Р. П. Федоренко

ВВЕДЕНИЕ В ВЫЧИСЛИТЕЛЬНУЮ ФИЗИКУ

Рекомендовано Государственным комитетом Российской Федерации по высшему образованию в качестве учебного пособия для студентов высших учебных заведений, обучающихся по направлениям «Математика», «Физика», специальностям «Математика», «Прикладная математика», «Физика»

Москва
Издательство
Московского физико-технического института
1994

ББК 22.31
Ф33
УДК 519.63 (075.8)

Издание выпущено в счет дотации, выделенной Комитетом РФ по печати

Рецензенты:

кафедра вычислительной математики механико-математического факультета Московского государственного университета им. М. В. Ломоносова (зав. кафедрой академик РАН *Н. С. Бахвалов*),

д. ф.-м. н. *А. В. Забродин*

ФЕДОРЕНКО Р. П. Введение в вычислительную физику: Учеб. пособие: Для вузов. — М.: Изд-во Моск. физ.-техн. ин-та, 1994. — 528 с. ISBN 5-7417-0002-0

Посвящено описанию методов приближенного решения задач математической физики, возникающих в различных областях. Изложение основных понятий и средств численного анализа доводится до описания специальных алгоритмов решения важных прикладных задач, разработка которых продолжается в настоящее время. Приближенные решения сложных задач получаются как общими средствами вычислительной математики, так и специфическими для данного узкого класса задач приемами, которые позволяют обходить существенные трудности в современной вычислительной работе и делают расчеты посильными для ЭВМ.

Для студентов и аспирантов факультетов прикладной математики и физико-технических специальностей вузов с достаточно высоким уровнем преподавания математики, а также для научных работников, специализирующихся в области применения численных методов в научных исследованиях.

Табл. 24. Ил. 66. Библиогр.: 165 назв.

Федеральная целевая программа книгоиздания в России

Ф $\frac{1604030000-006}{1Т4(03)-94}$ Инф. письмо

ISBN 5-7417-002-0

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

Предлагаемая вниманию читателя книга написана на основе двух курсов лекций, в течение ряда лет читавшихся студентам Московского физико-технического института. Им соответствуют две части книги. Первая часть содержит основы вычислительной математики (такой семестровый курс слушают студенты всех факультетов). Вторая часть соответствует годовому курсу вычислительной физики (на факультете общей и прикладной физики).

Почему книга называется «Введение в вычислительную физику», а не «Методы вычислительной математики», например? Это объясняется характером будущей работы слушателей. Для них вычислительная математика в первую очередь будет инструментом научных исследований, а не их предметом. Методы приближенных вычислений излагаются в книге не как самостоятельная научная дисциплина, а как набор средств, позволяющих продвинуться в исследовании тех или иных прикладных проблем физики, химии, аэромеханики и т.п. Это соответствует характеру образования, получаемого в Московском физико-техническом институте, и научному стилю Института прикладной математики им. М. В. Келдыша. Работа автора в этом институте определила его понимание науки, называемой «вычислительная математика», и нашла отражение как в содержании книги, так и в характере изложения.

Не следует думать, что физик-вычислитель обречен лишь на пассивное использование средств, развиваемых математиками. История развития методов приближенных вычислений ясно показывает большую, можно сказать, определяющую, роль решения именно частных прикладных задач, для которых известные методы оказываются неэффективными в силу каких-то специфических особенностей. Наличие важных приложений оправдывает выделение такого (неестественно вырожденного с общематематической точки зрения) класса задач в самостоятельный объект, заслуживающий отдельного углубленного изучения, а привлечение содержательной интуиции и неформализованных знаний той прикладной области, в которой возникла задача, помогает понять ее специфику и разработать эффективный метод решения. Эти

же знания существенно используются для контроля приближенных решений задачи. В этой связи с практикой — сила физика, позволяющая ему часто решающим образом влиять на развитие вычислительной математики. Однако в этом же и его слабость: нередко такой специалист воспринимает свою задачу слишком обособленной, понятной только ему и не имеющей никакого отношения к общематематической теории.

В современных методах приближенных вычислений можно выделить методы, имеющие широкое применение и уже ставшие достоянием математической теории, и методы, развитые для специальных, но важных в приложениях классов задач. Этому делению и соответствуют две части книги. Первая часть по содержанию близка к традиционным курсам численных методов, однако отбор материала, внимание, уделяемое тем или иным вопросом, и характер изложения определяются в первую очередь местом, которое эти вопросы занимали в практике автора и его коллег по Институту прикладной математики. В частности, относительно небольшое место отведено таким сильно развитым разделам, как теория интерполяции и квадратурные формулы, а вычислительные методы общей линейной алгебры совсем не отражены в книге. Это объясняется обилием стандартных программ и руководств, отражающих развитие теории соответствующих разделов вычислительной математики.

В ходе изложения мы не стремимся к максимальной общности и безупречной строгости формулировок. Современный стиль изложения математических результатов требует четкой и полной формулировки всех используемых в доказательстве предположений о свойствах встречающихся функций. Среди них можно выделить две характерные группы. Первая группа условий строго оговаривает свойства общего, типичного характера, отсутствие которых с прикладной точки зрения является исключением, редко встречающимся вырождением. Вторая группа условий выделяет специальный, частный случай, рассмотрение которого оправдано наличием важных и интересных приложений. Именно такие условия мы считаем необходимым выделять, обсуждать и комментировать. Условия первой группы обычно используются «неявно». К ним, в частности, относятся предположения о гладкости функций. Вместо строгого оформления таких условий в тексте часто используется термин «гладкая функция», означающий функцию, ограниченную вместе со своими производными того порядка, который используется в выкладках. При этом предполагается, что функция является гладкой в той части пространства, с которой мы имеем дело при решении задачи (т.е. «там, где нам это нужно»).

Такое отступление от педантичного стиля современной математической литературы представляется соответствующим духу «вычислительной физики». В свое время, прослушав аккуратный университетский курс обыкновенных дифференциальных уравнений, в котором теорема существования была изложена со всеми необходи-

мыми предположениями, автор вынес впечатление, что решения этих уравнений если и существуют, то только в виде редкого счастливого исключения и на очень малом интервале времени, который иногда можно и продолжить. Таков эффект стиля, при котором все ограничения перечисляются «на равных правах» и не комментируются с некой не очень-то понятной и однозначной «прикладной» точки зрения.

Достаточно большое внимание уделяется прикладному комментарию к некоторым теоремам. Этим формируется своеобразное «прикладное мировоззрение» читателя, его будущие взаимоотношения с теоретическими исследованиями. Дело в том, что алгоритм приближенного решения сложной задачи математической физики практически никогда не бывает строго обоснованным в том смысле, который придает этому слову математика. Полезные теоретические результаты, как правило, относятся к выделенным из него идеализированным фрагментам. Использование строгих результатов в практической вычислительной работе — своего рода искусство, в котором результат оправдывает средства. Это характеризует «вычислительную физику» как науку, в значительной мере экспериментальную. Ее взаимоотношения с чистой теорией достаточно сложны и неоднозначны.

Вторая часть книги точнее соответствует содержанию термина «вычислительная физика». В ней собраны описания методов приближенного решения частных задач, имеющих, однако, важные области приложения в современной науке и технике. Каждый параграф посвящен одной из таких задач. Принят следующий способ изложения. Вначале дается замкнутая математическая формулировка задачи, указывается ее «прикладное происхождение». Физическая терминология используется для «оживления» изложения, но никакого физического обоснования постановки задачи не проводится — это дело физика, а не вычислителя. При этом указываются те особые обстоятельства, которые делают задачу нестандартной, требующей разработки специальных вычислительных методов.

Затем описывается метод приближенного решения, оказавшийся достаточно эффективным. Основное внимание уделяется именно тем деталям метода, которые учитывают специфическую нестандартность данной задачи и которым метод обязан своей эффективностью. Попутно обсуждаются те трудности, с которыми сталкиваются при стандартном подходе к задаче (формально не только возможном, но иногда даже строго обоснованном). В некоторых случаях приводятся и обсуждаются характерные численные результаты. Стандартные детали вычислительной методики описываются бегло, а иногда и совсем опускаются.

Материал второй части книги несет двойную нагрузку. Во-первых, описываемые задачи достаточно интересны в приложениях и опыт их успешного решения представляет прямой интерес в связи с задачами именно этого типа. Во-вторых, разработка эффективного

алгоритма частной задачи обычно связана с использованием приемов, имеющих более широкое, выходящее за рамки данной задачи значение. Автор предпочитает знакомить читателя с такими приемами на примерах конкретных задач, в которых они были использованы с большим эффектом. Есть и другой путь — выделить эти приемы как отдельные самостоятельные сюжеты, дать абстрактное описание ситуаций, в которых их применение целесообразно. Подобный способ изложения представляется нам чуждым духу вычислительной физики.

Отметим некоторые технические детали изложения. Текст книги разбит на параграфы, каждый из которых имеет свою нумерацию формул. При ссылке на формулы другого параграфа используется двойной номер (параграфа и формулы). Впрочем, автор стремился свести к минимуму подобные ссылки. В тексте опускаются и библиографические ссылки. Этот недостаток компенсируется библиографическим комментарием, тем более необходимым, что во многих местах излагаются результаты, еще не вошедшие прочно в учебную литературу и часто освещенные лишь в журнальных, а то и ротапринтных публикациях. Курсивом в тексте выделены общеупотребительные термины вычислительной математики.

Оба курса, на основе которых написана эта книга, читались по предложению академика О. М. Белоцерковского, много сделавшего для внедрения компьютерных наук в «систему физтеха». Пользуюсь случаем высказать Олегу Михайловичу свою искреннюю благодарность.

Автор должен отметить и неоценимое влияние, оказанное на него коллегами по Институту прикладной математики им. М. В. Келдыша. Нет возможности упоминать их здесь, автор постарался должным образом отметить их вклад в развитие предмета книги в библиографическом комментарии к списку литературы. Это будет полезно и для будущих историков науки, которым рано или поздно предстоит изучать историю становления отечественной вычислительной математики.

ОСНОВЫ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ

§ 1. Решение систем нелинейных уравнений

В самых различных задачах вычислительной физики часто возникает необходимость решать системы нелинейных уравнений. Такую систему будем записывать кратко в виде

$$f(x) = 0, \tag{1}$$

имея в виду, что x есть точка n-мерного пространства, т.е. $x = \{x_1, x_2, \ldots, x_n\}$, а f — n-мерная вектор-функция, т.е. $f = \{f_1, f_2, \ldots, f_n\}$. Таким образом, (1) есть система n уравнений с n неизвестными:

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_n) &= 0, \\ \ldots\ldots\ldots\ldots & \\ f_n(x_1, x_2, \ldots, x_n) &= 0. \end{aligned} \tag{2}$$

Конечно, когда речь идет о нелинейных уравнениях в общем случае, мы не имеем ни теорем о существовании, ни теорем о единственности решения. Тем не менее, имея дело с системой (1), мы предполагаем, что искомое решение существует. Оно, быть может, не единственно, и методы, которые будут рассмотрены ниже, не имеют целью найти все решения; обычно достаточно будет какого-то одного. Более того, мы предполагаем, что из каких-то содержательных соображений нам известно примерное расположение этого решения, некоторая не очень большая область, в которой оно находится. Таким образом, лучше говорить не о решении системы нелинейных уравнений, а об уточнении имеющегося весьма грубого приближения к некоторому решению. Ниже это будет должным образом конкретизировано.

Метод Ньютона. Основная вычислительная конструкция, применяемая для решения системы (1), по традиции приписывается Ньютону, хотя теоретические исследования этого алгоритма были выполнены лет на сто позже (Фурье, Коши). Основу метода состав-

ляет фундаментальная для вычислительной математики конструкция — метод итераций (последовательных приближений) и линеаризация уравнений.

В методе Ньютона, начиная с некоторого начального приближения x^0, последовательно находятся точки $x^1, x^2, \ldots, x^k, \ldots$ таким образом, что $\lim_{k\to\infty} f(x^k) = 0$, а $\lim x^k = x^*$, где x^* — решение системы (1).

Нужно только иметь в виду, что вычислительную математику интересует не только факт сходимости, но и скорость сходимости. Метод Ньютона особенно ценен тем, что обеспечивает очень высокую, как говорят, «квадратичную» скорость сходимости (точный смысл этого термина выяснится позже, после доказательства соответствующей теоремы).

Рассмотрим стандартный шаг итерационного процесса метода Ньютона. Пусть имеется некоторое уже найденное приближение x^k; следующее приближение x^{k+1} ищем в виде $x^{k+1} = x^k + \delta x$, где δx — малая поправка, уточняющая x^k. Для ее определения выпишем уравнение $f(x^k + \delta x) = 0$. Само по себе оно не проще исходного уравнения (1), но, используя предположение о малости δx, его можно *линеаризовать*, т.е. использовать разложение f по δx с точностью только до членов первого порядка:

$$f(x^k + \delta x) = f(x^k) + f_x(x^k)\,\delta x + O(\|\delta x\|^2).$$

Пренебрегая членами $O(\|\delta x\|^2)$, получаем линеаризованное уравнение для δx:

$$f(x^k) + f_x(x^k)\,\delta x = 0, \tag{3}$$

которое уже решается, и можно выписать его явное решение:

$$\delta x = -f_x^{-1}(x^k)\,f(x^k).$$

Итак, алгоритм метода Ньютона (МН) имеет следующую форму:

1) имеется некоторое уже найденное приближение x;
2) вычисляются вектор $f(x)$ и матрица $f_x(x)$;
3) решается система (линейных) уравнений (3);
4) пересчитывается приближенное решение $x := x + \delta x$.

Далее процесс повторяется циклически до получения достаточно малой величины $\|f(x)\|$.

Прежде чем перейти к теоретическому исследованию, рассмотрим некоторые связанные с методом вопросы.

1. Что такое f_x? Это есть производная вектор-функции по векторному аргументу. Точный смысл f_x определяется первым членом

ряда Тейлора

$$f(x+\delta x) \approx f(x) + f_x(x)\,\delta x.$$

Это — векторная форма записи отрезков ряда Тейлора для всех n функций:

$$f_1(x_1, \ldots, x_n) + \frac{\partial f_1}{\partial x_1}\delta x_1 + \frac{\partial f_1}{\partial x_2}\delta x_2 + \ldots + \frac{\partial f_1}{\partial x_n}\delta x_n,$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$f_n(x_1, \ldots, x_n) + \frac{\partial f_n}{\partial x_1}\delta x_1 + \frac{\partial f_n}{\partial x_2}\delta x_2 + \ldots + \frac{\partial f_n}{\partial x_n}\delta x_n.$$

Отсюда следует расшифровка:

$$f_x = \begin{pmatrix} \dfrac{\partial f_1}{\partial x_1} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \cdots\cdots & & \cdots\cdots \\ \dfrac{\partial f_n}{\partial x_1} & \cdots & \dfrac{\partial f_n}{\partial x_n} \end{pmatrix}.$$

2. В теоретических оценках будем использовать явную форму

$$\delta x = -f_x^{-1} f,$$

хотя практически обычно находят матрицу f_x, формируют систему линейных алгебраических уравнений (3) и решают ее с помощью стандартной программы. Так поступают в том случае, когда размерность задачи n сравнительно невелика. В дальнейшем мы встретимся с ситуациями, когда n настолько велико, что применение стандартных методов решения линейных систем невозможно или по крайней мере нерационально. Обычно в таких ситуациях матрица f_x имеет специальную структуру, и часто удается построить специальные методы решения, существенно более эффективные, чем общие.

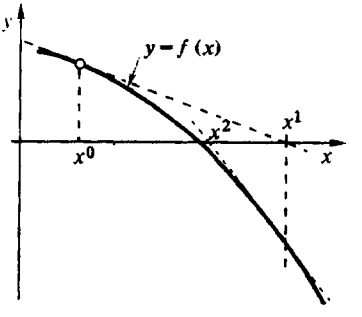


Рис. 1

При $n = 1$ метод Ньютона известен как «метод касательных». Этот термин поясняет рис. 1. Линеаризация состоит в том, что кривая $y = f(x)$ заменяется касательной, проведенной в уже найденной точке x^k, а в качестве следующего приближения x^{k+1} берется пересечение касательной с осью абсцисс. На рис. 1 показано несколько таких итераций и, естественно, возникает предположение о том, что процесс должен быть достаточно эффективным. Это и подтверждается более точным исследованием.

Сходимость метода Ньютона. Докажем теорему о квадратичной сходимости метода.

Теорема 1. Пусть x^* — решение системы (1). Предположим, что в некоторой окрестности x^*:

а) $f(x)$ является гладкой функцией в том смысле, что существуют ее производные до второго порядка и имеет место оценка $\|f_{xx}(x)\| \leqslant C_2$;

б) отображение $x \to f(x)$ равномерно невырождено в том смысле, что $f_x^{-1}(x)$ существует и ограничена: $\|f_x^{-1}(x)\| \leqslant C_1$.

Тогда, если начальное приближение x^0 достаточно близко к x^*, метод Ньютона сходится и имеет квадратичную скорость сходимости.

Точный аналитический смысл выражений «квадратичная скорость сходимости», «x^0 достаточно близко к x^*», «$\|f_{xx}\| \leqslant C_2$» выяснится в процессе доказательства.

Доказательство. Будем следить за эволюцией в процессе итераций величины $\|f(x^k)\|$ (нормы невязки). Установим связь между $\|f(x^{k+1})\|$ и $\|f(x^k)\|$:

$$f(x^{k+1}) = f(x^k) + f_x(x^k)\,\delta x + O(\|\delta x\|^2).$$

Используя выражение $\delta x = -f_x^{-1}(x^k)\,f(x^k)$ и оценку $\|O(\|\delta x\|^2)\| \leqslant C_2\,\|\delta x\|^2$ (именно в этом смысле понимается предположение а) теоремы), получаем

$$f(x^{k+1}) = O(\|\delta x\|^2) = O(\|f_x^{-1}\,f\|^2),$$

откуда

$$\|f(x^{k+1})\| \leqslant C_2\,\|f_x^{-1}\,f\|^2 \leqslant C_2 C_1^2\,\|f(x^k)\|^2.$$

Обозначая $r_k = \|f(x^k)\|$, имеем основное соотношение:

$$r_{k+1} \leqslant C r_k^2, \quad \text{где } C = C_2 C_1^2.$$

Эта оценка порождает следующую цепочку:

$$r_1 \leqslant C r_0^2, \qquad r_2 \leqslant C r_1^2 \leqslant C^3 r_0^4, \qquad r_3 \leqslant C r_2^2 \leqslant C^7 r_0^8.$$

Без труда угадываем общую форму:

$$r_k \leqslant C^{-1}(C r_0)^{2^k}.$$

Именно эту формулу (с показателем 2^k) имеют в виду, когда говорят о квадратичной скорости сходимости.

Теперь можно более точно указать, насколько хорошим должно быть начальное приближение x^0, чтобы процесс заведомо сходился. Очевидно, для этого достаточно выполнения неравенства

$$C\|f(x^0)\| \leqslant q < 1.$$

Можно уточнить и вид области, в которой предполагаются выполненными сформулированные в условиях теоремы оценки производных. Такой областью может быть область, выделенная неравенством $\|f(x)\| < 1/C$. В самом деле, если x^0 лежит в области $C\|f(x^0)\| \leqslant q < 1$, то $\|f(x^1)\| \leqslant q^2/C$ и т.д., т.е. все последующие приближения x^k лежат в этой области.

Условие $\|f_x^{-1}(x)\| \leqslant C_1$ существенно. Оно гарантирует взаимную однозначность (в некоторой области) отображения $x \to f(x)$, что, как известно, очень важно для существования и единственности решения системы $f(x) = 0$.

Модификация метода Ньютона. Метод Ньютона, являясь весьма эффективным средством уточнения сравнительно хорошего начального приближения, может расходиться, если x^0 — слишком грубое приближение к искомому решению. В схему алгоритма были

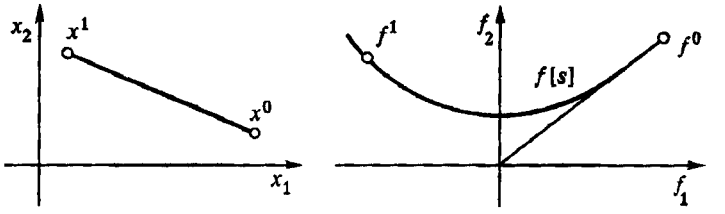


Рис. 2

внесены изменения, имеющие целью ослабить требования к начальному приближению и сделать сходимость не столь зависящей от его выбора. Идею такой модификации поясним, начав с геометрической интерпретации метода Ньютона в двумерном случае ($n = 2$).

Итак, пусть решается система

$$f_1(x_1, x_2) = 0, \qquad f_2(x_1, x_2) = 0.$$

На рис. 2 изображены плоскости (x_1, x_2) и (f_1, f_2). Точка x^0 отображается в точку f^0. В этой точке отображение $x \to f(x)$ линеаризуется, т.е. заменяется отображением

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \to \begin{pmatrix} f_1^0 + \dfrac{\partial f_1}{\partial x_1}(x_1 - x_1^0) + \dfrac{\partial f_1}{\partial x_2}(x_2 - x_2^0) \\ f_2^0 + \dfrac{\partial f_2}{\partial x_1}(x_1 - x_1^0) + \dfrac{\partial f_2}{\partial x_2}(x_2 - x_2^0) \end{pmatrix},$$

и находится точка (x_1^1, x_2^1), в линейном отображении переходящая в точку $(0, 0)$. Однако в нелинейном отображении $x \to f(x)$ точка x^1 отображается не в нуль, а в f^1.

Изучим непрерывное движение от x^0 к x^1 по прямой. Обозначив $\delta x = x^1 - x^0$, рассмотрим отрезок прямой

$$x(s) = x^0 + s\,\delta x, \quad s \geqslant 0, \quad x(1) = x^1$$

(в расчетах обычно берут $x(s) = x^0 + s\,\delta x/\|\delta x\|$). Образ этого отрезка в нелинейном отображении есть кривая $f[s] \triangleq f(x(s))$ (см. рис. 2). В точке $s = 0$ она касается направления на точку $(0, 0)$. В самом деле, при достаточно малых s имеем

$$f[s] = f(x^0 + s\,\delta x) = f^0 + s\,f_x(x^0)\,\delta x + O(s^2).$$

В силу $f_x(x^0)\,\delta x = -f^0$ функция

$$f[s] = f^0 - s\,f^0 + O(s^2) = (1 - s)\,f^0 + O(s^2).$$

Другими словами, при малых s точка $f[s]$ движется почти (с точностью до $O(s^2)$) прямо в начало координат.

По мере увеличения s величины $O(s^2)$ возрастают, они могут стать определяющими и существенно отклонить траекторию $f[s]$ от желаемого движения в начало координат. Теперь очевидно, что нужно двигаться по $[x^0, x^1]$ до тех пор, пока точка $f[s]$ приближается к началу координат, т.е. шаг s^* определяется решением одномерной задачи минимизации. Ищется

$$\min_s \|f(x^0 + s\,\delta x)\|.$$

Точку минимума принято обозначать в виде

$$s^* = \arg\min_s \|f(x^0 + s\,\delta x)\|.$$

Итак, сформируем алгоритм модифицированного метода Ньютона (ММН):

1) имеется некоторая точка x;
2) вычисляются $f(x)$ и $f_x(x)$;
3) находится δx из системы $f + f_x\,\delta x = 0$;
4) определяется функция скалярного аргумента s:

$$F(s) \equiv \|f(x + s\,\delta x)\|;$$

5) находится

$$s^* = \arg\min_s F(s);$$

6) вычисляется следующее приближение:

$$x := x + s^*\,\delta x.$$

В приведенном алгоритме есть элемент, требующий уточнения, — это решение задачи $\min F$. Этой задаче посвящен § 26. Иногда используют совсем простую процедуру дробления шага. Сначала берут значение $s = 1$ (как в стандартном методе Ньютона). Если окажется, что $\|f(x + s\,\delta x)\| < \|f(x)\|$, то этот шаг и остается. В противном случае s заменяют на $s/2$ и снова сравнивают нормы. И т.д. — до получения соотношения $\|f(x + \delta x/2^p)\| < \|f(x)\|$.

