu)| \leqslant 2KC^2h^{2(k-m)}\left\|\sum_1^l c_i\lambda_i^{-1}u_i\right\|_k\|u\|_k \leqslant$$

$$\leqslant C'h^{2(k-m)}\left\|\sum_1^l c_i\lambda_i^{(k/2m)-1}u_i\right\|_0\left\|\sum_1^l c_i\lambda_i^{(k/2m)}u_i\right\|_0 \leqslant C'h^{2(k-m)}\lambda_l^{(k/m)-1}. \tag{44}$$

Так как $\sum c_i^2 = 1$, то эти соотношения справедливы для всех функций $u = \sum c_iu_i$ из e_i; крайний случай возникает при $c_l = 1$, поскольку этот коэффициент умножается на степень наибольшего из собственных значений λ_l. Предпоследний шаг при выводе (44) был более тонкий; здесь потребовалось неравенство $\|v\|_k \leqslant c\|L^{k/2m}v\|_0$. Для $k = m$ это не что иное, как условие эллиптичности $\|v\|_m^2 \leqslant c^2a(v, v)$, и выполняется оно для всех k, если коэффициенты дифференциального оператора L гладкие.

Другой член в σ_l^h более высокого порядка относительно h, так как в силу теоремы аппроксимации при $s = 0$

$$(u-Pu, u-Pu) \leqslant C^2[h^k + h^{2(k-m)}]^2\|u\|_k^2. \tag{45}$$

Поэтому, если заменить постоянную C' в (44) большей постоянной δ, то результат будет превышать σ_l^h для всех малых h. Если теперь взять h достаточно малым для того, чтобы обеспечить $\sigma_l^h \leqslant 1/2$, то

$$\lambda_l^h \leqslant \lambda_l(1-\sigma_l^h)^{-1} \leqslant \lambda_l(1+2\sigma_l^h) \leqslant \lambda_l + 2\delta h^{2(k-m)}\lambda_l^{k/m}.$$

Теорема доказана.

Подобные оценки ошибок в методе Рэлея — Ритца имеют длинную историю, особенно в советской научной литературе. Вайникко [9] дал весьма завершенную теорию, включающую оценки ошибок в собственных значениях и собственных функциях даже для несамосопряженных случаев. Вместе с теоремами об аппроксимации для метода конечных элементов его анализ приводит к оценкам, доказанным выше с помощью принципа минимакса для самосопряженных задач. Этот принцип можно заменить интегрированием по контуру в комплексной плоскости, а затем применить оценки Галёркина для стационарных задач (разд. 2.3). Интеграл от $(L - zI)^{-1}$ вдоль контура вокруг истинного собственного значения λ_l равен точно $2\pi i u_l^T u_l$; Брамбл и Осборн [Б25], а также Бабушка и Фикс вычислили ошибку, возникающую, если применить здесь метод конечных элементов.

Метод минимакса, который мы выбрали для самосопряженных задач из-за его простоты, применялся рядом авторов при исследовании конечных элементов; основные ссылки приведены в [Б15]. Мы привели более аккуратное доказательство для того, чтобы определить не только степень $h^{2(k-m)}$, но также и правильную зависимость от l:

$$\lambda_l^h - \lambda_l \sim ch^{2(k-m)}\lambda_l^{k/m}.$$

Последний множитель означает, что *гораздо труднее вычислять собственные значения более высоких порядков*. Это утверждение проверено экспериментально и появление множителя $\lambda_l^{k/m}$ подтверждено описанными в [К14] вычислениями, в которых использовались элементы пятой степени (они будут рассматриваться в разд. 8.4 при вычислении собственных значений L-образной мембраны), и следующими численными результатами, взятыми из [Л3]. При расчете квадратной пластины были применены бикубические эрмитовы элементы. Пластина подпиралась либо на всех сторонах (SSSS), либо только на двух, а другие две оставались свободными (SFSF); для каждой стороны берется 10 элементов (рис. 6.1). Так как $k = 4$ и $m = 2$, то, согласно теоретическим результатам, относительная ошибка равна

$$\frac{\lambda_l^h - \lambda_l}{\lambda_l} \sim ch^4\lambda_l.$$

Обобщенная задача на собственные значения $Lu = \lambda Bu$ исследуется точно так же. Предположим, что B — симметричный оператор порядка $2m' < 2m$, соответствующий квадратичной форме $(Bv, v) = b(v, v)$. Тогда истинное отношение Рэлея есть $R(v) = a(v, v)/b(v, v)$, а дискретное имеет вид $q^T Kq/q^T M_b q$, где матрица массы M_b построена относительно нового скалярного про-

изведения: $(M_b)_{jk} = b(\varphi_j, \varphi_k)$. Изменяется только последний член в σ_l^h, для которого оценка (45) в 0-норме заменяется на

$$b(u - Pu, u - Pu) \leqslant C^2 [h^{k-m'} + h^{2(k-m)}]^2 \| u \|_k^2.$$

Однако преобладает по величине первый член в σ_l^h, а именно $2b(u, u - Pu) \sim h^{2(k-m)}$, так что окончательная оценка в теореме не меняется.

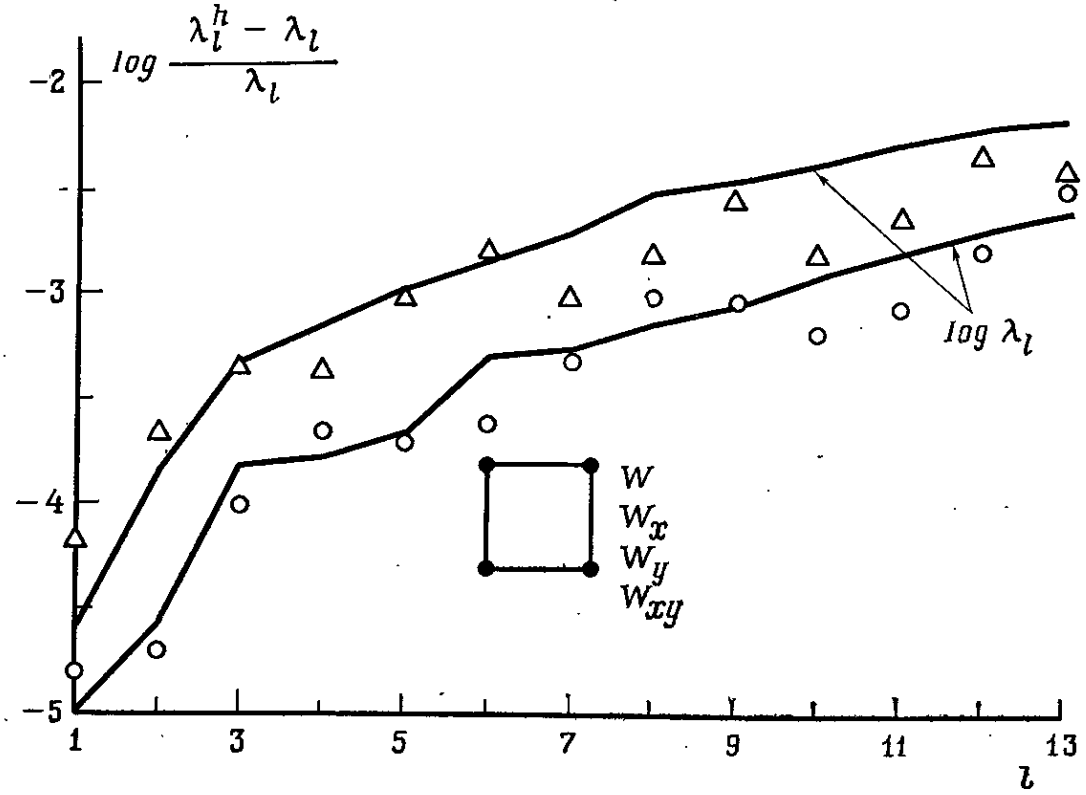


Рис. 6.1.
Ошибки собственных значений более высоких порядков для квадратной пластины. ○ — SFSF, △ — SSSS.

Влияние на вычисляемые собственные значения других приближений — изменение области или коэффициентов, численные квадратуры, несогласованность элементов — сравнимо в методе конечных элементов с влиянием этих возмущений на энергию в стационарных задачах. Предупреждаем лишь, что при замене области Ω многоугольником Ω^h ошибка в энергии должна измеряться на Ω. Поэтому ее порядок уже не $O(h^3)$, как было установлено в разд. 4.4 на многоугольнике: если все пробные функции равны нулю на $\Omega - \Omega^h$, то энергия по этой области полностью теряется; соответствующее возмущение, пропорциональное площади этой подобласти, есть $O(h^2)$.

Займемся собственными функциями. Хорошо было бы доказать, что их ошибки той же величины, что и в стационарных задачах $Lu = f$. Следует, конечно, ожидать, что появится некоторая зависимость от l и точность для собственных функций более

высоких порядков ухудшится. Наши рассуждения будут технически сложнее и снова будут опираться на два простых тождества, доказываемых в следующих леммах.

Лемма 6.3. *При нормировке* $(u_l, u_l) = (u_l^h, u_l^h) = 1$

$$a(u_l - u_l^h, u_l - u_l^h) = \lambda_l \|u_l - u_l^h\|_0^2 + \lambda_l^h - \lambda_l. \tag{46}$$

Доказательство. Заметим, что

$$\begin{aligned} a(u_l - u_l^h, u_l - u_l^h) &= a(u_l, u_l) - 2a(u_l, u_l^h) + a(u_l^h, u_l^h) = \\ &= \lambda_l - 2\lambda_l(u_l, u_l^h) + \lambda_l^h = \\ &= \lambda_l[2 - 2(u_l, u_l^h)] + \lambda_l^h - \lambda_l. \end{aligned}$$

Величина в квадратных скобках равна

$$2 - 2(u_l, u_l^h) = (u_l, u_l) - 2(u_l, u_l^h) + (u_l^h, u_l^h) = \|u_l - u_l^h\|_0^2,$$

так что тождество (46) доказано.

Теперь, когда доказано тождество (46), а еще раньше найдены оценки для $\lambda_l^h - \lambda_l$, осталось лишь оценить ошибку собственной функции в норме $\|u_l - u_l^h\|_0$; тогда ошибка в энергии будет непосредственно вытекать из полученных результатов.

Лемма 6.4. *Для всех* j *и* l

$$(\lambda_j^h - \lambda_l)(Pu_l, u_j^h) = \lambda_l(u_l - Pu_l, u_j^h). \tag{47}$$

Доказательство. Так как член $-\lambda_l(Pu_l, u_j^h)$ фигурирует в обеих частях тождества (47), достаточно показать, что

$$\lambda_j^h(Pu_l, u_j^h) = \lambda_l(u_l, u_j^h).$$

Так как u_j^h и u_l — собственные функции, обе части последнего равенства можно переписать в виде $a(Pu_l, u_j^h)$ и $a(u_l, u_j^h)$ соответственно. Доказываемое равенство вытекает из определения (38) проектора P.

Выражение (47) сходно с ошибкой отсечения в уравнении задачи на собственные значения; оно совпадает с $a(Pu_l, u_j^h) - \lambda_l(Pu_l, u_j^h)$.

Множество $u_1^h, \ldots, u_N^h$ образует ортогональный базис в S^h и, в частности,

$$Pu_l = \sum_{j=1}^{N} (Pu_l, u_j^h) u_j^h. \tag{48}$$

Тождества (47) и (48) можно интерпретировать так: из (47) видно, что коэффициент (Pu_l, u_j^h) мал, если λ_j^h не близко к λ_l, а тогда из (48) следует, что Pu_l близко к u_l^h. Это будет нашей стратегией при оценивании $u_l^h - Pu_l$ (а значит, и $u_l^h - u_l$), но чтобы процесс был строгим, удобно рассмотреть отдельно случаи различных и кратных собственных значений.

Если λ_l отлично от других собственных значений, то, согласно оценкам (43), найдется такая постоянная ρ, что для малых h

$$\frac{\lambda_l}{|\lambda_j^h - \lambda_l|} \leqslant \rho \quad \text{для всех } j. \tag{49}$$

Теперь приступим к вычислениям. Обозначим через β ведущий коэффициент (Pu_l, u_l^h) в (48) и оценим остальные члены

$$\begin{aligned} \|Pu_l - \beta u_l^h\|_0^2 &= \sum_{j \neq l} (Pu_l, u_j^h)^2 = \\ &= \sum_{j \neq l} \left(\frac{\lambda_l}{\lambda_j^h - \lambda_l}\right)^2 (u_l - Pu_l, u_j^h)^2 \leqslant \\ &\leqslant \rho^2 \sum_{j \neq l} (u_l - Pu_l, u_j^h)^2 \leqslant \\ &\leqslant \rho^2 \|u_l - Pu_l\|_0^2 \end{aligned} \tag{50}$$

(здесь учтено, что квадрат нормы равен сумме квадратов компонент). Итак, получили основную оценку:

$$\|u_l - \beta u_l^h\|_0 \leqslant \|u_l - Pu_l\|_0 + \|Pu_l - \beta u_l^h\|_0 \leqslant (1+\rho)\|u_l - Pu_l\|_0. \tag{51}$$

Ошибка собственной функции оценивается теперь через ошибку аппроксимации по Ритцу $u_l - Pu_l$; из (39) вытекает

$$\begin{aligned} \|u_l - \beta u_l^h\|_0 &\leqslant C'(1+\rho)\left[h^k + h^{2(k-m)}\right]\|u_l\|_k \leqslant \\ &\leqslant C''\left[h^k + h^{2(k-m)}\right]\lambda_l^{k/2m}. \end{aligned} \tag{52}$$

Это существенная часть нашей теоремы.

Теорема 6.2. *Если S^h — пространство метода конечных элементов степени $k-1$ и λ_l — отличное от других собственное значение, то для малых h*

$$\|u_l - u_l^h\|_0 \leqslant c\left[h^k + h^{2(k-m)}\right]\lambda_l^{k/2m}, \tag{53}$$

$$a(u_l - u_l^h, u_l - u_l^h) \leqslant c' h^{2(k-m)} \lambda_l^{k/m}. \tag{54}$$

Если λ_l — кратное собственное значение, то можно выбрать ортонормальные собственные функции u_j так, чтобы эти оценки

также были справедливы. Оценки (53) *и* (54) *неулучшаемы и согласуются с частным случаем — оценкой* (27) *для линейных элементов.*

Доказательство. Неравенство (53) по существу совпадает с доказанным неравенством (52). Осталось лишь показать, что множитель β близок к 1. Применим неравенство треугольника:

$$\|u_l\|_0 - \|u_l - \beta u_l^h\|_0 \leqslant \|\beta u_l^h\|_0 \leqslant \|u_l\|_0 + \|u_l - \beta u_l^h\|_0.$$

Если вспомнить, что u_l и u_l^h — единичные векторы, и выбрать их знаки так, чтобы $\beta \geqslant 0$, то последнее неравенство в выписанной цепочке равносильно $|\beta - 1| \leqslant \|u_l - \beta u_l^h\|_0$. Поэтому

$$\|u_l - u_l^h\|_0 \leqslant \|u_l - \beta u_l^h\|_0 + \|(\beta - 1) u_l^h\|_0 \leqslant 2\|u_l - \beta u_l^h\|_0.$$

Правая часть оценивается по (52), и неравенство (53) доказано: $c = 2C''$. Ошибка в энергии (54) немедленно следует из леммы 6.3. (Это рассуждение проще всех известных нам рассуждений, связанных с собственными функциями.)

Случай кратного собственного значения $\lambda_l = \lambda_{l+1} = \ldots = \lambda_{l+R}$ труднее, но различия несущественны. Как и в (49), здесь также найдется постоянная ρ, отделяющая эти собственные значения от аппроксимаций λ_j^h других собственных значений. Постоянная β становится матрицей порядка $R + 1$,

$$\beta_{ri} = (Pu_{l+r}, u_{l+i}^h), \quad 0 \leqslant i, \ r \leqslant R.$$

Длинная выкладка (50) теперь проделывается со всеми членами с номерами $j = l, \ldots, l + R$ в левой части, а не правой, и это дает

$$\left\| \sum_0^R \beta_{ri} u_{l+i}^h - u_{l+r} \right\|_0 \leqslant C\left[h^k + h^{2(k-m)}\right] \lambda_l^{k/2m}.$$

Обращая матрицу β, видим, что новые собственные функции U_{l+r} (линейные комбинации прежних) можно выбрать так, что

$$\|u_{l+r}^h - U_{l+r}\|_0 \leqslant C\left[h^k + h^{2(k-m)}\right] \lambda_l^{k/2m}.$$

Так как известно, что u_{l+r}^h ортонормальны, то без ущерба для оценки можно считать U_{l+r} также ортонормальными.

Эта теорема и ее доказательство (с заменой лишь повсюду скалярного произведения (u, v) на $b(u, v)$) применяются и к обобщенной задаче на собственные значения $Lu = \lambda Bu$.

6.4. ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ

Принцип Рэлея — Ритца привел к матричной задаче на собственные значения $KQ = \lambda MQ$, которую и надо теперь решить. Однако эта задача не тривиальна, и она почти не рассматривается в обычных учебниках по линейной алгебре. Эффективный алгоритм ее решения должен учитывать симметричность и положительную определенность матриц K и M, а также их разреженность. Последнее свойство, например, потерялось бы, если бы мы разложили M в произведение LL^T (исключение Холесского) и вычисляли собственные значения матрицы $L^{-1}K(L^{-1})^T$ с помощью обычного алгоритма. (Мы выбрали бы QR-алгоритм, начинающийся с преобразования исходной матрицы в треугольную, а не более старый метод Якоби.) Эта потеря разреженности не будет слишком серьезной в небольших задач, которые можно решать при наличии лишь оперативной памяти ЭВМ, но для больших систем этот подход не эффективен.

Мы предлагаем отыскивать собственные значения (точнее, несколько первых собственных значений, поскольку было бы бесполезно вычислять собственные значения высоких порядков, не имеющие физического смысла) непосредственно из уравнения $KQ = \lambda MQ$. Мы отказываемся от приближенного расчета матрицы M, поскольку диагональная матрица M ненамного лучше ленточной.

Есть также метод *экономизации*, уменьшающий порядок системы за счет того, что в вычислениях участвует лишь небольшое число *ведущих переменных*. Априори предполагается зависимость других *подчиненных переменных*, тем самым эти степени свободы исключаются [7]. Фрид описал эту идею так: поместим точечную нагрузку в узел z_j и обозначим через Φ_j решение стационарной задачи, построенное по методу конечных элементов; тогда эти функции Φ_j образуют базис (не локальный в отличие от базиса, образованного исходными функциями φ_j) для пространства пробных функций в экономизированной задаче. Можно ожидать, что эти функции достаточно хорошо представляют низкочастотные колебания, именно их и надо вычислять. Тем не менее у нас создается впечатление, что, по мере того как будут разрабатываться эффективные алгоритмы для исходной задачи, эта экономизация будет становиться менее необходимой и менее популярной.

Авторитетная библиография по алгоритмам решения задач на собственные значения в период до 1970 г. содержится в [21]. Прежде чем описать новый метод, мы хотим рассмотреть очень широко используемый метод из этого класса — *метод обратной итерации*, или *обратной степени*. В своей простейшей форме для задачи на собственные значения вида $Ax = \lambda x$ обратная итера-

ция заключается в решении линейной системы на каждом шаге: $Ay_{n+1} = x_n$. Тогда приближением к λ будет $\lambda_{n+1} = 1/\|y_{n+1}\|$, а новым приближением к x — нормализованный вектор $x_{n+1} = \lambda_{n+1} y_{n+1}$. Они были бы точными, если бы x_n был собственным вектором. Если представить себе, что начальный вектор x_0 разлагается по истинным собственным векторам v_j, т. е. $x_0 = \Sigma c_j v_j$, то в результате n обратных итераций каждая компонента увеличится в $(\lambda_j)^{-n}$ раз; вектор x_n пропорционален $\Sigma c_j (\lambda_j)^{-n} v_j$. Если λ_1 значительно меньше других собственных значений, то первая компонента станет преобладающей и x_n будет приближать единичный собственный вектор v_1. Сходимость подобна сходимости геометрической прогрессии со знаменателем λ_1/λ_2; ошибка $x_n - x$ имеет порядок $(\lambda_1/\lambda_2)^n$. Очевидно, что метод эффективнее, когда это отношение мало.

Один способ уменьшить это отношение состоит в том, чтобы сдвинуть начало координат, заменяя A на n-м шаге матрицей $A - \lambda_n I$. Это сдвигает все собственные значения матрицы A на одну и ту же величину λ_n. Если λ_n близко к истинному собственному значению λ, так что разность $\lambda - \lambda_n$ мала, то соответствующая компонента вектора y_{n+1} увеличивается на большой множитель $(\lambda - \lambda_n)^{-1}$. Этот процесс легко численно реализуем, даже хотя матрица $A - \lambda_n I$ почти вырождена. На самом деле полезно иметь несколько лучшее приближение к собственному значению, чем $\lambda_n = 1/\|y_n\|$. Например, отношение Рэлея $\lambda_n = (Ay_n, y_n)/(y_n, y_n)$ гораздо точнее. График этого отношения в окрестности точного собственного значения (где расположена стационарная точка) очень плоский, и алгоритм с этими улучшенными сдвигами обладает кубической сходимостью: $\lambda_{n+1} - \lambda \sim (\lambda_n - \lambda)^3$. С другой стороны, сдвиг $A \to A - \lambda_n I$ означает, что на каждой итерации надо снова выполнять исключение Гаусса: треугольные матрицы, на которые разлагается A, нельзя хранить и использовать в последующих итерациях, как это происходит в случае простой итерации $Ay_{n+1} = x_n$.

В обобщенной задаче на собственные значения проще всего было бы на каждом этапе решать задачу $Ky_{n+1} = Mx_n$, а затем, нормализуя y_{n+1}, получать x_{n+1}. Однако при этом матрица $K^{-1}M$ не будет симметричной. Поэтому если представить K в виде разложения Холесского $B \cdot B^T$, то можно использовать в итерационном процессе симметричную матрицу $B^{-1}M(B^T)^{-1}$. Но тогда в процесс вовлекаются неразреженные матрицы, которых мы хотели избежать, однако очевидно, что здесь итерация не начинается с их перемножения. Вместо этого для определения приближения u_{n+1} решаются уравнения $B^T v_{n+1} = u$, $Bw_{n+1} = Mv_{n+1}$ и затем нормализуется вектор w_{n+1}. Сходимость нормировочных множителей к λ_1 и векторов v_n (а не u_n!) к собственному вектору x_1 снова зависит от отношения λ_1/λ_2.

Очень полезный вариант обратной итерации — *блочно-степенной метод*, предложенный Бауэром и улучшенный Рутисхаузером, Петерсом и Уилкинсоном [1]). Его идея — одновременно искать несколько собственных значений, рассматривая в итерациях l приближенных собственных векторов. (Очевидно, что они должны спариваться по мере продолжения процесса, иначе получатся l различных приближений к одному и тому же основному колебанию.) Скорость сходимости приближений к λ_i равна λ_i/λ_{l+1}, и можно без труда вычислять кратные собственные значения.

Блочная итерация для уравнения $Ax = \lambda x$ выполняется следующим образом. Пусть в качестве столбцов матрицы P_0 размера $N \times l$ взяты l начальных векторов, предполагаемых ортонормальными. Первый шаг: решаем l уравнений $AZ_1 = P_0$. Затем, вместо нормализации отдельно каждого столбца из Z_1, мы ортонормализуем все l столбцов. Для этого образуем матрицу $Z_1^T Z_1$ размера $l \times l$ и найдем ее собственные значения μ_i^{-2} (первые приближения к $\lambda_1^{-2}, \ldots, \lambda_l^{-2}$) и соответствующие им собственные векторы w_i. Новое приближение P_1 есть произведение матрицы Z_1 и матрицы W со столбцами $w_1\mu_1, \ldots, w_l\mu_l$. Так как $P_1^T P_1 = W^T Z_1^T Z_1 W = I$, то столбцы матрицы P_1 ортонормальны (это приближенные собственные векторы матрицы A). На следующем шаге решаем уравнения $AZ_2 = P_1$ и т. д.

Описанный алгоритм воспроизводит точно собственные векторы $v_1, \ldots, v_l$ матрицы A. Точнее, пусть каждый столбец из P_0 записан в виде линейной комбинации векторов $v_1, \ldots, v_l$, т. е. $P_0 = VQ$, где Q — ортогональная матрица размера $l \times l$, а V — матрица размера $N \times l$, со столбцами $v_1, \ldots, v_l$. Отметим, что если Λ — диагональная матрица с элементами $\lambda_1, \ldots, \lambda_l$, то $V^T V = I$ и $AV = V\Lambda$.

Таким образом, первый шаг блочно-степенного метода дает $Z_1 = A^{-1}P_0 = A^{-1}VQ = V\Lambda^{-1}Q$, и $Z_1^T Z_1$ превращается в $Q^T\Lambda^{-2}Q$. Так как матрица Q ортогональна, то собственные значения μ_i^{-2} матрицы $Z_1^T Z_1$ равны элементам λ_i^{-2} диагональной матрицы Λ^{-2}. Другими словами, если $P_0 = VQ$, то в результате первого шага получим $P_1 = V$ и правильные собственные значения.

Изменения в алгоритме, вызванные решением задачи $AQ = \lambda MQ$ вместо $Ax = \lambda x$, описываются в последнем абзаце этой главы. Блочно-степенной метод в том виде, как он сформулирован здесь, легко поддается программированию.

Изложенные методы обратной степени очень просты и практичны, особенно когда достаточна небольшая точность. Но су-

---

[1]) В технической литературе он известен как *итерация в подпространстве*.

ществует также новый способ, основанный на более тонкой теореме о матрицах; наиболее успешно он применяется в задачах на собственные значения с ленточными матрицами $KQ = \lambda MQ$. Этот вариант предложен Петерсом и Уилкинсоном [П1, П2]; он опирается на следующую красивую теорему из теории матриц: *число собственных значений, меньших заданного* λ_0, *можно определить, подсчитывая количество отрицательных ведущих элементов при применении исключения Гаусса к матрице* $K - \lambda_0 M$.

Из этой теоремы вытекает алгоритм, использующий деление пополам. Допустим, что мы уже определили, что n_0 собственных значений меньше заданного λ_0. Тогда ведущие гауссовы элементы для $K - (\lambda_0/2)M$ дают n_1 собственных значений, меньших $\lambda_0/2$; остальные $n_0 - n_1$ собственных значений должны лежать между $\lambda_0/2$ и λ_0. Повторное деление пополам будет с большой точностью выделять любое собственное значение, но процесс требует исключения Гаусса на каждом шаге и потому займет много машинного времени. Его надо ускорить; это можно сделать, если использовать *значения* ведущих элементов (или их произведение, т. е. определитель $d(\lambda) = \det(K - \lambda M)$), а не только их знаки. Клаф, Бас и Парлетт предложили ускоренную итерацию метода секущих:

$$\lambda_{k+1} = \lambda_k - 2d(\lambda_k)\frac{\lambda_k - \lambda_{k-1}}{d(\lambda_k) - d(\lambda_{k-1})}.$$

Это вариант обычного метода Ньютона, в котором производная заменена разностным отношением; множитель 2 принят для ускорения. Как только λ_k станет близким к истинному значению λ, вычислитель может прекратить дальнейшую матричную факторизацию и вернуться к обычной обратной итерации. Мы можем подтвердить, что алгоритм Петерса — Уилкинсона очень успешно применялся в численных экспериментах, изложенных в гл. 8.