Докажем теорему о сходимости модифицированного метода Ньютона.

Теорема 2. Определим область Ω как множество точек, в которых $\|f(x)\| \leqslant \|f(x^0)\|$. Предположим, что:

а) $f(x)$ — гладкая функция и $\|f_{xx}(x)\| \leqslant C_2$, $x \in \Omega$;

б) отображение $x \to f(x)$ равномерно невырождено и $\|f_x^{-1}(x)\| \leqslant C_1$, $\forall\, x \in \Omega$;

в) Ω — ограниченная связная область.

Пусть x^k — точки, последовательно полученные, начиная с x^0, согласно модифицированному методу Ньютона, а r_k — соответствующие невязки ($r_k = \|f(x_k)\|$). Тогда в области Ω существует единственное решение x^* системы уравнений (т.е. $f(x^*) = 0$) и

$$\lim_{k\to\infty} x^k = x^*, \qquad \lim_{k\to\infty} r_k = 0.$$

Доказательство. Отметим очевидный факт: невязки r_k монотонно убывают, т.е. $r_0 \geqslant r_1 \geqslant \ldots \geqslant r_k \geqslant \ldots$ Следовательно, все $x^k \in \Omega$. Оценим величину убывания невязки r_k за один шаг, используя соотношение

$$f(x^k + s\,\delta x) = (1 - s)\,f(x^k) + s^2\,O(\|\delta x\|^2), \qquad \delta x = f_x^{-1} f(x^k).$$

Отсюда следует (при $s \in [0, 1]$)

$$r_{k+1} = \min_s \|f(x^k + s\,\delta x)\| \leqslant \min_{0\leqslant s\leqslant 1} \{(1 - s)\,\|f(x^k)\| + Cs^2 r_k^2\}.$$

Здесь мы оценили

$$\|O(\|\delta x\|^2)\| \leqslant C_2\,\|\delta x\|^2 \leqslant C_2\,\|f_x^{-1}\|^2\,\|f(x^k)\|^2.$$

Таким образом,

$$r_{k+1} \leqslant \min_{0\leqslant s\leqslant 1} \{(1 - s) r_k + Cs^2 r_k^2\}.$$

Вычислим минимум правой части (игнорируя пока ограничение $0 \leqslant s \leqslant 1$). Он достигается в точке $s^* = 1/(2Cr_k)$, а значение

минимума в этом случае есть $r_k - 1/(4C)$. Если $s^* \leqslant 1$, будем использовать эту оценку; если $s^* > 1$, оценим минимум значением в точке $s = 1$. В этом случае $r_{k+1} \leqslant Cr_k^2$. Так как при этом $s^* = 1/(2Cr_k) \geqslant 1$, то $Cr_k^2 < r_k/2$. Итак, в любом случае при переходе от x^k к x^{k+1} невязка убывает не меньше, чем на величину $\min\{1/(4C), r_k/2\}$.

Теперь допустим, что метод не сходится, т.е. $\lim x^k \neq x^*$ и $\lim r_k > 0$. По предположению Ω — ограниченная замкнутая область, т.е. последовательность $\{x^k\}_{k=0}^{\infty}$ имеет в Ω хотя бы одну точку сгущения $\tilde{x}$, причем $\tilde{r} = \|f(\tilde{x})\| \neq 0$. Тогда в силу непрерывности $r(x) = \|f(x)\| > \tilde{r}/2$ в некоторой ε-окрестности $\tilde{x}$. В эту окрестность попадает бесконечное число точек x^k; обозначим их x^{k_i} ($i = 1, 2, 3, \ldots$). Переход от x^{k_i} к x^{k_i+1} сопровождается падением невязки:

$$r_{k_i+1} \leqslant r_{k_i} - \min\{1/(4C), \tilde{r}/4\}.$$

Так как на остальных шагах невязка по меньшей мере не возрастает, получаем явное противоречие. Итак, в каждой точке сгущения $f(\tilde{x}) = 0$. Доказательство закончено.

Отметим важное обстоятельство: в условиях теоремы 2 по сравнению с условиями теоремы 1 отсутствует предположение о достаточной малости r^0 («количественное» предположение). Используются только «качественные» предположения о гладкости и невырожденности (взаимной однозначности) отображения $x \to f(x)$. Эти свойства очень важны для существования и единственности решения системы, которые в условия теоремы не включены. Они следуют из сформулированных предположений. Не вдаваясь в подробности, заметим, что если гладкая функция $\|f(x)\|^2$ в области Ω не обращается в нуль, то она достигает минимума, в котором все ее производные обращаются в нуль. Вычислим их:

$$\frac{\partial}{\partial x_j} \sum_{i=1}^{n} f_i^2(x) = 2 \sum_{i=1}^{n} f_i(x) \frac{\partial f_i(x)}{\partial x_j} = 0.$$

Если не все $f_i(x) = 0$, то $\det(f_x) = 0$, что противоречит одному из предположений.

Наконец, важно отметить, что в тех случаях, когда система $f(x) = 0$ имеет много решений, модифицированный метод Ньютона приводит к одному из них; к какому именно, это зависит от выбора начального приближения. Как говорят, каждое решение имеет свою «область притяжения» — совокупность точек x, стартуя из которых метод Ньютона приводит именно к этому решению.

Методы простых итераций. В некоторых ситуациях применение метода Ньютона может быть затруднено как из-за слишком трудоемкого вычисления матрицы f_x, так и из-за необходимости решать систему линейных уравнений. Поэтому наряду с надежным и эффективным методом Ньютона в вычислениях используются и более простые итерационные методы.

Рассмотрим простой пример, поясняющий суть дела. Решается система двух уравнений

$$f(x, y) = 0, \qquad \varphi(x, y) = 0. \tag{4}$$

Пусть функции f и φ таковы, что из уравнения $f(x, y) = 0$ при заданном y легко определяется x, а из уравнения $\varphi(x, y) = 0$ определяется y. Тогда можно построить итерационный процесс следующего вида. Если известны x^k, y^k, то следующее приближение вычисляется так:

1) из уравнения $f(x, y^k) = 0$ находится x^{k+1};

2) из уравнения $\varphi(x^{k+1}, y) = 0$ находится y^{k+1}; и т.д.

Проанализируем сходимость. Анализ таких процессов проводим в предположении, что x^k, y^k достаточно близки к решению x^*, y^*, т.е. полагаем

$$x^k = x^* + \delta x^k, \qquad y^k = y^* + \delta y^k.$$

Считая δx, δy малыми, линеаризуем уравнения итерационного процесса. Из

$$f(x^* + \delta x^{k+1}, y^* + \delta y^k) = 0, \qquad \varphi(x^* + \delta x^{k+1}, y^* + \delta y^{k+1}) = 0,$$

получаем линейные соотношения

$$f_x\, \delta x^{k+1} + f_y\, \delta y^k = 0, \qquad \varphi_x\, \delta x^{k+1} + \varphi_y\, \delta y^{k+1} = 0.$$

Обозначая $\delta z = (\delta x, \delta y)$, имеем векторное соотношение

$$\delta z^{k+1} = A\, \delta z^k, \quad \text{где } A = -\begin{pmatrix} f_x & 0 \\ \varphi_x & \varphi_y \end{pmatrix}^{-1} \begin{pmatrix} 0 & f_y \\ 0 & 0 \end{pmatrix}.$$

Сходимость обеспечивается при: а) достаточно хорошем начальном приближении x^0; б) при $\|A\| \leqslant q < 1$ (в этом случае $\|\delta z^k\| \leqslant q^k \|\delta z^0\|$).

Заметим, что между схемой простых итераций и методом Ньютона есть принципиальное отличие: сходимость метода Ньютона обеспечивается (при наличии хорошего приближения) чисто качественными факторами — гладкостью f и невырожденностью отображения $x \to f$. Для метода простых итераций требуется еще важное количественное условие: $\|A\| < 1$. При $\|A\| > 1$ метод может расхо-

диться в сколь угодно благополучном случае при сколь угодно хорошем начальном приближении.

Метод простых итераций в действительности объединяет необозримое количество итерационных методов, которые конструируются посвоему в том или ином конкретном случае. Например, можно одни и те же переменные, входящие в решаемое уравнение, брать один раз «с верхней» итерации, другой раз «с нижней». Поясним суть дела простым примером. Пусть имеется функция $F(x, \xi)$, а нужно решить уравнение

$$f(x) \equiv F(x, x) = 0.$$

Тогда можно построить итерационный процесс вида

$$F(x^{k+1}, x^k) = 0,$$

конечно, при условии, что из такого уравнения сравнительно легко находится x^{k+1} при известном x^k.

Анализ сходимости приводит к соотношению

$$F_x\, \delta x^{k+1} + F_\xi\, \delta x^k = 0, \quad \text{или} \quad \delta x^{k+1} = -F_x^{-1} F_\xi\, \delta x^k,$$

и сходимость (в окрестности решения) определяется нормой матрицы $F_x^{-1}F_\xi$: если она меньше единицы, процесс сходится; если она больше единицы, процесс расходится. Очевидно также, что если процесс сходится, т.е. существует $\lim\limits_k x^k = x^*$, то, переходя к пределу в соотношении $F(x^{k+1}, x^k) = 0$, получаем $F(x^*, x^*) = 0$.

Если в уравнении $f(x) = 0$ не удается выделить «разрешаемую» относительно x часть, можно ввести ее искусственно и очень просто, например преобразовав уравнение к виду

$$x - x + \alpha f(x) = 0$$

и построив итерационный процесс

$$x^{k+1} = x^k - \alpha f(x^k).$$

Строятся такие методы, как видим, легко, но сходимость их не гарантируется и является в известном смысле делом случая.

Заметим еще, что существуют теоремы, обосновывающие правомерность пренебрежения членами второго порядка: если метод сходится в теории «первого приближения», т.е. норма соответствующей матрицы $\|A\| \leqslant q < 1$, то при достаточно хорошем начальном приближении метод действительно сходится.

Метод Ньютона в специальных ситуациях. Часто приходится решать уравнение $f(x) = 0$ в специальной ситуации, когда функция задана не формулами, а алгоритмом, и достаточно сложным. Други-

ми словами, имеется программа, которая по заданному значению аргумента x вычисляет (после миллионов операций) значение f. Именно такую ситуацию изучает современный анализ, в котором термин «функция» (по традиции все-таки ассоциирующийся с такими понятиями, как «формула», «аналитическое выражение» и т.п.) вытесняется термином «отображение»: $x \to f$, где «$\to$» есть символ каких-то, быть может, очень сложных операций. В такой ситуации метод Ньютона, требующий использования матрицы f_x, должен быть дополнен алгоритмом ее вычисления.

Наиболее простым способом (естественно, простота покупается большим объемом вычислений) является численное дифференцирование. Пусть $f(x)$ есть $f(x_1, x_2, \ldots, x_n)$, а $h_1, h_2, h_3, \ldots$ — малые числа, «шаги численного дифференцирования». Приближенно можно положить

$$\frac{\partial f}{\partial x_i} = \frac{1}{h_i}[f(x_1, \ldots, x_{i-1}, x_i + h_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_i, \ldots, x_n)].$$

Таким образом, для вычисления всех частных производных нужно n раз вычислить значение f при возмущении поочередно аргументов.

Итак, (приближенное) вычисление частных производных функции n переменных по самой простой формуле (называемой формулой одностороннего дифференцирования) требует $(n+1)$-кратного вычисления функции. Существуют и другие формулы численного дифференцирования. Среди них особенно популярна формула «центральной разности»

$$\frac{\partial f}{\partial x_i} \approx \frac{1}{2h_i}[f(x_1, \ldots, x_i + h, \ldots) - f(x_1, \ldots, x_i - h, \ldots)].$$

Она, очевидно, более трудоемка: вычисление всех производных «стоит» $2n+1$ вычислений f. Естественно ожидать, что эта формула более точна. Вопросы о точности численного дифференцирования обсуждаются в следующем параграфе. Пока заметим лишь, что напрашивающийся ответ «чем меньше h, тем точнее численное дифференцирование» неверен.

Нормировка задачи. Практическое применение метода Ньютона в сложных задачах иногда приводит к очень медленной сходимости. В связи с этим возникает необходимость разбираться в причинах такого противоречия между обещаниями теории и реальными фактами. Правильно поняв причину, можно разработать приемы, существенно ускоряющие процесс решения. Один из них описывается ниже.

Начнем с простого примера. Применим схему модифицированного метода Ньютона для решения несложной системы уравнений

$$f(x, y) \equiv x^5 + y^4 - 2 = 0, \quad \varphi(x, y) \equiv (x-2)^3 + (y-2)^3 + 16 = 0.$$

Начальное приближение: $x^0 = 2.0$, $y^0 = 3.0$.

В табл. 1 представлены величины, подробно показывающие процесс решения задачи. Поясним обозначения: k — номер шага (итерации); x, y — текущие значения искомых величин; f, φ — значе-

Таблица 1

k	x	y	f	φ	r	s	α
0	2.0000	3.0000	111.000	17.000	112.294	0.11	40
1	2.0685	2.9380	110.379	16.826	111.654	0.108	42
2	2.1330	2.8680	109.801	16.656	111.057	0.094	48
4	2.2334	2.7274	108.906	16.398	110.134	0.064	57
5	2.2727	2.6589	108.607	16.306	109.824	0.047	52
7	2.3265	2.5471	108.246	16.199	109.451	0.021	46
8	2.3437	2.5057	108.138	16.170	109.340	0.015	43
10	2.3729	2.4284	108.008	16.130	109.206	0.012	33
12	2.3899	2.3781	107.946	16.113	109.141	0.004	25
14	2.4031	2.3358	107.914	16.130	109.109	0.0017	18
16	2.4108	2.3096	107.898	16.099	109.092	0.0010	15

ния функций в точке (x, y); $r = (f^2 + \varphi^2)^{1/2}$ — невязка; s — шаг спуска, найденный решением задачи одномерной минимизации; α — угол (в градусах) между векторами (f_x, f_y) и (φ_x, φ_y). Этот угол характеризует степень «невырожденности» отображения $(x, y) \to (f, \varphi)$ в данной точке.

Видно, что поиск решения проходит крайне неэффективно, шаг s очень далек от единицы, т.е. линеаризация уравнения работает на расстояниях, существенно меньших расстояния от текущей точки до искомого решения. Эволюция величины α указывает на приближение какой-то точки вырождения отображения. Однако начало процесса проходит очень медленно и при больших значениях α. Основная, видимо, причина — очень малые размеры области, в которой линейное приближение имеет хорошую точность. Точка (2.0, 3.0), однако, ничем не примечательна, и в данном простом примере можно проверить, что линеаризация f и φ достаточно точна на расстояниях, бо́льших смещения x и y за один шаг процесса.

Попробуем разобраться в ситуации. Для этого стоит посмотреть на систему уравнений метода Ньютона. В точке (2.0, 3.0) она имеет вид

$$\begin{pmatrix} f_x & f_y \\ \varphi_x & \varphi_y \end{pmatrix} \begin{pmatrix} \delta x \\ \delta y \end{pmatrix} = -\begin{pmatrix} f \\ \varphi \end{pmatrix} \equiv \begin{pmatrix} 80 & 108 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} \delta x \\ \delta y \end{pmatrix} = -\begin{pmatrix} 111 \\ 17 \end{pmatrix}.$$

(В других точках табл. 1 ситуация примерно та же самая.) Решение этой системы: $\delta x = 6.26$, $\delta y = -5.67$. (Противоречие между изменением x и y и шагом s в табл. 1 объясняется нормировкой направления движения.) Обратим внимание на характерную деталь: направление $(\delta x, \delta y)$ «почти ортогонально» вектору (f_x, f_y):

$$f_x\,\delta x + f_y\,\delta y = 80 \cdot 6.26 - 108 \cdot 5.67 \approx 500 - 611 = -111.$$

(Здесь 111 — действительно малая величина, ведь ее естественно относить к величине $500 + 611 = 1111$, т.е. в «безразмерных» единицах величина 111 мала в том же смысле, в каком 0.1 мала относительно 1.)

Итак, направление $(\delta x, \delta y)$ «почти совпадает» с касательной к линии уровня $f(x, y) = \text{const}$, а вдоль касательной приращение f определяется членами второго порядка, которые алгоритм игнорирует. Почему же алгоритм выбирает такое направление, т.е. его «не интересует» уменьшение величины $f = 111$, он в большей степени заинтересован в уменьшении относительно малой величины $\varphi = 17$? Возникает парадоксальное предположение: видимо, в точках (x^k, y^k) уравнение $f = 0$ уже почти выполнено, а уравнение $\varphi = 0$ — нет. Ведь из того, что $f = 111$, а $\varphi = 17$, еще ничего не следует. Откуда известно, как нужно сравнивать величины f и φ? Подобные вопросы всегда должны возникать перед вычислителем. Они приводят к требованию нормировки задачи.

В самом деле, не меняя существа дела, можно перейти к системе

$$\frac{1}{\varkappa_1} f(x, y) = 0, \qquad \frac{1}{\varkappa_2} \varphi(x, y) = 0.$$

Очевидно, что направление $(\delta x, \delta y)$ инвариантно относительно произвольного выбора «единиц измерения» $\varkappa_1$ и $\varkappa_2$. Но величина невязки $r = [(f/\varkappa_1)^2 + (\varphi/\varkappa_2)^2]^{1/2}$ и, следовательно, шаг s такой инвариантностью не обладают. При этом возникает проблема выбора «правильных» масштабов $\varkappa_1$, $\varkappa_2$. В своей практике автор в подобных ситуациях руководствовался правилом, условно названным «принципом равноправия»: масштабы нужно выбирать такими, чтобы одинаковые изменения x и y приводили к численно близким изменениям f и φ. Формулы для малых приращений f и φ показывают, что эта цель в известной мере будет достигнута (в окрестности данной точки (x, y)), если взять $\varkappa_1 = (f_x^2 + f_y^2)^{1/2}$, $\varkappa_2 = (\varphi_x^2 + \varphi_y^2)^{1/2}$.

Таким образом, мы приходим к модифицированному методу Ньютона с нормировкой. Алгоритм стандартного шага в точке (x, y) дополняется следующим: после вычисления производных и направления $(\delta x, \delta y)$ вычисляются «масштабы» $\varkappa_i$, и шаг s выбирается минимизацией масштабированной невязки. Эффект этого при-

ема иллюстрирует табл. 2. Обозначения в ней — те же, что и в табл. 1, только величины r_0 и r_1 означают величины масштабированной невязки в точке (x^k, y^k) и в следующей точке (x^{k+1}, y^{k+1}). Заметим, что теперь у нас нет единой невязки, которая убывала бы в процессе решения, чем в сущности и обеспечивается сходимость метода. В точке (x^{k+1}, y^{k+1}) есть две невязки: при масштабировании в точке (x^k, y^k) и при масштабировании в точке (x^{k+1}, y^{k+1}). Теорема о сходимости модифицированного метода Ньютона утрачивает силу, но зато сама сходимость стала существенно лучше.

Данные $f = 0$ и $\varphi = 0$ в последней строке табл. 2 означают, что эти величины не больше $5 \cdot 10^{-6}$. На основании вышесказанного естественно возникает вопрос: действительно ли это малая величина? Ответ на него несложен. Если вычислитель постулировал, что для

Таблица 2

k	x	y	f	φ	r_0	r_1	s
0	2.0000	3.0000	111.000	17.000	5.72	5.52	0.066
1	2.414	2.626	127.500	16.32	12.75	11.59	0.125
2	3.209	0.542	338.4	14.66	2.00	0.55	0.50
3	2.891	−0.389	200.0	3.07	0.60	0.0826	1.90
4	1.806	−0.578	17.3	−1.14	0.331	0.026	2.00
5	1.157	−0.460	0.115	0.516	0.031	0.00054	1.0
6	1.14265	−0.48663	0.004	−0.006	0.0006	0	1.0
7	1.14220	−0.48626	0	0			

x и y величина 10^{-5} является малой, то естественно считать малыми для f и φ изменения, порождаемые такими малыми δx и δy. Это опять-таки приводит нас к тому масштабированию, которое было использовано, и величины f, φ порядка $5 \cdot 10^{-6}$ следует тоже считать малыми.

Из табл. 2 видно, что в единицах $\varkappa_1$, $\varkappa_2$ величина $f = 111$ стала «малой» по сравнению с $\varphi = 17$. Допускается ее существенное увеличение ради уменьшения φ. Внешне большое значение $f = 338$ затем сравнительно малыми изменениями δx и δy доводится до нуля. Это есть следствие разной чувствительности f и φ к изменениям x, y, т.е. существенно разных величин их производных.

После того как два-три раза подряд минимизация r по s в модифицированном методе Ньютона приводит к значениям, близким к единице, переходят на обычный метод Ньютона, не тратя машинного времени на подбор s. Однако в данном примере, после получения

приближения, хорошего в смысле теоремы 1, сходимость оказывается столь быстрой, что этот прием не дал бы нам никакой экономии.

Другая, но по существу близкая, нормировка была предложена немецким математиком Дёфлхардом. В его варианте модифицированного метода Ньютона после вычисления $f(x)$ и матрицы $F_x(x)$ вычисляется матрица $A = f_x^{-1}$ и минимизируется невязка

$$r^2(s) = (A f(x + s\,\delta x),\ A f(x + s\,\delta x)).$$

Смысл этой конструкции станет ясен, если проанализировать поведение правой части при малых δ (в первом приближении):

$$A f(x + \delta) = A f(x) + A f_x(x)\,\delta = A f(x) + \delta.$$

Таким образом, в окрестности точки x невязка устроена очень просто:

$$r^2(\delta) \doteq C + (B, \delta) + (\delta, \delta).$$

Такая ситуация наиболее благоприятна для алгоритмов поиска минимума, а решение системы $f(x) = 0$ можно трактовать как поиск минимума $r^2 \equiv \|f(x)\|^2$.

Метод продолжения по параметру. Опишем в общих чертах другой прием, имеющий ту же цель — ослабить требования к выбору начального приближения и обеспечить надежную сходимость метода решения системы уравнений $f(x) = 0$. В литературе этот метод иногда называют «методом инвариантного погружения».

Рассмотрим семейство задач $F(x, \lambda) = 0$, где λ — скалярный параметр. Сконструируем это семейство так, что $F(x, 1) \equiv f(x)$, а при $\lambda = 0$ уравнение $F(x, 0) = 0$ легко решается или даже имеет явное решение. Это и есть «погружение» исходной задачи в семейство задач. Формально построить такое погружение часто не представляет труда. Вот, например, самый простой способ:

$$F(x, \lambda) \equiv (1 - \lambda)x + \lambda f(x). \tag{5}$$

Уравнение $F(x, 0) = 0$ имеет очевидное решение $x = 0$. Пусть $x(\lambda)$ есть решение уравнения (5). Последовательно решается серия задач. Имея решение $x(\lambda)$, меняем λ на $\lambda + \Delta\lambda$ и решаем уравнение $F(x, \lambda + \Delta\lambda) = 0$ тем или иным итерационным методом. Используем $x(\lambda)$ как хорошее начальное приближение ($\Delta\lambda$, естественно, считаем малым изменением). Таким образом можно (при благоприятном ходе событий) добраться до $\lambda = 1$ и получить решение исходной задачи.