Дадим теперь теоретическое обоснование этого алгоритма. Оно опирается на классическую теорему, известную как *закон инерции Сильвестра: если две вещественные симметричные матрицы* A *и* D *связаны конгруэнтным преобразованием* $A = BDB^T$, *где* B *— любая невырожденная матрица, то у одной из них столько же отрицательных, положительных и нулевых собственных значений, сколько у другой.*

Доказательство особенно просто, когда матрица D невырожденна, т. е. не имеет нулевых собственных значений. Пусть B_θ — семейство невырожденных матриц, содержащее все матрицы от единичной (при $\theta = 0$) до матрицы B (при $\theta = 1$). (Мы не можем гарантировать, что конкретная матрица $B_\theta = \theta B + (1 - \theta)I$ будет невырожденной, но подходящую матрицу B_θ построить нетрудно.) Матрицы $B_\theta D B_\theta^T$ всегда симметричны и невырожденны. Поэтому их собственные значения $\lambda(\theta)$ вещественны, по-

следовательно изменяются вместе с θ и никогда не пересекают нуль. Следовательно, число собственных значений по каждую сторону от нуля остается одним и тем же как при $\theta = 1$, так и при $\theta = 0$. Другими словами, A и D имеют одинаковое число как отрицательных, так и положительных собственных значений. Если матрица D окажется вырожденной, то наши рассуждения можно провести для матриц $D \pm \varepsilon I$, а затем устремить $\varepsilon \to 0$.

Применим этот закон инерции к исключению Гаусса. Если матрица A, как и в разд. 1.5, представлена в виде LDL^T, где L — нижняя треугольная матрица с единицами на диагонали, а D — диагональная матрица из ведущих элементов, то знаки *ведущих элементов определяют знаки собственных значений.* (В случае когда две строки в процессе исключения меняются местами, что необходимо, если один из предыдущих ведущих элементов оказался равным нулю, то для сохранения симметричности матрицы надо поменять местами два соответствующих столбца. Такая перестановка двух строк и двух столбцов есть конгруэнтное преобразование, осуществляемое матрицей перестановки B. Поэтому в данном случае закон инерции все еще применим и после таких перестановок ведущие элементы исключения Гаусса все еще правильно определяют знаки собственных значений.)

Описанный закон можно применить также к обобщенной задаче на собственные значения $KQ = \lambda MQ$. Положим $A = K - \lambda_0 M$ и подсчитаем число отрицательных ведущих элементов в исключении Гаусса. Мы утверждаем, что оно равно числу n_0 собственных значений данной задачи $KQ = \lambda MQ$, меньших λ_0. В самом деле, перепишем рассматриваемую задачу в виде $M^{-1/2}KM^{-1/2}(M^{1/2}Q) = \lambda(M^{1/2}Q)$, так что n_0 — число собственных значений матрицы $M^{-1/2}KM^{-1/2}$, меньших λ_0. Оно равно числу отрицательных собственных значений матрицы $M^{-1/2}KM^{-1/2} - \lambda_0 I$. Но по закону инерции (снова! берем эту последнюю матрицу в качестве D, а $B = M^{1/2}$) оно совпадает с числом отрицательных собственных значений матрицы $A = K - \lambda_0 M$. Поэтому n_0 легко определяется из исключения Гаусса, примененного к A. Собственные векторы получаются в результате одного или в крайнем случае двух шагов обратной итерации с матрицей $C = K - \lambda_{\text{вычисл}} M$. В работе [П2] приведены хорошие начальные приближения к собственному вектору.

Эксперименты по сравнению алгоритмов для вычисления собственных значений все время продолжаются. Клаф и Бас нашли, что способы вычисления определителя (один из вариантов — метод Петерса — Уилкинсона) особенно эффективны, когда ширина ленты невелика: для каждого собственного значения надо проделать в среднем пять факторизаций на треугольные матрицы, а стоимость каждой итерации пропорциональна квадрату ширины ленты. Наибольший успех давал блочно-степенной метод.

Клаф и Бас представили алгоритм в следующей эквивалентной форме. Исходя из матрицы X_{n-1}, состоящей из l ортонормальных столбцов, в качестве которых взяты приближения к собственным векторам, решаем уравнение $KY_n = MX_{n-1}$. Затем решаем l-мерную задачу на собственные значения

$$(Y_n^T K Y_n) Q = \nu (Y_n^T M Y_n) Q. \tag{55}$$

Приближенными собственными значениями будут ν_i, а новая матрица X_n, состоящая из приближенных собственных векторов, равна произведению матрицы Y_n на квадратную матрицу порядка l, образованную из собственных векторов Q задачи (55). Бас и Парлетт детально изучили две вычислительные задачи: решение небольшой задачи (55), для которой они используют исключение типа Якоби, как только матрицы становятся почти диагональными, и выбор начальной матрицы X_0. При выборе l допускается компромисс — при больших l требуется мало итераций, но каждая из них довольно дорога. При вычислении первых p собственных значений они брали $l = \min(2p, p+8)$ и обнаружили, что восемь итераций дают отличные результаты. Этот способ эффективен даже для задач, слишком больших, чтобы работать только с оперативной памятью ЭВМ, и его можно с успехом применять к задачам на собственные значения, возникающим в методе конечных элементов.

7 ЗАДАЧИ С НАЧАЛЬНЫМИ УСЛОВИЯМИ

7.1. МЕТОД ГАЛЁРКИНА—КРАНКА—НИКОЛСОНА ДЛЯ УРАВНЕНИЯ ТЕПЛОПРОВОДНОСТИ

До сих пор мы обсуждали только стационарные задачи — задачи для эллиптических уравнений и на собственные значения. Метод Галёркина достаточно гибок, чтобы применить его и к задаче с начальными условиями; в этой главе мы рассмотрим для нее приближения по методу конечных элементов. Для общего вида областей и для задач, сравнительно медленно развивающихся (малые числа Рейнольдса в случае течения жидкости), конечные элементы все еще обладают важными преимуществами перед конечными разностями. Для задач о распространении волн с большими скоростями мы укажем некоторые их недостатки.

Хорошо известна основная теорема для конечных разностей: формально согласованный метод сходится тогда и только тогда, когда он устойчив. Это утверждение дословно применимо и к приближениям по Галёркину. В самом деле, единственное отличие состоит в том, что для конечных элементов зачастую легче проверить устойчивость и согласованность, чем для конечных разностей. Мы покажем далее, как метод Галёркина моделирует устойчивость дифференциального уравнения, а сейчас повторим то, что скрыто за предположением согласованности: *метод Галёркина согласован, если подпространства S^h плотны в пространстве допустимых функций.* Это означает, что каждую допустимую функцию v можно сколь угодно хорошо приблизить пробными функциями v^h при $h \to 0$. Теоремы аппроксимации гл. 3 устанавливают как раз это свойство; они дают даже *степень аппроксимации*, переходящую в степень согласованности, или *порядок точности*, уравнений Галёркина.

Эти идеи можно проиллюстрировать на уравнении теплопроводности

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = f(x, t), \quad 0 < x < \pi, \quad t > 0. \tag{1}$$

Это параболическое дифференциальное уравнение описывает процесс передачи тепла в стержне; $u = u(x, t)$ — температура в точке x в момент времени $t > 0$; $f(x, t)$ — член источника тепла. Как и в предыдущих наших примерах, мы налагаем крае-

вое условие Дирихле в точке $x = 0$ и естественное краевое условие при $x = \pi$:

$$u(0, t) = \frac{\partial u}{\partial x}(\pi, t) = 0. \tag{2}$$

Физически первое условие означает, что температура левого конца стержня поддерживается равной 0. С другой стороны, условие Неймана означает, что правый конец стержня изолирован: нет градиента температуры через точку $x = \pi$. Чтобы закончить постановку задачи, выбираем начальную температуру:

$$u(x, 0) = u_0(x), \quad 0 \leqslant x \leqslant \pi. \tag{3}$$

Эта классическая формулировка задачи о передаче тепла чревата трудностями. Например, условия (2) и (3) противоречивы, если $u_0 \neq 0$ при $x = 0$ или $\partial u_0/\partial x \neq 0$ при $x = \pi$. Кроме того, f может быть точечным источником, сосредоточенным в некоторой точке x_0, и тогда уравнение (1) в этой точке не выполняется. Во всех этих случаях лежащая в основе физическая задача все еще имеет смысл и состоит в нахождении распределения температуры, соответствующего начальной температуре u_0 и источнику тепла f. Поэтому, как и в стационарном случае, мы ищем вторую, интегральную формулировку задачи.

Так как здесь не работает естественный принцип минимума энергии (уравнение не самосопряженное), мы возвращаемся к задаче в слабой форме: при каждом $t > 0$

$$\int_0^\pi (u_t - u_{xx} - f)\, v\, dx = 0. \tag{4}$$

Для стационарного случая при $u_t = 0$ и $f = f(x)$ эта формулировка совпадает с прежней формулировкой Галёркина.

Чтобы достичь большей симметрии между пробной функцией и функцией $u(x, t)$, проинтегрируем $-u_{xx}v$ по частям. Слагаемое $u_x v$, возникающее при интегрировании, естественным образом обращается в нуль при $x = \pi$, и мы снова приходим к главному условию $v(0) = 0$, т. е. к пространству $\mathscr{H}_E^1$:

$$\int_0^\pi (u_t v + u_x v_x - fv)\, dx = 0 \quad \text{для всех} \quad v \in \mathscr{H}_E^1. \tag{5}$$

Это равенство служит отправным пунктом для получения аппроксимаций по методу конечных элементов. Если задано N-мерное подпространство S^h в $\mathscr{H}_E^1$, то *принцип Галёркина состоит в следующем: найти функцию* $u^h(x, t)$, *принадлежащую*

при каждом $t > 0$ *подпространству* S^h *и удовлетворяющую при всех* $v^h \in S^h$ *уравнению*

$$\int_0^\pi (u_t^h v^h + u_x^h v_x^h - f v^h)\, dx = 0. \tag{6}$$

Отметим, что временна́я переменная непрерывна: формулировка Галёркина (или, точнее, Фаэдо — Галёркина) подразумевает дискретизацию по пространственным переменным и приводит к системе обыкновенных дифференциальных уравнений от временно́й переменной. Эти уравнения и подлежат численному решению. Чтобы записать полученную задачу в операторной форме, выберем в пространстве пробных функций S^h базис $\varphi_1, \ldots, \varphi_N$ и разложим неизвестное решение по базисным функциям:

$$u^h(x, t) = \sum_1^N Q_j(t)\, \varphi_j(x).$$

Оптимальные весовые коэффициенты Q_j определяются согласно принципу Галёркина (6): для $k = 1, \ldots, N$

$$\int_0^\pi \left(\sum_j \left(\frac{\partial Q_j}{\partial t} \varphi_j \varphi_k + Q_j \frac{\partial \varphi_j}{\partial x} \frac{\partial \varphi_k}{\partial x} \right) - f\varphi_k \right) dx = 0. \tag{7}$$

Так как каждую функцию v^h можно разложить по базисным функциям φ_k, то принцип Галёркина достаточно применить только к базисным функциям. В результате получается система N обыкновенных дифференциальных уравнений с N неизвестными $Q_1(t), \ldots, Q_N(t)$; краевые условия уже включены в эти уравнения. Начальное условие $u = u_0$ все же надо учесть, и для этого есть несколько возможностей. С математической точки зрения естественным выбором аппроксимации начального условия u_0^h будет наилучшее приближение к u_0 по методу наименьших квадратов: u_0^h принадлежит S^h и удовлетворяет уравнению $(u_0^h, v^h) = (u_0, v^h)$ для всех v^h, т. е.

$$\sum Q_{j0} \int \varphi_j \varphi_k\, dx = \int u_0 \varphi_k\, dx, \quad k = 1, \ldots, N.$$

На практике это означает, что надо вычислять интегралы в правой части и обращать матрицу массы в левой части. Поэтому часто будет эффективнее взять в качестве начального условия интерполянт функции u_0, т. е. положить $u_0^h = (u_0)_I$. При этом порядок точности не изменится.

Запишем теперь уравнения Галёркина (7) в векторном обозначении при $Q = (Q_1, \ldots, Q_N)$. Замечательно то, что *эти урав-*

нения содержат в точности те же матрицы и вектор нагрузки, что и в стационарной задаче; коэффициент при Q' — это матрица массы M, а коэффициент при Q — матрица жесткости K:

$$MQ' + KQ = F(t). \tag{8}$$

Компоненты правой части равны $F_k = \int f(x, t)\,\varphi_k(x)\,dx$. Если бы краевые условия были неоднородными (зависящими от времени или нет), их влияние также было бы заключено в F.

Естественно спросить: почему конечные элементы не используются также и по временно́й переменной? Конечно, можно было бы попытаться применить их, но это не даст особого успеха. С математической точки зрения вполне разумно изучить дискретизацию в два этапа: сначала исследовать ошибку метода конечных элементов $u(x, t) - u^h(x, t)$, а затем ошибку в u^h, возникающую при решении обыкновенных дифференциальных уравнений. По временно́й переменной геометрия области не вызывает трудностей, которые надо было бы преодолевать с помощью метода конечных элементов, и на самом деле непосредственное применение принципа Галёркина может связать все временные слои и уничтожить главное свойство *распространения вперед по времени*. Мы не видим причин отказываться от этой дополнительной гибкости конечных разностей.

Сначала стоит упомянуть *метод наложения колебаний* — конкурента общепринятого метода конечных разностей. Его идея проста и заключается в разложении начального значения u_0 и функции источника f по естественным гармоникам задачи — собственным функциям u_l^h из гл. 6. Вся вычислительная работа сводится к задаче на собственные значения. Вперед по времени передаются только более низкие собственные значения; Никкелл предполагает, что, если функция f не очень насыщена высокими гармониками, хороших результатов можно достичь с менее чем 30 гармониками из 1000.

Общепринятая разностная схема по времени, связывающая все колебания, должна бороться с чрезмерной *жесткостью* уравнения (8); число обусловленности матрицы $M^{-1}K$ легко может превысить 1000, так что колебания затухают с совершенно разными скоростями. Если Δt выбрать надлежащим образом, то схема «правила трапеций» (Кранк — Николсон, Наймарк β) будет автоматически отфильтровывать бесполезные высокие гармоники. Конечно, эти схемы *неявны*, но таково же и дифференциальное уравнение Галёркина: матрицу M в (8) нельзя обратить, не разрушив ее ленточной структуры. (Приближенный расчет матрицы M обсуждается в конце главы.) В одном отношении неявность уравнения не такое уж серьезное препятствие при рассмотрении параболических уравнений (например, уравнения

теплопроводности), поскольку требования устойчивости для явных разностных схем в любом случае суровы: временно́й шаг должен быть ограничен сверху $\Delta t \leqslant Ch^2$, иначе в разностном уравнении будут неустойчивости экспоненциального характера. Напротив, неявная схема может быть абсолютно устойчивой; величина Δt ограничена лишь требованиями к точности, а не устойчивостью. Это различие между неявными и явными схемами естественно для уравнения теплопроводности, где *скорость распространения бесконечна*: как бы ни был мал временной шаг, температура Q^{n+1} в любой точке x_0 зависит от температуры Q^n во всех точках среды. Эта зависимость отражается в том, что матрица M^{-1} в уравнении Галёркина не разрежена, и в явном разностном уравнении для устойчивости требуется, чтобы $\Delta t \sim h^2$.

Исследуем *схему Кранка — Николсона*, записанную симметричным образом относительно $(n + 1/2)\Delta t$ и потому имеющую по времени второй порядок точности:

$$M\frac{Q^{n+1}-Q^n}{\Delta t}+K\frac{Q^{n+1}+Q^n}{2}=\frac{f^{n+1}+f^n}{2}. \tag{9}$$

Приближение Q^{n+1} определяется из соотношения (9), если переписать его в виде

$$\left(M+\frac{K\,\Delta t}{2}\right)Q^{n+1}=\left(M-\frac{K\,\Delta t}{2}\right)Q^n+\frac{(f^{n+1}+f^n)\,\Delta t}{2}.$$

При практическом вычислении можно матрицу в левой части представить, применив исключение Гаусса, в виде произведения LL^T, где L — нижняя треугольная матрица Холесского, а затем вычислять Q^{n+1} на каждом шаге с помощью двух подстановок:

$$LQ^{n+1/2}=\left(M-\frac{K\,\Delta t}{2}\right)Q^n+\frac{(f^{n+1}+f^n)\,\Delta t}{2},\qquad L^TQ^{n+1}=Q^{n+1/2}.$$

Если коэффициенты в задаче зависят от времени (или нелинейные), то в строгой теории Галёркина матрицы M и K должны пересчитываться на каждом шаге. Весьма вероятно, что для получения матрицы жесткости, приближенно правильной, без пересчитывания каждого интеграла, обязательно найдется возмущенный вариационный принцип, приводящий к некоторому гибридному методу конечных элементов и конечных разностей. В больших задачах точный процесс отыскания Q^{n+1} может оказаться слишком дорогим; итерационный подход к построению приближения для Q^{n+1} (возможно, исходящий из Q^n как из начального приближения) может быть более эффективным. Дуглас и Дюпон [Д8, Д11] предложили для нелинейных задач несколько итерационных способов, позволяющих решать на каждом временно́м шаге большую нелинейную систему. Их анализ

полностью оправдывает эти модификации чистого метода Галёркина.

В двух других разделах данной главы проверяются устойчивости и ожидаемые скорости сходимости для параболических и гиперболических уравнений. В последнем разделе при рассмотрении простого уравнения $u_t = u_x$ появляются, наконец, неожиданности.

7.2. УСТОЙЧИВОСТЬ И СХОДИМОСТЬ ДЛЯ ПАРАБОЛИЧЕСКИХ ЗАДАЧ

Здесь возможны два подхода к исследованию. Один — более явный и понятный; другой — более общий. В первом подходе каждую собственную функцию исследуют во времени во всех трех уравнениях — в дифференциальном уравнении в частных производных для u, обыкновенном дифференциальном для u^h и конечно-разностном для Q^h. Если коэффициенты и краевые условия в уравнениях не зависят от времени (стационарный случай), то этот простой подход крайне успешен, а с учетом точных границ ошибок в собственных функциях, выведенных в предыдущей главе, рассуждения становятся совершенно элементарными[1]). В нестационарном случае исследование технически сложнее, но параболические уравнения так сильно диссипативны, что можно полностью объяснить эффекты временной зависимости (и даже нелинейности). Во втором подходе, основанном на энергетических неравенствах для каждого момента времени, это объяснение становится сравнительно простым.

Рассмотрим параболическое уравнение $u_t + Lu = f$, где L — некоторый эллиптический оператор, изученный в предыдущих главах. Он имеет порядок $2m$ (наиболее распространен случай $m = 1$), его коэффициенты могут зависеть от x. Предположим сначала, что $f \equiv 0$ и начальная функция u_0 разлагается в ряд по ортонормальным собственным функциям:

$$u_0 = \sum_1^\infty c_j u_j(x), \qquad c_j = \int_\Omega u_0(x)\, u_j(x)\, dx.$$

Каждая собственная функция затухает во времени с присущей ей скоростью, и решение меняется по закону

$$u(t, x) = \sum c_j e^{-\lambda_j t} u_j(x). \tag{10}$$

Для $t > 0$ это решение принадлежит допустимому пространству $\mathscr{H}_E^1$. Даже если начальная функция u_0 разрывна, легко заметить, что с течением времени функция u становится все более глад-

[1]) Этот подход применим также к анализу наложения колебаний.

кой; производные в любой положительный момент времени удовлетворяют соотношению

$$\left\|\left(\frac{\partial}{\partial t}\right)^k u\right\|_0^2 = \left\|\sum c_j(-\lambda_j)^k e^{-\lambda_j t} u_j\right\|_0^2 = \sum c_j^2 \lambda_j^{2k} e^{-2\lambda_j t}.$$

Экспоненциальные множители делают эту сумму конечной (то же справедливо и для пространственных производных). Эти нормы с ростом t монотонно уменьшаются, в частности

$$\sum c_j^2 e^{-2\lambda_j t} \leqslant e^{-2\lambda_1 t} \sum c_j^2, \quad \text{или} \quad \|u(t)\|_0 \leqslant e^{-\lambda_1 t} \|u_0\|_0. \tag{11}$$

Основная частота $\lambda_1 > 0$ дает скорость убывания решения.

Уравнение Галёркина $MQ' + KQ = 0$ на самом деле несколько *более* устойчиво, чем уравнение, которое оно приближает. Предположим, что начальный вектор Q_0 разлагается по дискретным собственным функциям Q_j матрицы $M^{-1}K$, или, что эквивалентно, начальная функция u_0^h — по приближенным собственным функциям u_j^h:

$$Q_0 = \sum_1^N d_j Q_j \quad \text{или} \quad u_0^h = \sum_1^N d_j u_j^h.$$

Тогда решением в более поздний момент t будет

$$Q(t) = \sum d_j Q_j \exp\left(-\lambda_j^h t\right) \quad \text{или} \quad u^h(x, t) = \sum d_j u_j^h(x) \exp\left(-\lambda_j^h t\right).$$

Поэтому скорость убывания равна λ_1^h, что несколько больше, чем λ_1:

$$\|u^h(x, t)\|_0 \leqslant \|u_0^h\|_0 \exp\left(-\lambda_1^h t\right) \leqslant \|u_0^h\|_0 \exp\left(-\lambda_1 t\right).$$

Наконец, схема Кранка — Николсона также устойчива — без ограничений на величину Δt. Разностный оператор на каждом шаге имеет те же собственные векторы, что и матрица $M^{-1}K$, так как он задается равенством

$$\frac{M + K\,\Delta t}{2} Q^{n+1} = \frac{M - K\,\Delta t}{2} Q^n,$$

или

$$Q^{n+1} = \left(I + \frac{M^{-1}K\,\Delta t}{2}\right)^{-1} \left(I - \frac{M^{-1}K\,\Delta t}{2}\right) Q^n,$$

Решением после n шагов будет $Q^n = \sum d_j \left(\mu_j^h\right)^n Q_j$, где *коэффициент усиления* μ_j^h равен

$$\mu_j^h = \frac{1 - \lambda_j^h \Delta t/2}{1 + \lambda_j^h \Delta t/2}. \tag{12}$$

Так как каждое число λ_j^h неотрицательно, то $|\mu_j^h| < 1$, и потому схема Кранка — Николсона автоматически устойчива. Затухание основной собственной функции u_1^h регулируется ее коэффициентом усиления μ_1^h, сравнимым с истинным коэффициентом усиления $\exp(-\lambda_1^h \Delta t)$ уравнения Галёркина на каждом временно́м шаге:

$$\mu_1^h \cong 1 - \lambda_1^h \Delta t + \frac{1}{2}(\lambda_1^h \Delta t)^2 - \frac{1}{4}(\lambda_1^h \Delta t)^3 \ldots,$$

$$e^{-\lambda_1^h \Delta t} \cong 1 - \lambda_1^h \Delta t + \frac{1}{2}(\lambda_1^h \Delta t)^2 - \frac{1}{6}(\lambda_1^h \Delta t)^3 \ldots.$$

Так как μ_1^h меньше, то эта компонента решения убывает в конечно-разностном уравнении несколько быстрее. Различие между μ_1^h и $\exp(-\lambda_1^h \Delta t)$ имеет порядок Δt^3, отражающий точность второго порядка схемы Кранка — Николсона.

В конечно-разностном случае *высокочастотные компоненты не затухают со все более высокими скоростями.* При $\lambda_j^h \to \infty$ коэффициент усиления μ_j^h сходится к -1, и веса при высоких частотах изменяют знак на каждом временно́м шаге. Этого нет в уравнении Галёркина или в полностью неявной разностной схеме. В последней $\mu_j^h = (1 + \lambda_j^h \Delta t)^{-1}$ и очевидно, что $\mu_j^h \to 0$ при $\lambda_j^h \to \infty$. Для схемы Кранка — Николсона, однако, коэффициент μ_j^h станет отрицательным и начнет расти по абсолютной величине при $\lambda_j^h \geqslant 2/\Delta t$. Наивысшая частота, которую сетка может «удержать», равна $\lambda_N^h \sim ch^{-2m}$; она обычно превышает $2/\Delta t$. Поэтому очень высокие частоты (возможно, присутствующие лишь в малом количестве) действительно ослабляются менее сильно, чем умеренные. Если бы это представило какую-нибудь трудность, то, как и в гиперболических задачах, можно было бы добавить простой диссипативный член.

Скорость сходимости легко определить из разложения по собственным функциям:

$$u = \sum_1^\infty c_j e^{-\lambda_j t} u_j(x), \qquad u^h = \sum_1^N d_j e^{-\lambda_j^h t} u_j^h(x). \tag{13}$$

Здесь два источника ошибки: *начальная ошибка и развивающаяся ошибка.* Независимо от того, выбирается ли u_0^h как наилучшее приближение к истинной функции u_0 или просто как ее интерполянт, начальная ошибка $u_0 - u_0^h$ будет порядка h^k и, подобно любому другому начальному условию, будет уменьшаться со временем, как $e^{-\lambda_1 t}$. Кроме того, появляется ошибка, когда u_0^h берется в качестве начального условия в обоих уравне-

ниях. В точном уравнении она разлагается по собственным функциям u_j, а в уравнении Галёркина — по u_j^h и по-разному развивается во времени. Из предыдущей главы мы знаем, что

$$\lambda_j^h - \lambda_j \sim h^{2(k-m)}\lambda_j^{k/m}, \quad \|u_j^h - u_j\|_0 \sim h^k \lambda_j^{k/2m}.$$

(В случае $k < 2m$, не обычном в практике, h^k следует заменить на $h^{2(k-m)}$.) Эти оценки показывают, что разность в весах равна лишь $c_j - d_j = \int u_0^h (u_j - u_j^h)\, dx \sim h^k$. Поэтому, сравнивая функции u и u^h, представленные в виде (13), видим, что развивающаяся ошибка имеет порядок h^k. Ошибка в производных, затухающая опять со скоростью $e^{-\lambda_1 t}$, будет обычного порядка h^{k-s}.

Подчеркнем, что этот способ отыскания границ ошибок очень прост. С такой же легкостью он приводит к ошибкам порядка $O(\Delta t^2)$ в схеме Кранка — Николсона. Может показаться, что его простота требует, чтобы для возможности применения оценок ошибок в собственных значениях и собственных функциях из предыдущей главы пространственная часть задачи была самосопряженной, но на самом деле это предположение несущественно. Действительно, существует простая формула, полностью обходящая теорию собственных значений, — она относит развивающуюся ошибку к основным оценкам стационарных задач. При одной и той же начальной функции u_0^h в обоих уравнениях различие в их решениях в момент t описывается формулой

$$u - u^h = \frac{1}{2\pi i} \int_C e^{zt} (u_z - u_z^h)\, dz. \tag{14}$$

Здесь z — комплексное число, u_z — решение стационарной (несамосопряженной) задачи $(L + z)u_z = u_0^h$ и u_z^h — его аппроксимация по Галёркину. (Действительно, u_z и u_z^h — преобразования Лапласа функций $u(t)$ и $u^h(t)$ соответственно, а интеграл (14) обращает преобразование Лапласа; контур интегрирования C проходит вдоль двух лучей $z = \pm(\pi/2 + \varepsilon)$ в левой полуплоскости, так что экспонента e^{zt} дает сходящийся интеграл.) Из этой формулы, приводящей к разложениям по собственным функциям в самосопряженном случае с дискретным спектром, и из стационарных ошибок, установленных в теореме 2.1, непосредственно получаем выражение для развивающейся ошибки в момент времени t, она имеет ожидаемый порядок h^k даже для несамосопряженных уравнений.

Чтобы закончить обсуждение этого первого подхода, заметим, что с помощью принципа Дюамеля можно исследовать и неоднородный случай $f \neq 0$. Согласно этому принципу, функция источника f в момент τ действует подобно начальному условию в момент $t - \tau$. Если сначала решение u связано с u_0 по неко-

торому закону $u(t) = E(t)u_0$, то в неоднородном случае эта зависимость от исходных данных определяется формулой

$$u(t) = E(t)u_0 + \int_0^t E(t-\tau) f(\tau)\, d\tau.$$

Если воспользоваться разложениями по собственным функциям, то эта формула принимает вид

$$u(t) = \sum u_j(x) \left[c_j e^{-\lambda_j t} + \int_0^t f_j(\tau) e^{-\lambda_j (t-\tau)}\, d\tau \right],$$

$$f_j(\tau) = \int_\Omega f(x, \tau) u_j\, dx.$$

Ошибка $u - u^h$ будет все еще порядка h^k, однако если источник f действует в течение всего времени, то коэффициент затухания $e^{-\lambda t}$ будет сказываться уже не очень сильно; будут еще ошибки, совершенные в моменты τ, настолько близкие к t, что их затухание и не начиналось.

Испробуем теперь другую идею — второй подход к параболическим задачам, упомянутый в начале раздела: оценим в каждый момент времени t *скорость изменения* ошибки $u - u^h$. Вместо того чтобы исследовать разложения u и u^h по собственным функциям при $t \geqslant 0$, ошибку в момент $t + dt$ (или, для случая разностного уравнения, в момент $t + \Delta t$) определяем из ошибки в момент t. Этот метод впервые предложили Дуглас и Дюпон [Д8, Д11]. Они опирались на ранние работы Шварца и Вендроффа [Ш1] и Прайса и Варги [П10]; Вилер, Денди и др. совсем недавно провели дополнительные исследования.

Предположим, что в каждый момент времени t дифференциальное уравнение и его аппроксимация по Галёркину записаны в вариационной форме:

$$(u_t, v) + a(u, v) = (f, v) \quad \text{для всех} \quad v \in \mathscr{H}_E^1, \tag{15}$$

$$(u_t^h, v^h) + a(u^h, v^h) = (f, v^h) \quad \text{для всех} \quad v^h \in S^h. \tag{16}$$

В стационарном случае энергетическое скалярное произведение $a(v, w)$ не зависит от времени; оно возникает из (Lv, w). Будем по-прежнему обозначать через P проектор Ритца, отображающий пространство допустимых функций $\mathscr{H}_E^1$ на его подпространство S^h и определяемый, как и в (6.38), соотношением $a(u - Pu, v^h) = 0$ для всех v^h.

Представим ошибку по методу Галёркина в виде $u - u^h = (u - Pu) + (Pu - u^h)$. Величина $u - Pu$ известна из теории

аппроксимации (6.39), а энергетические неравенства, необходимые для оценки $Pu - u^h$, вытекают из следующего тождества.

Лемма 7.1. *Если* $e(t) = Pu(t) - u^h(t)$, *то*

$$(e_t, e) + a(e, e) = (Pu_t - u_t, e). \tag{17}$$

Доказательство. Так как e принадлежит S^h, можно в (15) положить $v = e$, в (16) $v^h = e$ и затем вычесть из первого уравнения второе:

$$(u_t - u_t^h, e) + a(u - u^h, e) = 0.$$

Применив тождество $a(u - Pu, e) = 0$, видим, что второе слагаемое равно $a(e, e)$; после перегруппировки членов получаем

$$(Pu_t - u_t^h, e) + a(e, e) = (Pu_t - u_t, e).$$

Осталось отождествить первое слагаемое со скалярным произведением (e_t, e), т. е. убедиться, что $Pu_t = (Pu)_t$ (проектор Ритца P коммутирует с оператором дифференцирования $\partial/\partial t$). Это справедливо только в стационарном случае. Если бы энергетическое скалярное произведение зависело от t, то член $(Pu_t - (Pu)_t, e)$ также появился бы в тождестве и его надо было бы оценить. Если оператор L достаточно гладкий по временной переменной, то эта трудность чисто техническая (подробности мы опускаем). Очевидно, что в стационарном случае P не зависит от времени: дифференцируя тождество $(u - Pu, v^h) = 0$, имеем $(u_t - (Pu)_t, v^h) = 0$ для всех v^h, так что $(Pu)_t$ совпадает с Pu_t. Лемма доказана.

Из этого тождества легко найти скорость изменения ошибки. Первое слагаемое в (17) можно переписать в виде

$$(e_t, e) = \frac{\partial}{\partial t}\frac{(e, e)}{2} = \|e\|_0 \frac{\partial}{\partial t}\|e\|_0.$$

Так как λ_1 — минимум отношения Рэлея, то слагаемое $a(e, e)$ не меньше, чем $\lambda_1 \|e\|_0^2$. Наконец, правая часть тождества ограничена величиной $\|u_t - Pu_t\|_0 \cdot \|e\|_0$. Сокращая на общий множитель $\|e\|_0$, получаем неравенство

$$\frac{\partial}{\partial t}\|e\|_0 + \lambda_1 \|e\|_0 \leqslant \|u_t - Pu_t\|_0. \tag{18}$$

Умножим (18) на $e^{\lambda_1 \tau}$ и проинтегрируем по τ от 0 до t:

$$e^{\lambda_1 t}\|e(t)\|_0 - \|e(0)\|_0 \leqslant \int_0^t e^{\lambda_1 \tau}\|u_t(\tau) - Pu_t(\tau)\|_0 \, d\tau. \tag{19}$$

Независимо от того, выбирается u_0^h как интерполянт функции u_0 или как ее аппроксимация по методу наименьших квадратов, начальная ошибка $e(0) = Pu_0 - u_0^h$ имеет порядок $Ch^k \| u_0 \|_k$. Из неравенства (19) немедленно следует основная теорема, дающая правильный порядок сходимости, хотя формулируемые оценки не обязательно будут самыми точными.

Теорема 7.1. *Пусть S^h — пространство метода конечных элементов степени $k-1$. Тогда ошибка аппроксимации по методу Галёркина удовлетворяет неравенствам*

$$\| u(t) - u^h(t) \|_0 \leqslant \| u(t) - Pu(t) \|_0 + \| e(t) \|_0 \leqslant$$

$$\leqslant Ch^k \left[\| u(t) \|_k + e^{-\lambda_1 t} \| u_0 \|_k + \int_0^t e^{\lambda_1 (\tau - t)} \| u_t(\tau) \|_k \, d\tau \right]. \quad (20)$$

Таким образом, ошибка имеет тот же порядок h^k, что и в стационарных задачах, и при отсутствии функции источников она убывает так же быстро, как и основное колебание.

Оценка, которую дает эта теорема, совпадает с оценкой h^k, установленной ранее с помощью разложений по собственным функциям. Усредненную оценку ошибки по энергии получаем, интегрируя непосредственно тождество (17):

$$2 \int_0^t a(e(\tau), e(\tau)) \, d\tau \leqslant \| e(0) \|_0^2 - \| e(t) \|_0^2 + 2 \int_0^t | (Pu_t - u_t, e) |^2 \, d\tau.$$

Согласно только что доказанной теореме, правая часть последнего неравенства имеет порядок h^{2k}. Это говорит о том, что e в энергетической норме пренебрежимо мала по сравнению с $u - Pu$ и ошибка приближения по Галёркину $u - u^h = u - Pu + e$ удовлетворяет соотношению

$$a(u - u^h, u - u^h) \sim a(u - Pu, u - Pu) \leqslant C^2 h^{2(k-m)} \| u \|_k^2. \quad (21)$$

На этом мы заканчиваем исследование ошибок для параболических уравнений; ничего неожиданного в результатах нет. У нас создалось впечатление, что, как и в стационарных задачах, метод конечных элементов особенно эффективен при вычислениях на *грубых сетках* с большим шагом h. В этой ситуации физику процесса часто точнее отражает принцип Галёркина, на котором основан метод конечных элементов, чем предположения о близости разностных отношений к производным. Однако так как в методе конечных элементов приходится вычислять интегралы,

то за это надо платить затратой времени. Возможно, в конце концов появится удовлетворительная комбинация конечных элементов и конечных разностей.

7.3. ГИПЕРБОЛИЧЕСКИЕ УРАВНЕНИЯ

Естественно попробовать применить метод конечных элементов также и к гиперболическим задачам. Действительно, формулировку принципа Галёркина можно обобщить так (уравнение $M(u) = 0$ заменить из $(M(u^h), v^h) = 0$ для всех v^h), чтобы метод конечных элементов можно было проверить на большем количестве примеров. Некоторые предварительные испытания уже ведутся, но выводов еще нет (и, может быть, никогда не будет, поскольку характер большинства численных экспериментов недостаточно ясен). По-видимому, можно ожидать лишь договоренности по некоторым общим направлениям.

Математически можно гарантировать[1]) одно свойство, заключающееся в том, что если энергия в точной задаче сохраняется, то она сохраняется и в методе Галёркина и, если в точной задаче она со временем уменьшается, она уменьшается и в аппроксимации Галёркина. Это легко заметить для системы первого порядка $u_t + Lu = 0$. Скорость изменения энергии $(u, u) = \int u^2\, dx$ можно вычислить, умножая уравнение на u и интегрируя по пространственной переменной:

$$(u_t, u) + (Lu, u) = \frac{\partial}{\partial t}\frac{(u, u)}{2} + (Lu, u) = 0.$$

Если $(Lu, u) \equiv 0$, то рассматриваемое уравнение называется *консервативным* ((u, u) не изменяется со временем); если же $(Lu, u) \geqslant 0$ для всех возможных состояний u, то оно называется *диссипативным* (энергия (u, u) уменьшается). Параболические уравнения строго диссипативны, так как в этом случае (Lu, u) — положительно определенная форма от m-х производных от u. Гиперболические уравнения либо консервативны, либо в лучшем случае слегка диссипативны; энергия может «вытекать» наружу (но только очень медленно) на границах области. Поэтому исследование этих уравнений гораздо тоньше. Пусть в методе Галёркина Q означает проектор из $\mathscr{H}^0$ на подпространство S^h, т. е. Qu — наилучшее приближение к u в S^h по методу наименьших квадратов, в то время как ранее рассматриваемая проекция Pu была наилучшим приближением по норме энергии деформации $a(v, v)$. Тогда приближение по Галёркину u^h опреде-

[1]) При условии, что тестовое пространство V^h совпадает с пробным пространством S^h.

ляется проекцией дифференциального уравнения на подпространство S^h:

$$Q(u_t^h + Lu^h) = 0.$$

Так как требуется, чтобы функция u^h принадлежала S^h, то автоматически $u^h = Qu^h$, и уравнение Галёркина можно записать в более симметричном виде

$$u_t^h + QLQu^h = 0, \qquad u_0^h \in S^h.$$

Другими словами, точный порождающий оператор L заменяется на QLQ. Но тогда консервативное уравнение остается консервативным, $(QLQu, u) = (LQu, Qu) = 0$, а диссипативное — диссипативным: $(QLQu, u) = (LQu, Qu) \geqslant 0$. Соответствующие нелинейные операторы называются *монотонными* (разд. 2.4), и справедлив тот же результат.

Интересно, что свойство консервативности не всегда желательно, в частности в *нелинейных* гиперболических уравнениях. Простейший пример — закон сохранения $u_t = (u^2)_x$. В решениях этих задач могут быть самопроизвольные разрывы (скачки) и сохранение энергии теряется, даже хотя некоторые другие законы сохранения массы и момента выполняются. В уравнении Галёркина этих скачков, по-видимому, нет совсем и приближенное уравнение остается консервативным — отсюда следует, что сходимость к истинному решению невозможна. В методе конечных разностей обычный прием состоит в том, чтобы рассеять энергию с помощью искусственной вязкости; по-видимому, это будет необходимо и для конечных элементов.

Гиперболические уравнения могут появляться в двух формах — в виде системы первого порядка по времени, скажем $w_t + Lw = f$ с вектором неизвестных w, либо как уравнение второго порядка $u_{tt} + Lu = f$. Начнем со второго случая, в котором L — эллиптический оператор; типичным примером служит волновое уравнение $u_{tt} - c^2 u_{xx} = 0$. Слабая форма этого уравнения такова:

$$(u_{tt}, v) + a(u, v) = (f, v) \quad \text{для} \quad v \in \mathcal{H}_E^1, \qquad t > 0. \tag{22}$$

В аппроксимации Галёркина u и v заменяются на u^h и v^h; это означает, что $u^h = \sum Q_j(t)\,\varphi_j(x)$ определяется из уравнения

$$\left(\sum Q_j''\varphi_j, \varphi_k\right) + \left(\sum Q_j\varphi_j, \varphi_k\right) = (f, \varphi_k) \quad \text{для} \quad k = 1, \ldots, N, \; t > 0. \tag{23}$$

Это снова обыкновенное дифференциальное уравнение от временно́й переменной. Здесь участвуют те же матрицы массы и жесткости, но уравнение уже второго порядка:

$$MQ'' + KQ = F(t). \tag{24}$$

Исходными значениями будут приближения (из S^h) к точному перемещению $u_0(x)$ и точной начальной скорости $u_0'(x)$. Поведение решений в данном случае совершенно отличается от поведения в параболическом случае, где вместо Q'' стоит Q'. Там (при $F \equiv 0$) решение очень быстро затухает; каждый собственный вектор входит в решение вместе с экспонентой $e^{-\lambda_j t}$ и разрывы немедленно исчезают. В гиперболическом случае показатель степени меняется на $\pm i\lambda_j t$ и Q *скорее осциллирует, чем затухает.* Решение не более гладко, чем начальные значения, и разрывы распространяются сколь угодно долго по времени.

Для одномерного волнового уравнения и линейных элементов аппроксимация Галёркина имеет вид

$$\frac{Q''_{j+1} + 4Q''_j + Q''_{j-1}}{6} = c^2 \frac{Q_{j+1} - 2Q_j + Q_{j-1}}{h^2}.$$

Отметим снова связанную форму этих уравнений, автоматически приводящую к неявному разностному уравнению. Из экспериментов Клафа и др. и теоретических рассмотрений Фуджи, доложенных им на Втором японо-американском семинаре, можно сделать вывод, что если матрица M заменяется (в подходящем процессе ее приближенного расчета) диагональной, то *потери в точности не будет.* Но это верно не для всех степеней элементов: для членов, не продифференцированных по x, приближенный расчет неявно использует элементы низких степеней (большей частью кусочно постоянные), и это приводит к потере общей точности, *когда другие члены в уравнении обрабатываются с высокой точностью.*

Фуджи дал также полезный анализ устойчивости разностных аппроксимаций (по временно́й переменной) уравнения (24) в методе конечных элементов. Предположим, например, что члены Q'' заменяются центральными разностными отношениями второго порядка $(\Delta t)^{-2}(Q^{n+1} - 2Q^n + Q^{n-1})$. Из теории конечных разностей хорошо известно, что величина Δt должна быть ограничена, или же вычисляемые приближения Q^n будут экспоненциально расти вместе с n. Для одномерного волнового уравнения условия устойчивости процесса вычислений имеют вид $c\,\Delta t \leqslant h/\sqrt{3}$ для согласованной матрицы массы M и $c\Delta t \leqslant h$ — для диагональной матрицы, полученной при приближенном расчете матрицы M. (Тонг [T6] заметил в последнем случае дополнительную устойчивость.) Фуджи исследовал и другие конечно-разностные схемы, а также гиперболические уравнения более общего вида для краевых задач с начальными условиями, в том числе и уравнения упругости.

Аппроксимация по Галёркину обладает двумя важными свойствами: *сохранение энергии* (если $f = 0$) *и сходимость.* Чтобы

измерить энергию в гиперболической задаче второго порядка, сложим потенциальную и кинетическую энергии:

$$E(t) = \frac{1}{2}[(u_t, u_t) + a(u, u)].$$

Для волнового равнения эта энергия равна $^1/_2 \int (u_t^2 + c^2 u_x^2)\, dx$. Величина $E(t)$ не зависит от времени, так как при $v = u_t$ в (22)

$$\frac{dE}{dt} = (u_{tt}, u_t) + a(u, u_t) = 0. \tag{25}$$

Для волнового уравнения соотношение (25) принимает вид

$$\frac{dE}{dt} = \int (u_t u_{tt} + c^2 u_x u_{xt})\, dx = \int u_t (u_{tt} - c^2 u_{xx})\, dx = 0.$$

Сохранение энергии в уравнении Галёркина можно проверить тем же способом:

$$E^h(t) = \frac{1}{2}[(u_t^h, u_t^h) + a(u^h, u^h)],$$

$$\frac{dE^h}{dt} = (u_{tt}^h, u_t^h) + a(u^h, u_t^h) = 0.$$

Таким образом, подобно точному уравнению, приближенное уравнение только нейтрально устойчиво.

Дадим набросок доказательства сходимости, вытекающего из тождества, аналогичного доказанному в лемме 7.1: при $e = Pu - u^h$

$$(e_{tt}, e_t) + a(e, e_t) = ([Pu - u]_{tt}, e_t). \tag{26}$$

Левая часть есть производная от энергии $E(t, e)$ в e. Энергия ошибки не обладает свойством сохранения, но правая часть в (26) меньше, чем

$$\| (Pu - u)_{tt} \|_0 \| e_t \|_0 \leqslant Ch^k \sqrt{E(t, e)}.$$

Итак, $E' \leqslant Ch^k \sqrt{E}$. Интегрируя от 0 до t, получаем

$$E^{1/2} \leqslant E_0^{1/2} + \frac{Ch^k t}{2}.$$

Начальная ошибка E_0 будет порядка $h^{2(k-1)}$, и такой же будет энергия в $u - Pu$. Поэтому *энергия ошибки Галёркина* $u - u^h = u - Pu + e$ *имеет оптимальный порядок* $h^{2(k-1)}$. Это справедливо даже при больших значениях $t \sim 1/h$, если только начальные данные гладкие.

Рассмотрим теперь тривиальный, но интересный пример $u_t = u_x$. Само уравнение не слишком увлекательно; оно описывает распространение волны влево с единичной скоростью

$u(x, t) = u_0(x + t)$. Искажения волны нет и очевидно, что энергия системы первого порядка $\int_{-\infty}^{\infty} u^2\,dx$ сохраняется. Приближением по Галёркину в каждый момент времени будет $(u_t^h, v^h) = (u_x^h, v^h)$, оно также сохраняет энергию. При линейных элементах $u^h(t, x) = \sum u_j(t)\,\varphi_j(x)$, где φ_j — функция-крышка в узле jh, это приближение принимает вид

$$\frac{u'_{j+1} + 4u'_j + u'_{j-1}}{6} = \frac{u_{j+1} - u_{j-1}}{2h}. \tag{27}$$

Очевидно, что уравнение (27) снова неявное — это серьезный недостаток для гиперболических задач. Ошибка аппроксимации равна $O(h^2)$, т. е. равна обычной скорости сходимости для линейных элементов.

Дюпон [Д12] вычислил соответствующую скорость сходимости для кубических пробных функций, и оказалось, *что ожидаемая степень h^4 просто не появляется.* Ошибка $u - u^h$ имеет порядок $O(h^3)$, т. е. на порядок больше наилучшего приближения к u с помощью кубических элементов. Расчеты Дюпона, встреченные с удивлением и, возможно, даже с некоторым недоверием, проводились для эрмитовых кубических элементов со значениями u и u_x в каждом узле в качестве неизвестных. Для кубических сплайнов его вычисления показали ошибку $O(h^4)$. Поэтому скорость сходимости зависит не только от степени конечных элементов. Действительно, более широкое пространство эрмитовых кубических элементов дало худшую аппроксимацию, чем его подпространство кубических сплайнов. Частично это объясняется так. Вычислим ошибку аппроксимации в общем случае, подставляя истинное решение уравнения $u_t + Lu = 0$ в уравнение Галёркина $u_t^h + QLQu^h = 0$. В нашем случае $L = -\partial/\partial x$ и Q — проектор на подпространство S^h. Ошибка аппроксимации равна

$$Lu - QLQu = (I - Q)\,Lu + QL\,(I - Q)\,u.$$

Первое слагаемое в правой части представляет собой ошибку приближения функции $Lu = -u_x$ по методу наименьших квадратов. Если S^h имеет степень $k - 1$ и функция u гладкая, то эта ошибка обычного порядка h^k. Но решающую роль играет второе слагаемое $QL(I - Q)u$. Так как $L(I - Q)u$ есть производная от ошибки приближения по методу наименьших квадратов, то это ошибка в норме пространства $\mathcal{H}^1$, и она не может быть лучше чем h^{k-1}. Вопрос заключается в следующем: аннулируется ли $L(I - Q)u$ при действии последнего проектора Q? Мы полагаем, что это происходит в случае линейных элементов на равномер-

ной сетке, но не обычных кубических элементов. Так как такое уничтожение надо считать исключительным явлением, *скорость сходимости в общем случае для гиперболических систем первого порядка будет* h^{k-1}, *а не* h^k. Эту скорость для широкого класса гиперболических систем установил Лесен.

Чтобы понять, как происходит это уничтожение, применим $QL(I-Q)$ к полиномам низших степеней, не входящих тождественным образом в рассматриваемое подпространство (x^2 — для линейных элементов и x^4 — для кубических). В линейном случае $(I-Q)x^2$ есть функция ошибки, изображенная на рис. 3.3 в разд. 3.2. Ее производной будет *пилообразная функция*; на каждом подынтервале $L(I-Q)x^2$ линейно изменяется от -1 до $+1$. Наилучшее кусочно линейное непрерывное приближение такой функции — тождественный нуль: $QL(I-Q)x^2 = 0$, и уничтожение произошло. В кубическом случае $L(I-Q)x^4$ оказывается эрмитовой кубической функцией, и последнее действие проектора Q ее не изменяет; уничтожения здесь нет, и ошибка приближения по Галёркину имеет порядок $\|u-u^h\|_0 \sim h^3$.

В некотором смысле показатель $k-1$ можно было бы и предвидеть. Если волновое уравнение $u_{tt} = c^2 u_{xx}$ сводится к системе первого порядка, то вектор неизвестных образуется из первых производных u_t и cu_x:

$$\begin{pmatrix} u_t \\ cu_x \end{pmatrix}_t = \begin{pmatrix} 0 & c \\ c & 0 \end{pmatrix} \begin{pmatrix} u_t \\ cu_x \end{pmatrix}_x.$$

Поэтому обычная энергия $\|u_t\|_0^2 + \|cu_x\|_0^2$ в векторе неизвестных вдвое больше энергии $E(t)$, равной сумме кинетической и потенциальной энергий. Так как ошибка этой энергии для одного уравнения была порядка $h^{2(k-1)}$, то показатель $k-1$ и будет как раз тем показателем, которого следовало бы ожидать для системы.

С точки зрения практики, оценки ошибок при $h \to 0$ подчиняются задаче достижения приемлемой точности при разумных затратах. Для гиперболических уравнений мы не уверены, что это достигается наиболее эффективно с помощью метода конечных элементов. Конечная скорость распространения в истинном решении означает, что возможны явные конечно-разностные уравнения с шагом Δt того же порядка, что и h, и известно, как для повышения устойчивости ввести искусственную вязкость. Для конечного элемента разностные уравнения будут неявными и почти полностью консервативными. (Они могут быть явными, только если мы *приближенно* рассчитываем матрицу массы или, как предложил Равьяр, выбираем узловые точки в качестве точек ξ_i в численных квадратурах.) Сохраняя массу при прибли-

женном расчете или, как в [Т9], используя полиноминальные пробные функции низших степеней для матрицы массы элемента, мы получаем согласованное разностное уравнение с типичным условием Куранта для численной устойчивости. (Устойчивость не будет абсолютной, как для неявного чистого процесса Галёркина, описанного ранее в этой главе.)

Одно из важных преимуществ метода конечных элементов в гиперболических задачах, которое в будущем надо сохранить в конечно-разностных схемах, заключается в систематическом достижении точности даже на криволинейных границах.

8 ОСОБЕННОСТИ

8.1. УГЛЫ И ПОВЕРХНОСТИ РАЗДЕЛА

Возможно, наиболее характерное свойство эллиптических краевых задач, подобных задаче

$$-\nabla\cdot(q\nabla u)+qu=f \text{ в } \Omega, \quad u=0 \text{ на } \Gamma, \tag{1}$$

состоит в гладкости решения u, если граница Γ и исходные данные p, q, f гладкие. Действительно, *лемма Вейля* утверждает, что функция u аналитична в любой подобласти Ω_1 из Ω, если только p, q, f аналитичны в Ω_1. Аналогичное утверждение справедливо «вплоть до границы», при условии что сама граница области Ω аналитична.

Поэтому особенности могут появляться, только когда граница или некоторые из исходных данных не будут гладкими. К сожалению, эти случаи встречаются часто, например в задачах механики разрушения, и при наличии особенностей продолжение исследований методом конечных элементов на равномерной сетке даст совершенно неудовлетворительные результаты. Как и в разностных аппроксимациях, эффективным приемом работы с особенностями оказалось локальное сгущение сетки (в том смысле, который обсуждался в предыдущих главах). Однако о природе особенностей, возникающих в эллиптических задачах, известно много и специальная форма вариационного метода стимулирует нас к использованию этой информации в приближении Ритца-Галёркина. Данная глава и посвящается этой задаче. Мы начнем с выявления аналитической формы особенностей, которые могут возникнуть.

Рассмотрим сначала случай негладких границ. Пусть уравнение Лапласа

$$-\Delta u=f \text{ в } \Omega, \quad u=0 \text{ на } \Gamma, \tag{2}$$

задано на области Ω с углом (рис. 8.1). Для формулировки основных идей предположим, что функция f аналитична в замкнутой области $\bar{\Omega}$, а граница Γ аналитична всюду, кроме точки P. Тогда по лемме Вейля функция u аналитична всюду в Ω, кроме точки P, и вблизи этой точки мы ищем u. В частности, мы будем

изучать поведение u в секторе

$$\Omega_0 = \{(r,\ \theta) \mid 0 < r < r_0, \quad 0 < \theta < \alpha\pi\} \subset \Omega, \tag{3}$$

где (r, θ) — полярные координаты с центром в точке P. Слабая форма уравнения (2) имеет вид

$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega fv \quad \text{для всех} \quad v \quad \text{в} \quad \mathscr{H}_E^1(\Omega). \tag{4}$$

Если функция v выбирается равной нулю в окрестности угловой точки и вне сектора Ω_0, то (4) сводится к

$$0 = \int_{\Omega_0} [\nabla u \cdot \nabla v - fv] = \int_0^{r_0} r\,dr \int_0^{\alpha\pi} \left[\frac{\partial u}{\partial r}\frac{\partial v}{\partial r} + r^{-2}\frac{\partial u}{\partial \theta}\frac{\partial v}{\partial \theta} - fv\right] d\theta. \tag{5}$$

Далее, так как u аналитична вне угла и $u(r, 0) = u(r,\ \alpha\pi) = 0$,

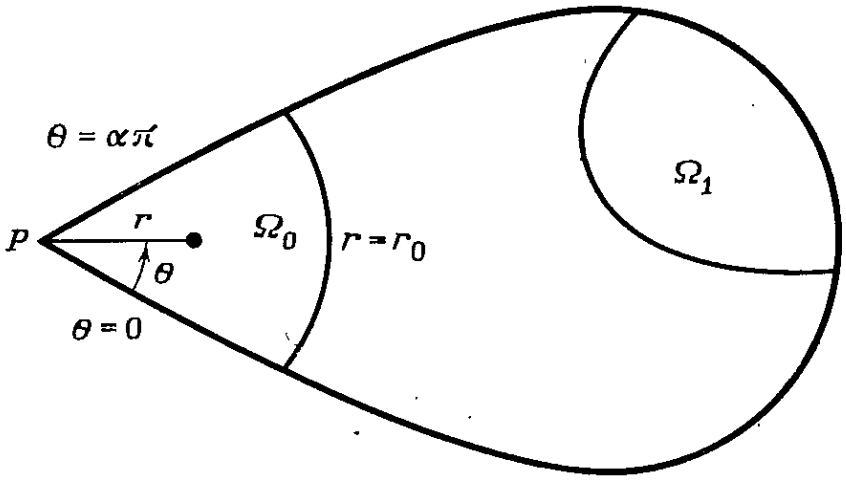


Рис. 8.1.
Область с углом $\alpha\pi$.