Нетрудно оформить эту конструкцию в виде задачи Коши для системы обыкновенных дифференциальных уравнений, т.е. вычислить производную $dx/d\lambda$. В самом деле, дифференцируя по λ, получаем

$$0 = \frac{d}{d\lambda} F(x, \lambda) = F_x(x, \lambda)\,\frac{dx}{d\lambda} + F_\lambda(x, \lambda),$$

откуда

$$\frac{dx}{d\lambda} = -F_x^{-1}(x, \lambda)\, F_\lambda(x, \lambda).$$

Связь с методом Ньютона достаточно прозрачная. Сведение к задаче Коши иногда считают исчерпывающим решением проблемы, поскольку многие полагают эту задачу самой простой в вычислительной математике. Это мнение (ошибочное в столь общей форме, как мы увидим в дальнейшем) основано на том факте, что для решения задачи Коши существуют не только строго обоснованные алгоритмы, но даже стандартные программы и можно, не имея представления о том, как они работают, просто обращаться к ним.

Реализуя метод продолжения по параметру, часто сталкиваются с тем, что график $x(\lambda)$ имеет S-образную форму, т.е. имеется несколько ветвей решения уравнения $F(x, \lambda) = 0$. Отслеживая одну из них, достигают некоторой точки λ_1, за которую данная ветвь $x(\lambda)$ не продолжается, — решения уравнения $F(x, \lambda + \Delta\lambda) = 0$, близкого к $x(\lambda)$, не существует. Внешне это проявляется в вырождении матрицы F_x, т.е. $\det F_x \to 0$. Для продолжения решения задачи нужно двигаться по λ в обратном направлении, перейдя, однако, на другую ветвь функции $x(\lambda)$.

Технически это реализуется следующим образом: n-мерный вектор x и скалярный параметр λ объединяются в единый $(n+1)$-мерный вектор $y = \{y_1, \ldots, y_{n+1}\}$. Специальным образом на каждом шаге процесса продолжения выбирается одна из компонент y_i, которой дается предписанное приращение; остальные находятся решением системы n нелинейных уравнений. Выбор номера этой ведущей компоненты основан на анализе предшествующих шагов. Рекомендуется выбирать в качестве ведущей ту компоненту y_i, эволюция которой в процессе продолжения не дает оснований предположить возможное изменение направления ее движения. Признаком приближения точки поворота для компоненты y_i может служить, например, уменьшение ее приращения за один шаг.

§ 2. Численное дифференцирование

Практическое применение приведенных в § 1 формул численного дифференцирования связано с необходимостью выбора подходящего шага h. Возникающие здесь проблемы рассмотрим для простоты на примере функции только одного переменного $f(x)$. Нас интересует погрешность формулы численного дифференцирования. Тривиальный («школьный») ответ «чем меньше h, тем точнее формула численного дифференцирования» основан на известном соотношении

$$\frac{df}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}.$$

Однако, как мы увидим, при реальных вычислениях ситуация сложнее.

Пусть функция $f(x)$ гладкая, но, работая на ЭВМ, мы имеем дело не с $f(x)$, а с ее машинным представлением

$$f^{\text{м}}(x) = f(x)(1 + \varepsilon(x)),$$

где $\varepsilon(x)$ — относительная погрешность вычисления f. Разумеется ε зависит от x, но для всех интересующих нас x пусть имеется оценка $|\varepsilon(x)| \leqslant \varepsilon \ll 1$. Величина ε может быть связана хотя бы с конечным числом знаков в представлении f в памяти ЭВМ ($\varepsilon \approx 10^{-12}$ на БЭСМ-6, $\varepsilon \approx 10^{-7}$ на ЕС, $\varepsilon \approx 10^{-16}$ на ЕС при двойной точности). Но если $f(x)$ вычисляется достаточно сложно, погрешность ε может быть и существенно большей величиной, не всегда допускающей хорошую оценку.

Таким образом, используя численное дифференцирование, вычисляем

$$\frac{f^{\text{м}}(x+h) - f^{\text{м}}(x)}{h} = \frac{f(x+h) - f(x)}{h} + f\,\frac{\varepsilon_1 - \varepsilon_2}{h}.$$

Разложим в ряд Тейлора:

$$f(x + h) = f + h\,f_x + \frac{h^2}{2} f_{xx} + O(h^3).$$

Итак,

$$\frac{f^{\text{м}}(x+h) - f^{\text{м}}(x)}{h} = f_x(x) + \frac{h}{2} f_{xx} + O(h^2) + O\left(\frac{2\varepsilon}{h} f\right).$$

Погрешность численного дифференцирования состоит из двух частей. Первая из них связана с заменой оператора дифференцирования оператором конечной разности. Она имеет величину $O(f_{xx}h)$ и стремится к нулю при $h \to 0$. Степень h в погрешности аппроксимации называют *порядком аппроксимации*. Вторая часть погрешности связана с неточностью вычисления f. Она имеет величину $O(\varepsilon f/h)$ и при $h \to 0$ стремится к бесконечности. Полная погрешность численного дифференцирования есть сумма погрешностей аппроксимации и округления: $0.5hf_{xx} + 2\varepsilon f/h$.

Легко вычислить шаг h_0, при котором полная погрешность минимальна. Очевидно, $h_0 = \sqrt{4|\varepsilon f/f_{xx}|}$. Это грубая оценка: ведь h_0 вычисляется через ε и f_{xx}, точные значения которых обычно неизвестны. Существуют достаточно надежные алгоритмы численного дифференцирования, основанные на вычислении большего числа значений f. По этим значениям вычисляются варианты разностной производной, из сравнений которых между собой отбирается наиболее достоверное значение. Надежность подобных алгоритмов оплачивается большим объемом вычислений. Часто, например при реше-

нии уравнений методом Ньютона, нет нужды в особенно высокой точности численного дифференцирования и не требуется очень точно определять h_0. В частности, автор иногда использовал простой способ проверки того, является ли данное h подходящим для численного дифференцирования. Этот способ основан на подсчете числа «сокращающихся знаков». Поясним его на примере численного дифференцирования функции e^x в точке $x = 1$. Пусть e^x вычисляется с

Таблица 3

h	0.1	0.01	0.001	0.0001	0.00001
e^{1+h}	3.004166	2.745601	2.721001	2.718554	2.718309
e^1	2.718282	2.718282	2.718282	2.718282	2.718282
$e^{1+h}-e^1$	0.285884	0.027319	0.002719	0.000272	0.000027
$(e^x)_x$	2.85884	2.7319	2.719	2.72	2.7
k	1	2	3	4	5

семью верными знаками. Вычислим ее разностную производную с шагом h, равным 0.1, 0.01, 0.001, 0.0001. В табл. 3 пояснения требует только последняя строка: k — это число сократившихся главных знаков.

Таким образом, наилучший результат получается при сокращении половины знаков. В общем случае число «сократившихся знаков» при численном дифференцировании можно оценивать величиной

$$\lg \frac{|f(x+h)-f(x)|}{|f(x+h)|+|f(x)|}.$$

Заметим, что в этом простом примере мы сталкиваемся с одной из самых грозных опасностей в приближенных вычислениях: если результат получается при вычитании двух очень близких друг к другу величин, происходит резкая потеря точности, относительная погрешность результата сильно возрастает. Это явление носит название «сокращение знаков» и доставляет массу неприятностей.

Оценим погрешность формулы центральных разностей:

$$\frac{df}{dx} \approx \frac{f^{\text{м}}(x+h)-f^{\text{м}}(x-h)}{2h} = \frac{f(x+h)-f(x-h)}{2h} + O\left(\frac{\varepsilon}{h} f\right).$$

Разлагая в ряд Тейлора $f(x+h)$ и $f(x-h)$, получаем

$$\frac{f^{\text{м}}(x+h)-f^{\text{м}}(x-h)}{2h} = f_x(x) + \frac{h^2}{3} f''' + O(h^3) + O\left(\frac{\varepsilon}{h} f\right).$$

Главный член погрешности аппроксимации имеет второй по h порядок, а оптимальный шаг

$$h_0 \approx (|3\varepsilon f/2f'''|)^{1/3}.$$

Таким образом, формула центральной разности точнее формулы односторонней разности, но требует большего шага h (при одинаковых шагах сокращается больше знаков). Например,

$$e^{1.001} - e^{0.999} = 2.721001 - 2.715565 = 0.005436, \quad (e^x)_x \approx 2.718$$

(т.е. получается та же точность, что и в формуле односторонней разности). При $h = 0.01$ имеем

$$e^{1.01} - e^{0.99} = 2.745601 - 2.691234 = 0.054367, \quad (e^x)_x \approx 2.71835.$$

В дальнейшем нам часто придется использовать формулу приближенного вычисления второй производной

$$\frac{d^2f}{dx^2} \approx \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}.$$

Оценим точность этой формулы. Очевидно, погрешность округления есть $(4\varepsilon/h^2)f$. Вычислим погрешность аппроксимации, используя разложение в ряд Тейлора $f(x+h)$ и $f(x-h)$. После простых вычислений найдем полную погрешность численного дифференцирования:

$$\frac{h^2}{12} f^{\mathrm{IV}} + \frac{4\varepsilon}{h^2} |f|,$$

и оптимальный шаг:

$$h_0 \approx (|48\varepsilon f/f^{\mathrm{IV}}|)^{1/4}.$$

Вообще, из всех полученных выше формул для h_0 видно, что «наилучший шаг» тот, при котором погрешности аппроксимации и округления совпадают (близки). Вычисление второй производной методом численного дифференцирования требует еще большего шага.

Нетрудно построить разностные формулы вычисления производных третьего и четвертого порядков:

$$\frac{d^3f}{dx^3} \approx \frac{f(x+h) - 3f(x) + 3f(x-h) - f(x-2h)}{h^3},$$

$$\frac{d^4f}{dx^4} \approx \frac{f(x+2h) - 4f(x+h) + 6f(x) - 4f(x-h) + f(x+2h)}{h^4}.$$

Пользоваться ими нужно, конечно, с большой осторожностью по причинам, понятным из приведенного выше обсуждения.

§ 3. Интерполяция функций

Приведем некоторые начальные сведения из теории интерполяции. Этот классический аппарат вычислительной математики в последние годы стал развиваться и использоваться в несколько ином направлении (по сравнению с его назначением в трудах классиков). Мы постараемся дать представление и об этих новых аспектах аппарата интерполяции.

Естествознание и, особенно, математическая физика обычно имеют дело с задачами, сформулированными в терминах функций: нужно найти некоторую функцию $f(t)$, удовлетворяющую тем или иным условиям, уравнениям. Произвольная («измеримая») функция полностью определяется «континуумом» информации. К счастью, мы не имеем дела со столь общими объектами, нас интересуют более узкие классы функций.

Непрерывная функция определяется «счетной» информацией: достаточно знать ее лишь на счетном множестве точек, всюду плотном на том интервале (множестве), где она нас интересует. Однако при реализации расчетов на ЭВМ мы располагаем конечным множеством чисел, причем и числа-то имеют конечное число знаков. Таким образом, мы располагаем лишь конечной информацией о функции и, следовательно, наши знания о решении какой-то задачи принципиально не полны. Естественно возникает вопрос о способах представления функции на ЭВМ, о потере информации, о возможно более рациональных способах представления специальных классов функций.

Ограничимся одним способом представления функций — сеточным или, иначе, табличным. Это связано с той особой ролью, которую играет сеточный метод в рассматриваемых нами методах приближенного решения задач математической физики. Начнем с классической задачи интерполяции. Пусть имеется некоторая функция $f(t)$, заданная на интервале $[0, T]$.

Первая задача состоит в том, чтобы сопоставить этой функции конечный набор чисел, по которому ее можно будет восстановить (конечно, с той или иной точностью). Задачу будем решать очень просто. Введем на $[0, T]$ некоторую *сетку*

$$t_0 < t_1 < t_2 < \ldots < t_N, \qquad t_n \in [0, T], \quad n = 0, 1, \ldots, N.$$

В частности, ради простоты будем использовать *равномерную* сетку с *шагом* $\tau = T/N$: $t_n = n\tau$.

В качестве конечномерного представителя функции $f(t)$ используем таблицу чисел

$$\{f_n\}_{n=0}^{N}, \qquad \text{где } f_n = f(t_n).$$

Оператор, сопоставляющий функции f такую таблицу, играет большую роль в современных методах приближенных вычислений. Ему

присвоено особое наименование «оператор ограничения на сетку» (Restriction) и стандартное обозначение R_s, где индекс s — символ сетки $\{t_n\}_{n=0}^N$. Существуют и другие способы составления таблиц, представляющих функцию f. Например, можно составить таблицу пар чисел $\{f_n, f'_n\}$ — значений $f(t_n)$ и производных $f'(t_n)$, но мы ограничимся самым простым способом.

Теперь возникает следующая задача: по таблице $\{f_n\}$ восстановить непрерывную функцию. Разумеется, это будет какая-то другая функция $\tilde{f}(t)$ и надо оценить «потерю информации», т.е. величину $|f(t) - \tilde{f}(t)|$ при $t \in [0, T]$. Это восстановление неоднозначно, оно осуществляется тем или иным *оператором интерполяции* (обозначим его I), а потеря информации, как легко догадаться, зависит от сетки, типа оператора I и свойств гладкости функции f. Итак, мы имеем дело со схемой

$$f(t) \xrightarrow{R_s} \{f_n\}_{n=0}^N \xrightarrow{I} \tilde{f}(t). \tag{1}$$

Ниже мы рассмотрим некоторые конкретные формы оператора интерполяции I.

Кусочно-линейная интерполяция. Это простейший вариант I, рассчитанный на функции f с небольшим запасом гладкости. Сам аппарат очень прост: точки (t_n, f_n) соединяются отрезками прямых

$$\tilde{f}(t) = \frac{(t - t_n) f_{n+1} + (t_{n+1} - t) f_n}{t_{n+1} - t_n}, \quad t \in [t_n, t_{n+1}].$$

Таким образом, функция $\tilde{f}(t)$ рассматривается как аппроксимация функции $f(t)$ и следует оценить погрешность $|f(t) - \tilde{f}(t)|$. Проблемы такого сорта возникли в классической математике, когда появилась необходимость работать с некоторыми специальными функциями (sin, ln, exp, функции Бесселя и т.д.), а естественным способом описания функций были таблицы. В наше время способом описания многих функций стали алгоритмы их вычисления, «запаянные» в процессорах карманных, например, калькуляторов.

Итак, предположим, что функция $f(t)$ всего лишь удовлетворяет условию Липшица с постоянной C:

$$|f(t) - f(t')| \leqslant C|t - t'|, \quad \forall\, t,\ t' \in [0, T]. \tag{2}$$

В этом случае погрешность интерполяции оценивает следующая теорема.

Теорема 1. Для любых $t \in [0, T]$ погрешность

$$|f(t) - \tilde{f}(t)| \leqslant C\tau/2, \quad \text{где } \tau = \max (t_{n+1} - t_n).$$

Эта оценка неулучшаема в классе (2).

Доказательство. Пусть $t \in [t_k, t_{k+1}]$. Тогда, вводя обозначение $h = t_{k+1} - t_k$, представим t в виде

$$t = t_k + \alpha h, \quad \alpha \in (0, 1).$$

Очевидно, $\tilde{f}(t) = \alpha f_{k+1} + (1 - \alpha) f_k$. Проведем оценку:

$$|f(t) - \tilde{f}(t)| = |\alpha f_{k+1} + (1 - \alpha) f_k - \alpha f(t) - (1 - \alpha) f(t)| \leqslant$$
$$\leqslant \alpha |f_{k+1} - f(t)| + (1 - \alpha) |f_k - f(t)|.$$

Но $f_{k+1} = f(t_k + h)$, поэтому

$$|f_{k+1} - f(t)| = |f(t_k + h) - f(t_k + \alpha h)| \leqslant C(1 - \alpha) h.$$

Аналогично

$$|f_k - f(t)| \leqslant C\alpha h.$$

Итак,

$$|f(t) - \tilde{f}(t)| \leqslant 2\alpha(1 - \alpha) Ch \leqslant Ch/2.$$

Тот же аппарат кусочно-линейной интерполяции имеет более высокую точность, если функция $f(t)$ имеет ограниченную вторую производную.

Теорема 2. Пусть $|f''(t)| \leqslant C$. Тогда

$$|f(t) - \tilde{f}(t)| \leqslant C\tau^2/2, \quad t \in [0, T],$$

и эта оценка неулучшаема.

Доказательство этой теоремы непосредственно следует из теоремы 3 (см. ниже). Пример функции, на которой достигается эта оценка, предоставим построить читателю.

Кусочно-линейная интерполяция послужит нам поводом для введения некоторых полезных объектов. С сеткой $\{t_n\}$ можно связать набор стандартных функций — интерполяционный базис, состоящий из функций $\varphi^0(t), \varphi^1(t), \ldots, \varphi^N(t)$. (Правильнее было бы использовать обозначения типа $\varphi_N^n(t, \{t_n\}_{n=0}^N)$, содержащие все определяющие базис величины, но мы этого делать не будем.) Каждая функция $\varphi^n(t)$ сопоставляется своему узлу сетки t_n и определяется следующим образом: в узлах сетки $\varphi^n(t_k) = \delta_k^n$, в остальных точках она вычисляется кусочно-линейной интерполяцией (рис. 3). Используя этот базис, можно представить $\tilde{f}$ в форме

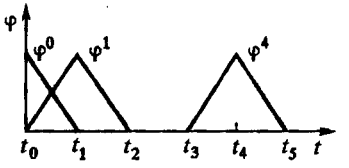


Рис. 3

$$\tilde{f}(t) = \sum_{n=0}^{N} f_n \varphi^n(t).$$

Аппарат кусочно-линейной интерполяции можно трактовать и как способ непрерывного восполнения сеточной функции до функции, определенной при всех $t \in [0, T]$, и как способ конечномерной аппроксимации некоторого функционального пространства — в данном случае, пространства непрерывных функций, имеющих кусочно-непрерывную первую производную. Наконец, функции базиса $\varphi^n(t)$ можно рассматривать и как простейший пример так называемых «конечных элементов». Это один из весьма важных и широко используемых в современных численных методах объектов, позволяющих моделировать (аппроксимировать) те или иные функциональные пространства. Ниже мы обсудим это подробнее.

Рассмотрев интерполяцию функций с малым запасом гладкости, обратимся к аппарату, напротив, рассчитанному на очень гладкие функции.

Интерполяционный полином. Итак, пусть имеется сетка $\{t_n\}_{n=0}^N$ и сеточная функция $\{f_n\}_{n=0}^N$, являющаяся ограничением некоторой гладкой функции $f(t)$ на сетку. Через точки (t_n, f_n) проведем полином степени N. Другими словами, построим полином $L(t)$ степени N, коэффициенты которого (их $N+1$) определяются из $(N+1)$-го условия:

$$L(t_n) = f_n, \quad n = 0, 1, \ldots, N. \tag{3}$$

Аккуратное обозначение этого полинома есть, очевидно, $L_N(t; \{t_n\}, \{f_n\})$, но мы будем вспоминать список аргументов только тогда, когда это потребуется по существу дела. Пока аргументы N, $\{t_n\}$, $\{f_n\}$ фиксированы, мы их опускаем. Вопросы существования и единственности интерполяционного полинома рассматриваются в анализе (определитель Ван-дер-Монда), мы их решим по ходу дела, написав явно выражение для L. Сначала построим базис из функций $\varphi_N^n(t)$ (аккуратнее, $\varphi_N^n(t; \{t_k\})$). Функции $\varphi_N^n(t)$ — полиномы степени N, каждый из которых сопоставлен со своим узлом сетки t_n таким образом, что $\varphi_N^n(t_k) = \delta_k^n$. Легко угадать явное выражение для $\varphi_N^n(t)$:

$$\varphi_N^n(t) = \prod_{i \neq n} (t - t_i)/(t_n - t_i).$$

(Произведение берется по всем индексам, кроме $i = n$.)

Имея интерполяционный базис, можно написать явное выражение интерполяционного полинома (в так называемой форме Лагранжа):

$$L_N(t; \{t_n\}, \{f_n\}) = \sum_{i=0}^{N} f_i \varphi_N^i(t; \{t_n\}). \tag{4}$$

Выполнение условий (3) очевидным образом следует из (4). Если записать L в общем виде: $L(t) = a_N t^N + \ldots + a_0$, то условия (3) превращаются в $N+1$ линейное уравнение для коэффициентов $a_0, a_1, \ldots, a_N$. Таблица $\{f_n\}$ определяет правую часть этой системы. Так как для любой такой правой части решение (в форме Лагранжа) существует, то оно в силу известных теорем линейной алгебры и единственно.

Переходя к оценке погрешности интерполяции, введем *остаточный член* интерполяционного полинома

$$R_N(t; \{t_n\}, \{f_n\}) \equiv f(t) - L_N(t; \{t_n\}, \{f_n\}). \tag{5}$$

Точка t может находиться как внутри интервала $[0, T]$ (и тогда говорят об *интерполяции*), так и вне его (и тогда употребляют термин *экстраполяция*). Обозначим $a = \min(t, t_0)$ и $b = \max(t, t_N)$. Таким образом, $t \in [a, b]$. Основу для конкретных оценок $|f(t) - L(t)|$ составляет следующая лемма.

Лемма (об остаточном члене). Пусть функция $f(t)$ на $[a, b]$ имеет $N+1$ ограниченную производную. Тогда

$$R_N(t) = \frac{1}{(N+1)!}(t - t_0)(t - t_1)\ldots(t - t_N)\, f^{(N+1)}(\xi), \tag{6}$$

где ξ — некоторая (зависящая от t, $\{t_n\}$ и f) точка, о которой известно только, что $\xi \in [a, b]$.

Доказательство. Считая, что t не совпадает ни с одним узлом сетки (при $t = t_n$ соотношение (6) очевидным образом выполнено), рассмотрим функцию одного переменного

$$F(x) \equiv f(x) - L(x) - R_N(t)\frac{(x - t_0)(x - t_1)\ldots(x - t_N)}{(t - t_0)(t - t_1)\ldots(t - t_N)}. \tag{7}$$

Об этой функции нам известно, что на $[a, b]$ она имеет по меньшей мере $N+1$ непрерывную производную, поскольку ее имеет функция $f(x)$, а остальные два слагаемых правой части (7) — полиномы от x.

Далее, функция $F(x)$ на $[a, b]$ имеет не менее $N+2$ нулей. Мы их просто укажем: очевидно, точки $x = t_n$ $(n = 0, 1, \ldots, N)$ — нули $F(x)$ в силу того, что $f(t_n) = L(t_n)$, а третье слагаемое правой части (7) обращается в нуль. В силу определения (5) остаточного члена $(N+2)$-м нулем $F(x)$ является точка $x = t$. Между каждыми двумя нулями непрерывно дифференцируемой функции имеется хотя бы один нуль ее производной. Таким образом, на $[a, b]$ имеется по крайней мере $N+1$ нуль F'.

Применяя это рассуждение последовательно к F'', F''', ..., установим, что существует точка $\xi \in [a, b]$ такая, что $F^{(N+1)}(\xi) = 0$.

Вычислим $(N+1)$-ю производную правой части (7), учитывая, что $L^{(N+1)}(x) \equiv 0$, произведение $\prod (x - t_n) = x^{N+1} + Ax^N + \ldots$ и его $(N+1)$-я производная равна $(N+1)!$:

$$F^{(N+1)}(\xi) - R_N(t) \frac{(N+1)!}{(t-t_0)(t-t_1)\ldots(t-t_N)} = 0. \qquad (8)$$

Из (8) непосредственно следует утверждение леммы (6).

Выражение (6) для остаточного члена позволяет получить серию более конкретных результатов. Приведем некоторые из них.

Теорема 3 (о точности интерполяции на равномерной сетке). Пусть сетка равномерна, т.е. $t_n = n\tau$, $\tau = T/N$, $t \in [0, T]$. Тогда

$$|f(t) - L(t)| \leqslant \frac{\tau^{N+1}}{N+1} C, \quad C = \max_{t \in [0, T]} |f^{(N+1)}(t)|. \qquad (9)$$

Доказательство. Положим $t = t_k + \alpha\tau$ ($\alpha \in (0, 1)$, $k \in 0, N-1$). Рассмотрим выражение $\prod_{n=0}^{N} (t - t_n)$, входящее в формулу (6). Очевидно,

$$t - t_n = k\tau + \alpha\tau - n\tau = (k + \alpha - n)\tau.$$

Таким образом,

$$\prod_{n=0}^{N} (t - t_n) = \tau^{N+1} \prod_{n=0}^{N} (k + \alpha - n).$$

Для оценки $\prod (k + \alpha - n)/(N+1)!$ воспользуемся табл. 4, в которой приведены значения $|k + \alpha - n|$ и соответствующим образом расположенные мажорирующие их множители из $N!$. Видно, что оцениваемое произведение состоит из множителей, не превосходящих единицы, и дополнительного множителя $1/(N+1)$. Отсюда следует (9).

Посмотрим, как изменяется оценка при переходе к задаче экстраполяции. Она ухудшается при удалении точки t от «носителя информации» — интервала $[t_0, t_N]$. Анализ, аналогичный только что приведенному, показывает, что:

1) при $t \in [t_N, t_N + \tau]$

$$|R_N(t)| \leqslant \tau^{N+1} \max_{t_0 \leqslant t' \leqslant t} |f^{(N+1)}(t')|;$$

2) при $t \in [t_N + \tau, t_N + 2\tau]$

$$|R_N(t)| \leqslant (N+2)\tau^{N+1} \max_{t_0 \leqslant t' \leqslant t} |f^{(N+1)}(t')|;$$

3) при $t \in [t_N + 2\tau, t_N + 3\tau]$

$$|R_N(t)| \leqslant \frac{(N+2)(N+3)}{1 \cdot 2} \tau^{N+1} \max_{t_0 \leqslant t' \leqslant t} |f^{(N+1)}(t')|;$$

и т.д.

Таким образом, оценка ухудшается как за счет появления множителей, так и за счет увеличения оценки $|f^{(N+1)}|$ при расширении интервала. Хотя это — оценка сверху, она правильно отражает общую тенденцию: экстраполяция функции менее надежна, чем интерполяция, и ее точность резко падает по мере удаления от носителя информации. (Этим, в частности, объясняется ненадежность

Таблица 4

n	$k-k$	...	$k-2$	$k-1$	k	$k+1$	$k+2$	...	N
$\lvert k+\alpha-n\rvert$	$k+\alpha$	...	$2+\alpha$	$1+\alpha$	α	$1-\alpha$	$2-\alpha$	...	$N-k-\alpha$
$N!$	$k+1$	...	3	2	1	1	$2+k$	...	N

разного рода прогнозов, многие из которых основаны на той или иной экстраполяции измеренных в прошлом значений прогнозируемой величины.)

Итак, интерполяционный полином на равномерной сетке дает существенно более высокую точность, чем кусочно-линейная интерполяция. Их погрешности суть $(T/N)^{N+1} \max |f^{(N+1)}|/(N+1)$ и, в лучшем случае, $(T/N)^2 \max |f''|/2$ соответственно. Правда, прямое сравнение этих величин невозможно, ведь оценки включают значения разных производных.

Наилучший выбор сетки. Рассмотрим следующую задачу. Можно ли повысить точность интерполяции за счет рационального размещения узлов сетки на интересующем нас интервале $[0, T]$? Если об интерполируемой функции мы не знаем ничего, кроме того что она $N+1$ раз непрерывно дифференцируема, единственным средством улучшить оценку остаточного члена является выбор узлов сетки $\{t_n\}$ решением характерной задачи на минимакс:

$$\min_{t_0, t_1, \ldots, t_N} \left\{ \max_{0 \leqslant t \leqslant T} \left| \prod_{n=0}^{N} (t - t_n) \right| \right\}.$$

Эта задача, известная как задача о построении полинома, наименее уклоняющегося от нуля на заданном интервале $[0, T]$ (с нормировкой «коэффициент при старшей степени t равен единице»), была решена (по другому поводу) П. Л. Чебышёвым.

Таким образом, «наилучшей» является сетка, узлы которой совпадают с корнями полинома Чебышева степени N на интервале $[0, T]$. Для этих корней имеются простые явные формулы. Чтобы не отвлекаться, приведем необходимые нам сведения о полиномах Чебышева в конце параграфа. Заметим только, что это один из важнейших объектов численного анализа, и в дальнейшем мы еще не раз встретимся с задачами, в которых полиномы Чебышева окажутся очень полезными для построения эффективных вычислительных методов. Сетки с узлами, равными корням полинома Чебышева, играют большую роль в разных вопросах и называются *чебышевскими*.

Чебышевская сетка является «наилучшей» с точки зрения целого класса функций, выделенного, например, условием $|f^{(N+1)}(t)| \leqslant C$. Для конкретной же функции $f(t)$ может оказаться лучшей своя индивидуальная сетка, и такие сетки часто используются в расчетах. Построение индивидуальной сетки требует гораздо большей информации о строении функции. При построении сетки используются простые соображения: например, сетка должна быть гуще там, где функция «устроена» более сложно, имеет резкие градиенты, совершает частые колебания. Там же, где функция очень гладкая, сетку можно взять более редкой.

Собственно говоря, приведенные соображения тоже связаны с оценкой производных, но если эти производные принимают на $[0, T]$ существенно разные значения, имеет смысл (если это технически возможно) разбить интервал $[0, T]$ на части с разными оценками производной и на каждой части строить сетку по-своему. Например, один из практических принципов построения сетки — «принцип равномерности погрешности». При использовании кусочно-линейной интерполяции для гладкой функции этот принцип рекомендует расставлять узлы таким образом, чтобы величина $(t_{n+1} - t_n)^2 C_{n+1/2}$, где $C_{n+1/2}$ — оценка $|f''|$ на $[t_n, t_{n+1}]$, была более или менее равномерной, не зависящей от n. Рисунок 4 поясняет, что примерно имеется в виду под индивидуальной сеткой.

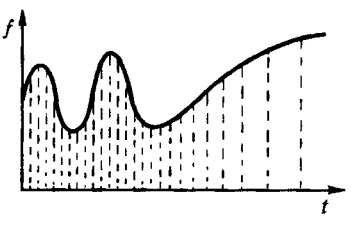


Рис. 4

Сравним точности интерполяций на равномерной и чебышевской сетках. Мы приведем только результат, сравнив оценки величины

$$\max_{0 \leqslant t \leqslant T} \frac{1}{(N+1)!} \prod_{n=0}^{N} |t - t_n| \tag{10}$$

для двух типов сеток. В случае равномерной сетки, как было показано выше, эта величина оценивается числом $(T/N)^{N+1}/(N+1)$. В

случае чебышевской сетки, как следует из сведений, приведенных в конце параграфа, для (10) имеем оценку $(T/4)^N/(N+1)!$. Используя формулу Стирлинга $N! \approx (N/e)^N$, получаем выигрыш примерно в $(4/e)^N$ раз. Это само по себе может быть и не очень много. Однако чебышевские сетки обладают и более важными преимуществами перед равномерными. Перейдем к их обсуждению.

Устойчивость интерполяционного полинома относительно погрешностей вычисления f. Рассмотрим следующий вопрос: пусть таблица $\{f_n\}$ вычислена не точно, а с погрешностями, связанными хотя бы с конечной разрядностью машинной арифметики. Как повлияет неточность вычисления f на интерполяционный полином? Итак, пусть полином вычисляется по таблице $\{f_n + \delta_n\}$, где δ_n — погрешность, относительно которой нам известна только оценка $|\delta_n| \leqslant \delta$. Тогда «машинный» интерполяционный многочлен связан с истинным очевидным соотношением

$$L_N(t; \{t_n\}, \{f_n + \delta_n\}) = L_N(t; \{t_n\}, \{f_n\}) + L_N(t; \{t_n\}, \{\delta_n\}). \quad (11)$$

Это следствие линейности полинома по f_n.

Второе слагаемое правой части (11) есть погрешность, связанная с неточностью вычисления f. Она и подлежит оценке. Естественно в качестве меры погрешности взять величину

$$\max_{|\delta_n| \leqslant \delta} \{ \max_{0 \leqslant t \leqslant T} |L_N(t; \{t_n\}, \{\delta_n\})| \},$$

т.е. оценить ее при самой неблагоприятной комбинации погрешностей δ_n в пределах заданной точности. В силу линейности по δ_n можно заменить эту величину на $\delta\eta(\{t_n\})$, где

$$\eta(\{t_n\}_{n=0}^N) = \max_{|\delta_n| \leqslant 1} \{ \max_{0 \leqslant t \leqslant T} |L_N(t; \{t_n\}, \{\delta_n\})| \}. \quad (12)$$

Величина $\eta(\{t_n\}_{n=0}^N)$ является характеристикой сетки и оценивает чувствительность интерполяционного многочлена к погрешностям в f_n.

В теории интерполяции такие характеристики были вычислены и оказалось, что $\eta \approx 2^N$ для равномерной сетки и $\eta \approx \ln N$ для чебышевской. Величину $\eta(\{t_n\})$ называют также нормой интерполяционного полинома на данной сетке. Этот термин связан с тем, что если определить $\|\{f_n\}\| \equiv \max_n |f_n|$, то при $t \in [0, T]$ имеем

$$|L_N(t; \{t_n\}, \{f_n\})| \leqslant \|\{f_n\}\| \, \eta(\{t_n\}).$$

Мы выяснили важное обстоятельство: интерполяционный полином на равномерной сетке очень чувствителен к погрешностям вычисления f_n ($\eta \approx 2^N$); полином же на чебышевской сетке слабо реагирует на эти погрешности: его погрешность мало отличается от погрешности вычисления f — только не очень большим множителем $\ln N$. Конечно, эти факторы существенны при высоких степенях интерполяционного полинома ($N = 10, 20, \ldots$). При низких же степенях интерполяционного полинома ($N \approx 5$) это еще не очень существенное различие.

Оценка (12) приводит к еще одному важному достоинству чебышевских сеток.

Устойчивость интерполяционного полинома относительно априорной информации. К априорной информации мы отнесем предположение о наличии у интерполируемой функции f нужного числа производных. Поставим следующий вопрос: пусть для функции $f(t)$ на сетке $\{t_n\}_{n=0}^N$ построен интерполяционный полином степени N, а фактически функция f на $[0, T]$ не имеет требуемой теорией $(N+1)$-й производной. Например, пусть какая-то ее производная невысокого порядка терпит разрыв. Как это скажется на качестве аппроксимации такой функции интерполяционным полиномом $L_N(t; \{t_n\}, \{f_n\})$?

Оказывается, ответ существенно связан с характером сетки. С. Н. Бернштейном было показано, что полиномы, интерполирующие на равномерной сетке функцию $f(t) = |t|$ с ростом N не только не сходятся к интерполируемой функции, но

$$\max_{-1 \leqslant t \leqslant 1} |L_N(t) - |t|| \to \infty \quad \text{при} \quad N \to \infty.$$

Совсем иначе и совершенно замечательно ведут себя интерполяционные полиномы с чебышевской сеткой. Для того чтобы объяснить это их свойство, нам понадобятся некоторые факты из так называемой «конструктивной теории функций». Введем полином степени N, наилучшим образом аппроксимирующий данную функцию $f(t)$ на $[0, T]$, т.е. полином, реализующий

$$\min_{P_N} \{\max_{0 \leqslant t \leqslant T} |P_N(t) - f(t)|\} = E_N(f).$$

Здесь min берется по всем полиномам степени не выше N. Полином наилучшей аппроксимации существует, но его фактическое построение является очень трудной задачей. Такие полиномы находятся только сложными итерационными алгоритмами так называемой *негладкой оптимизации* и требуют большого объема вычислений.

Таким образом, *полином наилучшего приближения* к $f(t)$ на $[0, T]$ является объектом по существу не конструктивным, его использование в практической вычислительной работе в общем

случае просто невозможно. Однако теоретические его свойства хорошо изучены. В частности, известна асимптотика величины $E_N(f)$ при $N \to \infty$. Она оказалась однозначно связанной с гладкостью $f(t)$. Несущественно огрубляя формулировки, приведем основные факты.

Если $f(t)$ имеет на $[0, T]$ ограниченную r-ю производную, то $E_N(f) = O(1/N^r)$. И наоборот, если $E_N(f)$ имеет такую асимптотику, то f_n имеет ограниченную r-ю производную. Хотелось бы иметь конструктивный метод построения полиномов, дающих аппроксимацию функции $f(t)$, близкую к наилучшей. Оказывается, таким является интерполяционный полином на чебышевской сетке! Этот факт мы сейчас докажем.

Теорема 4. Пусть $L_N(t; \{t_n\}, \{f_n\})$ — интерполяционный полином для $f(t)$ на интервале $[0, T]$ с чебышевской сеткой. Тогда

$$|L_N(t) - f(t)| \leqslant (1 + \ln N) E_N(f), \quad t \in [0, T],$$

т.е. такой полином дает почти наилучшую для полиномов степени N аппроксимацию функции $f(t)$ на $[0, T]$.

Доказательство. Пусть $P_N(t)$ — полином наилучшей аппроксимации $f(t)$. Тогда $f_n = P_N(t_n) + \delta_n$, причем погрешности δ_n удовлетворяют условиям $|\delta_n| \leqslant E_N(f)$. Обозначая $p_n = P_N(t_n)$, имеем

$$L_N(t; \{t_n\}, \{f_n\}) = L_N(t; \{t_n\}, \{p_n + \delta_n\}) = \\ = L_N(t; \{t_n\}, \{p_n\}) + L_N(t; \{t_n\}, \{\delta_n\}).$$

Но $L_N(t; \{t_n\}, \{p_n\}) = P_N(t)$, так как интерполяционный полином степени N, построенный для полинома степени не выше N, в точности совпадает с интерполируемым (при любой сетке).

Для второго слагаемого имеем оценку

$$|L_N(t; \{t_n\}, \{\delta_n\})| \leqslant \ln N \, \|\{\delta_n\}\| = \ln N \, E_N(f).$$

Используя представление $P_N(t) = f(t) + \delta(t)$, где $|\delta(t)| \leqslant E_N(f)$, получаем

$$L_N(t; \{t_n\}, \{f_n\}) = f(t) + \delta(t) + L_N(t; \{t_n\}, \{\delta_n\}).$$

Для погрешности аппроксимации имеем оценку

$$|\delta(t) + L_N(t; \{t_n\}, \{\delta_n\})| \leqslant (1 + \ln N) E_N(f).$$

Теорема доказана. Таким образом, интерполяционный полином на чебышевской сетке (напомним, что он легко выписывается в явном виде, так как корни полиномов Чебышева вычисляются по простым формулам) замечательным образом адаптируется к фактическим свойствам гладкости интерполируемой функции.

Дифференцирование интерполяционного многочлена. Из того, что две функции близки, как известно, еще не следует близость их производных. Однако в рассматриваемом случае ситуация более благополучная: $L(t)$ аппроксимирует $f(t)$, а производные $L(t)$ — производные $f(t)$, хотя точность аппроксимации, конечно, падает с ростом порядка производной. Ограничимся здесь лишь указанием на следующий факт: производная $L_N(t)$ является интерполяционным полиномом степени $N-1$ для функции $f'(t)$, построенным на сетке, отличающейся от исходной. Это почти очевидно: между каждыми двумя нулями функции $f(t)-L_N(t)$, например между t_n и t_{n+1}, имеется хотя бы один нуль производной. Обозначим его t'_n.

Итак, в точках $\{t'_n\}_{n=0}^{N-1}$ полином $L'_N(t)$ совпадает с $f'(t)$. Точность аппроксимации производных f производными L_N падает как за счет понижения порядка полинома, так и за счет возможного появления в новой сетке неравномерного распределения шагов $t'_{n+1}-t'_n$.

Сплайн-интерполяция. В последние годы большое распространение в прикладных исследованиях получил новый, достаточно точный вид интерполяции, требующий, однако, от функции существенно меньшего запаса гладкости, чем интерполяционный полином. Происхождение этого аппарата интерполяции и сам термин «сплайн» связывают с техническим приемом чертежников. При необходимости провести непрерывную кривую через сеточный график $\{t_n, f_n\}_{n=0}^{N}$ на бумаге наносят точки (t_n, f_n), около каждой из которых втыкают рядом друг с другом две булавки, и через образовавшийся «коридор» пропускают тонкую, гибкую и упругую стальную линейку («сплайн»). Форма, которую принимает эта линейка (вдоль нее и проводят требуемую линию) решает, как мы увидим, задачу гладкой интерполяции.

Математическое исследование объяснило популярность описанного приема, заменяющего работу с лекалом. Полученная кривая оказывается дважды непрерывно дифференцируемой, а это свойство очень ценится в технике. Пример, который приводили в годы учебы автора и, верно, приводят сейчас: железнодорожный путь должен быть кривой с непрерывной второй производной. В противном случае в местах разрыва второй производной при движении поезда возникает «удар», разрушающий и рельсы, и колеса.

Упомянутое исследование читатель без труда проведет сам. Пусть $y(t)$ — форма, которую принял «сплайн». Теория упругости определяет эту форму требованием минимума энергии упругого со-

стояния. Таким образом, $y(t)$ определяется решением вариационной задачи

$$\min_{y(\cdot)} \int_0^T [y''(t)]^2\, dt \quad \text{при} \quad y(t_n) = f_n, \quad n = 0, 1, \ldots, N.$$

Здесь $y(\cdot)$ — функция $y(t)$, рассматриваемая целиком как точка функционального пространства, а $y(t)$ — в данном случае число, являющееся значением функции в точке t.

Элементарное вариационное исчисление сразу дает результат: внутри интервалов $[t_n, t_{n+1}]$ функция $y(t)$ удовлетворяет «уравнению Эйлера» $d^4y/dt^4 = 0$ и условиям трансверсальности, которые во внутренних узлах $t_1, t_2, \ldots, t_{N-1}$ сводятся к непрерывности первых и вторых производных. Из уравнения Эйлера следует, что $y(t)$ является кубическим многочленом: на каждом интервале $[t_n, t_{n+1}]$ имеется свой кубический многочлен, и все они гладко сопрягаются друг с другом.

Итак, с алгоритмической точки зрения сплайн $y(t)$ определяется таблицей

$$\{a_{n+1/2}, b_{n+1/2}, c_{n+1/2}, d_{n+1/2}\}_{n=0}^{N-1}$$

и формулой вычисления

$$y(t) = a_{n+1/2}\, t^3 + b_{n+1/2}\, t^2 + c_{n+1/2}\, t + d_{n+1/2}, \quad t \in [t_n, t_{n+1}].$$

Построение сплайна по таблице $\{t_n, f_n\}_{n=0}^N$ требует, таким образом, вычисления $4N$ коэффициентов. Для этого мы имеем следующие уравнения. Каждый кубический полином на концах своего интервала $[t_n, t_{n+1}]$ принимает заданные значения f_n и f_{n+1}. Это, очевидно, дает $2N$ линейных уравнений. В каждой внутренней точке сетки $t_1, t_2, \ldots, t_{N-1}$ имеем два условия, приравнивая правые и левые значения первой и второй производных, т.е. еще $2(N-1)$ линейных уравнений. Оставшиеся два уравнения получаем из условий трансверсальности при $t = 0$ и $t = T$.

Мы не будем здесь выписывать конкретного вида уравнений, не будем излагать и метода их решения (при больших N это самостоятельная проблема, требующая применения специфического алгоритма). Соответствующие вопросы давно решены, алгоритмы разработаны, описаны в многочисленных руководствах и включены в математическое обеспечение ЭВМ как стандартные программы. Мы ограничимся лишь общими сведениями о сплайнах. Этот аппарат был существенно обобщен и развит. В частности, разработаны методы двумерного гладкого восполнения функций, т.е. построения на основе таблицы $\{f_{k,m}\}$, заданной на

двумерной сетке узлов $\{x_k, y_m\}$ $(k = 0, 1, \ldots, K, \quad m = 0, 1, \ldots, M)$, гладкой интерполирующей поверхности $\tilde{f}(x, y)$, точки которой достаточно легко вычисляются. Сплайны нашли широкое применение при описании пространственных форм разных изделий. Они часто сообщаются (из конструкторских бюро на производства) таблицами (сеточными функциями) с указанием, что недостающие точки поверхности можно получить сплайн-интерполяцией.

Интерполяция конечными элементами. Опишем аппарат восполнения функции, заданной на сетке, до некоторой непрерывной функции той или иной степени гладкости (т.е. имеющей заданное число непрерывных производных). Первоначально интерполяция рассматривалась как способ приближенного (но достаточно «дешевого» по затратам вычислительной работы) вычисления функций, «точные» значения которых в узлах сетки можно было найти каким-либо очень «дорогим» алгоритмом. В настоящее время методы интерполяции рассматриваются и используются в несколько ином аспекте — как способы конечномерной аппроксимации тех или иных функциональных пространств.

Поясним сказанное. Многие задачи математической физики ставятся следующим образом. В некоторой заданной области изменения независимых переменных (для определенности, x и y) надо найти функцию из заданного функционального пространства W, удовлетворяющую некоторым уравнениям (включая краевые условия). Пространство W обычно определяется как множество функций, имеющих ограниченные (в той или иной норме) производные. Хорошая математическая теория данного класса задач стремится определить пространство W возможно более узким, в котором, однако, решение еще существует (сужение пространства обычно достигается увеличением порядка ограниченных производных).

Приближенное решение ищется в подходящем конечномерном пространстве, и его нужно строить минимальным, но содержащим функцию, достаточно близкую к искомому решению исходной задачи. Возникает задача аппроксимации различных функциональных пространств, причем требуется именно хорошая аппроксимация, определяемая возможно меньшим количеством информации. В этом смысле классическая теорема Вейерштрасса о возможности сколь угодно точно аппроксимировать произвольную («измеримую») функцию полиномом достаточно высокой степени практически мало полезна. Если функция не слишком гладкая, такая аппроксимация требует слишком высокой степени полинома и оказывается нерациональной.

Теперь рассмотрим одну, интересную саму по себе задачу математической физики — задачу Пуассона. В некоторой заданной об-

ласти G в плоскости (x, y) нужно найти функцию $u(x, y)$, удовлетворяющую уравнению

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y),$$

где f — заданная «правая часть». Кроме того, функция u должна принимать заданные значения (для простоты, $u = 0$) на границе области. В такой постановке решение следовало бы искать в классе W дважды непрерывно-дифференцируемых функций. Однако в этом слишком узком пространстве решение задачи не всегда существует. Решения, называемые «классическими» (т.е. имеющие непрерывные те производные, которые входят в уравнение, и удовлетворяющие уравнению в прямом смысле слова), существуют при ограничениях на правую часть, слишком стеснительных для практики и часто не выполняющихся.