то u при каждом фиксированном $r > 0$ можно разложить в ряд

$$u(r,\ \theta) = \sum_{j=1}^{\infty} u_j(r)\,\varphi_j(\theta), \tag{6}$$

где

$$\varphi_j(\theta) = \sqrt{\frac{2}{\alpha\pi}}\sin \nu_j\theta, \qquad \nu_j = \frac{j}{\alpha}. \tag{7}$$

Коэффициенты Фурье $u_j(r)$ определяются из соотношения (5): пусть $v = \psi(r)\varphi_j(\theta)$; в силу свойства ортогональности

$$\int_0^{\alpha\pi} \frac{\partial \varphi_j}{\partial \Im}\frac{\partial \varphi_l}{\partial \theta}\,d\theta = \nu_l^2 \int_0^{\alpha\pi} \varphi_j(\theta)\,\varphi_l(\theta)\,d\theta = \nu_l^2\delta_{jl}.$$

Тогда после интегрирования по частям первого члена по r соотношение (5) принимает вид

$$\int_0^{r_0} dr \left[\frac{d}{dr}\left(r\frac{du_j}{dr}\right) - \nu_j^2 r^{-1} u_j - r f_j(r)\right] \psi(r) = 0, \tag{8}$$

где

$$f_j(r) = \int_0^{\alpha\pi} f \varphi_j(\theta)\, d\theta.$$

Так как (8) справедливо при любых ψ, то выражение в квадратных скобках должно равняться нулю. Это дает основное дифференциальное уравнение для $u_j(r)$. Разлагая $f_j(r)$ в ряд $\Sigma f_{jl} r^l$, получаем общее решение этого уравнения

$$u_j(r) = \alpha_j r^{\nu_j} + \beta_j r^{-\nu_j} + \sum_{l=0}^{\infty} f_{jl} \left[(l+2)^2 - \nu_j^2\right]^{-1} r^{l+2}, \tag{9}$$

где всякий раз, когда $\nu_j^2 = (l+2)^2$, мы условимся о замене

$$\left[(l+2)^2 - \nu_j^2\right]^{-1} r^{l+2} \quad \text{на} \quad \left[2(\nu_j + 1)\right]^{-1} r^{\nu_j} \ln r. \tag{10}$$

Кроме того, мы обрасываем член $r^{-\nu_j}$ и полагаем $\beta_j = 0$, поскольку в противном случае этот член будет обладать бесконечной энергией. Другая постоянная α_j выбирается так, чтобы при $r = r_0$ (9) было точным коэффициентом Фурье функции $u(r_0, \theta)$. Вообще, решение около угла имеет вид (Леман [Л2])

$$u(r, \theta) = \sum_{j=1}^{\infty} \alpha_j r^{\nu_j} \varphi_j(\theta) + \sum_{j=1}^{\infty} \sum_{l=0}^{\infty} f_{jl} \left[(l+2)^2 - \nu_j^2\right]^{-1} \varphi_j(\theta)\, r^{l+2}. \tag{11}$$

Предположим на минуту, что число $1/\alpha$ не целое. Тогда из (11) следует, что главный член в особенности функции u есть

$$r^{\nu_1} \sin \nu_1 \theta = r^{1/\alpha} \sin \frac{\theta}{\alpha}.$$

Заметим, что эта особенность становится более выраженной при увеличении угла $\alpha\pi$, и если угол в точке P больше развернутого, т. е. $\alpha > 1$, то неограничены даже первые производные от u. Наихудший случай связан с разрезом в области, направленным внутрь нее (это одна из задач, которую мы численно исследуем в последнем разделе). Полный внутренний угол вокруг точки P, находящейся в конце разреза (см. рис. 8.3), равен 2π, и решение ведет себя как $r^{1/2} \sin(\theta/2)$.

Легко определить степень гладкости u около любой такой особенности. Непосредственной проверкой можно убедиться, что $r^\nu \sin \nu\theta$ обладает «почти» $1 + \nu$ производными в среднем квадратичном смысле (и только ν производными в поточечном —

чтобы предсказать любую разумную сходимость, нам совершенно необходима теория приближений в среднем квадратичном). Таким образом, для любого числа $\beta < \nu_1 = 1/\alpha$ решение имеет $1+\beta$ дробных производных. При входящем угле (где область Ω не выпукла и $\alpha > 1$) u принадлежит $\mathscr{H}^1$, но не $\mathscr{H}^2$. В окрестности разреза u не принадлежит вовсе и $\mathscr{H}^{3/2}$.

Когда $1/\alpha$ — целое число, первая сумма в (11) аналитична. Однако (за исключением случая $\alpha = 1$, при котором Γ — прямая в окрестности точки P) нельзя сделать вывод о том, что функция u аналитична в Ω_0; *обычно во второй сумме появляются логарифмические особенности.* Например, если $\alpha = 1/2$, так что Γ образует прямой угол в P, то $\nu_j^2 = (l+2)^2 = 4$ при $j = 1$ и $l = 0$. Этот случай служит примером, когда требуется замена (10) и в решении u появляется член $(f_{10} r^2 \ln r \cdot \sin 2\theta)/6$. Заметим, что он принадлежит $\mathscr{H}^2$, но не $\mathscr{H}^l$ для любого $l > 2$.

С помощью аналогичных рассуждений можно установить характер поведения решения около углов и в других задачах. Более сложные вычисления показывают, что в задаче (1) с переменными коэффициентами особенность все еще имеет вид (11). Более общо, особенность в задаче $2m$-го порядка определяется главной частью оператора. Главный член в u равен $r^\lambda \varphi_\lambda(\theta)$, где φ_λ — гладкая функция от θ, а λ — собственное значение вспомогательной задачи, обе они зависят от краевых условий. Рекомендуем читателю ознакомиться с работами Келлога [К3], включающей также трехмерный случай, Кондратьева [К10] и с технической литературой [И1, П5, У2, Х4].

Рассмотрим теперь особенность второго вида, возникающую в случае, когда граница гладкая, а одна или несколько исходных данных не гладкие. Такая особенность обычно появляется в *задачах с поверхностью раздела*, простым примером которых служит задача

$$-\nabla \cdot (p \nabla u) = f \text{ в } \Omega, \quad u = 0 \text{ на } \Gamma, \tag{12}$$

где Ω — область, изображенная на рис. 8.2. Коэффициент p мы полагаем кусочно постоянным:

$$p = \begin{cases} p_1 & \text{в } \Omega_1, \\ p_2 & \text{в } \Omega_2. \end{cases} \tag{13}$$

Классическая формулировка задачи состоит в том, чтобы дифференциальное уравнение (12) удовлетворялось отдельно в Ω_1 и Ω_2, а функции u и $p\partial u/\partial \nu$ были непрерывными при переходе через поверхность раздела Γ (ν — нормаль к Γ). Таким образом (см. рис. 8.2),

$$p_1 \frac{\partial u}{\partial \nu}\Big|_{\Gamma_-} = p_2 \frac{\partial u}{\partial \nu}\Big|_{\Gamma_+}. \tag{14}$$

Как отмечалось в разд. 3.1, слабой формой этого уравнения будет

$$\int_\Omega p\nabla u\cdot\nabla v=\int_\Omega fv \quad \text{для всех} \quad v \quad \text{в} \quad \mathcal{H}_E^1,$$

и (14) — *естественное краевое условие.* Ему *не* обязаны удовлетворять пробные функции в методе Ритца. В самом деле, было бы очень трудно удовлетворять этому условию в угловой точке на Г.

За исключением случая, когда поверхность раздела прямая ($\alpha = 1$) или $p_1 = p_2$, решение u в точке P (рис. 8.2) имеет особенность, и функцию u в окрестности этой точки ищем в виде, аналогичном (11). Следуя Биркгофу [Б12] и Келлогу [К2], введем периодическую систему Штурма — Лиувилля

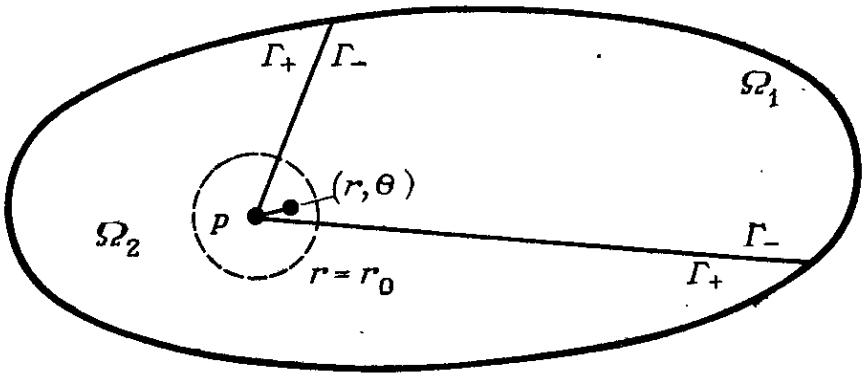


Рис. 8.2.
Угловая поверхность раздела.

$$-\frac{d}{d\theta}\left(p\frac{d\varphi}{d\theta}\right)=\lambda p\varphi, \quad p=p(\theta)=\begin{cases} p_1, & \text{если} \quad 0<\theta<\alpha\pi, \\ p_2, & \text{если} \quad \alpha\pi<\theta<2\pi. \end{cases} \tag{15}$$

Требуется, чтобы собственные функции $\varphi(\theta)$ были периодическими, $\varphi(\theta)=\varphi(\theta+2\pi)$, и на поверхности раздела удовлетворяли условиям

$$\lim_{\theta\downarrow 0}\left[p_1\frac{d\varphi}{d\theta}(\theta)\right]=\lim_{\theta\downarrow 0}\left[p_2\frac{d\varphi}{d\theta}(-\theta)\right], \tag{16}$$

$$\lim_{\theta\downarrow 0}\left[p_1\frac{d\varphi}{d\theta}(\alpha\pi-\theta)\right]=\lim_{\theta\downarrow 0}\left[p_2\frac{d\varphi}{d\theta}(\alpha\pi+\theta)\right]. \tag{17}$$

Существует бесконечная последовательность положительных собственных значений $\lambda_j=\nu_j^2$, и соответствующие им собственные функции $\varphi_j(\theta)$ ортогональны:

$$\int_0^{2\pi} p(\theta)\varphi_j'(\theta)\varphi_l'(\theta)\,d\theta=\nu_j^2\int_0^{2\pi} p(\theta)\varphi_j(\theta)\varphi_l(\theta)\,d\theta=\delta_{jl}\nu_j^2.$$

Для каждого фиксированного числа $r > 0$ решение $u(r, \theta)$ удовлетворяет условиям скачка (16)—(17) и потому

$$u(r, \theta) = \sum_{j=1}^{\infty} u_j(r)\,\varphi_j(\theta), \qquad u_j(r) = \int_0^{2\pi} p(\theta)\,u(r, \theta)\,\varphi_j(\theta)\,d\theta. \tag{18}$$

Теперь рассуждения проводим так же, как и в случае особенностей на границе. Подставляем (18) в дифференциальное уравнение для коэффициентов Фурье $u_j(r)$. Эти уравнения можно решить точно и найти для u выражение, аналогичное (11), но только в этом случае показатели степеней $\{\nu_j\}$ будут квадратными корнями из собственных значений системы (15), а $\{\varphi_j\}$ — соответствующими собственными функциями.

Для этой простой задачи можно вывести точные формулы для собственных функций и собственных значений; в более сложных задачах потребуются численные методы, и мы для иллюстрации рассмотрим частный случай $\alpha = {}^1/_2$. Собственные функции задачи (15) распадаются на две группы симметричности: симметричные относительно $\theta = 0$ и антисимметричные. В первом случае собственные функции имеют вид

$$\varphi_\nu(\theta) = \begin{cases} \cos \nu\theta & \text{для } |\theta| < \pi/4, \\ \alpha_\nu \cos \nu(\pi - \theta) & \text{для } |\theta| > \pi/4. \end{cases} \tag{19}$$

Постоянная α_ν выбирается так, чтобы выполнялись условия (16), (17) на поверхности раздела:

$$\alpha_\nu = -\left[p_1 \sin\frac{\nu\pi}{4}\right] \Big/ \left[p_2 \sin\frac{3\nu\pi}{4}\right].$$

Собственное значение ν находим, подставляя (19) в (15); можно показать, что или $\operatorname{tg}(\nu\pi/4) = 0$, т. е. $\nu = 4n$, или $\nu = \pm\nu_1 \pm 4n$, где ν_1 — наименьший положительный корень из

$$\left[3 - \operatorname{tg}^2\frac{\nu\pi}{4}\right] \Big/ \left[1 - 3\operatorname{tg}^2\frac{\nu\pi}{4}\right] = -\frac{p_1}{p_2}. \tag{20}$$

Аналогичные формулы справедливы и для класса собственных функций с нечетной симметрией.

Главной особенностью в решении поэтому будет $r^\nu\varphi_\nu(\theta)$, где $\nu = \nu_1$ лежит между 0 и 2. Заметим, что $\nu_1 = 1$ в (20) тогда и только тогда, когда $p_1 = p_2$. В противном случае u принадлежит лишь дробному пространству $\mathcal{H}^{1+\beta}$, $\beta < \nu_1$, и, в частности, не принадлежит $\mathcal{H}^2$ при $\nu_1 < 1$.

Нетрудно распространить эти исследования на случай, когда в P встречается любое число поверхностей раздела. У коэффициента $p(\theta)$ в (15) будет несколько разрывов при условии скачка вида (16) на каждом из них. Келлог заметил, что при пересечении двух прямых поверхностей раздела главный показатель

ν_1 можно сделать сколь угодно малым, если подходящим образом выбрать коэффициенты P_i в каждом из четырех квадрантов. Поэтому особенность может оказаться очень суровой, и без дополнительных условий на пробные функции во всех таких исключительных точках метод конечных элементов будет давать разочаровывающие результаты.

8.2. СИНГУЛЯРНЫЕ ФУНКЦИИ

Разложения, приведенные в разд. 8.1, предлагают модификацию пространств метода конечных элементов, позволяющую улучшить аппроксимацию сингулярных решений. Предположим, что мы можем построить такие независимые функции $\psi_1, \ldots, \psi_s$, что при подходящих (но неизвестных) коэффициентах $c_1, \ldots, c_s$ функция $u - \sum c_i\psi_i$ будет гладкой, скажем будет принадлежать пространству $\mathscr{H}^k$. Тогда почему бы не добавить $\psi_1, \ldots, \psi_s$ к пространству метода конечных элементов S^h? Идея очевидна и заключается в том, чтобы около особенности аппроксимировать u сингулярными функциями $\psi_1, \ldots, \psi_s$ при обычных конечных элементах в других частях области. В результате необходимо определить сингулярные функции только в локальной подобласти около каждой особенности. Поэтому как для углов, так и для поверхностей раздела возьмем

$$\psi_j(r, \theta) = \begin{cases} r^{\nu_j} \sin \nu_j\theta & \text{при} \quad 0 \leqslant r \leqslant r_0, \\ p_j(r) \sin \nu_j\theta & \text{при} \quad r_0 \leqslant r \leqslant r_1, \\ 0 & \text{при} \quad r_1 \leqslant r. \end{cases} \tag{21}$$

Переходные точки r_0 и r_1 фиксируются (независимо от h) и полиномы p_j выбираются так, чтобы коэффициент r^{ν_j} гладко переходил в нуль. Если мы хотим, чтобы пробные функции ψ_i принадлежали $\mathscr{C}^{k-1}$ (и потому $\mathscr{H}^k$), то степень полиномов выбирается равной $2k-1$ и они определяются с помощью эрмитовых условий [1])

$$\frac{d^l}{dr^l}[p_j(r) - r^{\nu_j}]\Big|_{r=r_0} = \frac{d^l}{dr^l} p_j(r)\Big|_{r=r_1} = 0, \quad l = 0, 1, \ldots, k-1. \tag{22}$$

Например, пусть для решения уравнения Лапласа в области с разрезом используются кубические элементы ($k = 4$). В этом случае показатели равны $\nu_j = j - 1/2$, и, согласно (11), суще-

---

[1]) Можно поступать и так: умножать сингулярную функцию $r^\nu \sin \nu\theta$ непосредственно на полином $q(r)$, гладко переходящий в нуль. Тем самым условие в r_0 заменяется условием $q^{(l)}(0) = \delta_{0,l}$.

ствуют такие постоянные $\alpha_1, \ldots, \alpha_4$, что

$$u - \sum_{j=1}^{3} \alpha_j \psi_j = \alpha_4 r^{7/2} \sin\frac{7\theta}{2} + \ \ldots \ + \text{аналитические члены.}$$

Левая часть равенства принадлежит $\mathcal{H}^4$, и потому ее можно аппроксимировать кубическими элементами с оптимальным порядком точности h^4. Итак, наилучшая из возможных степень аппроксимации достигается, если ввести в рассмотрение три сингулярные функции ψ_j.

Более общо, предположим, что мы построили такие функции $\psi_1, \ldots, \psi_s$, что $\left(u - \sum c_i\psi_i\right) \in \mathcal{H}^k$ (при подходящих постоянных $c_1, \ldots, c_s$). Кроме того, пусть S^h — обычное пространство метода конечных элементов степени $k-1$. Тогда если S_s^h — пространство, натянутое на S^h вместе с $\psi_1, \ldots, \psi_s$, то для сингулярной функции u найдется такое приближение $U^h \in S_s^h$, что

$$\|u - U^h\|_l \leqslant Ch^{k-l} \left\| u - \sum_{i=1}^{s} c_i\psi_i \right\|_k. \tag{23}$$

Это ясно: так как функция $v = u - \sum c_i\psi_i$ принадлежит $\mathcal{H}^k$, ее можно приблизить интерполянтом v^h из пространства метода конечных элементов S^h. Тогда $U^h = v^h + \sum c_i\psi_i$.

Эта дополнительная точность аппроксимации, полученная за счет добавления сингулярных функций к пробному пространству, отразится на дополнительной точности приближения u^h методом Ритца — Галёркина. Исследование этого вопроса отложим до следующего раздела, а сейчас остановимся на вычислительных проблемах, которые сопровождают сингулярные функции.

При применении этих идей к фактическим вычислениям (даже когда вид особенностей известен) возникают две трудности. Первая — вычисление скалярных произведений, включающих сингулярные функции, а вторая — обращение матрицы жесткости. Для преодоления первой трудности известно много приемов [Ф6, Ф7], использующих специальный вид сингулярных функций. Вообще говоря, наиболее сингулярную часть — радиальную зависимость в интегралах энергии $a(\psi_i, \psi_j)$ — можно найти аналитически. Интегрирование по θ и, если уж совсем необходимо, вычисление интегралов $a(\varphi_i, \psi_j)$, содержащих только одну сингулярную функцию, можно провести с помощью численного интегрирования высокого порядка. Квадратурные формулы низкой точности здесь совершенно неприемлемы.

Обращение матрицы жесткости составляет более серьезную проблему. Добавление сингулярных функций уничтожает ленточную структуру матриц и может привести к дополнительным

операциям в методе исключений, а также к требованию дополнительной памяти ЭВМ. К тому же под вопросом и обусловленность матрицы жесткости. Сингулярные функции могут приближаться другими конечными элементами, и, следовательно, базис для S_s^h будет «почти линейно зависимым».

На практике обе трудности можно устранить за счет правильного упорядочения неизвестных. Пусть $\varphi_1, \ldots, \varphi_N$ — базис для S^h, а $\psi_1, \ldots, \psi_s$ — сингулярные функции. Тогда упорядочим неизвестные так, чтобы составляющие по функциям $\psi_1, \ldots, \psi_s$ были последними, т. е. чтобы вектор неизвестных имел вид $(Q_1 \ldots Q_N P_1 \ldots P_s) = (QP)$. Поэтому матрицей жесткости для S_s^h будет

$$K_s = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix},$$

где K_{11} — обычная матрица жесткости для S^h, а другие блоки содержат сингулярные функции. Элементами K_{22} служат энергетические скалярные произведения $a(\psi_i, \psi_j)$.

Фаддеев и Фаддеева [22] разлагают K_s в произведение треугольных матриц, которые мы запишем в блочном виде:

$$L = \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix}, \quad U = \begin{bmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix}.$$

Таким образом,

$$K_{11} = L_{11}U_{11}, \quad K_{12} = L_{11}U_{12}, \quad K_{21} = L_{21}U_{11}, \quad K_{22} = L_{21}U_{12} + L_{22}U_{22}.$$

Очевидно, что L_{11} и U_{11} — сомножители обычной матрицы жесткости K_{11}, соответствующей S^h. Следовательно, достаточно хранить ленты матрицы K_{11} и гораздо меньшие матрицы K_{12}, K_{21} и K_{22}. (Отметим, что в симметричном случае $K_{21}^T = K_{12}$.) Дополнительная память ЭВМ равна всего лишь $sN + s^2$, что на несколько порядков меньше по сравнению с требованиями к памяти для K_{11}. Кроме того, на факторизацию и вычисление неизвестных Q и P с помощью обратной подстановки уходит только $O(w^2N)$ операций, где w — ширина ленты матрицы K_{11}. Это столько же, сколько занимает решение уравнения $K_{11}Q = F_1$. Действительно, основная часть работы заключается в факторизации. Эффекты численной неустойчивости изолируются уже в матрицах меньшего размера, и за ошибками округления относительно легко осуществить контроль. Обычно реальное беспокойство вызывает лишь процесс построения матрицы $K_{22} - L_{21}U_{12}$, и зачастую желательно осуществлять последнее умножение с высокой точностью.

Эти идеи применимы также к задачам на собственные значения. Методы деления пополам и (блочной) обратной итерации

требуют представления матриц массы и жесткости в виде LU; если используется окаймление Фаддева — Фаддеевой, то при добавлении сингулярных функций дополнительные проблемы в отношении памяти ЭВМ и численной устойчивости не возникают.

8.3. ОШИБКИ ПРИ НАЛИЧИИ ОСОБЕННОСТЕЙ

Пусть L — самосопряженный эллиптический оператор $2m$-го порядка при однородных краевых условиях и $a(v, w)$ — соответствующее скалярное произведение в энергетическом пространстве $\mathscr{H}_E^m$. Если задача $Lu = f$ имеет поверхность раздела или особенности на границе, аналогичные описанным в разд. 8.1, то оценки ошибок, полученные в разд. 3.4, уже не справедливы. В этом разделе мы модифицируем проведенные ранее исследования по нахождению правильных скоростей сходимости при наличии особенностей.

С помощью разложений, аналогичных выведенным в разд. 8.1, запишем точное решение в виде суммы сингулярных функций и гладкой функции:

$$u = \sum_{i=1}^{s} c_i \psi_i + w. \tag{24}$$

Каждая сингулярная функция ψ_i принадлежит $\mathscr{H}^\sigma$ для некоторого $\sigma > m$ и не зависит от f; в случае углов на границе она зависит только от геометрии области Ω, а в задачах с поверхностями раздела — от коэффициентов оператора L. Можно воспользоваться либо функциями $\psi_i = r^{\nu_i}\varphi_i(\theta)$, как в (11), либо функциями (21); все что здесь нужно, это сохранить правильное поведение решения около точки P. С другой стороны, гладкая функция w и коэффициенты $c_1, \ldots, c_s$ зависят от f. Согласно основополагающей работе Кондратьева [K10], можно не только гарантировать, что $w \in \mathscr{H}^k$ (рассматривая достаточно много функций ψ_i), но даже оценить величину w:

$$\| w \|_k \leqslant C \| f \|_{k-2m}, \qquad \max_{1 \leqslant i \leqslant s} | c_i | \leqslant C \| f \|_0. \tag{25}$$

Заметим, что если Ω и коэффициенты оператора L гладкие, то соответствующие сингулярные функции $\psi_1, \ldots, \psi_s$ равны нулю и, следовательно, (25) переходит в обычную оценку решения через исходные данные задачи.

Вычислим сначала скорость сходимости метода конечных элементов, *не применяя специальные приемы* — сгущение сетки и использование сингулярных функций. Оценить ошибку по энергии деформации не трудно. Как всегда, u^h — ближайшая к u пробная функция, и, если $u \in \mathscr{H}^k$, эта ошибка имеет порядок $h^{2(k-m)}$. Однако в общем случае u принадлежит всего лишь неко-

торому менее гладкому пространству $\mathscr{H}^\sigma$ и скорость сходимости по энергии понижается до $h^{2(\sigma-m)}$. Вероятно, эта ошибка слишком большая.

Предположим, что для того, чтобы без привлечения сингулярных функций оценить ошибку в перемещениях, мы, как и в разд. 1.6 и 3.4, попытаемся воспользоваться приемом Нитше. Как и ранее, в качестве правой части g во вспомогательной задаче возьмем $u - u^h$. Решение этой задачи обозначим через z: $Lz = u - u^h$. Тогда проводимые ранее рассуждения можно повторить здесь без изменений, за исключением одного решающего момента: оценка $\|z\|_{2m} \leqslant C\|u - u^h\|_0$ уже может не быть верной. Действительно, согласно (24), может оказаться, что z будет содержать ненулевые компоненты по сингулярным составляющим $\psi_1, \ldots, \psi_s$, и, следовательно, мы должны довольствоваться более слабым неравенством

$$\|z\|_\sigma \leqslant C\|u - u^h\|_0,$$

вытекающем из (25) (с заменой u на z). Действие этого более слабого неравенства состоит в том, что надо пожертвовать множителем $h^{2m-\sigma}$ и оптимальной оценкой будет

$$\|u - u^h\|_0 \leqslant C\left[h^{r+\sigma-2m} + h^{2(k-m)}\right]\|u\|_r, \qquad r = \min(k, \sigma).$$

Например, в эксперименте с кручением при наличии трещины, описанном в следующем разделе, $\sigma = {}^3/_2$. В результате *при любом выборе элемента* ошибка в наклонах равна $O(h^{1/2})$, а в перемещении — $O(h)$.

Это ошибка на всей области Ω. Однако так как эллиптические уравнения всегда обладают сильным эффектом сглаживания внутри Ω, то вдали от особенности можно надеяться на лучшее. В самом деле, если бы вопрос заключался в обычной аппроксимации с помощью кусочных полиномов по методу наименьших квадратов, то, по-видимому, не было бы нежелательного влияния особенностей: если функция u в подобласти Ω' обладает k производными, то даже без специальных приемов наилучшее приближение по методу наименьших квадратов на Ω дает точность порядка h^k в Ω' [Н6]. Для уравнений второго порядка это уже не так; сказывается некоторое влияние особенностей. Однако показатель степени у h все еще лучше внутри Ω', чем около особенности. Предположим, например, что в области около угла решение ведет себя как r^α. Тогда, согласно Нитше и Шатцу, ошибка в энергетической норме на Ω', которую можно отнести за счет особенности, имеет порядок $h^{2\alpha}$. Для области с разрезом это означает, что ошибка в $\mathscr{H}^1$ вдали от особенности имеет порядок h (и $h^{1/2}$ на всей области Ω).

Теперь перейдем к важному вопросу — скорости сходимости при введении в пробное пространство сингулярных функций. Для

оценки в энергетической норме отметим, что по построению сингулярное пространство S_s^h содержит по крайней мере одну функцию U^h, удовлетворяющую (в соответствии с (23)) соотношению

$$a(u-U^h,\ u-U^h)\leqslant Ch^{2(k-m)}\left\|u-\sum_{1=1}^{s}c_i\psi_i\right\|_k^2. \tag{26}$$

Поэтому *та же оценка справедлива и для* $u-u^h$.