Приемлемым оказалось следующее расширение W, при котором сама задача трансформировалась в вариационную: найти функцию $u(x, y)$, непрерывную и имеющую кусочно-непрерывные первые производные, из условия

$$\min_{u(\cdot) \in W} \iint_G (u_x^2 + u_y^2 - 2fu)\, dx\, dy.$$

При численном решении задачи Пуассона возникает задача моделирования, аппроксимации указанного выше функционального пространства W. На ней мы и продемонстрируем важную в вычислительных методах технику интерполяции конечными элементами.

Построим сначала триангуляцию области G, т.е. покроем ее сетью треугольников, каждые два из которых либо совсем не пересекаются, либо имеют только одну общую вершину, либо — общую сторону. Можно говорить, что G покрыта сеткой точек, каждая из которых находится только в вершине упоминавшихся выше треугольников (рис. 5). Построение триангуляции — не такая уж простая задача, особенно при большом числе узлов сетки. Выполнение этой работы вручную иногда становится просто непосильным: ведь в современных расчетах число узлов достигает тысяч, десятков тысяч. А если иметь в виду «триангуляцию» трехмерной области (покрытие ее сетью тетраэдров), работа становится почти невыполнимой. Поэтому процесс триангуляции нужно алгоритмизировать.

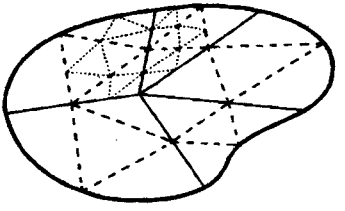


Рис. 5

Один из популярных алгоритмов состоит в следующем. Задается (вручную) грубая триангуляция области, включающая сравнительно небольшое число треугольников. Каждый треугольник разбивает-

ся на четыре: к имевшимся ранее вершинам треугольников добавляются точки на серединах сторон начальных треугольников. Этот процесс повторяется, и после нескольких таких «циклов» получается достаточно густая сетка, область триангулирована достаточно мелкими треугольниками. Заметим, что некоторые треугольники могут быть и криволинейными, но конечная триангуляция состоит только из обычных треугольников. Это, конечно, приводит к триангуляции не исходной области G, а некоторой ее аппроксимации. На рис. 5 показан процесс построения триангуляции (линии, возникающие на разных его этапах, имеют разные обозначения).

Рассмотрим один из треугольников и сеточную функцию, определенную в его вершинах, занумерованных индексами (1, 2, 3). Сеточная функция — это три числа u_1, u_2, u_3. Теперь решим задачу «интерполяции» такой функции, т.е. построим функцию, определенную внутри треугольника. Построим в нем интерполяционный базис из трех функций $\varphi_1(x, y)$, $\varphi_2(x, y)$, $\varphi_3(x, y)$, линейных по x и y, равных единице в «своей» вершине треугольника и нулю в остальных двух. Тогда интерполяция внутри треугольника выполняется по очевидной формуле

$$u(x, y) = \sum_{i=1}^{3} u_i \varphi_i(x, y). \tag{13}$$

Пусть $\{u^n\}_{n=1}^N$ — сеточная функция, определенная во всех узлах триангуляции. Используя в каждом треугольнике интерполяцию (13), получаем в G функцию $u(x, y)$. Она, очевидно, непрерывна и имеет кусочно-непрерывные (а точнее, кусочно-постоянные) первые производные. Каждый треугольник, оснащенный своим базисом, называют *конечным элементом*.

Технически оказывается удобнее задавать конечные элементы в виде стандартного треугольника в плоскости параметров (ξ, η) с вершинами в точках (0, 0), (1, 0), (0, 1), оснащенного стандартным базисом из функций

$$\varphi_1(\xi, \eta) = 1 - \xi - \eta, \qquad \varphi_2(\xi, \eta) = \xi, \qquad \varphi_3(\xi, \eta) = \eta.$$

Для конкретного треугольника с вершинами (x_1, y_1), (x_2, y_2), (x_3, y_3) легко построить линейное отображение его в стандартный треугольник:

$$\xi = a_1 + a_{11}x + a_{12}y, \qquad \eta = a_2 + a_{21}x + a_{22}y,$$

и обратное к нему.

Легко вычисляются, например, производные интерполированной функции:

$$\frac{\partial u}{\partial x} = \frac{\partial u}{\partial \xi}\frac{\partial \xi}{\partial x} + \frac{\partial u}{\partial \eta}\frac{\partial \eta}{\partial x} = (u_2 - u_1)\, a_{11} + (u_3 - u_1)\, a_{21}.$$

Они используются при решении задач Пуассона. Интегрирование по элементарному треугольнику сводится к интегрированию по стандартному треугольнику (в плоскости (ξ, η)) с последующим умножением на определитель преобразования (постоянный в силу линейности отображения).

Таким образом, конечные элементы позволили построить некоторое подпространство из того пространства W (непрерывных функций с кусочно-непрерывными производными), в котором имеется решение задачи. Этот аппарат оказался очень удобным для решения задач в областях не слишком простых форм. Он составляет основу одного из самых гибких алгоритмов решения задач математической физики — *метода конечных элементов* (МКЭ).

Обобщения метода конечных элементов связаны с решением не таких уж простых задач. В некоторых задачах (для бигармонического уравнения, например) естественным пространством W, которое следует аппроксимировать для приближенного решения, является пространство непрерывных функций с непрерывными первыми и кусочно-непрерывными вторыми производными. Процедура гладкого восполнения внутри отдельного треугольника должна быть такой, чтобы совпадали не только проинтерполированные в двух соседних треугольниках функции на общей границе, но и их первые производные. В некоторых задачах (гидродинамика несжимаемой жидкости) желательно построить аппарат интерполяции вектор-функции с нулевой дивергенцией, и т.п. Построение таких «элементов» ценится в науке настолько, что им присваивают имена их конструкторов (элемент Оргириса и т.д.).

Необходимо подчеркнуть, что работа с конечными элементами требует решения ряда достаточно сложных алгоритмических проблем. Многие вещи, о которых так легко было говорить, ссылаясь на геометрическую картинку (см. рис. 5), не так-то просто реализовать в ЭВМ. В самом деле, пусть задана последовательность координат всех точек сетки $\{x_n, y_n\}_{n=1}^{N}$ (не следует забывать, что N могут быть порядка 10^2, 10^3, а иногда 10^4). Эта информация полностью определяет сетку. Но попробуйте решить такую задачу. Составьте группы по три точки, образующие все элементы триангуляции. Не сомневаюсь, что каждый справится с этой работой, но каков будет объем вычислений? Поэтому работать с такой «минимальной» информацией практически невозможно.

В процессе формирования триангуляции необходимо формировать и дополнительную, избыточную информацию, позволяющую быстро решать те задачи, которые возникают при реализации метода конечных элементов. Например, можно сформировать последовательность троек чисел $\{m_{1,n}; m_{2,n}; m_{3,n}\}$ номеров точек сетки, образующих вершины m-го элемента. Допустим, такая информационная таблица уже есть. Попробуйте с ее помощью решить следующую за-

дачу. Пусть задан номер n одной из точек сетки. Найдите номера k элементов, одной из вершин которых является точка (x_n, y_n); в некоторых случаях желательно, чтобы эти номера были расположены в порядке следования элементов при обходе n-го узла по часовой, например, стрелке.

Нетрудно понять, что если такая задача относится к числу «массовых» (решаемых многократно) при реализации метода конечных элементов, имеет смысл решить ее один раз и запасти в памяти ЭВМ соответствующую таблицу. Пусть и эта информация уже есть. Попробуйте решить такую задачу. Дана точка (x, y) в области G, даны значения $\{u^n\}$ функции в узлах сетки. Вычислите $u(x, y)$. Иными словами, нужно найти номер того элемента, внутрь которого попала точка (x, y), и воспользоваться формулой интерполяции (13). Но как найти этот номер, не перебирая всех элементов? Здесь очень полезной может оказаться информация, возникающая (и сохраняемая в памяти ЭВМ) в том процессе генерирования триангуляции, который был описан выше. Просматривают все треугольники первичной, «ручной» триангуляции (а их немного) и определяют, в каком из них находится точка (x, y), затем просматривают только четыре треугольника следующего уровня триангуляции, и т.д. Но чтобы это можно было сделать, надо вырабатывать и сохранять соответствующую информацию (в виде некоторых таблиц).

Задача вычисления $u(x, y)$ является, как нетрудно понять, одной из типичных, наиболее массовых задач при содержательном истолковании решения, полученного методом конечных элементов в виде таблицы $\{u^n\}$ значений функции в узлах сетки. Ведь нас интересует, что происходит в той или иной точке именно исходного геометрического пространства (x, y).

Выше были приведены характерные геометрические задачи, которые легко решаются, если мы имеем перед собой чертеж триангуляции, на котором около каждого узла указан его номер n, а в каждом треугольнике — его номер m. Алгоритмизация, т.е. перевод этих «интуитивно» очевидных способов решения на чисто цифровой способ задания и обработки информации, — увлекательное и часто очень непростое занятие, особенно если строятся не просто принципиально верные алгоритмы, а, например, оптимальные по числу операций.

Реализация метода конечных элементов, не содержащая на первый взгляд серьезных трудностей, в действительности требует решения большого числа вспомогательных задач (некоторые из них были указаны выше). Кроме того, метод конечных элементов требует и достаточно больших объемов оперативной памяти ЭВМ. Этим в известной мере объясняется тот факт, что развитие и широкое внедрение в расчетную практику метода конечных элементов произошло в США, хотя основополагающие теоретические работы в этой области принадлежат чешскому математику М. Зламалу.

Некоторые сведения о полиномах Чебышева. Как это принято в теории аппроксимации, будем рассматривать стандартный интервал изменения независимого переменного $[-1, 1]$. Полиномом Чебышева, или полиномом наименее уклоняющимся от нуля, степени p называют полином, реализующий

$$\min \{ \max_{-1 \leqslant t \leqslant 1} |T_p(t)| \}.$$

Здесь min берется по всем полиномам степени p, нормированным условием: коэффициент при t^p равен единице. Иногда эту задачу трактуют и как задачу наилучшей (в норме C) аппроксимации функции t^p полиномом степени не выше $p-1$. Чебышевым была указана явная формула

$$T_p(t) = \frac{1}{2^{p-1}} \cos \{p \arccos t\}. \tag{14}$$

Правая часть, несмотря на тригонометрическую форму представления, в действительности является именно полиномом от t степени p.

Если нас интересует полином, наименее уклоняющийся от нуля на произвольном интервале $[a, b]$, следует сделать замену переменных

$$x = \frac{a+b}{2} + \frac{b-a}{2} t, \quad t = \frac{2}{b-a} \left(x - \frac{a+b}{2} \right),$$

которая переводит интервал $-1 \leqslant t \leqslant 1$ в $a \leqslant x \leqslant b$. Очевидно,

$$\tilde{T}_p(x) = \left(\frac{b-a}{2} \right)^p T_p \left[\frac{2}{b-a} \left(x - \frac{a+b}{2} \right) \right].$$

Множитель $[(b-a)/2]^p$ введен для сохранения нормировки: коэффициент при x^p в $\tilde{T}_p(x)$ равен единице.

Легко вычислить корни полинома Чебышева, используя его тригонометрическое представление (14):

$$t_k = \cos \frac{2k-1}{2p} \pi.$$

Для $k = 1, 2, \ldots, p$ получаем разные корни. Их значения имеют простую геометрическую интерпретацию: полуокружность единичного радиуса нужно разделить на $2p$ равных частей и из каждой нечетной точки деления опустить перпендикуляр. Отметим, что плотность корней повышается на концах интервала $[-1, 1]$. Корни чебышевского полинома на произвольном интервале $[a, b]$ суть

$$x_k = \frac{a+b}{2} + \frac{b-a}{2} \cos \frac{2k-1}{2p} \pi.$$

Полиномы Чебышева являются хорошим базисом в пространстве функций, заданных на каком-то интервале, для определенности на

[−1, 1]. Базис — важное понятие в приближенных методах. Напомним, что это система функций, обладающая свойством полноты, т.е. другие функции можно сколь угодно точно представлять конечными суммами (линейными агрегатами) функций базиса с числовыми коэффициентами.

Кроме свойства полноты, с практической точки зрения важна «цена» (в числе операций) вычислений функций базиса. С этой точки зрения удобен степенной базис $\{1, t, t^2, \ldots, t^n, \ldots\}$. Он является полным. Полином с коэффициентами $a_0, a_1, \ldots, a_n$ вычисляется достаточно просто. Обозначая частичные суммы через

$$s_k = \sum_{i=k}^{n} a_i t^{i-k},$$

полином s_0 вычисляем рекуррентно:

$$s_n = a_n, \qquad s_{k-1} = a_{k-1} + ts_k,$$

т.е. за n сложений и n умножений.

К сожалению, степенной базис обладает серьезным дефектом: он *плохо обусловлен*. Плохими являются такие базисы, элементы которых хотя и линейно независимы, но очень «похожи» друг на друга. Рисунок 6 поясняет, что имеется в виду. Точка A в плохом базисе имеет представление $a_1\varphi_1 + a_2\varphi_2$ с очень большими, противоположными по знаку и близкими друг к другу по модулю коэффициентами a_1, a_2. Вычисление такой суммы сопровождается уже знакомым нам неприятным явлением — сокращением знаков.

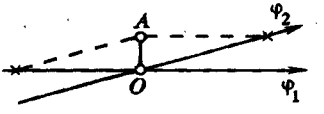


Рис. 6

Итак, плохой базис — это «сплюснутый» базис. Если читатель потрудится «нарисовать» графики функций $t^{20}, t^{21}, \ldots, t^{30}$, едва ли он отличит их друг от друга, и это заставит его насторожиться. Если читатель сможет найти аппроксимацию какой-либо нормальной функции полиномом высокого порядка, он увидит, что коэффициенты растут очень быстро при повышении степени (т.е. при уменьшении погрешности аппроксимации). При не таких уж больших степенях (при p, равных 20, 30, 40) они достигают столь больших величин, что вычисление полинома на ЭВМ, имеющей 10—15 десятичных знаков в мантиссе машинного числа, оказывается невозможным из-за полной потери точности.

Вспомним, что совсем не так ведут себя коэффициенты разложения какой-либо функции по такому хорошему базису, как тригонометрический: коэффициенты Фурье «честно» убывают в соответствии со степенью гладкости функции. Хорошими базисами являются ортогональные или близкие к ним. Это одна из причин, определяющих

большую роль различных ортогональных систем функций в приближенных методах. Полиномы Чебышева с этой точки зрения хороши. Они образуют ортогональную (правда, в специальной метрике с весом) систему функций:

$$\int_{-1}^{1} T_k(t)\, T_m(t)\, \frac{dt}{\sqrt{1-t^2}} = \delta_m^k.$$

Полиномы Чебышева близки к известному хорошему базису — тригонометрическому и, в сущности, совпадают с ним с точностью до замены независимого переменного $x = \arccos t$. Дело стало только за «ценой» вычисления $T_p(t)$. Но и здесь ситуация достаточно благоприятная: существует удобная рекуррентная формула, позволяющая очень дешево вычислить в какой-то точке t последовательность $T_0(t)$, $T_1(t)$, ...:

$$T_0(t) = 1, \quad T_1(t) = t, \quad \ldots, \quad T_{k+1}(t) = 2t\, T_k(t) - T_{k-1}(t).$$

Таким образом, вычисление $\sum_{k=0}^{p} a_k\, T_k(t)$ стоит, как нетрудно подсчитать, примерно $2p$ умножений и $2p$ сложений.

§ 4. Вычисление определенных интегралов

Речь пойдет об одной из самых распространенных в анализе операций — вычислении определенного интеграла

$$\int_a^b f(t)\, dt.$$

Весьма общий подход состоит в том, чтобы аппроксимировать функцию $f(t)$ какой-то другой функцией $\tilde{f}(t)$, для которой интеграл вычисляется аналитически. Итак, строим $f(t)$ с оценкой

$$|f(t) - \tilde{f}(t)| \leqslant \varepsilon, \quad \forall\, t \in [a, b],$$

и полагаем приближенно

$$\int_a^b f(t)\, dt \approx \int_a^b \tilde{f}(t)\, dt$$

с очевидной оценкой погрешности $\varepsilon(b-a)$.

Введем на $[a, b]$ сетку $\{t_k\}_{n=0}^N$ и таблицу $\{f_n\}_{n=0}^N$, являющуюся ограничением подынтегральной функции f на сетку. Рассмотрим несколько простых вариантов построения $\tilde{f}$, приводящих к широко распространенным формулам.

1. Функция $\tilde{f}(t)$ строится как кусочно-линейная интерполяция $\{f_n\}_{n=0}^N$ на равномерной сетке с шагом $\tau = (b-a)/N$. Очевидно, что

$$\int_a^b \tilde{f}(t)\,dt = \sum_{n=0}^{N-1} 0.5\,(f_n + f_{n+1})\,\tau = \\ = \tau\,(0.5\,f_0 + f_1 + f_2 + \ldots + f_{N-1} + 0.5\,f_N).$$

Эта формула известна как *формула трапеций*. Формулы такого сорта $\left(\sum C_n f_n\right)$ называют *механическими квадратурами*, C_n — коэффициентами (весами) квадратуры, t_n — ее узлами.

Точность формулы трапеций зависит от гладкости f. Если $f \in \mathrm{Lip}(C)$, то $|f(t) - \tilde{f}(t)| \leqslant 0.5C\tau$ и погрешность формулы трапеций не превосходит $0.5C\tau(b-a)$. Если f на $[a, b]$ имеет вторую производную, ограниченную числом C, то линейная интерполяция на каждом малом интервале есть интерполяционный полином первой степени: $|f(t) - \tilde{f}(t)| \leqslant 0.5\tau^2 C$ и погрешность формулы трапеций не превосходит $0.5\tau^2 C(b-a)$.

2. Еще более популярна формула Симпсона. Она так же строится на основе равномерной сетки, содержащей четное число интервалов. Находится таблица $\{f_n\}_{n=0}^{2N}$ и $\tilde{f}(t)$ строится как кусочно-квадратичная интерполяция, т.е. на каждой паре интервалов $(t_{2n}, t_{2n+1}, t_{2n+2})$ по значениям $f_{2n}, f_{2n+1}, f_{2n+2}$ строится интерполяционный полином Лагранжа второй степени.

Несложные выкладки дают

$$\int_{t_{2n}}^{t_{2n+2}} \tilde{f}(t)\,dt = \frac{\tau}{3}\,(f_{2n} + 4f_{2n+1} + f_{2n+2}).$$

Суммируя для $n = 0, 1, \ldots, N-1$, получаем

$$\int_a^b \tilde{f}(t)\,dt = \frac{\tau}{3} \sum_{n=0}^{N-1} (f_{2n} + 4f_{2n+1} + f_{2n+2}) = \\ = \frac{\tau}{3}\,(f_0 + 4f_1 + 2f_2 + 4f_3 + \ldots + 2f_{2N-2} + 4f_{2N-1} + f_{2N}).$$

Формула легко запоминается, ее точность легко оценивается. Если функция f имеет третью производную и $|f'''(t)| \leqslant C$, то $|\tilde{f}(t) - f(t)| \leqslant C\tau^3/3$ и погрешность формулы Симпсона не превосходит $C\tau^3(b-a)/3$, $\tau = (b-a)/2N$.

Однако теоретические оценки не очень популярны среди практиков. Если нужно вычислить интеграл с погрешностью ε, то мало кто

сначала оценит третью производную функции f и вычислит шаг сетки $\tau = [(3\varepsilon/C(b-a))]^{1/3}$. Дело, конечно, в том, что сама оценка завышена и тем более завышена константа C, особенно если функция f задана сложным алгоритмом. Поступают иначе. Вычисляя интеграл с небольшим числом узлов ($N = 2 \div 3$), получают число S_N; вычисляя интеграл с удвоенным N, получают S_{2N}. Если модуль $|S_{2N} - S_N| < \varepsilon$, число S_{2N} считают ответом с требуемой точностью. В противном случае вычисляют еще S_{4N} и сравнивают $|S_{4N} - S_{2N}| \leqslant \varepsilon$, и т.д. Нужно иметь в виду, что для гладких функций f часто интеграл вычисляется очень точно при неожиданно малом числе узлов.

Таблица 5

t_n	f_n	$N=1$	$N=2$	$N=4$
0	1.00000	1	1	1
0.25	1.2840254			4
0.5	1.6487213		4	2
0.75	2.1170000			4
1.0	2.7182818	4	2	2
1.25	3.4903425			4
1.5	4.4816891		4	2
1.75	5.7546026			4
2.0	7.3890561	1	1	1

Поясним на конкретном примере полезный прием, позволяющий заметно повысить точность ответа, когда известны $S_N, S_{2N}, \ldots$ (это так называемая экстраполяция Ричардсона). Вычислим, например,

$$\int_0^2 e^t\,dt = e^2 - 1 = 6.389056098.$$

В табл. 5 представлены $\{f_n\}$ и коэффициенты квадратур Симпсона C_n для N, равных 1, 2, 4. По таблице легко вычисляются значения

$$S_1 = 6.42072776, \quad S_2 = 6.391210176, \quad S_3 = 6.3891937.$$

Относительные погрешности этих величин суть 0.5, 0.03, 0.002 %. Для многих инженерных приложений даже погрешность 0.5 % считается малой.

Экстраполяция Ричардсона. Вычисление значений S_N для нескольких N открывает возможность в качестве «бесплатного приложения» получить значительно более точное значение интеграла. Это достигается простой процедурой экстраполяции полученных значений. Идея очень проста. Пусть для величины S имеются приближенный метод вычисления с малым параметром τ и теоретическая оценка $S = S_\tau + C\tau^p + o(\tau^p)$. Постоянная C, конечно, неизвестна, порядок же погрешности p известен.

Если вычислены две величины $S_{2\tau}$ и S_τ, мы имеем два следующих соотношения:

$$S = S_{2\tau} + C(2\tau)^p + o(\tau^p), \qquad S = S_\tau + C\tau^p + o(\tau^p).$$

Из них можно найти S с точностью до $o(\tau^p)$, исключив главный член погрешности. Умножая второе уравнение на 2^p и вычитая из него первое, получаем

$$S = (2^p S_\tau - S_{2\tau})/(2^p - 1) + o(\tau^p).$$

В табл. 6 представлены результаты: значения S_τ, вычисленные по формуле Симпсона, и значения S, полученные экстраполяцией Ричардсона по значениям S_1, $S_{0.5}$ и $S_{0.5}$, $S_{0.25}$. Для всех величин приведена погрешность (абсолютная).

Выше было указано, что погрешность формулы Симпсона не превосходит $o(\tau^3)$, поэтому экстраполяция проводилась сначала при $p = 3$. Результат оказался обескураживающим: экстраполяция не только не повысила точности, но и дала заметно худший результат. В чем же дело? Причина кроется в очень важном обстоятельстве, которое никогда не следует забывать, применяя подобную экстраполяцию. Обычно порядок погрешности p устанавливается на основе теоретиче-

Таблица 6

Шаг τ	S_τ	Погрешность	$S(p=3)$	$S(p=4)$	Погрешность
1.	6.42072776	0.032			
0.5	6.39121012	0.0022	6.38699	6.389242	$2\cdot10^{-4}$
0.25	6.3891937	0.00014	6.388905	6.3890593	$3\cdot10^{-6}$

ских оценок, которые, как правило, дают завышенное значение. Завышение оценки может влиять как на коэффициент при τ^p, так и на саму степень. Фактически точность может быть более высокой. Те же соображения могут быть использованы в условиях, когда не только коэффициент C в оценке погрешности, но и степень p не известны. Разумеется, теперь надо иметь три приближенных значения, чтобы исключить член $C\tau^p$ с двумя неизвестными. Формулы выводятся просто:

$$S_\tau = S + C\tau^p + o(\tau^p),$$

$$S_{2\tau} = S + 2^p C\tau^p + o(\tau^p),$$

$$S_{4\tau} = S + 2^p 2^p C\tau^p + o(\tau^p).$$

Пренебрегая членами $o(\tau^p)$, имеем

$$(S_{4\tau} - S_{2\tau})/(S_{2\tau} - S_\tau) \approx 2^p.$$

Используя числа табл. 6, получаем $2^p \approx 14.64 = 2^{3.87}$. Поскольку порядок точности формул обычно является целым числом, положим $p = 4$, относя разницу на счет величин $O(\tau^p)$. Если теперь проделать экстраполяцию при $p = 4$, получим числа, представленные в той же табл. 6. Эффект уточнения явно виден и не нуждается в комментариях. Однако читатель должен усвоить и другой урок: применение этой экстраполяции требует определенной осторожности, особенно в сложных задачах, в которых нет полной ясности с главным членом асимптотики погрешности.

Что касается формулы Симпсона, то более аккуратный анализ показывает, что она на самом деле имеет четвертый порядок точности. Это дополнительное повышение ее точности — следствие симметричности формулы. Симметричность часто и в других случаях приводит к повышению точности. Например, формула численного дифференцирования (односторонняя, несимметричная) и формула центральной разности (симметричная) имеют первый и второй порядки точности соответственно, хотя они одинаковы по трудоемкости (каждая «стоит» двух вычислений функции).

Несколько слов о чувствительности формулы Симпсона по отношению к возможной неточности априорного предположения о функции f.

Пусть на самом деле функция $f(t)$ лишь кусочно-гладкая, для простоты, имеет один разрыв на $[a, b]$, а в остальном гладкая. Нетрудно понять, что если разрыв попал на четный узел сетки, точность формулы сохраняется. При произвольном положении разрыва испортится только значение интеграла по той паре шагов сетки, в которую попал разрыв. Этот интеграл и вычисленный по Симпсону не имеют между собой ничего общего кроме того, что оба суть $O(\tau)$. Следовательно, точность вычисления интеграла станет $O(\tau)$ вместо ожидаемой $O(\tau^4)$.

Вычисление интегралов с особенностями. Рассмотрим простой пример — приближенное вычисление интеграла

$$\int_0^1 \frac{\cos t}{\sqrt{t}}\, dt.$$

Подынтегральная функция обращается в бесконечность при $t = 0$, но интеграл существует. Попытка его прямого вычисления по формуле Симпсона сразу же приведет к неудаче: первое слагаемое f_0

обращается в бесконечность. Грамотный студент легко находит выход: надо вычислить интеграл

$$\int_{\varepsilon}^{1} \frac{\cos t}{\sqrt{t}}\, dt, \qquad \varepsilon > 0;$$

при достаточно малом ε он приближает нужное значение. И даже оценка связанной с этим погрешности легко вычисляется: это величина

$$\int_{0}^{\varepsilon} \frac{1}{\sqrt{t}}\, dt = 2\sqrt{\varepsilon}.$$

Некоторые даже догадаются взять в качестве приближенного решения сумму

$$2\sqrt{\varepsilon} + \int_{\varepsilon}^{1} \frac{\cos t}{\sqrt{t}}\, dt,$$

где интеграл по $[\varepsilon, 1]$ вычисляется, например, по формуле Симпсона.

В вычислительной математике многие проблемы допускают решения на уровне студенческой грамотности. Нередко на основе подобных соображений работы пишутся людьми далеко не студенческого возраста, знающими о проблемах этой науки с чужих слов и не «чувствующих» такого важнейшего фактора, как число операций (речь идет, конечно, о более сложных задачах). И здесь приведенный выше, вполне правильный ответ не устраивает профессиональных вычислителей именно в силу нерационально большого объема вычислений.

Аккуратное вычисление интеграла с особенностью может быть выполнено гораздо более экономными средствами. Это достигается с помощью приема *регуляризации*, или *выделения особенности*. Поясним его в более общей ситуации. Пусть требуется вычислить

$$\int_{0}^{1} \frac{f(t)}{\sqrt{t}}\, dt,$$

где $f(t)$ — гладкая функция. Регуляризация состоит в том, что проделывается тождественное преобразование

$$\int_{0}^{1} f(t)\, t^{-1/2}\, dt = \int_{0}^{1} [f(t) - \varphi(t)]\, t^{-1/2}\, dt + \int_{0}^{1} \varphi(t)\, t^{-1/2}\, dt.$$

Функция $\varphi(t)$ выбирается такой, чтобы первый интеграл правой части не содержал особенности и при небольшом объеме вычислений достаточно точно определялся хотя бы по формуле Симпсона. Второй интеграл особенность содержит, но вычисляется аналитически.

В данном случае цель будет достигнута, если в качестве $\varphi(t)$ взять отрезок ряда Тейлора $f(t)$ в точке $t=0$. Это приводит к вычислению

$$\int_0^1 \frac{f(t)-f(0)-tf'(0)}{\sqrt{t}}\,dt + f(0)\int_0^1 \frac{1}{\sqrt{t}}\,dt + f'(0)\int_0^1 \sqrt{t}\,dt.$$

В примере, с которого мы начали ($f(t)=\cos t$), приходим к вычислению

$$\int_0^1 (\cos t - 1)\,t^{-1/2}\,dt + \int_0^1 t^{-1/2}\,dt.$$

Второе слагаемое есть 2, первое вычислим по формуле Симпсона: сначала с шагом 0.5, что даст значение 1.807967, затем с шагом 0.25, что даст значение 1.808850. Эти вычисления «стоили» всего четырех вычислений подынтегральной функции. Поучительно сравнить их с тем, сколько вычислений этой функции потребуется при «студенческом» рецепте (для достижения такой же точности).

Вычисление интегралов от быстроосциллирующих функций. Начнем с простого примера. Пусть требуется вычислить

$$\int_0^\pi e^{-t}\sin kt\,dt$$

при большом значении k, например $k=100$. Интегралы типа

$$\int f(t)\sin kt\,dt,$$

где $f(t)$ — гладкая функция, часто приходится вычислять в некоторых разделах физики. Сложность задачи состоит в том, что подынтегральная функция совершает большое число колебаний. Вычисление интеграла по стандартной формуле Симпсона, конечно, возможно, но требует сетки с очень малым шагом: каждая волна должна быть описана некоторым числом узлов сетки, а волн много.

Дело осложняется еще и тем, что вычисление должно проводиться с высокой точностью, так как результат есть сумма большого числа близких величин с противоположными знаками (интегралов от отдельных волн подынтегральной функции), происходит сильное сокращение знаков и для обеспечения точности остатка (результата) отдельные слагаемые должны вычисляться с существенно более высокой точностью. Для вычисления подобных интегралов используется следующий прием: гладкая функция $f(t)$ аппроксимируется некоторой другой гладкой функцией $\tilde{f}(t)$, такой, чтобы интеграл от $\tilde{f}(t)\sin kt$ вычислялся аналитически.

Итак, дело сводится к тождественному преобразованию

$$\int_0^\pi f(t)\sin kt\,dt = \int_0^\pi \tilde{f}(t)\sin kt\,dt + \int_0^\pi [f(t)-\tilde{f}(t)]\sin kt\,dt.$$

Второе слагаемое является малым и отбрасывается. Правда, если оценить отбрасываемую величину, опираясь только на оценку типа $|f(t)-\tilde{f}(t)| \leqslant \varepsilon$, т.е. в данном случае величиной $\pi\varepsilon$, ничего хорошего (даже если ε — точная оценка погрешности аппроксимации) не получится, так как величина $\pi\varepsilon$ может оказаться значительно большей интересующего нас интеграла. На самом деле погрешность существенно меньше. Это ведь интеграл от гладкой функции, не превосходящей ε, умноженной на быстроосциллирующую функцию. Естественно ожидать, что погрешность будет во столько раз меньше результата, во сколько раз $|f-\tilde{f}|$ меньше f. При $f(t)=e^{-t}$ интеграл вычисляется аналитически. Поучительно вычислить его приближенно, заменив функцию e^{-t} ее интерполяционным полиномом всего лишь второй степени. Интересно, совпадает ли точность результата с ожиданиями? На этой идее построена формула механической квадратуры, аналогичная формуле Симпсона. Интервал интегрирования разбивается на четное число шагов длиной τ, на каждой паре шагов функция заменяется интерполяционным полиномом второй степени $L_2(t)$, интегралы от $L_2(t)\sin kt$ вычисляются аналитически, полученные выражения суммируются.

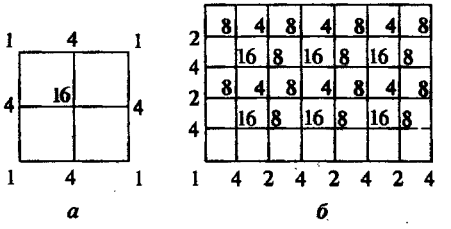


Рис. 7

Вычисление многомерных интегралов. Метод Монте-Карло. Рассмотрим задачу вычисления интеграла по многомерному кубу

$$\int_0^1\int_0^1\cdots\int_0^1 f(x_1, x_2, \ldots, x_n)\,dx_1\,dx_2\ldots dx_n.$$

Нетрудно и здесь построить формулы механических квадратур, аналогичные, например, формуле Симпсона. Проще всего такие формулы получить, используя процедуру повторного интегрирования, т.е. заменяя многомерный интеграл на равный ему

$$\int_0^1 dx_1\int_0^1 dx_2\ldots\int_0^1 f(x_1, x_2, \ldots, x_n)\,dx_n.$$

В двумерном случае для «элементарной ячейки» имеем коэффициенты (аналогичные коэффициентам 1, 4, 1 в одномерной формуле Симпсона), изображенные на рис. 7*а*. Суммируя интегралы по элементарным ячейкам, получаем коэффициенты, частично изображенные на рис. 7*б*. Эти коэффициенты поставлены около точек сетки, покрывающей двумерный квадрат. Вычисление интеграла по такой квадратурной формуле состоит в вычислении значений f в узлах сетки, умножении на соответствующий коэффициент и суммировании результатов.

Аналогичные формулы можно построить и в кубах бóльших размерностей, но пользы от этого мало. Дело в том, что при росте размерности объем вычислений катастрофически растет. Операция интегрирования, справедливо считающаяся одной из самых элементарных в анализе, практически является одной из самых сложных, точнее трудоемких. Поэтому описанные выше квадратуры, основанные на регулярных сетках, используются в практических вычислениях лишь для двумерных и трехмерных пространств.

При переходе к вычислению интегралов по кубам большей размерности применяется другой метод, получивший название *метода Монте-Карло*. Он состоит в том, что с помощью специальных быстро работающих алгоритмов генерируется последовательность «случайных» точек единичного n-мерного куба $x^1, x^2, \ldots, x^M \in R^n$, в каждой точке вычисляется значение $f_m = f(x^m)$, и в качестве приближенного значения интеграла принимают величину $M^{-1} \sum_m f_m$.

Что касается алгоритмов, генерирующих случайные точки (соответствующие программы называют «датчиками случайных чисел»), то их разработка — отдельная достаточно тонкая наука. Эти точки должны быть «равномерно распределены» в n-мерном кубе, т.е. они не должны «сбиваться в кучу» и оставлять в кубе «пустоты», в которые долго не попадают генерируемые точки.

Можно еще иначе пояснить требование равномерности распределения случайных точек. Выделим в кубе некоторую часть σ не слишком причудливой формы и не слишком малой меры. Тогда число точек последовательности $x^1, x^2, \ldots, x^M$, попавших в σ, должно быть близко к M mes σ и не только асимптотически (при $M \to \infty$), но и при конечных, не слишком больших M. Разумеется, если мера mes σ мала, такое свойство проявляется лишь при очень больших длинах последовательности случайных чисел (поскольку случайные числа выдает программный датчик, работающий детерминированно, их называют *псевдослучайными*).

Ограничимся этим поверхностным описанием, дающим самое общее представление о важном разделе вычислительной математики. Добавим еще несколько слов о вычислении интеграла от f по слож-

ной области Ω. Пусть область Ω есть часть единичного куба, выделенная неравенствами $g_j(x) \leqslant 0$, $j = 1, 2, \ldots, J$. Техника вычисления интеграла по Ω выглядит примерно так. Генерируется та же последовательность псевдослучайных чисел, равномерно распределенных в единичном кубе. Для каждой точки x^m проверяются неравенства $g_j(x^m) \leqslant 0$, $j = 1, 2, \ldots, J$. Если все они выполнены, т.е. $x^m \in \Omega$, вычисляется $f(x^m)$, которая прибавляется к накапливающейся сумме.

Кроме того, ведется подсчет числа попавших в Ω точек. Пусть вычислено M случайных точек, K из которых попало в Ω, и накоплена соответствующая сумма $\sum f(x^{m_k})$. Среднее значение f в области Ω можно определить величиной

$$\int\limits_{\Omega} \frac{f(x)}{\operatorname{mes}\Omega}\, dx.$$

Вместе с тем это среднее вычисляется методом Монте-Карло: $K^{-1}\sum f(x^{m_k})$. Что касается меры mes Ω, то она приближенно равна K/M.

Итак, имеем формулу

$$\int\limits_{\Omega} f(x)\, dx = \frac{1}{M}\sum_{k=1}^{K} f(x^{m_k}) + \varepsilon_k.$$

О погрешности ε_k известно, что это случайная величина, математическое ожидание ее модуля имеет оценку $C/\sqrt{K}$; константа C зависит от гладкости f.

Квадратуры высокой точности. В качестве $\tilde{f}$ можно взять интерполяционный полином $L_N(t)$, построенный на сетке $\{t_n\}_{n=0}^{N}$. За счет того или иного выбора узлов сетки можно получить соответствующие преимущества. Широко используются чебышевские сетки, обеспечивающие устойчивость чебышевских квадратур при больших N. Отметим еще гауссовы квадратуры. Они основаны на следующей идее. При любой сетке квадратура, в которой используется интерполяционный полином L_N, точна в классе полиномов степени не выше N. Можно так подобрать узлы сетки, что квадратура станет точной в классе полиномов степени не выше $2N + 1$. Узлы сеток и коэффициенты квадратурных формул вычислены для разных N на стандартном интервале $[-1, 1]$. Их можно найти в справочниках по методам приближенных вычислений (гауссовы узлы и коэффициенты, чебышевские и некоторые другие).

§ 5. Численное интегрирование задачи Коши для систем обыкновенных дифференциальных уравнений

Рассмотрим задачу Коши для системы дифференциальных уравнений. Требуется найти функцию $x(t)$, $0 \leqslant t \leqslant T$, удовлетворяющую уравнению

$$\frac{dx}{dt} = f(x, t), \quad x(0) = x_0, \quad 0 \leqslant t \leqslant T. \tag{1}$$

Здесь x — p-мерный вектор, $f(x, t)$ — заданная вектор-функция той же размерности, x_0 — заданная точка (данные Коши).

Как известно, при весьма простых и общих предположениях о гладкости f решение задачи (1) существует и единственно. Хорошо также известно, что найти явное аналитическое выражение для $x(t)$ удается в крайне редких случаях, и только для очень частных классов задач (например, линейных уравнений) имеются способы явного решения задачи Коши. Сама же задача (1) — одна из наиболее часто встречающихся в различных приложениях (в физике, механике, астрономии, биологии, экономике и т.д.). К таким задачам приходят, изучая движения планет и ракет, эволюцию биологических и экономических систем.

Рассмотрим методы приближенного интегрирования задачи Коши, начиная с самых простых и старых. Используем *метод сеток*, или, иначе, *метод конечных разностей*, являющийся одним из наиболее универсальных и общих (хотя и достаточно трудоемких) методов приближенного решения дифференциальных уравнений.

Начнем с основных объектов метода сеток. На интервале $[0, T]$, на котором ищется решение, введем покрывающую его дискретную *сетку* точек

$$0 = t_0 < t_1 < t_2 < \ldots < t_N = T, \qquad \text{или} \quad \{t_n\}_{n=0}^{N}.$$

Ради простоты в дальнейшем будем в основном использовать равномерную сетку с шагом $\tau = T/N$: $t_n = n\tau$.

Приближенное решение будем искать в виде *сеточной функции*, т.е. в виде функции дискретного аргумента n; обозначим ее как $\{x_n\}_{n=0}^{N}$. Напомним, что каждый x_n есть p-мерный вектор. С содержательной точки зрения x_n будет представлять (приближенно) значение искомой функции $x(t)$ в узле t_n: $x_n \approx x(t_n)$.

Сеточная функция $\{x_n\}$ не может удовлетворять никакому дифференциальному уравнению, и нужно построить какие-то другие уравнения, из которых можно было бы найти функцию $\{x_n\}$ и притом так, чтобы она была приближенным решением исходной задачи.

Такие уравнения (так называемые *разностные уравнения*) строятся очень просто: входящие в дифференциальные уравнения производные заменяются соответствующими разностями. Это можно сделать разными способами, и они приводят к разным уравнениям. Например,

$$\frac{x_{n+1}-x_n}{\tau} = f(x_n, t_n) \qquad (\textit{явная схема} \text{ Эйлера}), \tag{2}$$

$$\frac{x_{n+1}-x_n}{\tau} = f(x_{n+1}, t_{n+1}) \quad (\textit{неявная схема} \text{ Эйлера}), \tag{3}$$

$$\frac{x_{n+1}-x_{n-1}}{2\tau} = f(x_n, t_n) \qquad (\text{схема с центральной разностью}). \tag{4}$$

Ограничимся пока этими простыми примерами, ниже мы рассмотрим и более сложные схемы.

Первая проблема, с которой мы сталкиваемся при решении разностных уравнений, — это фактическое определение x_n. Во всех случаях значения x_n определяются последовательно, слева направо. Особенно просто вычисляются x_n при использовании явной схемы Эйлера:

$$x_{n+1} = x_n + \tau f(x_n, t_n), \qquad n = 0, 1, \ldots, N-1.$$

Здесь x_0 известно (данные Коши). В правой части этой формулы используются уже найденные значения x_n.

В случае неявной схемы Эйлера ситуация несколько сложнее. Пусть значение x_n уже найдено. Тогда x_{n+1} находится из уравнения

$$x_{n+1} = x_n + \tau f(x_{n+1}, t_{n+1}), \tag{5}$$

которое является нелинейным относительно неизвестного x_{n+1}. Правда, это слабая нелинейность, так как перед f стоит малый множитель τ, что делает уравнение (5) не таким уж сложным. Интуитивно ясно, что x_{n+1} мало отличается от x_n, т.е. x_n есть очень хорошее начальное приближение для какого-либо итерационного метода определения x_{n+1} (см. § 1).

В схеме (4) мы сталкиваемся с характерным явлением: формальные порядки дифференциального и разностного уравнений не совпадают. Под формальным порядком мы понимаем число произвольных постоянных в общем решении, или, если угодно, число дополнительных данных, полностью определяющих решение. В схемах Эйлера (2) и (3), как и в дифференциальном уравнении, достаточно задать x_0, чтобы все остальные значения $x_1, x_2, \ldots$ определялись однозначно. В схеме (4) ситуация иная: только задав x_0 и x_1, мы определим $x_2 = x_0 + \tau f(x_1, t_1)$ и т.д.

Значение x_0 определяется постановкой задачи, x_1 формально можно задать каким угодно. Фактически, конечно, если мы стремимся получить хорошие результаты, x_1 нужно задавать достаточно аккуратно. Например, приемлемым (но не лучшим) является $x_1 = x_0$. Лучше (и правильнее) вычислять x_1 по какой-либо более простой разностной схеме, например по явной схеме Эйлера: $x_1 = x_0 + \tau f(x_0)$.

Обсудим некоторые вопросы, естественно возникающие при построении и использовании разностных схем. Первый вопрос: чем следует руководствоваться при построении разностных уравнений? Интуитивно ясно, что разностные уравнения (2)—(4) имеют прямое отношение к исходному дифференциальному уравнению хотя бы потому, что в пределе (при $\tau \to 0$) они переходят в (1). Но это соображение нужно четко оформить и дать ему количественное выражение. Таковым будет фундаментальное в вычислительной математике понятие аппроксимации.

Второй вопрос. Пусть разностное уравнение явно «соответствует» дифференциальному. Но значит ли это, что его решение $\{x_n\}$ в какой-то мере аппроксимирует решение дифференциального уравнения? Так мы приходим к другому фундаментальному понятию — к сходимости разностной схемы. Правильнее и аккуратнее было бы сказать: сходимость решения разностного уравнения к решению дифференциального при $\tau \to 0$. Но мы будем использовать «жаргонное» выражение «сходимость».

Перейдем к аккуратному оформлению этих понятий. Рассмотрим следующие математические объекты:

1) дифференциальное уравнение

$$\frac{d\mathscr{X}}{dt} = f(\mathscr{X}, t), \quad \mathscr{X}(0) = x_0$$

и его решение $\mathscr{X}(t)$;

2) ограничение функции $\mathscr{X}(t)$ на сетку $\{t_n\}$, т.е. сеточную функцию $\{\mathscr{X}_n\}_{n=0}^N$, где $\mathscr{X}_n = \mathscr{X}(t_n)$;

3) разностное уравнение (например, явную схему Эйлера или еще какую-нибудь) и его решение $\{x_n\}$.

Мы будем иметь дело с последовательностью сеток, соответствующих уменьшающимся шагам $\tau \to 0$, так что при стремлении к полной аккуратности следовало бы каждый объект пометить еще индексом τ: $\mathscr{X}_n^\tau$, x_n^τ и т.д. Этот индекс τ неявно везде следует иметь в виду.

Определение. Говорят, что разностное решение сходится, если

$$\lim_{\tau \to 0} \|\mathscr{X}_n^\tau - x_n^\tau\| = 0, \quad n = 0, 1, \ldots, N = T/\tau.$$

Если установлена оценка

$$\|\mathscr{X}_n^\tau - x_n^\tau\| \leqslant C\tau^p, \quad n = 1, 2, \ldots, N,$$

где C не зависит от τ и n, то говорят, что установлен p-й порядок сходимости, а схема имеет p-й порядок точности.

Установление сходимости, а лучше, и порядка сходимости (еще лучше, с хорошей оценкой константы C) есть основная цель теоретического обоснования метода приближенного решения. Этому вопросу посвящен § 7.

Схема Адамса. Опишем общую конструкцию схем численного интегрирования, достоинством которой является ее экономичность. Каждый шаг интегрирования требует только одного вычисления правой части f, в то же время порядок точности метода может быть (формально) любым желаемым. В методах Рунге—Кутты (они описаны в § 7) число вычислений f на шаг равно порядку точности метода.

Итак, пусть задача решается на равномерной сетке, значения x_n (и все предшествующие x_{n-1}, x_{n-2}, ..., x_0) уже найдены. По значениям $f_i = f(x_{n-p+i})$ для $i = 0, 1, \ldots, p$ (p определяет порядок точности метода) в узлах $t^i = t_{n-p+i}$ построим интерполяционный полином степени p:

$$L_p(t, \{t^i\}, \{f_i\}) = \sum_{i=0}^{p} l_p^i(t)\, f_i.$$

Его можно применить и для экстраполяции функции $f[x(t)]$ на интервале $[t_n, t_n + \tau = t_{n+1}]$.