Переходя к оценкам в $\mathscr{H}^0$, рассмотрим снова вспомогательную задачу $Lz=u-u^h$. Решающий момент теории Кондратьева состоит в том, что z можно записать в виде $\sum_{i=1}^{s} d_i\psi_i+v$, где v принадлежит $\mathscr{H}^k$:

$$\|v\|_{2m}\leqslant C\|u-u^h\|_0. \tag{27}$$

В силу гладкости функции v можно приблизить ее пробной функцией v^h из S^h с точностью $O(h^{k-m})$ в энергетической норме. (Как обычно для конечных элементов, мы предполагаем, что $k\geqslant 2m$.) С другой стороны, функция $\sum d_i\psi_i$ принадлежит нашему сингулярному пространству S_s^h, так что $z^h=\sum d_i\psi_i+v^h$ тоже принадлежит S_s^h. Поэтому

$$\|z-z^h\|_m=\|v-v^h\|_m\leqslant C'h^m\|v\|_{2m}\leqslant C''h^m\|u-u^h\|_0. \tag{28}$$

Повторяя (несколько усложненное) рассуждение из разд. 4.4, находим, что

$$\|u-u^h\|_0\leqslant C'''h^k\left\|u-\sum_{i=1}^{s}c_i\psi_i\right\|_k\leqslant C^{(IV)}h^k\|f\|_{k-2m}. \tag{29}$$

Это означает, что *за счет включения достаточного числа сингулярных функций можно получить ту же скорость сходимости, что и для задач с гладкими решениями.* Аналогичный вывод можно сделать и для случая сгущения сетки, если шаг сетки выбирается как среднее $\bar{h}=N^{-1/2}$, где N — размерность пробного пространства. Здесь уже сингулярные функции не принадлежат пространству, но, согласно замечаниям в конце разд. 3.2, подходящее сгущение сетки позволяет аппроксимировать их с порядком $\bar{h}^k$. При этой оценке, как и ранее, действует прием Нитше.

Очевидно, что все эти теоретические предсказания надо тщательно проверить. В сложных физических задачах выявление особенностей, а значит, и их включение в специальные пробные функции, может оказаться чрезвычайно сложным. Поэтому такие построения можно выполнять, только если выгоды будут соответственно большими. Даже сгущение сетки вносит определенные усложнения, хотя оно обычно намного проще, чем построение сингулярных функций. Все, что можно здесь сделать, это

опробовать каждый из указанных методов на ряде упрощенных физических задач и привести полученные результаты. Взглянув на рисунки в следующем разделе, читатель сможет предсказать наше заключение: высокая точность оплачивается или машинным временем при применении простых методов, или временем программирования при применении более сложных. Цены меняются вместе с задачей. Но почти наверняка читатель знает это из своего собственного опыта.

8.4. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Закончим эту главу тремя примерами, вытекающими из физики: (1) вычисление жесткости и деформации упругой балки квадратного сечения с трещиной при кручении; (2) расчет в одногрупповом диффузионном приближении критичности идеализированного квадратного ядерного реактора, состоящего из однородной квадратной активной зоны, окруженной квадратным отражателем; (3) вычисление основной частоты колеблющейся L-образной мембраны.

В задаче о кручении мы проводим численно сравнение локального сгущения сетки и использования сингулярных функций. Если заданы конечные элементы, то скорости сходимости при применении этих методов совпадают и эффективность их главным образом зависит от числа неизвестных, которые требуется найти. С другой стороны, в реакторной задаче мы меньше будем заботиться об особенностях и сосредоточим внимание на эффективных методах, устраняющих трудности, вызванные наличием поверхности раздела. Наконец, мы рассмотрим L-образную мембрану, так как она издавна служила моделью эллиптической задачи с особенностью. В самом деле, специальные методы, разработанные для этой задачи, давали чрезвычайно точные приближения к вибрационным частотам. Мы сравним эти результаты с результатами, полученными по методу конечных элементов.

Дифференциальное уравнение, описывающее задачу о кручении (1), можно представить в нормализованной форме:

$$-\Delta u = 1 \quad \text{в} \quad \Omega, \quad u = 0 \quad \text{на} \quad \Gamma. \tag{30}$$

Область Ω изображена на рис. 8.3; граница Γ включает в себя трещину P_1P. Наши разложения около точки P из разд. 8.1 сводятся к

$$u(r, \theta) = \sum_{j=1}^{\infty} c_j r^{\nu_j} \sin \nu_j \theta + \text{аналитические члены},$$

$$\nu_j = \frac{2j-1}{2}.$$

Таким образом, доминирующий член в особенности в P равен $r^{1/2}\sin(\theta/2)$. Коэффициент при нем

$$c_1 = \lim_{r \to 0} r^{-1/2} [u(r, \pi) - u(0, \pi)] \qquad (31)$$

имеет большое практическое значение в инженерных расчетах: он широко используется как мера кручения, которое балка может вынести, прежде чем сломается; называется он *коэффициентом интенсивности напряжений* [И1]. Отметим, что из-за множителя $r^{1/2}$ решение u обладает производными (в среднем квадратичном) до порядка 3/2.

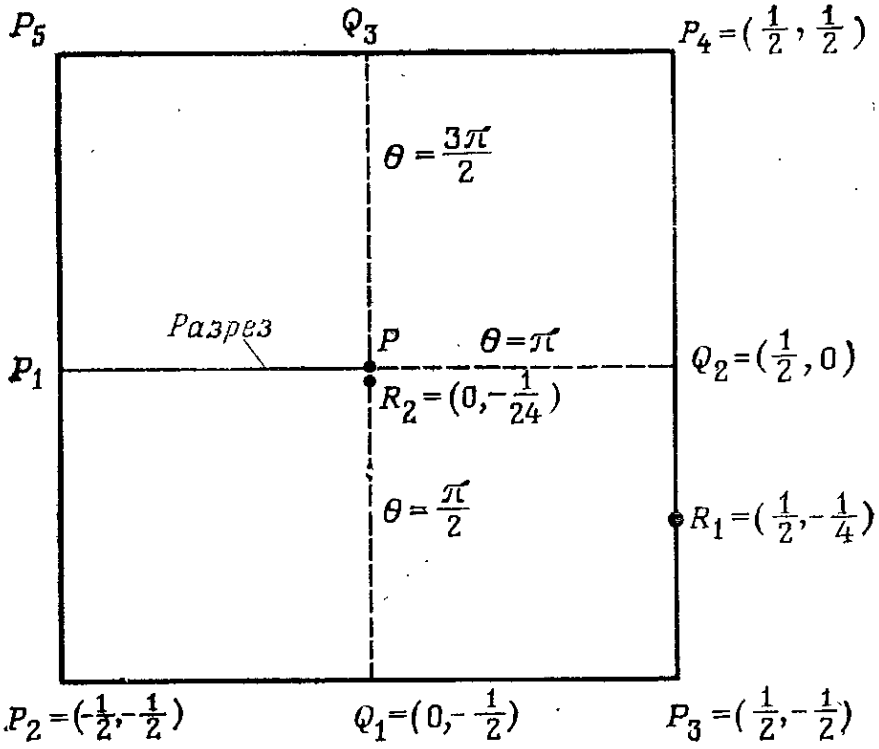


Рис. 8.3.
Область с разрезом; внутренний угол равен 2π.

В задаче, в том виде как она сформулирована здесь, в каждом из углов P_2, P_3, P_4 и P_5 также есть особенности вида $\rho^2 \ln \rho$. Так как эти особенности сравнительно невелики, их можно устранить, просто изменяя краевые условия задачи условиями

$$u = 0 \text{ на } PP_1,\ P_2P_3,\ P_4P_5;\quad \frac{\partial u}{\partial \nu} = 0 \text{ на } P_1P_2,\ P_1P_5,\ P_3P_4, \qquad (32)$$

где ν — нормаль к Г. С помощью методики разд. 8.1 можно проверить, что решение u новой задачи с краевыми условиями (32) аналитично вне точки P [1]).

[1]) Согласно принципу Сен-Венана, указанное изменение краевых условий не затрагивает особенность в P; это очевидно также из наших рассуждений в разд. 8.1.

Мы будем вычислять приближения в четырех различных пространствах; в первых трех случаях будем рассматривать равномерную сетку с шагом h и использовать сингулярные функции, построенные, как и в разд. 8.2, из $r^{\nu_j}\sin\nu_j\theta$, $\nu_j=(2j-1)/2$. Эти пространства таковы:

1. S_L^h — пространство непрерывных функций на равномерной сетке, на каждой подобласти представляющих собой билинейные полиномы

$$a+bx+cy+dxy.$$

Это пространство содержит сингулярную функцию ψ_1, построенную из $r^{1/2}\sin(\theta/2)$; так выполняется свойство аппроксимации (26) с $k=2$.

2. S_H^h — *бикубическое эрмитово пространство* (разд. 1.8) вместе с сингулярными функциями ψ_1, ψ_2 и ψ_3, так что (26) выполняется с $k=4$.

Эти пространства входят в категорию пространств *узловых конечных элементов* (в том смысле, который мы обсуждали ранее), и потому, быть может, стоит рассмотреть также и пример пространства, рассматриваемого в *абстрактном методе конечных элементов*. Пространство бикубических сплайнов, принадлежащих классу $\mathscr{C}^2$ в Ω, вероятно, подходит лучше всего, но в данной задаче с этим пространством работать трудно. Неприятности связаны с главным краевым условием $u=0$ на трещине PP_1. Бикубический сплайн, равный нулю на этой прямой, будет в точке P таким, что не сможет как следует аппроксимировать истинное решение u, все производные которого сингулярны. Чтобы обойти эту трудность, потребуем, чтобы сплайны при переходе через прямые PQ_2, PQ_3 и PQ_4 были всего лишь непрерывными ($\mathscr{C}^0$), пространство таких сплайнов обычно называют *сплайн-лагранжевым*:

3. S_{SL}^h состоит из кусочно бикубических полиномов класса $\mathscr{C}^2$ всюду, кроме прямых PQ_2, PQ_3 и PQ_4. На них производная по нормали допускает разрывы. Как и в S_H^h, сюда включаются сингулярные функции ψ_1, ψ_2 и ψ_3, так что (26) выполняется с $k=4$.

Преимущество сплайн-лагранжева пространства состоит в том, что здесь в каждом узле всего лишь по одному неизвестному, за исключением прямых PQ_1, PQ_2, PQ_3, где их в узле три. Таким образом, размерность пространства S_{SL}^h примерно в 4 раза меньше размерности эрмитова пространства S_H^h и фактически та же, что и у пространства кусочно билинейных эле-

ментов S_L^h. Между прочим, базис для S_{SL}^h сразу получается из обычных сплайновых формул, если рассматривать прямые PQ_1, PQ_2 и PQ_3 как *тройное слияние линий узлов.*

Наше последнее пространство состоит из треугольных элементов и использует градуированную сетку с максимальной стороной треугольника, равной h, и минимальной, равной δ.

4. $S_L^{h,\delta}$ — пространство непрерывных функций, которые сводятся к *линейным полиномам* на каждом треугольнике. Мы предполагаем, что $\delta = Q(h^2)$, так что (26) выполняется с $k = 2$ [1]).

Из оценок, установленных в разд. 8.3, вытекает, что $\left(\int |u - u^h|^2\right)^{1/2}$ имеет порядок $O(h^4)$ для кубических пространств и $O(h^2)$ для S_L^h и $S_L^{h,\delta}$. Однако, так как решение u в замкнутом виде не известно, непосредственно измерить эту ошибку нельзя. Поэтому мы брали также элементы пятой степени с несколькими сингулярными функциями. Точнее, мы проводили вычисления со сплайнами пятой степени, принадлежащими классу $\mathscr{C}^4$ всюду в Ω, кроме прямых PQ_1, PQ_2 и PQ_3, т. е. рассматривали сплайн-лагранжево пространство пятой степени, дополненное шестью сингулярными функциями. При работе с этим пространством приближения очень быстро сходятся: скорость сходимости имеет порядок h^6 и приближенные решения для $h = 1/6$, $1/8$ и $1/10$ отличаются лишь в седьмом знаке. Первые шесть цифр брались в качестве значения истинного решения u.

На рис. 8.4 и 8.5 показаны ошибки в точках R_1 и R_2 в зависимости от усредненного шага $\bar{h} = N^{-1/2}$ [2]). Анализ их позволяет выбрать наиболее «эффективное» пространство, дающее заданную ошибку при наименьшем числе неизвестных параметров. Пространство S_L^h в этом отношении оказывается эффективнее пространства треугольных элементов $S_L^{h,\delta}$, использующего локальное сгущение сетки. Для последнего пространства мы брали $h = 1/4$, $\delta = 1/16$ и $h = 1/8$, $\delta = 1/64$ согласно правилу в п. 4. Таким образом, чтобы получить сходимость $h^{3/2}$, требуется много дополнительных элементов, однако при S_L^h достаточно одного дополнительного неизвестного — коэффициента при сингулярной функции ψ_1. Пространства кубических элементов S_H^h, S_{SL}^h значительно превосходили по эффективности S_L^h; пространство наи-

---

[1]) В вычислениях переход от шага h к $\delta = h^2$ достигался последовательным делением треугольников пополам.

[2]) Хотя приведенные выше оценки ошибок не дают информации относительно скорости поточечной сходимости, она для данной задачи, как это можно проверить по графикам, оказывается того же порядка.

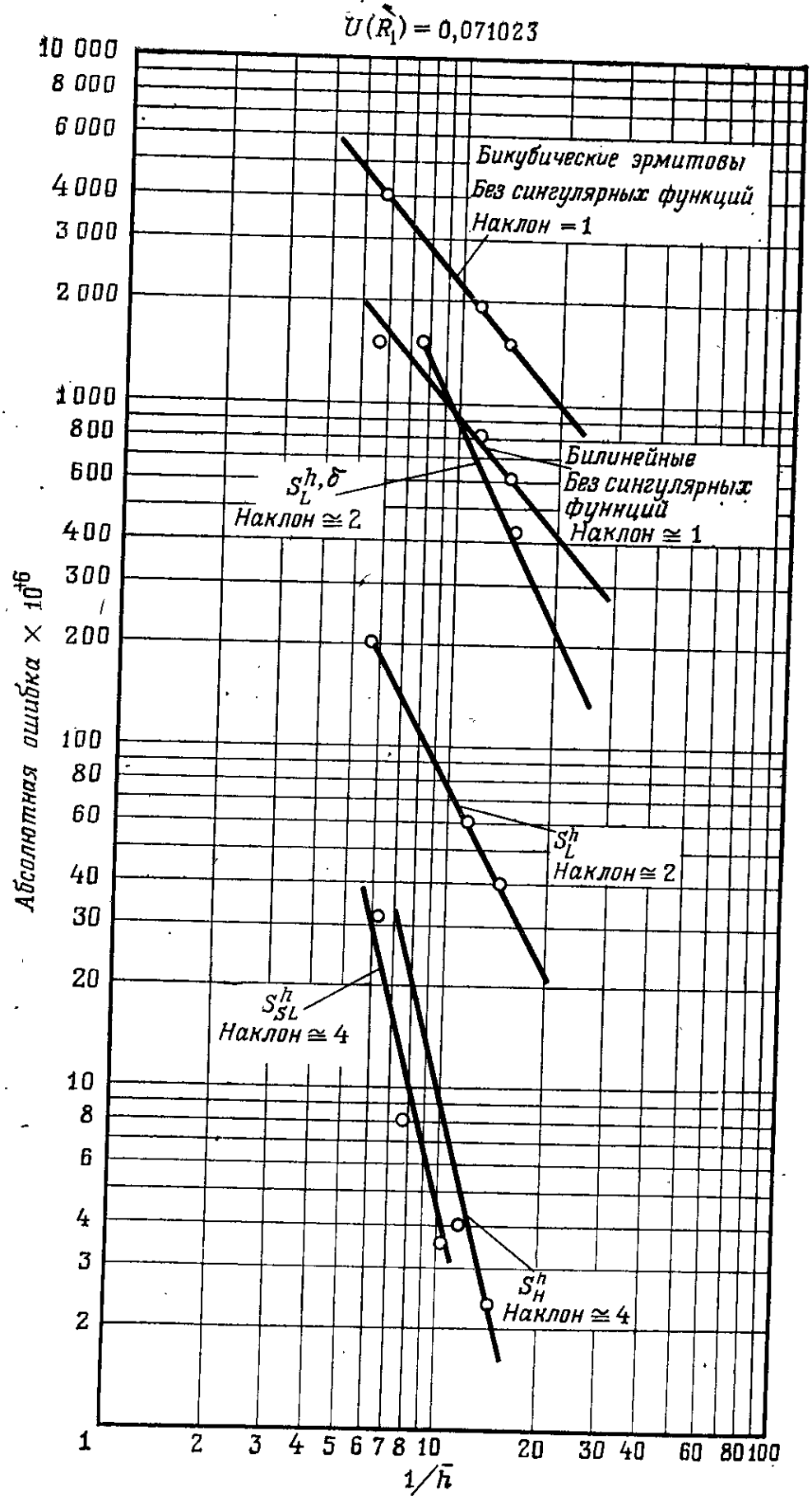

$U(\hat{R}_1) = 0{,}071023$
Бикубические эрмитовы
Без сингулярных функций
Наклон = 1
Билинейные
Без сингулярных
фунний
Наклон ≅ 1
$S_L^{h,\delta}$
Наклон ≅ 2
S_L^h
Наклон ≅ 2
S_{SL}^h
Наклон ≅ 4
S_H^h
Наклон ≅ 4
Абсолютная ошибка × 10^{+6}
10 000
8 000
6 000
4 000
3 000
2 000
1000
800
600
400
300
200
100
80
60
40
30
20
10
8
6
4
3
2
1
2 3 4 5 6 7 8 10 20 30 40 60 80 100
$1/\bar{h}$

Рис. 8.4.

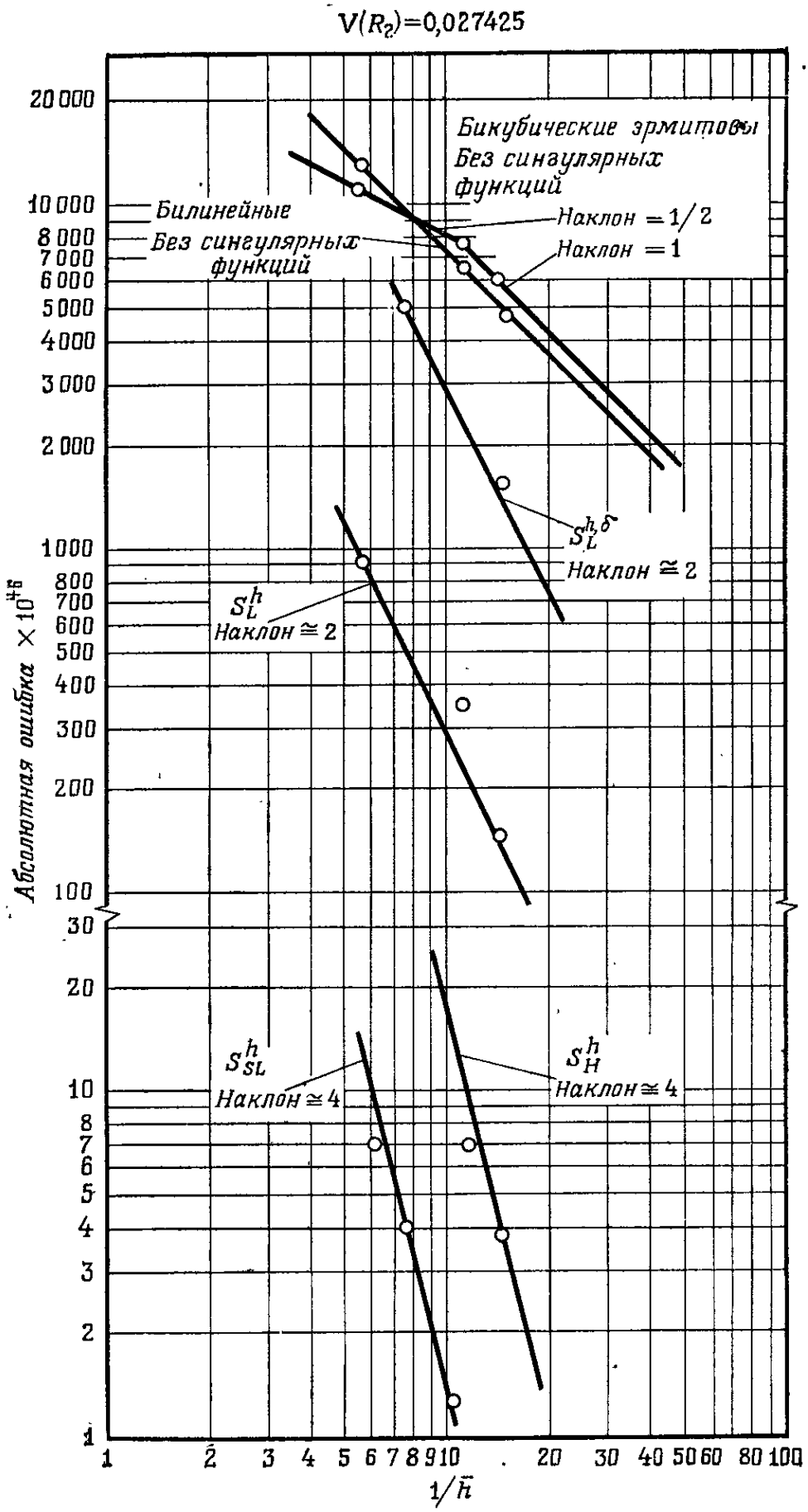

V(R2)=0,027425
Бикубические эрмитовы
Без синэулярных
функций
Наклон =1/2
Наклон =1
Билинейные
Без синэулярных
функций
$S_L^{h,\delta}$
Наклон ≅2
S_L^h
Наклон ≅2
S_{SL}^h
Наклон ≅4
S_H^h
Наклон ≅4
Абсолютная ошибка ×10^4δ
1/h
20 000
10 000
8 000
7 000
6 000
5 000
4 000
3 000
2 000
1000
800
700
600
500
400
300
200
100
30
20
10
8
7
6
5
4
3
2
1
1 2 3 4 5 6 7 8 9 10 20 30 40 50 60 80 100

Рис. 8.5.

меньшей размерности, а именно сплайн-лагранжево пространство S^h_{SL}, оказывается наиболее эффективным из всех.

На рисунках приводятся (возросшие) ошибки для пространств метода конечных элементов *без сингулярных функций.* На этих графиках интересно отметить три момента. Во-первых, естественно, улучшение, достигаемое с сингулярными функциями (например, в R_1 при $h = 1/4$ относительная ошибка без сингулярных функций равна примерно 40%, и она падает до 0,1% при добавлении их). Во-вторых, при отсутствии сингулярных функций поточечные ошибки максимальны вблизи P. В частности, эти ошибки имеют порядок $O(h^{1/2})$ вблизи P и $O(h)$ в остальных точках [1]). В-третьих, обычные эрмитовы кубические элементы оказываются *хуже*, чем простейшие линейные элементы. Кубические элементы слишком гладки, чтобы справиться с особенностью.

Наконец, на рис. 8.6 мы приводим приближения к коэффициенту интенсивности напряжений (31). Для пространств S^h_L, S^h_H и S^h_{SL} мы берем коэффициент c_1 при сингулярной функции ψ_1, так как

$$c_1 = \lim_{r \to 0} r^{-1/2} [u^h(r, \pi) - u^h(0, \pi)].$$

Для пространства $S^{h,\delta}_L$ эта величина равна нулю, так что надо выбрать некоторое разностное отношение

$$c_1 = \frac{u^h(\xi_h, \pi) - u^h(0, \pi)}{\xi_h^{1/2}}.$$

Поскольку ошибки приближения методом конечных элементов без сингулярных функций довольно быстро растут при приближении к точке P, выбор $\xi_h = O(h)$ так же хорош, как и любой другой. Эти ошибки представлены на рис. 8.6. Выводы здесь те же, что и в предыдущих экспериментах, но только теперь из-за дополнительной ошибки, возникающей в разностном отношении, метод сгущения сетки становится даже менее конкурентоспособным.

Наш второй пример — одногрупповой двухзонный реактор, описываемый уравнением

$$-\nabla \cdot (p \nabla u) + qu = \lambda \rho u. \tag{33}$$

---

[1]) Исследования, аналогичные проведенным в разд. 8.3, показывают, что среднеквадратичная ошибка в этом случае есть $O(h)$; поэтому площадь области около P, где поточечная ошибка есть $O(h^{1/2})$, должна быть довольно мала и должна стремиться к нулю при $h \to 0$. Это объясняет странное поведение ошибки при линейных функциях; R_2 при малом h находится в пределах пограничного слоя.

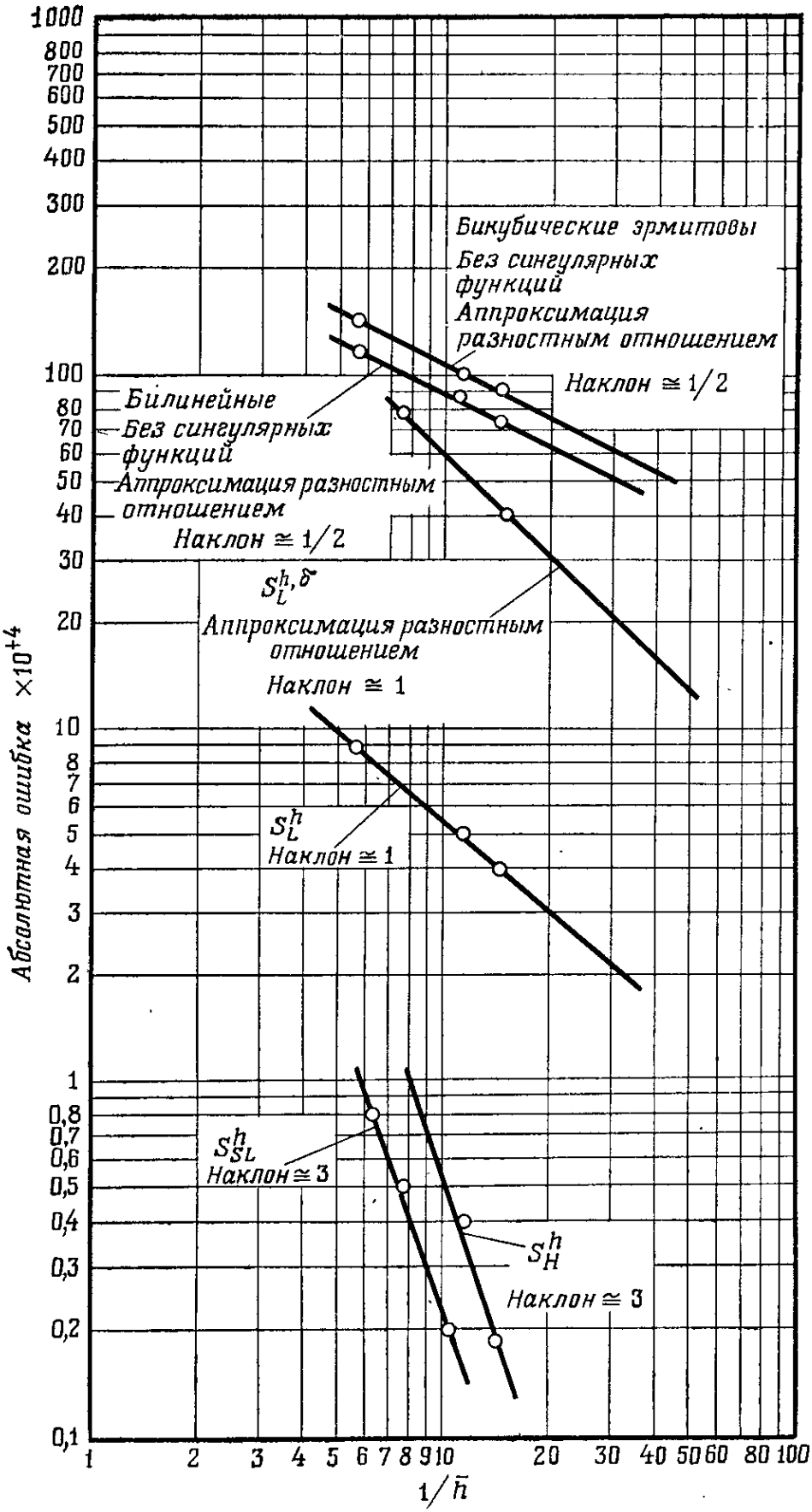


Рис. 8.6.

Коэффицнент нитенсивности напряжений равен 0,1917.

Это дифференциальное уравнение выполняется в активной зоне Ω_1 и отражателе Ω_2 (см. рис. 8.7); мы требуем, чтобы энергетический поток u равнялся нулю на внешней границе:

$$u = 0 \quad \text{на} \quad \Gamma_2. \tag{34}$$

Коэффициенты p, q и ρ постоянны в каждой зоне. Таким образом, кривая Γ, разделяющая Ω_1 и Ω_2, является поверхностью

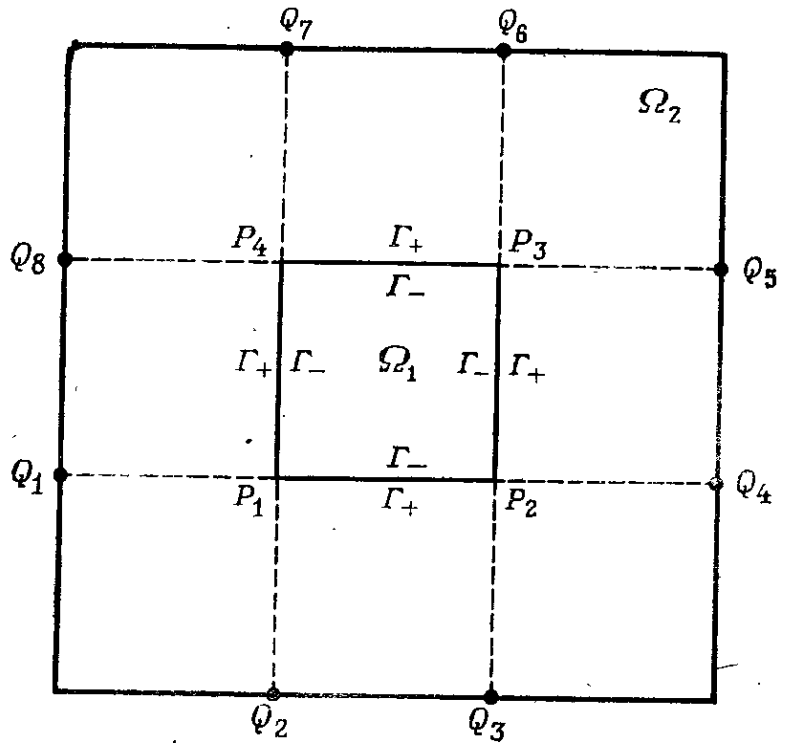


Рис. 8.7.

Квадратная активная зона, окруженная квадратным отражателем.

раздела зон, и мы требуем, чтобы u и $p\partial u/\partial \nu$ при переходе через Γ были непрерывными:

$$p\frac{\partial u}{\partial \nu}\Big|_{\Gamma_+} = p\frac{\partial u}{\partial \nu}\Big|_{\Gamma_-}, \qquad u|_{\Gamma_+} = u|_{\Gamma_-}. \tag{35}$$

Из всех величин, которые надо вычислить, важнее всех наименьшее собственное значение λ, измеряющее критичность реактора.

Независимо от того, имеет граница угол или нет, наличие поверхностей раздела основательно меняет наилучший выбор подходящего пространства метода конечных элементов S^h. Так как u обладает разрывными на Γ производными, использование кусочных полиномов, принадлежащих классу $\mathscr{C}^1$ при переходе через Γ, обычно приводит к плохой аппроксимации. Более того, использование пробных функций, удовлетворяющих условию скачка (35), приводит к дополнительным трудностям в углах P_j, $j = 1, 2, 3, 4$. Если мы заставим пробные функции удовлетворять условиям скачка вдоль P_1P_2, то их влияние будет сказываться и на отрезке Q_1P_1, где решение гладкое.

Благодаря своей удобной геометрии, сплайн-лагранжево пространство, подобное описанному в задаче о кручении, оказывается наиболее подходящим. Идея заключается в разбиении области Ω на квадраты так, чтобы поверхности раздела зон лежали на линиях сетки. Пространство пробных функций S^h состоит из кусочных бикубических полиномов, принадлежащих $\mathscr{C}^2$ всюду, кроме прямых Q_1Q_4, Q_2Q_7, Q_3Q_6 и Q_5Q_8, при переходе через которые они лишь непрерывны. Условием скачка (35), являющимся *естественным краевым условием*, пренебрегаем и позволяем пробным функциям претерпевать любые разрывы-скачки в их производных по направлению нормали при переходе через поверхности раздела. Так как метод Галеркина дает в некотором смысле наилучшее приближение, то скачки на поверхностях раздела, по-видимому, можно будет вычислить удовлетворительно!

Приближения к первому собственному значению при использовании этого пространства приведены на рис. 8.8 для случая

$$p_1 = 5, \quad p_2 = 1, \quad q = 0, \quad \rho = 1.$$

Приближенные значения λ^h обладают приемлемой точностью, но *они очень медленно сходятся к* λ. В частности,

$$\lambda - \lambda^h = O(h^{2\nu}), \quad \nu = 0{,}78. \tag{36}$$

Это объясняется тем, что первые производные собственной функции u в точках P_1, P_2, P_3 и P_4 не ограничены. Действительно, исследования в разд. 8.1 показывают, что доминирующий член в особенности имеет вид $r^\nu \varphi_\nu(\theta)$, где φ_ν — периодическая функция от θ и $\nu = 0{,}78$. Поэтому ошибка в собственном значении не может быть лучше чем $h^{2\nu}$. Включая $r^\nu\varphi_\nu$ в пространство пробных функций, можно увеличить скорость сходимости примерно до $h^{6-2\nu}$. Это также подтверждается численными результатами.

Из-за наличия особенности может показаться неожиданным, что значение λ^h при равномерной сетке и без сингулярных функций довольно точное. Это прямо противоположно ситуации в задаче о кручении и объясняется тем, что коэффициент при $r^\nu\varphi_\nu$ совсем мал; график, построенный по вычисленным значениям функции u^h, показывает, что она фактически постоянна в активной зоне Ω_1. Мы проверили это свойство (вытекающее из физики процесса), вычисляя также критическое собственное значение для случая [1])

$$p_1 = 500, \quad p_2 = 1, \quad q = 0, \quad \rho = 1.$$

Это собственное значение сдвинуто лишь до 5,582 и потому фактически не зависит от p_1; вклад из внутреннего квадрата Ω_1 в отношение Рэлея почти равен нулю.

---

[1]) Здесь ν примерно равно 2/3.

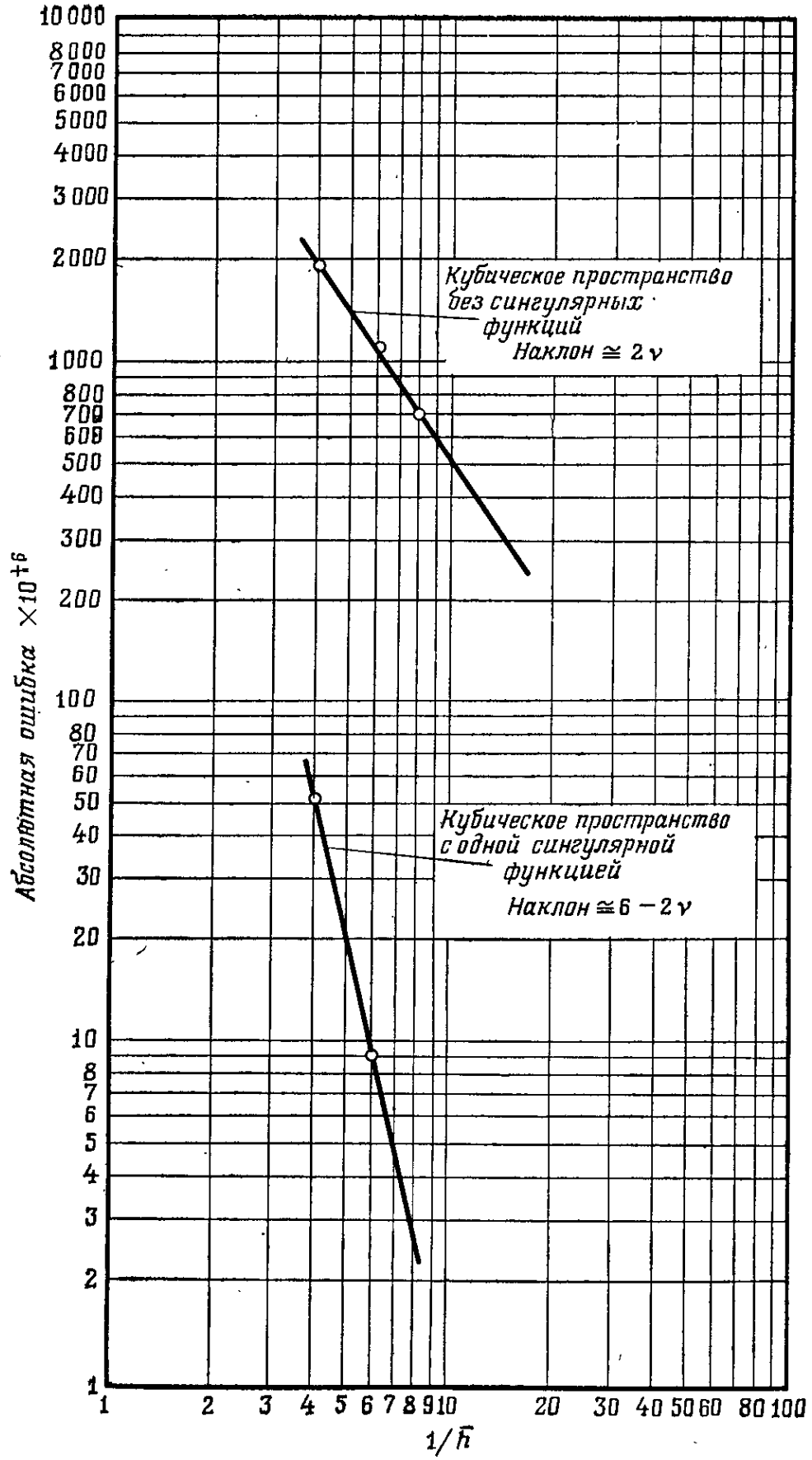


Рис. 8.8.

$p_1 = 5$, $p_2 = 1$, $\nu \cong 0{,}784$, $\lambda \cong 5{,}5822736$.

С физической точки зрения эта «слабая» связь с «сильной» особеиностью вполне удовлетворительна. Действительно, в отличие от задачи о кручении, особенность в P не имеет физического смысла — уравнение (33) в этой области нужно заменить уравнением переноса. Поэтому, казалось бы, привлечение сингулярных функций для таких задач не обязательно и не стоит затрачивать усилия на сгущение сетки. Этот вывод применим *не* ко всем задачам с поверхностями раздела. В разд. 8.1 упоминалось, что

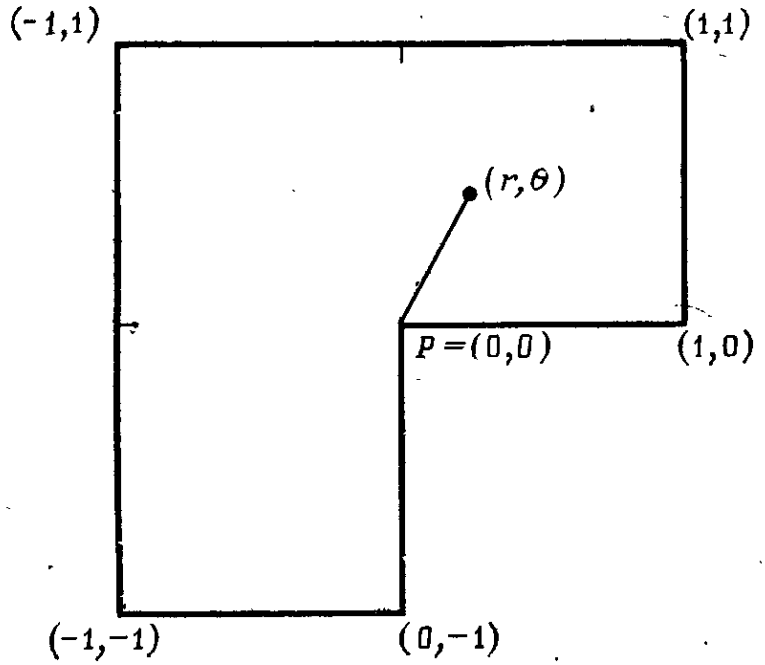


Рис. 8.9.
L-образная мембрана.

пересечение поверхностей разделов может создать доминирующий член в особенности любого порядка $r^{\varepsilon}\varphi_{\varepsilon}(\theta)$ [1]). Представляется правдоподобным, что малое значение ε, как и в задаче о кручении, может привести к плохим аппроксимациям и сингулярные функции или локальное сгущение сетки будут необходимы для получения приемлемых результатов.

Наш последний пример — извечная L-образная мембрана, «перегруженная», но тем не менее эффективная модель (рис. 8.9). Требуется найти собственные значения задачи $-\Delta u = \lambda u$ в Ω, $u = 0$ на Γ. Отметим (разд. 8.1), что собственные функции вблизи входящего угла P ведут себя, как

$$u(r, \theta) = \sum_{j=1}^{\infty} \sum_{l=0}^{\infty} c_{jl} r^{\nu_j + 2l} \sin \nu_j \theta, \quad \nu_j = \frac{2j}{3} \tag{37}$$

(плюс аналитические члены).

[1]) Для одногруппового двухзонного реактора собственные значения совпадают с корнями уравнения (20) и $\varepsilon \geqslant 2/3$.

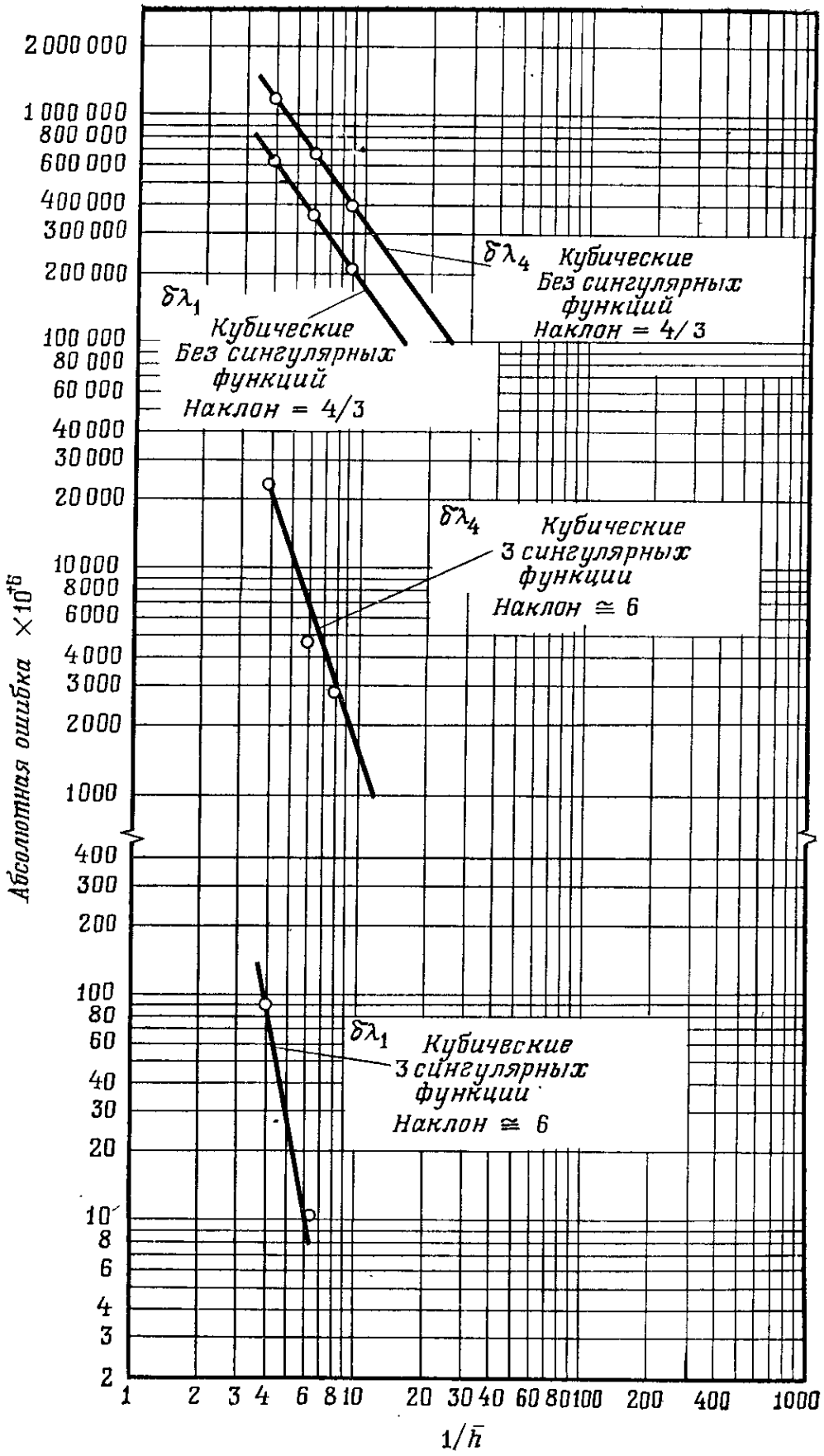


Рис. 8.10.

$\lambda_1 = 9{,}639723844$, $\lambda_4 = 41{,}474516$.

Наиболее точные аппроксимации получили Фокс, Генричи и Молер [Ф11]. Их идея состоит в том, чтобы брать линейную комбинацию решений уравнения $\Delta u + \lambda u = 0$, в данном случае комбинацию функций $K_\nu = J_\nu(\sqrt{\lambda} r) \sin \nu\theta$, $\nu = \nu_j$, и, минимизируя ее на границе Г, определить неизвестные коэффициенты. Этот метод в некотором смысле двойствен методу Галёркина: он оперирует с точными решениями и приближает краевые условия. Но здесь важен класс используемых функций. Известно, что собственные функции u можно очень точно приблизить линейной комбинацией функций K_ν, и потому метод Галёркина при том же классе функций дает приближения той же точности. Строго говоря, функции K_ν недопустимы, так как они не удовлетворяют главным краевым условиям. Однако опубликованные в [Ф7] результаты вычислений с близкими к ним функциями $(1 - x^2)(1 - y^2) r^{\nu+2l} \sin \nu\theta$ сравнимы с результатами Фокса, Генричи, Молера.

Эти вычисления отражают замечательную точность, которую можно достичь, если точно известны правильные пробные функции, причем не всегда кусочно полиномиальные! (Потребовалось бы 14 сингулярных функций, объединенных с бикубическими элементами, чтобы улучшить описанные выше результаты.) Однако нас больше интересует хорошая точность при простых модификациях обычной программы метода конечных элементов. Поэтому мы вычислили первое и четвертое собственные значения, используя пространство кубических сплайнов с сингулярными функциями и без них. На рис. 8.10 показано, как резко ускоряется медленная сходимость ($h^{4/3}$) в последнем случае при введении трех сингулярных пробных функций. Без сомнения, построение этих специальных функций делает всю программу более эффективной.

Из анализа всех описанных нами численных экспериментов заключаем, что даже для грубых сеток и для задач с особенностями предсказываемые теорией скорости сходимости четко воспроизводятся вычислениями. В технической литературе описывается много численных экспериментов, приводящих к тому же выводу. Это означает, что наша цель — проанализировать этапы метода конечных элементов и объяснить его успех — в основном достигнута. Мы надеемся, что этот анализ послужит теоретической основой дальнейшего развития метода. Простота и удобство полиномиальных элементов уже были ясны, а достигаемая ими точность теперь подтвердилась математически.

СПИСОК ЛИТЕРАТУРЫ

Число публикаций по методу конечных элементов, как в технической литературе, так и в литературе по численному анализу, продолжает расти так быстро, что любая попытка составить полный список становится невозможной. Библиография, включая 170 работ, опубликованных до 1971 г., содержится в прекрасном обзоре

Зенкевич (Zienkiewicz O. C.), The finite element method: from intuition to generality, *Appl. Mech. Rev.*, **23** (1970), 249—256.

При подготовке данной книги мы пользовались следующими монографиями (за исключением тех, которые появились в 1972 г.).

1. Агмон (Agmon S.), Lectures on elliptic boundary value problems, Van Nostrand Reinhold, New York, 1965.
2. Аржирис (Argyris J. H.), Energy theorems and structural analysis, Butterworth, London, 1960.
3. Аржирис (Argyris J. H.), Recent advances in matrix methods of structural analysis, Pergamon Press, Elmsford, N. Y., 1964.
4. Вайнберг М. М., Вариационные методы исследования нелинейных операторов, Гостехиздат, М., 1956.
5. Визер (Visser M.), The finite element method in deformation and heat conduction problems, Delft, Holland, 1968.
6. Дизей, Абель (Desai C., Abel J.), Introduction to the finite element method, Van Nostrand Reinhold, New York, 1972.
7. Зенкевич О., Метод конечных элементов в технике, изд-во «Мир», М., 1975.
8. Зенкевич, Холистер (Zienkiewicz O. C., Holister G. S., eds.), Stress analysis, Wiley, New York, 1965.
9. Красносельский М. А., Вайникко Г. М. и др., Приближенное решение операторных уравнений, изд-во «Наука», 1969.
10. Лионс Ж.-Л., Некоторые методы решения нелинейных краевых задач, изд-во «Мир», М., 1972.
11. Лионс Ж.-Л., Мадженес Э., Неоднородные граничные задачи и их приложения, изд-во «Мир», М., 1971.
12. Михлин С. Г., Вариационные методы в математической физике, изд-во «Наука», М., 1970.
13. Михлин С. Г., Проблема минимума квадратичного функционала, Гостехиздат, М. — Л., 1952.
14. Михлин С. Г., Численная реализация вариационных методов, изд-во «Наука», М., 1966.
15. Нечас (Nečas J.), Les méthodes directes en théorie des équations elliptiques, Academia, Prague, 1967.
16. Обэн Ж.-П., Приближенное решение эллиптических краевых задач, изд-во «Мир», М., 1977.

17. Оден Дж., Конечные элементы в нелинейной механике сплошных сред, изд-во «Мир», М., 1976.
18. Пржемииицкий (Przemieniecki J. S.), Theory of matrix structural analysis, McGraw-Hill, New York, 1968.
19. Синж (Synge J. L.), The hypercircle in mathematical physics, Cambridge University Press, New York, 1957.
20. Уилкинсон (Wilkinson J. H.), Rounding errors in algebraic processes, Prentice-Hall, Englewood Cliffs, N. J., 1963.
21. Уилкинсон, Райнш (Wilkinson J. H., Reinsch C.), Linear algebra, Springer-Verlag, Berlin, 1971.
22. Фаддеев Д. К., Фаддеева В. Н., Вычислительные методы линейной алгебры, Физматгиз, М., 1960.
23. Холанд, Белл (ред.) (Holand I., Bell K., eds.), Finite element methods in stress analysis, Tapir, Trondheim, Norway, 1969.

Полезно перечислить некоторые недавно прошедшие конференции и симпозиумы, на которых обсуждался метод конечных элементов. Их опубликованные записки содержат много ценных статей. Те из них, на которые мы ссылались в тексте, приведены, кроме того, в списке отдельных статей; при этом указана соответствующая конференция, например Wright-Patterson Conference II.

1. Proceedings of the First Conference on Matrix Methods in Structural Mechanics, Wright-Patterson AFB, Ohio, 1965.
2. Proceedings of the Second Conference on Matrix Methods in Structural Mechanics, Wright-Patterson AFB, Ohio, 1968.
3. Proceedings of the Third Conference on Matrix Methods in Structural Mechanics, Wright-Patterson AFB, Ohio, 1971.
4. Proceedings IUTAM Symposium, High Speed Computing of Elastic Structures, Liège, Belgium, 1970.
5. Numerical Solution of Partial Differential Equations (SYNSPADE, University of Maryland), ed. B. Hubbard, Academic Press, New York, 1971.
6. The Mathematical Foundations of the Finite Element Method (University of Maryland at Baltimore), Academic Press, New York, 1973.
7. Numerical Solution of Field Problems in Continuum Physics, ed. G. Birkhoff, Duke University, SIAM-AMS Proceedings, Vol. 2, 1970.
8. SMD Symposium on Computer-Aided Engineering, ed. G. L. M. Gladwell, University of Waterloo, May 1971.
9. Finite Element Techniques in Structural Mechanics, eds. H. Tottenham and C. Brebbia, Southampton University Press, 1970.
10. Conference on Variational Methods in Engineering, Southampton University, England, 1972.
11. Conference on the Mathematics of Finite Elements and Applications, Brunel University, England, 1972.
12. Proceedings of the International Symposium on Numerical and Computer Methods in Engineering, University of Illinois, 1971.
13. Proceedings of the American Nuclear Society Meeting, Boston, 1971.
14. Proceedings of the First International Conference on Nuclear Reactor Structures, Berlin, 1971.
15. Proceedings of the First Symposium on Naval Structural Mechanics, eds. J. H. Goodier and H. J. Hoff, Pergamon Press, 1960.
16. Proceedings of the Symposium on Finite Element Techniques, ISD, Stuttgart, 1969.
17. Recent Advances in Matrix Methods of Structural Analysis and Design, eds. J. T. Oden, R. H. Gallagher, and Y. Yamada, University of Alabama Press, 1971. (Первый Японо-американский семинар в Беркли, 1972.)
18. Application of Finite Element Methods to Stress Analysis Problems in Nuclear Engineering, ISPRA, Italy, 1971.

19. Conference on Computer Oriented Analysis of Shell Structures, Lockheed Palo Alto Research Laboratories, Palo Alto, Calif., 1970.
20. Symposium on the Application of Finite Element Methods in Civil Engineering, Vanderbilt University, 1969.
21. Symposium on Application of the Finite Element Method in Stress Analysis, Swiss Society of Architects and Engineers, Zurich, 1970.
22. National Symposium on Computerized Structural Analysis and Design, George Washington University, 1972.
23. On General Purpose Finite Element Computer Programs, ed. P. V. Marcal, Amer. Society of Mechanical Engineers.
24. Proceedings of the NATO Advanced Study Institute, Lisbon, 1971.
25. Computational Approaches in Applied Mechanics, ASME Joint Computer Conference, Chicago, 1969.

Приведем список статей, на которые мы ссылались, а также и много других работ, на которые можно было бы сослаться. Здесь еще раз уместно повторить, что поиск нужной литературы часто приводит к выявлению громадного числа важных работ. В частности, это относится к техническим статьям. Приводимый список указывает журналы, где хорошо представлен метод конечных элементов; дополненный записками конференций, он поможет проделать значительную аналитическую и теоретическую работу по данному методу.

A1. Айронс (Ironns B. M.), Engineering applications of numerical integration in stiffness methods, *AIAA J.*, **4** (1966), 2035—2037. (Русский перевод: *Ракетная техника и космонавтика*, 1966.)
A2. Айронс (Irons B. M.), Roundoff criteria in direct stiffness solutions, *AIAA J.*, **6** (1968), 1308—1312. (Русский перевод: *Ракетная техника и космонавтика*, 1968.)
A3. Айронс (Irons B. M.), Economical computer techniques for numerically integrated finite elements, *Int. J. Num. Meth. Eng.*, **1** (1969), 201—203.
A4. Айронс (Irons B. M.), A frontal solution program for finite element analysis, *Int. J. Num. Meth. Eng.*, **2** (1970), 5—32.
A5. Айронс (Irons B. M.), Quadrature rules for brick based finite elements, *AIAA J.*, **9** (1971), 293—294. (Русский перевод: *Ракетная техника и космонавтика*, 1971.)
A6. Айронс, де Оливьера, Зенкевич (Irons B. M., de Oliveira E. A., Zienkiewicz O. C.), Comments on the paper: Theoretical foundations of the finite element method, *Int. J. Solids Struct.*, **6** (1970), 695—697.
A7. Айронс, Раззак (Irons B. M., Razzaque A.), A new formulation for plate bending elements (рукопись), 1971.
A8. Алман (Allman D. J.), Finite element analysis of plate buckling using a mixed variational principle, Wright-Patterson III, 1971.
A9. Андерхеген (Anderheggen E.), A conforming triangular finite element plate bending solution, *Int. J. Num. Meth. Eng.*, **2** (1970), 259—264.
A10. Аржирис, Брёнлунд, Григер, Сёренсен (Argyris J. H., Brönlund O. E., Grieger T., Sörensen M.), A survey of the application of finite element methods to stress analysis problems with particular emphasis on their application to nuclear engineering problems, ISPRA Conference, 1971.
A11. Аржирис, Келси (Argyris J. H., Kelsey S.), Modern fuselage analysis and the elastic aircraft, Butterworth, London, 1963.
A12. Аржирис, Фрид (Argyris J. H., Fried I.), The LUMINA element for the matrix displacement method, *J. Royal Aero Soc.*, 1968, 514—517.
A13. Аржирис, Фрид, Шарп (Argyris J. H., Fried I., Scharpf D. W.), The Hermes eight element for the matrix displacement method, *J. Royal Aero. Soc.*, 1968, 613—617.