Теперь используем очевидное тождество

$$x(t_n + \tau) = x(t_n) + \int_{t_n}^{t_n+\tau} f[x(t)]\, dt. \tag{6}$$

Заменяя $f[x(t)]$ интерполяционным полиномом и вычисляя интеграл, получаем формулу

$$x_{n+1} = x_n + \tau(a_0 f_p + a_1 f_{p-1} + \ldots + a_p f_0), \tag{7}$$

где $a_0, a_1, \ldots, a_p$ — некоторые универсальные (не зависящие от шага τ) числа. Они очевидным образом вычисляются через интегралы от базисных интерполяционных полиномов $l_p^i(t)$. При вычислении a_i делается замена переменных $t = t_n + \xi\tau$ и рассматривается стандартный шаг по ξ, равный единице.

Оценим погрешность аппроксимации, предполагая $f(x)$, а следовательно, и решение $x(t)$ достаточно гладкими. Погрешность экстраполяции $\|f[x(t)] - L_p(t)\| = O(\tau^{p+1})$ (см. § 3). При интегриро-

вании по $[t_n, t_{n+1}]$ в оценке погрешности вычисления интеграла появляется еще множитель τ. Переписывая (6), (7) в форме, дающей в пределе дифференциальное уравнение, получаем

$$\frac{x(t_n+\tau)-x(t_n)}{\tau} = a_0 f(x_n) + a_1 f(x_{n-1}) + \ldots + a_p f(x_{n-p}) + O(\tau^{p+1}).$$

Итак, порядок погрешности аппроксимации равен числу используемых в (7) точек $p+1$.

Неудобством метода является необходимость помнить некоторое число прошлых значений f_{n-i}. Это, конечно, мелочь, если не считать самого начала процесса интегрирования, когда прошлого просто нет. Приходится несколько первых шагов выполнять нестандартно, например методом Рунге—Кутты (см. § 7).

Метод Адамса является характерным примером схемы, формальный порядок которой превышает порядок дифференциального уравнения. Стандартный алгоритм начинает работать лишь при задании, кроме начальных данных x_0, еще и значений $x_1, x_2, \ldots, x_p$. Таким образом, общее решение разностного уравнения содержит больше, чем нужно, произвольных постоянных и, следовательно, какие-то лишние «решения».

Полезно иметь представление о том, во что переходят лишние решения в пределе при $\tau \to 0$. Рассмотрим простейшее уравнение $\dot{x} = ax$ и две схемы второго порядка — примитивную схему (4) и квалифицированную схему Адамса второго порядка:

$$\frac{x_{n+1}-x_n}{\tau} = \frac{3}{2} f(x_n) - \frac{1}{2} f(x_{n-1}), \quad f(x) \equiv ax.$$

В этом простом случае можно вычислить и проанализировать общие решения разностных уравнений. Они ищутся в стандартной для однородных разностных уравнений с постоянными коэффициентами форме $x_n = C_1 q_1^n + C_2 q_2^n$, где q_1, q_2 — корни характеристического уравнения, C_1, C_2 — постоянные, определяющиеся в данном случае начальными данными x_0 и x_1.

Характеристические уравнения получаем, подставляя q^n в уравнение. Для простейшей схемы (4) имеем

$$q^2 - 2a\tau q - 1 = 0, \quad \text{т.е.} \quad q_{1,2} = a\tau \pm \sqrt{1 + (a\tau)^2}.$$

Первый корень (при $|a\tau| \ll 1$):

$$q_1 = 1 + a\tau + \frac{1}{2}(a\tau)^2 + O(\tau^3) = e^{a\tau} + O(\tau^3),$$

$$q_1^n = e^{an\tau}(1 + O(\tau^2)), \quad n \approx O(1/\tau).$$

Это решение в пределе дает решение дифференциального уравнения. Второй корень:

$$q_2 \approx -1 + a\tau, \qquad q_2^n \approx (-1)^n e^{-an\tau},$$

т.е. сеточная функция никакого разумного предела не имеет (из-за множителя $(-1)^n = (-1)^{t/\tau}$). Это есть паразитическое решение, появившееся из-за превышения порядка разностного уравнения над порядком дифференциального.

В общем решении $x_n = C_1 q_1^n + C_2 q_2^n$, и для того чтобы схема имела второй порядок точности, нужно обеспечить соотношения $C_1 = x_0 + O(\tau^2)$, $C_2 = O(\tau^2)$.

Аналогичные выкладки для схемы Адамса приводят к характеристическому уравнению

$$q^2 = q + \frac{3}{2} a\tau q - \frac{1}{2} a\tau.$$

Его корни:

$$q_{1,2} = \frac{1}{2} + \frac{3}{4} a\tau \pm \sqrt{\frac{1}{4} + \frac{1}{4} a\tau + \frac{9}{16} a^2\tau^2}.$$

Предоставим читателю убедиться, что

$$q_2 \approx a\tau/2, \qquad q_1 = 1 + a\tau + a^2\tau^2/2 + O(\tau^3) = e^{a\tau}\,(1 + O(\tau^3)).$$

Таким образом, $q_1^{t/\tau} = e^{at}(1 + O(\tau^2))$, а паразитическое решение $q_2^n \approx (a\tau)^n$ очень быстро стремится к нулю (мы, конечно, считаем $|a\tau| \ll 1$, например $a\tau \approx 0.1$).

Итак, выбор x_1 в схеме Адамса должен обеспечить соотношение $C_1 = x_0 + O(\tau^2)$. Полезно проверить, что выбор $x_1 = x_0 + \tau f(x_0)$ обеспечивает требуемые соотношения. Более высокий интеллектуальный уровень схемы Адамса (по сравнению с примитивной схемой с центральной разностью) сказался в том, что паразитическое решение этого метода очень быстро убывает — как $(a\tau)^{t/\tau}$.

Численное интегрирование на ЭВМ. Представленный выше анализ погрешностей приводит к выводу, что точность численного интегрирования тем выше, чем меньше шаг τ. Это верно только до известного предела — до тех пор, пока погрешности округления, связанные с конечной разрядностью машинной арифметики, остаются пренебрежимо малыми.

Реальная расчетная формула (метода Эйлера, для определенности) в действительности при реализации на ЭВМ имеет вид

$$x_{n+1} = x_n + \tau f(x_n) + \varepsilon_n.$$

Величина ε_n обычно определяется погрешностью округления при машинном представлении x_n, т.е. имеет величину $10^{-r}|x|$ ($r = 12$ на БЭСМ-6, $r = 6 \div 7$ на ЕС, $r = 16$ на ЕС при двойной точности). Погрешность округления τf, если f вычисляется с машинной точностью, несущественна, так как обычно $|\tau f| \ll |x|$ (x мало меняется за шаг). Но если функция f вычисляется сложным алгоритмом, ее погрешность может определять величину ε_n.

Таким образом, машинная вычислительная формула имеет вид

$$x^{\text{м}}_{n+1} = x^{\text{м}}_n + \tau f(x^{\text{м}}_n)\left(1 + \tau\delta + \frac{\eta|x|}{\tau|f|}\right),$$

где δ — относительная погрешность вычисления f, η — относительная погрешность представления x. Можно трактовать эту формулу как точную с погрешностью вычисления f. И теперь ясно видно, что, начиная с некоторых малых величин, дальнейшее уменьшение τ приводит к падению точности.

Интегрирование уравнений высокого порядка. Пусть требуется проинтегрировать уравнение

$$\frac{d^4x}{dt^4} = f(x), \qquad x(0), \ldots, \dddot{x}(0) \text{ заданы.}$$

Не составляет труда построить разностное уравнение:

$$\frac{x_{n+2} - 4x_{n+1} + 6x_n - 4x_{n-1} + x_{n-2}}{\tau^4} = f(x_n).$$

Данные Коши позволяют вычислить значения x_0, x_1, x_2, x_3, по ним находим x_4, затем x_5 и т.д. Однако здесь конечная разрядность машинной арифметики имеет еще худшие последствия. Машинная расчетная формула имеет вид

$$x_{n+2} = 4x_{n+1} - 6x_n + 4x_{n-1} - x_{n-2} + \tau^4 f(x_n) + \eta|x|.$$

Таким образом, погрешность округления $\eta|x|$ эквивалентна погрешности вычисления f порядка $\eta|x|/\tau^4$. К счастью, есть простой выход — переход от уравнения четвертого порядка к системе уравнений первого порядка:

$$\dot{x}^1 = x^2, \qquad \dot{x}^2 = x^3, \qquad \dot{x}^3 = x^4, \qquad \dot{x}^4 = f(x^1).$$

Именно по этой причине теория численного интегрирования строится для систем уравнений первого порядка.

Замечание. Выше без определений были использованы некоторые понятия (аппроксимация, точность и т.п.). Смысл их достаточно прозрачен. Он уточняется в следующем параграфе.

§ 6. Абстрактная форма приближенного метода

Приближенное интегрирование задачи Коши послужит нам удобным примером, на котором можно будет ввести основные объекты общего приближенного метода и установить связи между ними. Настоящий параграф носит, так сказать, идеологический характер, в нем появляются фундаментальные понятия теории приближенных методов вычислений.

Итак, мы исходим из задачи, записанной в общей форме:

$$L(\mathscr{Z}) = \mathscr{F}. \tag{1}$$

Здесь $\mathscr{Z}$ — искомый элемент некоторого функционального пространства X, $\mathscr{F}$ — некоторый заданный элемент пространства F, L — оператор, отображающий X в F ($\mathscr{F}$ мы будем называть иногда «правой частью уравнения»). Приближенное решение задачи (1) тем или иным способом сводится к решению уравнения

$$L_s(x_s) = \mathscr{F}_s. \tag{2}$$

Здесь x_s — искомый элемент некоторого конечномерного пространства X_s, $\mathscr{F}_s$ — элемент другого конечномерного пространства F_s, L_s — оператор, отображающий X_s в F_s.

По существу (2) есть конечная система (вообще говоря, нелинейных) уравнений. Поясним смысл индекса s (символ «сетки» в обобщенном смысле слова). Наличие индекса s связано с тем, что в теории численных методов мы имеем дело не с одной задачей (2), а с бесконечной последовательностью задач, с целым семейством, s — параметр семейства (который может быть не только скалярным, но и векторным). При интегрировании задачи Коши в роли параметра выступает шаг сетки τ. Нас будет интересовать предельный переход при $s \to 0$, т.е. точное решение $\mathscr{Z}$ задачи (1) должно быть пределом решений систем (2) при $s \to 0$. Однако еще предстоит ввести процедуру сравнения $\mathscr{Z}$ и x_s, ведь это элементы разных пространств.

Следующий элемент приближенного метода — некоторый оператор P_s, отображающий X в X_s. Мы еще вернемся к обсуждению этого оператора. Можно вычислить элемент $\mathscr{Z}_s = P_s\mathscr{Z} \in X_s$ и подставить его в уравнение (2). Конечно, $\mathscr{Z}_s$ не удовлетворяет уравнению (2), и появляется новый важный объект — *невязка*, или *погрешность аппроксимации*,

$$r_s = L_s(\mathscr{Z}_s) - \mathscr{F}_s. \tag{3}$$

Теперь можно установить связь между уравнениями (1) и (2). Пока что у них не было ничего общего, кроме использования одинаковых букв (L, $\mathscr{F}$ и т.д.).

Определение 1. Говорят, что семейство задач (2) аппроксимирует уравнение (1), если

$$\|r_s\| \to 0 \quad \text{при } s \to 0. \tag{4}$$

Если, кроме того, установлена оценка

$$\|r_s\| \leqslant C_1 |s|^p \quad (C_1 \text{ не зависит от } s), \tag{5}$$

говорят, что аппроксимация имеет порядок p по s. В общем случае s есть набор малых параметров, а p — соответствующий набор показателей.

Отметим важное требование: оценка (5) — равномерная на семействе задач (2), т.е. C_1 — универсальная для всего семейства постоянная. Если имеет место факт аппроксимации, значит уравнения (1) и (2) уже имеют между собой много общего, так как решение исходной задачи (1) в некотором смысле является «почти решением» уравнения (2).

Дальнейшее основано на следующем соображении. Приближенное решение x_s и образ точного решения $\mathscr{Z}_s$ удовлетворяют близким уравнениям: одно — уравнению (2), а второе — почти такому же уравнению, но с мало измененной правой частью (тем меньшей, чем меньше s):

$$L_s(\mathscr{Z}_s) = \mathscr{F}_s + r_s.$$

Можно надеяться, что их решения x_s и $\mathscr{Z}_s$ мало отличаются друг от друга. Для того чтобы это было так, нужно предположить семейство задач (2) устойчивым в следующем смысле.

Определение 2. Говорят, что семейство задач (2) устойчиво, если из отношений

$$L_s(x_s) - \mathscr{F}_s = \xi_s, \qquad L_s(y_s) - \mathscr{F}_s = \eta_s$$

следует

$$\|x_s - y_s\| \leqslant C_2(\|\xi_s\| + \|\eta_s\|).$$

(Здесь подчеркнем равномерность оценки: C_2 не зависит от s.) И наконец, дадим еще одно определение.

Определение 3. Говорят, что приближенное решение x_s задачи (2) сходится при $s \to 0$ к решению исходной задачи (1), если

$$\|x_s - \mathscr{Z}_s\| \to 0 \quad \text{при } s \to 0.$$

Если установлена оценка

$$\|x_s - \mathscr{Z}_s\| \leqslant C|s|^q \quad (C \text{ не зависит от } s),$$

говорят, что сходимость имеет порядок q.

Выше мы ввели три фундаментальных понятия теории приближенных методов: аппроксимация, устойчивость и сходимость. Связь

между ними устанавливает теорема «аппроксимация + устойчивость ⇒ сходимость».

Теорема. Пусть приближенная задача (2) аппроксимирует исходную задачу (1) и семейство задач (2) устойчиво. Тогда приближенное решение x_s сходится к решению исходной задачи $\mathscr{Z}$. Если аппроксимация имеет порядок p по s, то и сходимость имеет тот же порядок.

Доказательство. Приближенное решение x_s находится из уравнения $L_s(x_s) - \mathscr{F}_s = 0$, а образ точного решения $\mathscr{Z}_s = P_s\mathscr{Z}$ — из уравнения $L_s(\mathscr{Z}_s) - \mathscr{F}_s = r_s$, причем в силу аппроксимации $\|r_s\| \leqslant C_1|s|^p$. Тогда из предположения об устойчивости немедленно следует

$$\|x_s - \mathscr{Z}_s\| \leqslant C_1C_2|s|^p. \tag{6}$$

«Доказательство» закончено. Заметим, однако, что здесь есть еще один вопрос: что нам дает оценка (6)? Ведь нас интересует связь между x_s (это то, что вычислитель имеет, если он умеет решать задачу (2)) и $\mathscr{Z}$ (это то, что его интересует). Непосредственно сравнивать элементы разных пространств мы не можем. Значит, для того чтобы сходимость (6) была содержательно ценным фактом, нужно предположить какие-то важные свойства оператора P_s. Грубо говоря, он должен быть в некотором смысле «обратимым», более того, еще и «непрерывно обратимым». Это означает, что по $\mathscr{Z}_s$ мы должны уметь восстанавливать $\mathscr{Z}$. Реально же мы будем восстанавливать функцию $\mathscr{Z}$ по приближенному решению x_s, мало отличающемуся от $\mathscr{Z}_s$. Выше не случайно были использованы кавычки, так как в строгом смысле слова операторы P_s просто необратимы. Тем не менее сделанные для P_s естественные предположения о «непрерывной обратимости» в дальнейшем приобретут некоторое обоснование.

Приведем простые примеры операторов P_s. Нас больше всего будет интересовать оператор «ограничения функции на сетку» R_s (см. § 3).

В задаче Коши такой оператор строится очень просто. Пусть $\{t_n\}_{n=0}^N$ — сетка на интервале $[0, T]$, а $\mathscr{Z}(t)$ — определенная на этом интервале функция. Тогда $P_s\mathscr{Z}$ определяется как таблица чисел $\{\mathscr{Z}_n\}_{n=0}^N$, где $\mathscr{Z}_n = \mathscr{Z}(t_n)$. (Здесь s — символ сетки, или, если угодно, ее шаг τ; если сетка неравномерная, то $|s| = \max\limits_n |t_{n+1} - t_n|$.) «Обратным» к R_s будет тот или иной оператор интерполяции, который по таблице $\{\mathscr{Z}_n\}$ строит непрерывную функцию $\tilde{\mathscr{Z}}(t)$ (разумеется, не

совпадающую с $\mathscr{Z}(t)$, но близкую к ней при выполнении некоторых предположений о $\mathscr{Z}(t)$).

Рассмотрим еще один пример оператора P_s. Пусть функция $\mathscr{Z}(t)$ разлагается в ряд Фурье:

$$\mathscr{Z}(t) = \sum_{k=-\infty}^{\infty} C_k\, e^{ik\pi t/T}.$$

Тогда $P_s\mathscr{Z}$ определяется как конечномерный вектор $\{C_k\}_{k=-N}^{N}$. (Здесь под малым параметром s можно понимать $1/N$.) «Обратный» к P_s оператор очевиден: это конечная сумма

$$\tilde{\mathscr{Z}}(t) = \sum_{k=-N}^{N} C_k\, e^{ik\pi t/T}.$$

Очевидно, что оба оператора P_s необратимы в строгом смысле слова (на всём пространстве функций). Но они «почти обратимы» на подпространстве гладких функций. Оператор ограничения на сетку $P_s = R_s$ в качестве «обратного» имеет тот или иной интерполяционный аппарат, и смысл термина «почти обратим на гладких функциях» разъяснен в § 3. Такое его обращение сопровождается потерей информации, зависящей от вида интерполяции и гладкости функций. Во втором примере оператор P_s обратим на подпространстве конечных сумм Фурье ($C_k = 0$ при $|k| > N$), почти обратим (с малой потерей информации) на подпространстве гладких функций (у которых часть ряда Фурье с $|k| > N$ пренебрежимо мала).

Замечание 1. Внимательный читатель, наверное, обратит внимание на то, что погрешность аппроксимации, как невязка в уравнениях приближенного метода при подстановке в них $P_s\mathscr{Z}$, не определена однозначно, так как эти уравнения можно записать в тривиально эквивалентных разных формах, от которых, однако, существенно меняется невязка. Например, уравнения метода Эйлера можно записать в таких формах:

$$\tau^{-1}(x_{n+1} - x_n) = f(x_n), \quad x_{n+1} = x_n + \tau f(x_n),$$

$$\tau\,[x_{n+1} - x_n - \tau f(x_n)] = 0,$$

и т.д. Это одно и то же, но невязка имеет оценки $O(\tau)$, $O(\tau^2)$, $O(\tau^3)$ соответственно.

В принципе, здесь нет ничего страшного — ведь в теорию входит произведение оценок аппроксимации и устойчивости. Ум-

ножая уравнение на τ, мы «выигрываем» в аппроксимации, но ровно столько же проигрываем в устойчивости. Однако принято устранять эту неоднозначность, выбирая из всех форм ту, которая в пределе $\tau \to 0$ переходит в решаемое дифференциальное уравнение, и относительно такой нормировки считать порядок аппроксимации. Не будем давать строгих определений для абстрактной формулировки задачи. Вышесказанного достаточно, чтобы в любом, практически, случае была выбрана «каноническая» форма записи уравнений приближенного метода (имеются в виду в основном методы конечно-разностного типа).

Замечание 2. При изложении абстрактной теории мы не обсуждали вопросов выбора норм, хотя все это становится содержательной теорией только при том или ином конкретном их выборе. В каком-то смысле была изложена «инвариантная» относительно выбора норм схема теории. В нее входят разные варианты обоснования численного метода, отличающиеся выбором норм. Содержательно такие теории не все равноценны, и доказательства свойств аппроксимации и устойчивости могут в одной и той же схеме приближенного решения задачи сильно отличаться.

Разумеется, мы заинтересованы в том, чтобы из относительно слабых предположений получить возможно более сильные оценки отклонения приближенного решения от точного. Поэтому хотелось бы иметь в основном ключевом свойстве схемы — устойчивости возможно более слабую норму для погрешностей аппроксимации и возможно более сильную норму в оценке $\|x_s - y_s\|$, например чтобы норма невязки $\|r_s\|$ была аналогом какой-то интегральной нормы, а $\|x_s - y_s\|$ — нормой типа C. Поэтому не стоит удивляться, встречая разные обоснования одного и того же метода приближенного решения какого-то класса задач.

Замечание 3. Использованная нами форма записи уравнения $L(\mathscr{Z}) = \mathscr{F}$ и приближенного метода $L_s(x_s) = \mathscr{F}_s$ не является универсальной. Можно записывать задачу в форме $L(\mathscr{Z}, \mathscr{F}) = 0$ и т.д. Предоставим читателю, если ему это покажется интересным, соответствующим образом скорректировать абстрактную форму записи приближенного метода, определения аппроксимации, устойчивости и т.п. Нам будет достаточно и такого уровня абстракции.

Замечание 4. Доказанная выше теорема «аппроксимация + устойчивость $\Rightarrow$ сходимость» применительно к методу конечных разностей установлена В. С. Рябеньким и А. Ф. Филипповым и носит их имя. В западной литературе аналогичная теорема называется теоремой Лакса. Впервые, видимо, теоремы такого типа в различных методах приближенного решения задач в функциональных пространствах доказывал Л. В. Канторович.

§ 7. Исследование сходимости методов Рунге–Кутты

Применим описанную выше абстрактную схему исследования к конкретному вопросу — к обоснованию сходимости метода конечных разностей для системы обыкновенных дифференциальных уравнений (5.1). Мы должны доказать аппроксимацию и устойчивость, а не принимать их как предположения. Начнем с аппроксимации для простейшей схемы — явного метода Эйлера (5.2). Прежде всего, нужно четко оформить разностные уравнения в абстрактной форме $L^\tau(x^\tau) = 0$. Под x^τ мы будем понимать совокупность $\{x_n\}$, $n = 0, 1, \ldots, N = T/\tau$. Оператор L^τ отображает сеточную функцию x^τ в такую же сеточную функцию, и надо определить все ее компоненты.

Положим

$$[L^\tau(x^\tau)]_n = \begin{cases} x_0 - x_0^*, & n = 0 \\ \dfrac{x_n - x_{n-1}}{\tau} - f(x_{n-1}, t_{n-1}), & n = 1, 2, \ldots, N. \end{cases}$$

Очевидно, теперь запись $L^\tau(x^\tau) = 0$ эквивалентна разностной задаче. Вычисление погрешности аппроксимации состоит в том, что в уравнения $L^\tau(x^\tau) = 0$ подставляется ограничение на сетку точного решения $\{\mathscr{X}_n\}$, $\mathscr{X}_n = \mathscr{X}(t_n)$. Получим невязку

$$r_n = \begin{cases} \mathscr{X}(0) - x_0^* = 0, & n = 0, \\ \dfrac{\mathscr{X}_n - \mathscr{X}_{n-1}}{\tau} - f(\mathscr{X}_{n-1}, t_{n-1}) = O(\tau), & n = 1, 2, \ldots, N. \end{cases}$$

Таким образом, можно сформулировать следующее утверждение.

Утверждение. Пусть решение дифференциального уравнения $\mathscr{X}(t)$ имеет две ограниченные производные (для этого достаточно, чтобы $f(x, t)$ имела ограниченные первые производные). Тогда явная схема Эйлера аппроксимирует дифференциальную задачу и имеет первый порядок аппроксимации.

Для неявной схемы (5.3) имеем, очевидно, тот же самый результат. Несколько сложнее вопрос с третьей из простейшей схем (5.4). Она может быть оформлена в абстрактной форме:

$$[L^\tau(x^\tau)]_n = \begin{cases} x_0 - x_0^*, & n = 0, \\ x_1 - x_1^*, & n = 1, \\ \dfrac{x_n - x_{n-2}}{\tau} - f(x_{n-1}, t_{n-1}), & n = 2, 3, \ldots, N. \end{cases}$$

Но величина x_1^* не входит в постановку задачи, ее нужно как-то определить. В зависимости от того, как это будет сделано, порядок аппроксимации будет разный.