А14. Ахмад, Айронс, Зенкевич (Ahmad S., Irons B. M., Zienkiewicz O. C.), Curved thick shell and membrane elements with particular reference to axisymmetric problems, Wright-Patterson II, 1968.
Б1. Бабушка (Babuška I.), Устойчивость областей определения..., *Czech. Math. J.*, **11**(86), (1961), 76—105, 165—203.
Б2. Бабушка (Babuška I.), Finite element method for domains with corners, *Computing*, **6** (1970), 264—273.
Б3. Бабушка (Babuška I.), Approximation by hill functions, Tech. Note 648, Univ. Maryland, 1970.
Б4. Бабушка (Babuška I.), Finite element method with penalty, Rept. BN-710, Univ. Maryland, 1971.
Б5. Бабушка (Babuška I.), Error bounds for the finite element method, *Numer. Math.*, **16** (1971), 322—333.
Б6. Бабушка (Babuška I.), The finite element method with Lagrangian multipliers, *Numer. Math.*, **20** (1973), 179—192.
Б7. Базелей, Чаиг, Айронс, Зенкевич (Bazeley G. P., Cheung Y. K., Irons B. M., Zienkiewicz O. C.), Triangular elements in plate bending — conforming and nonconforming solutions, Wright-Patterson I., 1965.
Б8. Бауер (Bauer F. L.), Optimally scaled matrices, *Numer. Math.*, **5** (1963), 73—87.
Б9. Бергер, Скотт, Стренг (Berger A., Scott R., Strang G.), Approximate boundary conditions in the finite element method, Sympos. Math. Conf., 1971—1972, v. 10, London — N. Y., 1972, pp. 295—313, (См. *РЖМ*, 1973, 9Б907.)
Б10. Биркгоф (Birkhoff G.), Piecewise bicubic interpolation and approximation in polygons, Approximation with Special Emphasis on Spline Functions, ed. I. Schoenberg, Academic Press, New York, 1969, 185—221.
Б11. Биркгоф (Birkhoff G.), Numerical solutions of elliptic equations, SIAM Regional Conference Series, Vol. 1, 1971.
Б12. Биркгоф (Birkhoff G.), Angular singularities of elliptic problems, *J. Approx. Th.*, **6** (1972), 215—230.
Б13. Биркгоф, де Бур (Birkhoff G., de Boor C.), Error bounds for spline interpolation, *J. Math. Mech.*, **13** (1964), 827—836.
Б14. Биркгоф, де Бур (Birkhoff G., de Boor C.), Piecewise polynomial interpolation and approximation, Approximation of Functions, ed. H. L. Garabedian, Elsevier, Amsterdam, 1965.
Б15. Биркгоф, де Бур, Шварц, Вендрофф (Birkhoff G., de Boor C., Swartz B., Wendroff B.), Rayleigh — Ritz approximation by piecewise cubic polynomials, *SIAM J. Num. Anal.*, **13** (1966), 188—203.
Б16. Биркгоф, Фикс (Birkhoff G., Fix G.), Rayleigh — Ritz approximation by trigonometric polynomials, *Indian J. Math.*, **9** (1967), 269—277.
Б17. Биркгоф, Фикс (Birkhoff G., Fix G.), Accurate eigenvalue computations for elliptic problems, Duke University SIAM-AMS Symposium, 1970.
Б18. Биркгоф, Шульц, Варга (Birkhoff G., Schultz M. H., Varga R.), Piecewise Hermite interpolation in one and two variables with applications to partial differential equations, *Numer. Math.*, **11** (1968), 232—256.
Б19. Блер (Blair J. J.), Bounds for the change in the solutions of second order elliptic PDEs when the boundary is perturbed, *SIAM J. Appl. Math.*, **24**, № 3 (1973), 277—285. (См. *РЖМ*, 1973, 11Б 260.)
Б20. Брамбл (Bramble J. H.), Variational methods for the numerical solution of elliptic problems, Lecture notes, Chalmers Institute of Technology, Göteborg, Sweden, 1971.
Б21. Брамбл, Гилберт (Bramble J. H., Hilbert S. R.), Estimation of linear functionals on Sobolev spaces with application to Fourier transforms and spline interpolation, *SIAM J. Num. Anal.*, **7** (1970), 113—124.

Б22. Брамбл, Гилберт (Bramble J. H., Hilbert S. R.), Bounds for a class of linear functionals with applications to Hermite interpolation, *Numer. Math.*, 16 (1971), 362—369.
Б23. Брамбл, Дюпон, Томе (Bramble J. H., Dupont T., Thomée V.), Projection methods for Dirichlet's problem in approximating polygonal domains with boundary value corrections, MRC Tech. Rept. 1213, Univ. Wisconsin, 1972.
Б24. Брамбл, Зламал (Bramble J. H., Zlamal M.), Triangular elements in the finite element method, *Math. Comp.*, 24 (1970), 809—821.
Б25. Брамбл, Осборн (Bramble J. H., Osborn J.), Rate of convergence estimates for nonselfadjoint eigenvalue approximations, MRC Tech. Rept. 1232, Univ. Wisconsin, 1972.
Б26. Брамбл, Шатц (Bramble J. H., Schatz A. H.), Rayleigh — Ritz — Galerkin methods for Dirichlet's problem using subspaces without boundary conditions, *Comm. Pure Appl. Math.*, 23 (1970), 653—675.
Б27. Брамбл, Шатц (Bramble J. H., Schatz A. H.), On the numerical solution of elliptic boundary value problems by least square approximation of the data, *SYNSPADE*, 1971, 107—133.
Б28. де Бур (de Boor C.), On local spline approximation by moments, *J. Math. Mech.*, 17 (1968), 729—736.
Б29. де Бур (de Boor C.), The method of projections as applied to the numerical solution of two-point boundary value problems using cubic splines, Thesis, University of Michigan, 1968.
Б30. де Бур, Фикс (de Boor C., Fix G.), Spline approximation by quasi-interpolants, *J. Approx. Theory*, 8, № 1 (1973), 19—45. (См. *РЖМ*, 1974, 2Б 1132.)
Б31. де Бур, Шварц (de Boor C., Swartz B.), Collocation at Gaussian points, Los Alamos Rept. 65—72, 1972.
Б32. Бэклунд (Bäcklund J.), Mixed finite element analysis of elastic and elastoplastic plates in bending, Chalmers Institute of Technology, Göteborg, Sweden, 1971.
В1. Вайникко Г. М., Асимптотические оцеики погрешности проекционных методов в проблеме собствеииых зиачений, *ЖВМ и МФ*, 4, № 3 (1964).
В2. Вайникко Г. М., О быстроте сходимости приближенных методов в проблеме собствеиных значений, *ЖВМ и МФ*, 7, № 5 (1967).
В3. Варга (Varga R. S.), Hermite interpolation and Ritz-type methods for two-point boundary value problems, in Numberical Solutions of Partial Differential Equations, ed. J. H. Bramble, Academic Press, New York, 1965.
В4. Варга Р., Фуикциональиый анализ и теория аппроксимаций в числением анализе, изд-во «Мир», М., 1974.
В5. де Вебек (Fraeijs de Veubeke B.), Displacement and equilibrium models in the finite element method, Chap. 9 of Stress Analysis, eds. O. S. Zienkiewicz and G. S. Holister, Wiley, New York, 1965.
В6. де Вебек (Fraeijs de Veubeke B.), A conforming finite element for plate bending, *Int. J. Solids Structures*, 4 (1968), 96—108.
В7. Вейнбергер (Weinberger H. F.), Variational Methods in boundary value problems, University of Minnesota, 1961.
В8. Видлуид (Widlund O. B.), Some recent applications of asymptotic error expiansions to finite-difference schemes, *Proc. Roy. Soc. Lond.*, A323 (1971), 167—177.
В9. Визер (Visser W.), A refined mixed-type plate bending element, *AIAA J.*, 7 (1969), 1801—1803. (Русский перевод: *Ракетная техника и космонавтика*, 1969.)
В10. Вильсон, Тейлор, Доерти, Габусси (Wilson E. L., Taylor R. L., Doherty W. P., Ghaboussi J.), Incompatible displacement models, University of Illinois Symposium, 1971.

Г1 Галлагер, Дхалла (Gallagher R. H., Dhalla A. K.), Direct flexibility finite element elastoplastic analysis, Berlin Symposium, 1971 (см. также статьи Галлагера в Vanderbilt and Japan-U. S. Symposia).
Г2 Гербольд, Шульц, Варга (Herbold R. J., Schultz M. H., Varga R. S.), Quadrature schemes for the numerical solution of boundary value problems by variational techniques, *Aequ. Math.*, **3** (1969), 96—119.
Г3 Герман (Hermann L. R.), Finite-element bending analysis for plates, *J. Eng. Mech. Div. ASCE*, **94** (1967), 13—25.
Г4 Гоэль (Goël J.-J.), Construction of basic functions for numerical utilization of Ritz's method, *Numer. Math.*, **12** (1968), 435—447.
Г5 ди Гульельмино (di Guglielmino F.), Construction d'approximations des espaces de Sobolev sur des réseaux en simplexes, *Calcolo*, **6** (1969), 279—331.
Д1 Деклу (Descloux J.), On the numerical integration of the heat equation, *Numer. Math.*, **15** (1970), 371—381.
Д2 Деклу (Descloux J.), On finite element matrices, *SIAM J. Num. Anal.*, **9** (1972), 260—265.
Д3 Демьянович Ю. К., Метод сеток для некоторых задач математической физики, *ДАН СССР*, **159**, № 2, 1964.
Д4 Демьянович Ю. К., Об аппроксимации и сходимости метода сеток в эллиптических задачах, *ДАН СССР*, **170**, № 1, 1966.
Д5 Денди (Dendy J.), Thesis, Rice University, 1971.
Д6 Джордж (George A.), Computer implementation of the finite element method, Thesis, Stanford University, 1971.
Д7 Джордж (George A.), Block elimination of finite element systems of equations (рукопись), 1971.
Д8 Дуглас, Дюпон (Douglas J., Dupont T.), Galerkin methods for parabolic problems, *SIAM J. Numer. Anal.*, **4** (1970), 575—626.
Д9 Дуглас, Дюпон (Douglas J., Dupont T.), A finite element collocation method for quasilinear parabolic equations (рукопись), 1972.
Д10 Дуглас, Дюпон (Douglas J., Dupont T.), неопубликованная рукопись по интерполяции коэффициентов, 1972.
Д11 Дуглас, Дюпон (Douglas J., Dupont T.), Galerkin methods for parabolic equations with nonlinear boundary conditions, *Numer. Math.*, **20** (1973), 213—237.
Д12 Дюпон (Dupont T.), Galerkin methods for first-order hyperbolics: an example, *SIAM J Numer. Anal.*, **10**, № 5 (1973), 890—899.
Д13 Дюпюи, Гоэль (Dupuis G., Goël J. J.), Eléments finis raffinés en élasticité bidimensionelle, *ZAMP*, **20** (1969), 858—881.
Д14 Дюпюи, Гоэль (Dupuis G., Goël J. J.), A curved element for thin elastic shells, Tech. Rept., Brown Univ., 1969.
Д15 Дюпюи, Гоэль (Dupuis G., Goël J. J.), Finite element with high degree of regularity, *Int. J. Num. Meth. Eng.*, **2** (1970), 563—577.
Ж1 Женишек (Ženišek A.), Polynomial approximations on tetrahedrons in the finite element method (рукопись), 1970.
Ж2 Женишек (Ženišek A.), Higher degree tetrahedral finite elements (рукопись), 1970.
Ж3 Женишек (Ženišek A.), Interpolation polynomials on the triangle, *Numer. Math.*, **15** (1970), 283—296.
З1 Зенкевич (Zienkiewicz O. C.), Isoparametric and allied numerically integrated elements — a review, University of Illinois Symposium, 1971.
З2 Зенкевич, Айронс и др. (Zienkiewicz O. C., Irons B. M. et al.), Isoparametric and associated element families for two and three-dimensional analysis, в сб. под ред. Холанда и Белла [23].
З3 Зенкевич, Тейлор, Ту (Zienkiewicz O. C., Taylor R. L., Too J. M.), Reduced integration technique in general analysis of plates and shells, *Int. J. Num. Meth. Eng.*, **3** (1971), 275—290.

34. Зламал (Zlamal M.), On the finite element method, *Numer. Math.*, **12** (1968), 394—409.
35. Зламал (Zlamal M.), A finite element procedure of the second order of accuracy, *Numer. Math.*, **14** (1970), 394—402.
36. Зламал (Zlamal M.), Curved elements in the finite element method I, *SIAM J. Num. Anal.*, **5**, № 3 (1973), 367—373.
37. Зламал (Zlamal M.), The finite element method in domains with curved boundaries, *Int. J. Num. Meth. Eng.*, в печати, 1972.
И1. Ирвин (Irwin G. R.), Fracture mechanics, Symposium on Naval Structural Mechanics, 1960.
К1. Карлсон, Холл (Carlson R. E., Hall C. A.), Ritz approximations to two-dimensional boundary value problems, *Numer. Math.*, **18** (1971), 171—181.
К2. Келлог (Kellogg B.), On the Poisson equation with intersecting interfaces, Tech. Note BN-643, Univ. Maryland, 1970.
К3. Келлог (Kellogg B.), Singularities in interface problems, *SYNSPADE*, (1971), 351—400.
К4. Клаф (Clough R. W.), Comparison of three-dimensional elements, Vanderbilt Symposium, 1969.
К5. Клаф, Точер (Clough R. W., Tocher J. L.), Finite element stiffness matrices for analysis of plates in bending, Wright-Patterson I., 1965.
К6. Клаф, Фелиппа (Clough R. W., Felippa C. A.), A refined quadrilateral element for analysis of plate bending, Wright-Patterson II, 1968.
К7. Крайсс (Kreiss H. O.), Difference approximations for ordinary differential equations, Computer Science Department, Uppsala University, 1971.
К8. Красносельский М. А., Сходимость метода Галёркина для нелинейных уравнений, *ДАН СССР*, **73**, № 6 (1950).
К9. Краточвил, Женишек, Зламал (Kratochvil J., Ženišek A., Zlamal M.), A simple algorithm for the stiffness matrix of triangular plate bending finite elements, *Int. J. Num. Meth. Eng.*, **3** (1971), 553—563.
К10. Кондратьев В. А., Краевые задачи для эллиптических уравнений в областях с коническими или угловыми точками, Труды Московского Матем. общ-ва, т. 16, 1967.
К11. Кукал (Koukal S.), Piecewise polynomial interpolations and their applications to partial differential equations, Czech. Sbornik VAAZ, Brno, 1970, 29—38.
К12. Купер (Cowper G. R.), CURSHL: a high-precision finite element for shells of arbitrary shape, National Research Council of Canada Report, 1972.
К13. Купер (Cowper G. R.), Gaussian quadrature formulas for triangles (рукопись), 1972.
К14. Купер, Коско, Линдберг, Олсон (Cowper G. R., Kosko E., Lindberg G. M., Olson M. D.), Static and dynamic applications of a high-precision triangular plate bending element, *AIAA J.*, **7** (1969), 1957—1965. (Русский перевод: *Ракетная техника и космонавтика*, 1969.)
К15. Курант (Courant R.), Variational methods for the solution of problems of equilibrium and vibrations, *Bull. Amer. Math. Soc.*, **49** (1943), 1—23.
Л1. Лаасонен (Laasonen P.), On the discretization error of the Dirichlet problems in a plane region with corners, *Ann. Acad. Scient. Fennicae*, **408** (1967), 3—15.
Л2. Леман (Lehman R. S.), Developments near an analytic corner of solutions of elliptic partial differential equations, *J. Math. Mech.*, **8** (1959), 727—760.
Л3. Линдберг, Олсон (Lindberg G. M., Olson M. D.), Convergence studies of eigenvalue solutions using two finite plate bending elements, *Int. J. Num. Meth. Eng.*, **2** (1970), 99—116.

M1. Маккарти, Стренг (McCarthy C., Strang G.), Optimal conditioning of matrices, *SIAM J. Num. Anal.*, **10**, N 2 (1973), 370—388.
M2. Маклей (McLay R. W.), Completeness and convergence properties of finite element displacement functions — a general treatment, *AIAA Paper 67—143, 5th Aerospace Science Meeting*, 1968.
M3. Маклей (McLay R. W.). On certain approximations in the finite-element method, *Trans. ASME*, 1971, 58—61. (Русский перевод: *Труды Американского общества инженеров-механиков, Прикладная математика*, сер. E, **38** (1971).)
M4. Маркал (Marcal P. V.), Finite element analysis with nonlinearities — theory and practice, First Japan-U. S. Seminar, 1971.
M5. Мартин (Martin H. C.), Finite elements and the analysis of geometrically nonlinear problems, First Japan-U. S. Seminar, 1971.
M6. Мелош (Melosh R. J.) (1966), Basis for derivation of matrices for the direct stiffness method, *AIAA J.*, **34** (1966), 153—170. (Русский перевод: *Ракетная техника и космонавтика*, 1966.)
M7. Миллер (Miller C.), Thesis, M. I. T., 1971.
M8. Митчел, Филипс и Уочпресс (Mitchell A. R., Phillips G., Wachpress R.), Forbidden shapes in the finite element method, *J. Inst. Math. Appl.*, 8 (1971).
M9. Михлин С. Г., Об устойчивости метода Ритца, *ДАН СССР*, **135**, № 1 (1960).
M10. Морли (Morley L. S. D.), A modification of the Rayleigh — Ritz method for stress concentration problems in elastostatics, *J. Mech. Phys. Solids*, **17** (1969), 73—82.
H1. Нитше (Nitsche J.), Ein Kriterium fur die Quasi-Optimalität des Ritzschen Verfahrens, *Numer. Math.*, **11** (1968), 346—348.
H2. Нитше (Nitsche J.), Bemerkungen zur Approximationsgüte bei projektiven Verfahren, *Math. Zeit.*, **106** (1968), 327—331.
H3. Нитше (Nitsche J.), Über ein Variationsprinzip zur Lösung von Dirichlet Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind, Abh. Math. Sem. Univ. Hamburg, 36, 1970.
H4. Нитше (Nitsche J.), Lineare Spline-Funktionen und die Methoden von Ritz für elliptische Randwertprobleme, *Arch. Rat. Mech. Anal.*, **36** (1970), 348—355.
H5. Нитше (Nitsche J.) (1971), A projection method for Dirichlet problems using subspaces with almost zero boundary conditions (рукопись), 1971.
H6. Нитше, Шатц (Nitsche J., Schatz A.), On local approximation of L_2-projections on spline-subspaces, *Appl. Anal.*, **2** (1972), 161—168.
O1. Обэн (Aubin J. P.), Approximation des espaces de distributions et des opérateurs différentiels, *Bull. Soc. Math. France*, **12** (1967).
O2. Обэн (Aubin J. P.), Evaluation des erreurs de troncature des approximations des espaces de Sobolev, *J. Math. Anal. Appl.*, **21** (1968), 356—368.
O3. Обэн, Бушар (Aubin J. P., Burchard H.), Some aspects of the method of the hypercircle applied to elliptic variational problems, *SYNSPADE*, 1971, 1—67.
O4. Оганесян Л. А., Сходимость вариационно-разностных схем при улучшенной аппроксимации границы, *ДАН СССР*, **170**, № 1, 1966.
O5. Оганесян Л. А., Руховец Л. А., Исследование скорости сходимости вариационно-разностных схем для эллиптических уравнений второго порядка в двумерной области с гладкой границей, *ЖВМ и МФ*, **9**, № 5, 1969.
O6. де Оливьера (de Oliveira E. A.), Theoretical foundations of the finite element method, *Int. J. Solids Struct.*, **4** (1968), 929—952.
O7. Олсон, Линдберг (Olson M. D., Lindberg G. M.), Dynamic analysis of shallow shells with a doubly-curved triangular finite element, *J. Sound Vibration*, **19** (1971), 299—318.

П1. Петерс, Уилкинсон (Peters G., Wilkinson J. H.), $Ax = \lambda Bx$ and the generalized eigenproblem, *SIAM J. Num. Anal.*, **7** (1970), 479—492.
П2. Петерс, Уилкинсон (Peters G., Wilkinson J. H.), Eigenvalues of $Ax = \lambda Bx$ with band symmetric A and B, *Comput. J.*, **14** (1971).
П3. Пиан (Pian T. H. H.), Finite element stiffness methods by different variational principles in elasticity, Duke University SIAM-AMS Symposium, 1970, 253—271.
П4. Пиан, Тонг (Pian T. H. H., Tong P.), Basis of finite element methods for solid continua, *Int. J. Num. Meth. Eng.*, **1** (1969), 3—28.
П5. Пиан, Тонг, Лук (Pian T. H. H., Tong P., Luk C. H.), Elastic crack analysis by a finite element hybrid method, Wright-Patterson III, 1971.
П6. Пирс, Варга (Pierce J. G., Varga R. S.), Higher order convergence results for the Rayleigh-Ritz method applied to eigenvalue problems I, *SIAM J. Num. Anal.*, **9** (1972), 137—151.
П7. Пойа (Pólya G.), Sur une interprétation de la méthode des différences finies qui peut fournir des bornes supérieures ou inférieures, *Comptes Rendus*, **235** (1952), 995—997.
П8. Прагер (Prager W.), Variational principles for elastic plates with relaxed continuity requirements, *Int. J. Solids Struct.*, **4** (1968), 837—844.
П9. Прагер, Синж (Prager W., Synge J. L.), Approximations in elasticity based on the concept of function space, *Quart. Appl. Math.*, **5** (1947), 241—269.
П10. Прайс, Варга (Price H. S., Varga R. S.), Error bounds for semidiscrete Galerkin approximations of parabolic problems, Duke University, SIAM-AMS Symposium, 1970, 74—94.
Р1. Рей, Раджаях (Rai A. K., Rajaiah K.), Polygon-circle paradox of simply supported thin plates under uniform pressure, *AIAA J.*, **6** (1967), 155—156. (Русский перевод: *Ракетная техника и космонавтика*, 1967.)
Р2. Рейд (Reid J. K.), On the construction and convergence of a finite-element solution of Laplace's equation, *J. Inst. Maths. Appl.*, **9** (1972), 1—13.
С1. Сандер (Sander G.), Application of the dual analysis principle, *IUTAM Symposium*, 1970.
С2. Сеа (Cea J.), Approximation variationelle des problemes aux limites, *Ann. Inst. Fourier*, **14** (1964), 345—444.
С3. Сиарле, Вагшаль (Ciarlet P. G., Wagschal C.), Multipoint Taylor formulas and applications to the finite element method, *Numer. Math.*, **17** (1971), 84—100.
С4. Сиарле, Равьяр (Ciarlet P. G., Raviart P. A.), General Lagrange and Hermite interpolation in R^n with applications to the finite element method, *Arch. Rat. Mech. Anal.*, **46** (1972), 177—199.
С5. Сиарле, Равьяр (Ciarlet P. G., Raviart P. A.), Interpolation theory over curved elements, with applications to finite element methods, *Comp. Meth. Appl. Mech. Eng.*, **1** (1972), 217—249.
С6. Сиарле, Шульц, Варга (Ciarlet P. G., Schultz M. H., Varga R. S.), Numerical methods of higher order accuracy for nonlinear boundary value problems, *Number. Math.*, **9** (1967), 394—430; *Numer Math.*, **13**, 51—77.
С7. Стреит (Strang G.), The finite element method and approximation theory, *SYNSPADE*, 1971, 547—584.
С8. Стренг (Strang G.), Approximation in the finite element method, *Numer. Math.*, **19** (1972), 81—98.
С9. Стренг (Strang G.), Variational crimes in the finite element method, Maryland Symposium, 1972.
С10. Стренг (Strang G.), Piecewise polynomials and the finite element method, *Bull. Amer. Math. Soc.*, **79** (1973), 1128—1137.
С11. Стренг, Бергер (Strang G., Berger A. E.), The change in solution due to change in domain, Proc. AMS Symposium on Partial Differential Equations, Berkeley, 1971.

С12. Стренг, Фикс (Strang G., Fix G.), A Fourier analysis of the finite element method, Proc. CIME Summer School, Italy, в печати, 1971.
Т1. Тейлор (Taylor R. L.), On completeness of shape functions for finite element analysis, *Int. J. Num. Meth. Eng.*, **4** (1972), 17—22.
Т2. Тёрнер, Клаф, Мартин, Топп (Turner M. J., Clough R. W., Martin H. C., Topp L. J.), Stiffness and deflection analysis of complex structures, *J. Aero Sciences*, **23** (1956).
Т3. Томе (Thomée V.), Elliptic difference operators and Dirichlet's problem, *Diff. Eqns.*, **3** (1964), 301—324.
Т4. Томе (Thomée V.), Polygonal domain approximation in Dirichlet's problem, MRC Tech. Rept. 1188, Univ. Wisconsin, 1971.
Т5. Тонг (Tong P.), Exact solution of certain problems by finite-element method, *AIAA J.*, **7** (1969), 178—180. (Русский перевод: *Ракетная техника и космонавтика*, 1969.)
Т6. Тонг (Tong P.), On the numerical problems of the finite element methods, Waterloo Conference, 1971.
Т7. Тонг, Пиан (Tong P., Pian T. H. H.), The convergence of finite element method in solving linear elastic problems, *Int. J. Solids Struct.*, **3**, 865—879.
Т8. Тонг, Пиан (Tong P., Pian T. H. H.), Bounds to the influence coefficients by the assumed stress method, *Int. J. Solids Struct.*, **6** (1970), 1429—1432.
Т9. Тонг, Пиан, Буччирелли (Tong P., Pian T. H. H., Bucciarelli L. L.), Mode shapes and frequencies by the finite element method using consistent and lumped masses, *J. Comp. Struct.*, **1** (1971), 623—638.
Т10. Треффц (Trefftz E.), Ein Gegenstück zum Ritzschen Verfahren, Second Congress Applied Mechanics, Zurich, 1926.
У1. Уайт, Митчелл (Wait R., Mitchell A. R.), Corner singularities in elliptic problems by finite element methods, *J. Comp. Physics*, 8 (1971), 45—52.
У2. Уильямс (Williams M. L.), Stress singularities resulting from various boundary conditions in angular corners of plates in extension, *J. Appl. Mech.*, 1952, 526—527.
Ф1. Фелиппа (Felippa C. A.), Refined finite element analysis of linear and nonlinear two-dimensional structures, Rept., Univ. California at Berkeley, 1966.
Ф2. Фелиппа (Felippa C. A.), Analysis of plate bending problems by the finite element method, SESM Rept. (Dept. Civil Eng.), Univ. California at Berkeley, 1969.
Ф3. Фелиппа, Клаф (Felippa C. A., Clough R. W.), The finite element method in solid mechanics, Duke University SIAM-AMS Symposium, 1970, 210—252.
Ф4. Фикс (Fix G.), Orders of convergence of the Rayleigh — Ritz and Weinstein — Bazley methods, *Proc. Nat. Acad.*, **61** (1968), 1219—1223.
Ф5. Фикс (Fix G.), Higher-order Rayleigh — Ritz approximations, *J. Math. Mech.*, **18** (1969), 645—658.
Ф6. Фикс, Гулати (Eix G., Gulati S.), Computational problems arising from the use of singular functions, Rept., Harvard Univ., 1971.
Ф7. Фикс, Гулати, Вакофф (Fix G., Gulati S., Wakoff G. I.), On the use of singular functions with the finite element method, *J. Comp. Physics*, to appear, 1972.
Ф8. Фикс, Назиф (Fix G., Nassif N.), On finite element approximations to timedependent problems, *Numer. Math.*, **19** (1972), 127—135.
Ф9. Фикс, Стренг (Fix G., Strang G.), Fourier analysis of the finite element method in Ritz — Galerkin theory, *Stud. Appl. Math.*, **48** (1969), 265—273.
Ф10. Филлипс, Филлипс (Phillips Z., Phillips D. V.), An automatic generation scheme for plane and curved surfaces by isoparametric coordinates, *Int. J. Num. Meth. Eng.*, 3 (1971), 519—528.

Ф11. Фокс, Генричи, Молер (Fox L., Henrici P., Moler C.), Aproximation and bounds for eigenvalues of elliptic operators, *SIAM J. Numer. Anal.*, **4** (1967), 89—102.
Ф12. Фредериксон (Frederickson P. O.), Generalized triangular splines, Math. Rept. 7, Lakehead Univ., Canada, 1971.
Ф13. Фрид (Fried I.), Condition of finite element matrices generated from nonuniform meshes, *AIAA J.*, **10** (1971), 219—221. (Русский перевод: *Ракетная техника и космонавтика*, 1971.)
Ф14. Фрид (Fried I.), Accuracy of finite element eigenproblems, *J. Sound Vibration*, 18 (1971), 289—295.
Ф15. Фрид (Fried I.), Basic computational problems in the finite element analysis of shells, *Int. J. Solids Struct.*, **7** (1971), 1705—1715.
Ф16. Фрид (Fried I.), Discretization and computational errors in high-order finite elements, *AIAA J.*, **9** (1971), 2071—2073. (Русский перевод: *Ракетная техника и космонавтика*, 1971.)
Ф17. Фрид (Fried I.), The l_2 and l_∞ condition numbers..., Conference at Brunel University, 1972.
Ф18. Фридрихс (Friedrichs K.), Die Randwert — und Eigenwertprobleme aus der Theorie der elastischen Platten, *Math. Ann.*, **98** (1928), 205—247.
Ф19. Фридрихс, Келлер (Friedrichs K. O., Keller H. B.), A finite difference scheme for generalized Neumann problems, в сб. «Numerical Solutions of Partial Differential Equations», под ред. Bramble J., Academic Press, New York, 1966.
Х1. Халм (Hulme B. L.), Interpolation by Ritz approximation, *J. Math. Mech.*, 18 (1968), 337—342.
Х2. Ханна, Смит (Hanna M. S., Smith K. T.), Some remarks on the Dirichlet problem in piecewise smooth domains, *Comm. Pure Appl. Math.*, **20** (1967), 575—593.
Х3. Харрик И. Ю., О приближении функций, обращающихся в нуль на границе области, функциями особого вида, *Матем. сб.*, **37**, № 2, 1955.
Х4. Хилтон, Хатчинсон (Hilton P. D., Hutchinson J.), Plastic intensity factors for cracked plates, *Eng. Fract.*, **3** (1971), 435—451.
Ч1. Чернука, Купер, Линдберг, Олсон (Chernuka M. W., Cowper G. R., Lindberg G. M., Olson M. D.), Finite element analysis of plates with curved edges, *Int. J. Num. Meth. Eng.*, **4** (1972), 49—65.
Ш1. Шварц, Вендрофф (Swartz B., Wendroff B.), Generalized finite difference schemes, *Math. Comp.*, **23** (1969), 37—50.
Ш2. Шёнберг (Schoenberg I. J.), Contributions to the problem of approximation of equidistant data by analytic functions, *Quart. Appl. Math.*, **4** (1946), 45—99, 112—141.
Ш3. Ван дер Шлюс (Van der Sluis A.), Condition, equilibration, and pivoting in linear algebraic systems, *Numer. Math.*, **15** (1970), 74—86.
Ш4. Шульц (Schultz M. N.), Rayleigh — Ritz methods for multidimensional problems, *SIAM J. Num. Anal.*, **6** (1969), 523—528.
Ш5. Шульц (Schultz M. H.), L^2 error bounds for the Rayleigh — Galerkin method, *SIAM J. Num. Anal.*, **8** (1971), 737—748.
Э1. Эргатодис, Айронс, Зенкевич (Ergatoudis I., Irons B., Zienkeiwicz O. C.), Curved isoparametric quadrilateral elements for finite element analysis, *Int. J. Solids Struct.*, **4** (1968), 31—42.
Я1. Ямамото, Токуда (Yamamoto Y., Tokuda N.), A note on convergence of finite element solutions, *Int. J. Num. Meth. Eng.*, **3** (1971), 485—493.

ДОПОЛНИТЕЛЬНАЯ ЛИТЕРАТУРА

1′ Бас (Bathe K.-J.), Solution methods for large generalized eigenvalue problems in structural engineering, Thesis, Univ. of Calif., Berkeley, 1971.

2′ Бас, Вильсон (Bathe K.-J., Wilson E. L.), Stability and accuracy analysis of direct integration methods, to appear, 1972.

3′ Бирман М. Ш., Соломяк М. З., Кусочно полиномиальные приближения функций классов W_p^α, *Матем. сб.*, **73 (115)**, № 3 (1967).

4′ Вандерграфт (Vandergraft J. S.), Generalized Rayleigh methods with applications to finding eigenvalues of large matrices, *Lin. Alg. and Applics.*, **4** (1971), 353—368.

5′ Вивер (Weaver W., Jr.), The eigenvalue problem for banded matrices, *Computers and Structures*, **1** (1971), 651—664.

6′ Голуб, Ундервуд, Уилкинсон (Golub G. H., Underwood R., Wilkinson J. H.), The Lanczos algorithm for the symmetric $Ax = \lambda Bx$ problem, в печати, 1972.

7′ Джером (Jerome J. W.), Topics multivariate approximation theory, Symposium on Approximation at Austin, Texas, 1973.

8′ Джонсон (Johnson C.), On the convergence of a mixed finite-element method for plate bending problems, *Numer. Math.*, в печати, 1972.

9′ Джонсон (Johnson C.), Convergence of another mixed finite-element method for plate bending problems, не опубликовано, 1972.

10′ Жиро (Girault V.), A finite difference method on irregular networks, *SIAM J. Numer. Anal.*, в печати, 1972.

11′ Клаф, Бас (Clough R. W., Bathe K.-J.), Finite element analysis of dynamic response, Second U. S.-Japan Seminar, (1972).

12′ Корнеев В. Г., О построении вариационно-разностных схем высокого порядка точности, *Вестник ЛГУ*, сер. матем., мех. и астр., **19** (1970).

13′ Маклеод, Митчелл (McLeod R., Mitchell A. R.), The construction of basis functions for curved elements in the finite element method, *J. Inst. Maths. Applics.*, **10** (1972), 382—393.

14′ Моут (Mote C. D.), Global-local finite element, *Int. J. Numer. Meth. Eng.*, **3** (1971), 565—574.

15′ Пиан, Тонг (Pian T. H., Tong P.), Finite element methods in continuum mechanics, *Adv. Appl. Mech.*, **12** (1972), 1—58.

16′ Финлейсон (Finlayson T.), The method of weighted residuals and variational principles, Academic Press, New York, 1972.

17′ Фрид, Ян (Fried I., Yang S. K.), Best finite elements distribution around a singularity, *AIAA J.*, **10** (1972), 1244—1246. (Русский перевод: *Ракетная техника и космонавтика*, 1972.)

18′ Фрид, Ян (Fried I., Yang S. K.), Triangular, 9 degrees of freedom, C^0 plate bending element of quadratic accuracy, *Q. Appl. Math.*, в печати, 1972.

19′ Фуджи (Fujii H.), Finite element schemes: stability and convergence, Second U. S.-Japan Seminar, 1972.

20′ Conference on Numerical Analysis, Royal Irish Academy, Dublin, 1972.

21′ Applications of the finite element method in geotechnical engineering, U. S. Army Engineers Symposium at Vicksburg, Mississippi, 1972.

УКАЗАТЕЛЬ ОБОЗНАЧЕНИЙ

Это нечто большее, чем просто указатель. Мы попытаемся в удобной форме сформулировать идеи, существенные для понимания трех из основных тем книги:

I. Нормы, функциональные пространства и краевые условия.
II. Энергетические скалярные произведения, эллиптичность и проектор Ритца.
III. Пространства конечных элементов и кусочное тестирование.

Определения в I и II более или менее стандартны, в III характерны для предмета конечных элементов.

I. Нормы, функциональные пространства и краевые условия

Норма есть мера величины функции ($\| u \|$) или расстояние между двумя функциями ($\| u - v \|$). Она удовлетворяет условию $\| cu \| = |c| \, \| u \|$ и неравенству треугольника $\| u + v \| \leqslant \| u \| + \| v \|$. Кроме того, в отличие от *полунормы*, норма равна нулю, только когда u — нулевая функция. Следовательно, если все решения u линейной задачи ограничены исходными данными ($\| u \| \leqslant C \| f \|$), то решения *единственны*: $u = 0$, если $f = 0$, или (согласно суперпозиции) $u_1 - u_2 = 0$, если u_1 и u_2 соответствуют одной и той же функции f.

Некоторые известные нормы:

(i) максимальная норма $=$ sup-норма $= L_\infty$-норма $= \sup\limits_{x \in \Omega} | u(x) |$;

(ii) L_2-норма $= \mathscr{H}^0$-норма $= \left(\int\limits_\Omega | u(x) |^2 \, dx_1 \ldots dx_n \right)^{1/2}$.

В дискретном случае, т. е. при рассмотрении векторов $u = (u_1, u_2, \ldots)$ вместо функций $u(x)$, интегралы заменяются соответствующими суммами. Одно из обобщений — это L_p-нормы: показатели 2 и $1/2$ заменяются на p и $1/p$. Неравенство треугольника выполняется (и мы имеем норму) и для $p \geqslant 1$. Эти пространства становятся интересными и ценными при решении нелинейных задач; в линейной теории мы не считаем их существенными.

Второе, очень важное обобщение заключается в том, чтобы при вычислении нормы рассматривать не только значения функции u, но и ее производные (стр. 15, 171). $\mathscr{H}^s$-норма состоит из $\mathscr{H}^0$- (или L_2-)норм всех частных производных

$$D^\alpha u = \left(\frac{\partial}{\partial x_1} \right)^{\alpha_1} \ldots \left(\frac{\partial}{\partial x_n} \right)^{\alpha_n} u \quad \text{порядка} \quad |\alpha| = \alpha_1 + \ldots + \alpha_n \leqslant s;$$

$$\| u \|^2_{\mathscr{H}^s} = \sum_{|\alpha| \leqslant s} \int | D^\alpha u |^2 \, dx_1 \ldots dx_n. \tag{1}$$

Полунорма $|u|_s$ включает лишь члены, порядок которых в точности равен $|\alpha| = s$; она равна нулю, если u — полином степени $s-1$. Квадраты вводятся в уравнение (1) для того, чтобы получить структуру скалярного произведения, т. е. чтобы сделать $\mathscr{H}^s$ гильбертовым пространством (см. II ниже). Важные применения имеют также и *дробные производные* (s — не целое число) (стр. 172, 301), но их определение значительно сложнее [1, 11 15], за исключением случая, когда Ω — все n-пространство и можно использовать преобразование Фурье:

$$\|u\|_s^2 = \int_{-\infty}^{\infty} |\hat{u}(\xi)|^2 (1+|\xi|^2)^s \, d\xi_1 \ldots d\xi_n.$$

Отрицательные нормы (отрицателеи индекс s, но не сама норма!) определяются по двойственности (стр. 28, 92, 198):

$$\|u\|_{-s} = \max_{v \in \mathscr{H}^s} \frac{\left|\int uv\right|}{\|v\|_s}.$$

Функции в этих «соболевских пространствах» в отличие от их норм обычно определяют, исходя из миожества сравнительно простых функций и затем пополняя его (стр. 21). В результате получается *банахово пространство*; оно содержит все предельные точки любой последовательности, для которой $|u_N - u_M| \to 0$ при $N, M \to \infty$.

Пополненное пространство будет зависеть от исходного множества простых функций. Если исходят из множества всех иепрерывных функций $\mathscr{C}^0$, то в максимальной норме это множество содержит все предельные функции. Если исходное множество включает также *кусочно непрерывные* функции, то окончательное пространство L_∞ намного шире (и его труднее описать). Подобная ситуация возникает вновь в связи с краевыми условиями: если на исходное множество всех бесконечно дифференцируемых функций не налагать никаких условий, то его пополнением будет все пространство $\mathscr{H}^s(\Omega)$; это допустимое пространство для *задачи Неймана* без условий на функции на границе. Если же требуется, чтобы каждая фуикция исходного множества обращалась в нуль на полосе около границы Γ (полоса для одной фуикции может быть меньше, чем для другой), то пополнением *в той же самой норме* будет пространство $\mathscr{H}_0^s$ функций, производные которых порядка менее чем s равны нулю на Γ (стр. 85). Это допустимое пространство для *задачи Дирихле*.

Укажем две группы важных теорем об этих пространствах, правда, вспомогательного характера для данной книги. Одна группа характеризуется *неравенством Соболева* (стр. 92, 170) и отвечает иа вопрос: принадлежат ли производиые порядка s_2 пространству L_{p_2}, если производные порядка s_1 (целого или нет) принадлежат L_{p_1}? Иными словами, какое функциональное пространство содержит другое функциональное пространство? (Соболев: $\mathscr{H}^s$ содержит $\mathscr{C}^q$ тогда и только тогда, когда $s - q > n/2$.) Вторую группу теорем составляют *теоремы о следах*: пусть u — функция в $\mathscr{H}^s$; насколько гладки ее граничные значения, рассматриваемые как функция на Γ? Грубо говоря, эта функция принадлежит $\mathscr{H}^{s-1/2}(\Gamma)$. Поэтому $\mathscr{H}^{m-1/2}$ — подходящее «пространство исходных данных» для неоднородного краевого условия $u = g$; оно соответствует пространству решений $\mathscr{H}_E^m$ для u.

Центральная задача для дифференциальных уравнеиий в частных производных (стр. 14—16) — привести пространство данных в соответствие пространству решений. Такое приведеиие не выполняется автоматически. Например, для уравнения $-\Delta u = f$ в L_∞-норме неверно, что $\|u_{xx}\| + \|u_{yy}\| \leqslant C\|f\|$,

и от этого страдает поточечная теория. $\mathscr{H}^s$-нормы *действительно* хороши для эллиптических задач любого порядка $2m$: $\|u\|_s \leqslant C\|f\|_{s-2m}$.

Для уравиения Эйлера $s = 2m$ берем f из $\mathscr{H}^0$ и ищем функцию u в $\mathscr{H}_B^{2m}$, удовлетворяющую всем m краевым условиям. Для вариационной задачи $s = m$: решение u лежит в допустимом пространстве $\mathscr{H}_E^m$, ограничением только главными краевыми условиями.

II. Энергетические скалярные произведения, эллиптичность и проектор Ритца

Линейные вариационные задачи формулируются в терминах квадратичных функционалов $I(v)$. Классический случай заключается в минимизации *потенциальной энергии* $I(v) = a(v, v) - 2(f, v)$ (здесь мы разделяем члены второго и первого порядков) на пространстве допустимых решений v. Член второго порядка описывает энергию деформации и связан с *энергетическим скалярным произведением*

$$a(v, w) = \frac{1}{4}(a(v + w, v + w) - a(v - w, v - w)). \tag{2}$$

В терминах скалярного произведения условием того, что u минимизирует $I(v)$, служит обращение в нуль первой вариации, т. е. так называемое *уравнение виртуальной работы*:

$$a(u, v) = (f, v) \quad \text{для всех допустимых} \quad v. \tag{3}$$

Интегрирование по частям заменяет эту *слабую форму* (или *форму Галёркина*) задачи дифференциальным уравнением Эйлера для u (порядка $2m$) без функции v и с *условиями скачка*, возникающими в результате интегрирования на любых разрывах.

Замечание. Скалярное произведение бнлинейно:

$$a(u + v, w + z) = a(u, w) + a(v, w) + a(u, z) + a(v, z);$$

это верио лишь в случае, когда энергия деформации имеет благоприятную форму для выполнения этого равенства. Если бы энергия зависела от точки максимума деформации $a(v, v) = \max|\operatorname{grad} v|^2$, то равенство нарушалось бы. Среди L_p-иорм только случай $p=2$ дает такое скалярное произведение (уравнение (2)); только в этом случае мы *имеем гильбертово пространство*. Свойство скалярного произведения распространяется и на пространство $\mathscr{H}^s$, а также на энергии деформации в теории линейной упругости и в других приложениях.

Задачи, не являющиеся самосопряженными, иачинаются непосредственно с решения уравнения виртуальной работы (3), а не с минимизации. Билинейная форма $a(u, v)$ уже не симметрична: $a(u, v) \neq a(v, u)$, и допускаются комплекснозначные функции. Тем не менее если *вещественная часть* формы $a(v, v)$ эллиптична (см. ниже), то результаты теории Галёркина (стр. 144) аналогичны результатам теории Ритца. Они совпадают в случае симметричности $a(u, v)$.

Разрешимость основного вариационного уравнения (3) гарантируется, если форма эллиптична: $\operatorname{Re} a(v, v) \geqslant \sigma\|v\|_m^2$ (также гарантируется разрешимость и соответствующего параболического уравиения). В случае *систем* уравнений с вектором неизвестных $u = (u_1(x), \ldots, u_r(x))$ (типичном для приложений) появляется несколько разновидностей эллиптичности. Одна из них [1, 11, 15] требует, чтобы собственные значения определенных матриц

порядка r имели положительные вещественные части; это слишком мало, чтобы гарантировать успешное применение метода Галёркина на подпространстве. Сильная эллиптичность — условие не на собственные значения, а на сами матрицы; еще более сильным условием будет условие, близкое к рассматриваемому выше для $\mathrm{Re}\, a(v, v)$, оно так же успешно применяется к системам, как и к одному уравнению; здесь v становится допустимым вектором функций. Для краевых условий перечень возможностей детально описан Келлогом в трудах симпозиума в Балтиморе [6]; в приложениях центральным моментом остается выделение главных условий и, следовательно, пространства допустимых функций и требование определенности $\mathrm{Re}\, a(v, v)$ на этом пространстве.

Метод Ритца заключается в минимизации функционала $I(v)$ на последовательности подпространств S^h. Основная теорема (стр. 54) устанавливает, что минимизирующая функция u^h есть *проекция* u на S^h, иными словами, u^h — *ближайшая к* u *функция по норме энергии деформации* $a(v, v)$. Поэтому, если каждое подпространство S^h содержится в следующем (как это предполагается в классическом методе Ритца и обычно выполняется в методе конечных элементов, когда новые элементы строятся в результате разбиения старых), сходимость по норме энергии деформации *монотонна* при $h \to 0$. Такова же и сходимость собственных значений. В теории Ритца это, возможно, и полезно, но не столь существенно: монотонность последовательности подпространств S^h предполагается дополнительно, так что монотонность сходимости — дополнительный вывод.

III. Пространства конечных элементов и кусочное тестирование

Обычное описание конечного элемента задает вид функции формы (пробный полином), расположение, а также параметр (значение функции v или некоторой производной $D_j v$), приписанный каждому узлу. В разд. 1.9 приведено несколько примеров. Их достаточно, чтобы получить представление о том, как подсчитывать матрицы элементов и составлять из них глобальные матрицы жесткости и массы K и M.

В разд. 2.1 по математическим причинам мы предпринимали еще один шаг при описании пространства пробных функций S^h: задавали множество базисных функций $\varphi_1, \ldots, \varphi_N$ пространства. Функция φ_j непосредственно ставилась в соответствие узлу z_j с конкретным узловым параметром $D_j v$. Если функции формы однозначно определяются значениями узловых параметров (как это должно быть!), то существует единственная пробная функция φ_j, для которой $D_j \varphi_j(z_j) = 1$, а все другие узловые параметры $D_i \varphi_j(z_i)$ равны нулю. Эти функции образуют базис, так как любую пробную функцию можно разложить по ее узловым параметрам:

$$v^h = \sum q_j \varphi_j, \quad q_j = D_j v^h(z_j);$$

q_j — весовые коэффициенты. Это немедленно приводит к определению *интерполянта* u_I для любой заданной функции u. Интерполянт предполагает те же узловые параметры, что и u, но внутри каждого элемента он служит одним из пробных полиномов: $u_I = \sum q_j \varphi_j$, где $q_j = D_j u(z_j)$. *Теоремы аппроксимации* из гл. 3 устанавливают, что интерполянт u_I близок к u в нормах, описанных ранее. Ошибка $u - u_I$ зависит от размеров элемента h_i и от степени $k - 1$ полноты функций формы (стр. 163).

Размерность пространства S^h пробных функций равна числу N базисных функций φ_j или свободных параметров q_j. Очевидно, что N зависит от числа элементов. Решающую роль играет число M *параметров в каждой вершине*; оно позволяет сравнивать порядки матриц K при рассмотрении двух конкури-

рующих элементов. Пусть d — число *степеней свободы* (коэффициенты функции формы) внутри каждого элемента; M меньше этого числа и зависит от условий непрерывности между элементами. (Наша гипотеза относительно пространства пробных функций на стр. 105, содержащего полиномы степени $k-1, \mathscr{C}^q$-непрерывные на треугольниках, состоит в том, что $M = (k-1-q)(k-1-2q)$.)

Перейдем к *кусочному тестированию*. До последнего времени оно едва ли было известно (по крайней мере под этим названием) даже специалистам. Появилось оно впервые в приложении к [Б7] в связи с необходимостью объяснить, почему треугольник Зенкевича в одной конфигурации давал сходящийся процесс, а в другой нет (стр. 206). Под своим официальным названием тестирование появилось в кратком комментарии [А6] к более ранней работе. На Симпозиуме в Балтиморе Айронс сделал довольно полный доклад на эту тему, однако достаточность тестирования для сходимости тогда вызывала сомнение.

Мы убеждены, что при разумных предположениях оно достаточно. Тестирование описывается на стр. 205; оно весьма просто при реализации. Напомним эквивалентную формулировку на стр. 209: если энергия деформации включает производные $D^m v$ порядка m, то все интегралы $\iint D^m \varphi_j$ должны вычисляться точно, даже если пренебречь членами на промежуточных границах между элементами в случае несогласованности или применить численное интегрирование. *Кусочное тестирование высокого порядка* требует, чтобы интегралы $\iint P_{n-m} D^m \varphi_j$ были точными для всех полиномов степени $n-m$. Это обобщение сделано Стренгом; оно дает порядок $h^{2(n-m+1)}$ для сходимости по энергии деформации.

Несколько слов о доказательстве. На стр. 211 и 219 задача сводится к оценке ошибки скалярного произведения $a_*(u, v^h) - a(u, v^h)$. (При численном интегрировании рассматриваются и линейные члены, включающие f.). Предположим, что решается модельная задача $-\Delta u = f$, для которой энергетическое скалярное произведение равно $a(u, v) = \iint u_x v_x + u_y v_y$. Успех при кусочном тестировании означает, что для любого линейного полинома P

$$a_*(u, \varphi_j) - a(u, \varphi_j) = a_*(u - P, \varphi_j) - a(u - P, \varphi_j). \tag{4}$$

Выбирая P близким к u на всем множестве E_j, на котором $\varphi_j \not\equiv 0$, и нормализуя базис так, чтобы $a(\varphi_j, \varphi_j) = 1$, видим, что правая часть в (4) меньше, чем $ch \| u \|_{2, E_j}$. Поэтому, если записать $v^h = \sum q_j \varphi_j$, то получим

$$\begin{aligned} |a_*(u, v^h) - a(u, v^h)| &\leqslant ch \sum \| u \|_{2, E_j} |q_j| \leqslant \\ &\leqslant ch \Big(\sum \| u \|^2_{2, E_j}\Big)^{1/2} \Big(\sum q_j^2\Big)^{1/2} \leqslant c' h \| u \|_2 \Big(\sum q_j^2\Big)^{1/2}. \end{aligned} \tag{5}$$

Если бы выполнялось неравенство

$$\sum q_j^2 \leqslant C^2 a_*\Big(\sum q_j \varphi_j, \sum q_j \varphi_j\Big), \tag{6}$$

то сходимость была бы доказана: Δ меньше, чем $c'Ch \| u \|_2$, и, согласно оценкам на стр. 211, ошибка в деформациях имеет порядок $O(h)$.

Результат правильный, но, к сожалению, неравенство (6) не выполняется. Оно равносильно ограниченности числа обусловленности матрицы жесткости K (или равномерной независимости φ_j в энергетической норме), что справедливо для матрицы массы (стр. 247), но не для K. Однако выход есть. В наших

расчетах можно пренебречь любой *согласованной* функцией φ_j, поскольку в этом случае разность в (4) тождественно равна нулю. Поэтому если пробное пространство можно было бы рассматривать как согласованное пространство, *к которому добавляются равномерно независимые несогласованные элементы*, то доказательство проходит. Это было очевидно для случая Вильсона, поскольку он исходил из обычных билинейных элементов и двух дополнительных несогласованных полиномов второй степени внутри каждого квадрата. (Такие внутренние степени свободы называются *безузловыми координатами.*) Так как квадраты никогда не перекрываются, то равномерная независимость выполняется автоматически.

Крузей и Равьяр недавно провели прекрасное исследование несогласованных элементов для соленоидальных течений. Их метод применяется ко всем элементам, выдерживающим тестирование: на каждой грани или поверхности между элементами интеграл от скачка несогласованности равен нулю. Этот метод Сиарле и Ласко формалнзовали и применили к плоским элементам для доказательства сходимости.

Непрерывное увеличение области приложений конечных элементов к уравнениям Навье — Стокса, задачам контроля, предсказанию землетрясений, нелинейным упругости и пластичности в грунтах и в металлах, а также к конструированию танкеров и реакторов обещает «счастливое будущее» как для математиков, так и для инженеров.

ИМЕННОЙ УКАЗАТЕЛЬ [1])

[1]) В скобках заключены номера страниц, на которых при ссылке на работу не указана фамилия ее автора. — *Прим. ред.*

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

ОГЛАВЛЕНИЕ

УВАЖАЕМЫЙ ЧИТАТЕЛЬ!

Ваши замечания о содержании книги, ее оформлении, качестве перевода и другие просим присылать по адресу: 129820, Москва, И-110, ГСП, 1-й Рижский пер., д. 2, издательство «Мир».

ИБ № 93

Г. Стренг, Дж. Фикс

ТЕОРИЯ МЕТОДА
КОНЕЧНЫХ ЭЛЕМЕНТОВ

Редактор Л. В. Штейнпресс
Художник В. И. Шаповалов
Технический редактор В. П. Сизова
Корректоры С. А. Денисова, В. И. Киселева
Сдано в набор 19.08.76.
Подписано к печати 18.01.77.
Вумага тип. № 3 $60\times90^{1}/_{16}=11$ бум. л.
22 печ. л.
Уч.-изд. л. 20,73 Изд. № 1/8755
Цена 1 р. 73 к. Зак. № 287

ИЗДАТЕЛЬСТВО «МИР»
Москва, 1-й Рижский пер., 2

Ордена Трудового Красного Знамени
Ленинградская типография № 2 им. Евгении Соколовой
Союзполиграфпрома при Государственном комитете
Совета Министров СССР
по делам издательств, полиграфии и книжной торговли
198052, Ленинград, Л-52,
Измайловский пр., 29