Почти очевидны следующие факты для погрешности аппроксимации:

$$r_0 = 0, \qquad r_n = O(\tau^2) \qquad \text{при } n = 2, 3, \ldots, N.$$

Оценим r_1 для нескольких способов определения x_1^*:

а) если x_1^* — какая угодно величина, то $r_1 = \mathscr{Z}(\tau) - x_1^* = O(1)$, схема имеет нулевой порядок аппроксимации и для расчетов непригодна;

б) если $x_1^* = x_0^*$, то $r_1 = \mathscr{Z}(\tau) - \mathscr{Z}(0) = O(\tau)$ и схема имеет первый порядок аппроксимации.

в) если $x_1^* = x_0^* + \tau f(x_0^*, t_0)$, то $r_1 = \mathscr{Z}(\tau) - x_0^* - \tau f(x_0^*, t_0) = O(\tau^2)$ и схема имеет второй порядок аппроксимации.

Разумеется, все это верно лишь в предположении, что решение дифференциальной задачи $\mathscr{Z}(t)$ имеет три ограниченных производных (для чего достаточно, чтобы $f(x, t)$ имела две ограниченных производных в интересующем нас диапазоне изменения x и t).

Подчеркнем, что обеспечение второго порядка аппроксимации третьей схемы потребовало достаточно аккуратного (хотя и не очень сложного в данном случае) определения дополнительных начальных данных x_1^*. Это типичное обстоятельство для схем, формальный порядок которых превышает формальный порядок дифференциального уравнения. Такие схемы используются именно с целью получить более высокий порядок аппроксимации, и недостаточно внимательное решение вопроса о дополнительных начальных данных может лишить схему желаемого порядка аппроксимации.

Устойчивость разностных схем. Рассмотрим схемы следующей формы:

$$\frac{x_{n+1} - x_n}{\tau} = F(x_n), \qquad n = 0, 1, \ldots, N-1, \quad N = T/\tau.$$

Как мы увидим в дальнейшем, такая форма охватывает важные классы схем Рунге—Кутты. Функция $F(x)$, конечно, связана с правой частью уравнения $f(x, t)$ (эта связь будет конкретизирована ниже). Все дальнейшее не претерпит никаких изменений, если вместо $F(x)$ будут использоваться функции $F(x, t, \tau)$, но ради простоты записи мы выбросим несущественные аргументы. Во все выкладки их при желании можно вписать механически, ничего не меняя.

Проверка устойчивости связана со сравнением решений двух «почти совпадающих» систем уравнений:

$$\frac{x_{n+1} - x_n}{\tau} = F(x_n) + \varepsilon_n', \quad x_0 = x_0^*,$$

$$\frac{y_{n+1} - y_n}{\tau} = F(y_n) + \varepsilon_n'', \quad y_0 = y_0^*.$$

Относительно ε_n' и ε_n'' предположим, что они ограничены общей константой: $\|\varepsilon_n'\| \leqslant \varepsilon$, $\|\varepsilon_n''\| \leqslant \varepsilon$. Еще раз подчеркнем, что на самом деле ниже речь пойдет не о двух системах, а о семействах систем с параметром τ. Чем меньше τ, тем больше уравнений в системах, а нас будут интересовать оценки, равномерные по τ. Ради простоты мы не пишем x_n^τ, $F(x_n^\tau, \tau)$ и т.д., но «скрытый» аргумент τ следует всегда иметь в виду.

Теорема. Пусть функция $F(x)$ удовлетворяет условию Липшица с постоянной C: $\|F(x) - F(y)\| \leqslant C\|x - y\|$. (Заметим, что это равномерная по τ оценка, C от τ не должна зависеть.) Пусть шаг τ мал: $C\tau \ll 1$. Тогда система разностных уравнений устойчива и имеет место оценка

$$\|x_n - y_n\| \leqslant e^{CT}\|x_0^* - y_0^*\| + 2\varepsilon e^{CT}/C, \quad \forall\, n \leqslant T/\tau. \tag{1}$$

(Таким образом мы докажем устойчивость разностных уравнений по начальным данным и правым частям.)

Доказательство. Перепишем уравнения в виде

$$\begin{aligned} x_{n+1} &= x_n + \tau F(x_n) + \tau\varepsilon_n', \\ y_{n+1} &= y_n + \tau F(y_n) + \tau\varepsilon_n''. \end{aligned} \tag{2}$$

Вычитая второе уравнение из первого, получаем

$$\|x_{n+1} - y_{n+1}\| \leqslant \|x_n - y_n\| + \tau\|F(x_n) - F(y_n)\| + 2\tau\varepsilon.$$

Используя условие Липшица, преобразуем оценку

$$\|x_{n+1} - y_{n+1}\| \leqslant (1 + C\tau)\|x_n - y_n\| + 2\tau\varepsilon.$$

Применим эту основную оценку последовательно:

$$\|x_1 - y_1\| \leqslant (1 + C\tau)\|x_0 - y_0\| + 2\tau\varepsilon,$$

$$\|x_2 - y_2\| \leqslant (1 + C\tau)\|x_1 - y_1\| + 2\tau\varepsilon \leqslant$$

$$\leqslant (1 + C\tau)^2\|x_0 - y_0\| + 2\tau\varepsilon[1 + (1 + C\tau)],$$

$$\|x_3 - y_3\| \leqslant (1 + C\tau)\|x_2 - y_2\| + 2\tau\varepsilon \leqslant$$

$$\leqslant (1 + C\tau)^3\|x_0 - y_0\| + 2\tau\varepsilon[1 + (1 + C\tau) + (1 + C\tau)^2].$$

Легко угадывается и доказывается по индукции общая формула:

$$\|x_n - y_n\| \leqslant (1 + C\tau)^n\|x_0 - y_0\| +$$

$$+ 2\tau\varepsilon[1 + (1 + C\tau) + \ldots + (1 + C\tau)^{n-1}],$$

или (после суммирования прогрессии)

$$\|x_n - y_n\| \leqslant (1 + C\tau)^n + 2\tau\varepsilon \frac{(1+C\tau)^n - 1}{1 + C\tau - 1} \leqslant$$
$$\leqslant (1 + C\tau)^n \left[\|x_0 - y_0\| + \frac{2\varepsilon}{C}\right]. \quad (3)$$

Заметим, что $n \leqslant T/\tau$; следовательно, $(1 + C\tau)^n \leqslant (1 + C\tau)^{T/\tau} \approx e^{CT}$ при $C\tau \ll 1$. Используя эту оценку в (3), получаем (1). Кстати, условие $C\tau \ll 1$ не следует считать очень жестким, так как, например, $1.5 \approx e^{0.4}$, а $1.1 \approx e^{0.095}$.

Теорема об устойчивости доказана. Применим ее к некоторым широко используемым на практике схемам.

Метод Эйлера с пересчетом. Переход от известного x_n к новому x_{n+1} делается в два этапа:

а) находится значение $x_{n+1/2} = x_n + 0.5\tau f(x_n, t_n)$;

б) вычисляется $x_{n+1} = x_n + \tau f(x_{n+1/2}, t_n + 0.5\tau)$.

Эту схему можно записать в общей форме:

$$x_{n+1} = x_n + \tau F(x_n, t_n, \tau),$$

где $F(x, t, \tau) \equiv f(x + 0.5\tau f(x), t + 0.5\tau)$. Легко проверить, что если f удовлетворяет условию Липшица (по x) с константой C, то и F удовлетворяет этому условию с несущественно большей константой $C(1 + C\tau)$. Таким образом, схема метода Эйлера с пересчетом устойчива.

Проверим порядок аппроксимации, т.е. оценим выражение

$$\frac{1}{\tau}(\mathscr{x}_{n+1} - \mathscr{x}_n) - F(\mathscr{x}_n, t_n, \tau).$$

Используем ряд Тейлора:

$$\mathscr{x}_{n+1} = \mathscr{x}(t_n + \tau) = \mathscr{x}_n + \tau\dot{\mathscr{x}}_n + \frac{\tau^2}{2}\ddot{\mathscr{x}}_n + O(\tau^3).$$

В силу уравнения $\dot{\mathscr{x}}_n = f(\mathscr{x}_n, t_n)$ имеем

$$\ddot{\mathscr{x}}_n = f_x\dot{\mathscr{x}}_n + f_t = f_x(\mathscr{x}_n, t_n) f(\mathscr{x}_n, t_n) + f_t(\mathscr{x}_n, t_n).$$

Итак,

$$\frac{1}{\tau}(\mathscr{x}_{n+1} - \mathscr{x}_n) = f + \frac{\tau}{2}[f_x f + f_t] + O(\tau^2).$$

С другой стороны,

$$F(\mathscr{x}_n, t_n, \tau) = f\left(\mathscr{x}_n + \frac{\tau}{2} f(\mathscr{x}_n, t_n), t_n + \frac{\tau}{2}\right) =$$
$$= f(\mathscr{x}_n, t_n) + \frac{\tau}{2} f_x(\mathscr{x}_n, t_n) f(\mathscr{x}_n, t_n) + \frac{\tau}{2} f_t(\mathscr{x}_n, t_n) + O(\tau^2).$$

Объединяя оба результата, получаем

$$r_{n+1} = \frac{1}{\tau}(\mathscr{X}_{n+1} - \mathscr{X}_n) - F(\mathscr{X}_n, t_n, \tau) = O(\tau^2).$$

Таким образом, схема Эйлера с пересчетом имеет второй порядок аппроксимации и в силу теоремы § 6 — второй порядок точности (в предположении, что $f(x, t)$ имеет две ограниченных производных и, следовательно, $\mathscr{X}(t)$ — три).

Методы Рунге–Кутты. Метод Эйлера с пересчетом является простейшим вариантом одной из наиболее распространенных в современной вычислительной практике схем численного интегрирования обыкновенных дифференциальных уравнений, объединяющей семейство методов с общим названием «методы Рунге—Кутты». Основу этих методов составляет ряд Тейлора. Связь $\mathscr{X}_{n+1} = \mathscr{X}(t_n + \tau)$ с $\mathscr{X}_n$ имеет форму

$$\mathscr{X}_{n+1} = \mathscr{X}_n + \tau\dot{\mathscr{X}}_n + \frac{\tau^2}{2}\ddot{\mathscr{X}}_n + \ldots + \frac{\tau^k}{k!}\mathscr{X}_n^{(k)} + O(\tau^{k+1}). \qquad (4)$$

Приближенное решение находится из того же выражения, но с выброшенным остаточным членом:

$$x_{n+1} = x_n + \tau\dot{\mathscr{X}}_n + \frac{\tau^2}{2}\ddot{\mathscr{X}}_n + \ldots + \frac{\tau^k}{k!}\mathscr{X}_n^{(k)}..$$

Чтобы иметь вычислительную схему, нужно дать выражение для производных. В принципе здесь нет серьезных проблем:

$$\dot{\mathscr{X}}_n = f(x_n, t_n), \qquad \ddot{\mathscr{X}}_n = f_x(x_n, t_n) f(x_n, t_n) + f_t(x_n, t_n)$$

и т.д. Однако аналитические выражения начинают катастрофически усложняться в результате дифференцирования, и этот путь оказывается крайне неудобным. **В методах Рунге—Кутты строится способ вычисления отрезка ряда Тейлора, требующий лишь вычисления** $f(x, t)$ в разных точках. Вот одна из распространенных схем.

Переход от (t_n, x_n) к (t_{n+1}, x_{n+1}) начинается с вычисления вспомогательных величин:

$$k_1 = f(x_n, t_n)\,\tau,$$

$$k_2 = f(x_n + 0.5k_1, t_n + 0.5\tau)\,\tau,$$

$$k_3 = f(x_n + 0.5k_2, t_n + 0.5\tau)\,\tau,$$

$$k_4 = f(x_n + k_3, t_n + \tau)\,\tau.$$

Затем делается собственно шаг интегрирования

$$x_{n+1} = x_n + (1/6)(k_1 + 2k_2 + 2k_3 + k_4).$$

Таким образом, один шаг требует четырехкратного вычисления правой части.

Мы не станем точно вычислять погрешность аппроксимации, это требует громоздких выкладок. Поясним, однако, основную идею. Если провести разложения k_1, k_2, k_3, k_4 по малым параметрам с достаточным числом членов и вычислить после этого выражение

$$x_n + (1/6)(k_1 + 2k_2 + 2k_3 + k_4),$$

то оно совпадет с отрезком ряда Тейлора (4) вплоть до членов порядка τ^4. Расхождение начнется только в членах $O(\tau^5)$.

Метод Рунге—Кутты можно записать в стандартной форме:

$$\frac{x_{n+1} - x_n}{\tau} = F(x_n, t_n, \tau),$$

где $F(x, t, \tau)$ — суперпозиция («многоэтажная») функций $f(x, t)$. Легко проверить, что константы Липшица для f и F почти (с точностью до $O(\tau)$) совпадают.

Что касается порядка аппроксимации, то, как уже было объяснено, он в данном случае четвертый. Существуют схемы Рунге—Кутты разных порядков, причем порядок аппроксимации равен числу вычислений правой части на один шаг процесса. Сказанного выше достаточно, чтобы сформулировать следующее утверждение.

Утверждение 1. Если функция $f(x, t)$ имеет ограниченные производные k-го порядка (следовательно, решение $\mathscr{X}(t)$ — производные $(k+1)$-го порядка), то метод Рунге—Кутты k-го порядка имеет k-й порядок аппроксимации и k-й порядок сходимости.

Однако стоит еще раз напомнить, что, кроме погрешности аппроксимации, есть еще погрешность машинного представления чисел, т.е. фактически после подстановки в разностные уравнения машинных чисел $\mathscr{X}^{\text{м}}_n$ получим

$$\frac{1}{\tau}(\mathscr{X}^{\text{м}}_{n+1} - \mathscr{X}^{\text{м}}_n) - F(\mathscr{X}^{\text{м}}_n, t_n, \tau) = O(\tau^k) + \frac{2}{\tau}|\mathscr{X}|\varepsilon,$$

и при $\tau < (|\mathscr{X}|\varepsilon)^{1/(k+1)}$ результаты численного интегрирования с уменьшением шага начнут ухудшаться, а не улучшаться.

Обратим внимание на то, что в теореме об устойчивости в оценке фигурирует крайне неприятный множитель e^{CT}. Конечно, оценки очень грубы, но простые примеры показывают, что в общем случае, если, кроме условия Липшица для f, других предположений не делать, эта оценка не улучшаема. Но в важных частных случаях она может быть намного улучшена, и это существенно, так как часто приходится проводить численное интегрирование в ситуации,

когда $\tau^k e^{CT}$ (k — порядок аппроксимации метода) есть величина много бо́льшая требуемой точности, а попытки достичь этой точности за счет уменьшения τ приводят к непосильному для ЭВМ объему вычислений.

Простые примеры более точных оценок мы сейчас получим. Рассмотрим два случая.

1. Пусть определяемая численным интегрированием система такова, что матрица $A(x) = \frac{1}{2}(F_x(x) + F^*(x))$ в нужном нам диапазоне изменения x строго отрицательна, т.е.

$$(A(x)\xi, \xi) \leqslant -a(\xi, \xi), \quad \forall\, \xi, x \ \ (a > 0).$$

Траекторию, в окрестности которой выполняется это условие, будем называть «устойчивой».

Утверждение 2. При интегрировании устойчивой траектории методом Рунге—Кутты k-го порядка аппроксимации погрешность приближенного решения есть $O(\tau^k)$ при всех $t > 0$.

Утверждение будет доказано, если мы получим оценку устойчивости разностной схемы, не содержащую экспоненциального множителя типа e^{CT}.

Доказательство. Оценим сначала норму

$$\|E + \tau F_x\|^2 = \sup_{\xi} \frac{((E+\tau F_x)\xi,\ (E+\tau F_x)\xi)}{(\xi, \xi)} =$$

$$= \sup_{\xi} \frac{(\xi, \xi) + 2\tau(A\xi, \xi) + \tau^2(F_x\xi, F_x\xi)}{(\xi, \xi)} \leqslant 1 - 2\tau a + O(\tau^2).$$

При достаточно малых τ можно пренебречь величиной $O(\tau^2)$ и пользоваться оценкой

$$\|E + \tau F_x(x)\| \leqslant 1 - C_1\tau, \quad C_1 > 0.$$

Теперь обратимся к оценке устойчивости. Как и раньше, из уравнений (2) имеем

$$\|x_{n+1} - y_{n+1}\| \leqslant \|(x_n - y_n) + \tau(F(x_n) - F(y_n))\| + 2\tau\varepsilon.$$

Оценим норму в правой части более аккуратно:

$$F(x) - F(y) = F(y + s(x-y))\Big|_{s=0}^{s=1} =$$

$$= \int_0^1 \frac{d}{ds} F(y + s(x-y))\, ds = \int_0^1 F_x(y + s(x-y))(x-y)\, ds.$$

Так как $x - y = \int_0^1 (x - y)\, ds$, то

$$\|(x - y) + \tau(F(x) - F(y))\| =$$

$$= \left\| \int_0^1 \{E + \tau F_x(y + s(x - y))\}(x - y)\, ds \right\| \leqslant$$

$$\leqslant \int_0^1 \|E + \tau F_x\|\, ds\, \|x - y\| \leqslant (1 - C_1\tau)\|x - y\|.$$

В итоге мы имеем оценку

$$\|x_{n+1} - y_{n+1}\| \leqslant (1 - C_1\tau)\|x_n - y_n\| + 2\tau\varepsilon.$$

Отсюда, как и раньше,

$$\|x_n - y_n\| \leqslant (1 - C_1\tau)^n \|x_0 - y_0\| + 2\tau\varepsilon \frac{1 - (1 - C_1\tau)^n}{1 - (1 - C_1\tau)}.$$

Таким образом,

$$\|x_n - y_n\| \leqslant \|x_0 - y_0\| + 2\varepsilon/C_1,$$

и эта оценка не ухудшается при всех $n > 0$. Имея такую оценку устойчивости, получаем утверждение о порядке сходимости, совпадающем с порядком аппроксимации при интегрировании на сколь угодно большом интервале времени (конечно, только при интегрировании устойчивой задачи).

2. Более интересный и тонкий результат можно получить для «не неустойчивых» систем, т.е. для систем, у которых матрица $A(x)$ (симметричная часть F_x) только неположительна, т.е. $(A\xi, \xi) \leqslant 0$, $\forall\, \xi, x$. В этом случае оценка $\|E + \tau F_x(x)\|$ находится так же, как это делалось выше:

$$\|E + \tau F_x(x)\| \leqslant 1 + C_2\tau^2.$$

Повторяя далее оценки для решений двух разностных уравнений, имеем

$$\|x_{n+1} - y_{n+1}\| \leqslant (1 + C_2\tau^2)\|x_n - y_n\| + 2\tau\varepsilon,$$

откуда уже известным способом получаем

$$\|x_n - y_n\| \leqslant (1 + C_2\tau^2)^n \|x_0 - y_0\| + 2\tau\varepsilon \frac{(1 + C_2\tau^2)^n - 1}{C_2\tau^2}.$$

Пусть теперь используется метод k-го порядка аппроксимации, т.е. $\varepsilon = O(\tau^k)$ $(k \geqslant 2)$. Тогда, как нетрудно понять, можно положить $n \leqslant 1/\tau^2$ (что соответствует времени $t_n = n\tau \approx C/\tau$); при этом $(1 + C_2\tau^2)^n \leqslant e^{CC_2}$. Таким образом, для $n \leqslant C_3/\tau^2$, $C_3 = O(1)$, имеем

$$\|x_n - y_n\| \leqslant e^{C_2C_3} \|x_0 - y_0)\| + O(\tau^{k-1})\, e^{C_2C_3}.$$

Итак, при интегрировании «не неустойчивой» системы методом k-го $(k \geqslant 2)$ порядка аппроксимации с шагом τ на интервале времени $0 \leqslant t \leqslant C_3/\tau$ численное решение имеет точность $O(\tau^{k-1})$. Разумеется, на конечном интервале времени $0 \leqslant t \leqslant T$ точность метода есть $O(\tau^k)$. В дальнейшем она понижается на один порядок. С такими задачами вычислители встречаются при расчете процессов, носящих характер «вращений», колебаний и т.п. Их приходится рассчитывать на длительных интервалах времени, так как физическое время, интересное для приложений, обычно бывает очень большим для системы в том смысле, что оно содержит большое число колебаний или оборотов. Поэтому приведенная выше оценка точности численного интегрирования очень важна.

Мы рассмотрели простые варианты теорем, дающих оценки погрешности численного интегрирования существенно более точные, чем стандартные (опирающиеся только на условие Липшица для правой части). Они доказаны при достаточно сильных предположениях о свойствах матрицы $f_x(x)$. Более тонкие теоремы должны основываться на более слабых предположениях. Но в любом случае такие оценки будут существенно опираться на свойства решений так называемого *уравнения в вариациях*:

$$\frac{d\,\delta x}{dt} = f_x[x(t)]\,\delta x + r(t).$$

Это линейное уравнение с переменными коэффициентами. Оно определено на исследуемой траектории $x(t)$ и описывает (в первом порядке) эволюцию возмущения траектории $x(t)$, вызванного малым возмущением правой части.

Возмущенная траектория удовлетворяет уравнению

$$\frac{d\tilde{x}}{dt} = f(\tilde{x}) + r(t), \quad r(t) \text{ — малое возмущение.}$$

Полагая $\tilde{x}(t) = x(t) + \delta x(t)$ (здесь $(x(t)$ — решение уравнения $\dot{x} = f(x)$), разлагая $f(x + \delta x)$ в ряд Тейлора и пренебрегая членами $O(\|\delta x\|^2)$, получаем уравнение в вариациях, играющее огромную роль в теории устойчивости и в близких к ней вопросах о точности численного интегрирования.

§ 8. Приближенное решение краевых задач для систем обыкновенных дифференциальных уравнений

Следующий по сложности (после задачи Коши) класс задач — это краевые задачи, в которых часть конечных условий задана на левом конце интервала времени, а часть — на правом. Краевые условия могут быть сформулированы вообще в терминах левых и правых концов траектории одновременно. Начнем с линейных краевых задач.

Итак, требуется найти решение линейной неоднородной системы обыкновенных дифференциальных уравнений с переменными коэффициентами

$$\frac{dx}{dt} = A(t)\,x + a(t), \qquad 0 \leqslant t \leqslant T. \tag{1}$$

Здесь x и a — p-мерные векторы, $A(t)$ — $p \to p$-матрица. Как известно, для выделения однозначной траектории требуется еще задать p конечных соотношений. Запишем их в общем виде:

$$C\,x(0) + D\,x(T) = f \tag{2}$$

(C, D — $p \to p$-матрицы, f — p-вектор).

Стандартный метод решения такой краевой задачи связан с основным результатом теории линейных систем: общее решение системы (1) задается явной конструкцией

$$x(t) = x^0(t) + \sum_{i=1}^{p} \alpha_i\, x^i(t), \tag{3}$$

где $x^0(t)$ — произвольное решение неоднородной системы, т.е. x^0 должно удовлетворять уравнению $\dot{x}^0 = A(t)x^0 + a(t)$ (краевые условия для x^0 какие угодно, вернее, те, которые нам по каким-то причинам удобны). В соотношении (3) $x^i(t)$ — это p линейно-независимых решений однородной системы, т.е. x^i удовлетворяет уравнению $\dot{x}^i = A(t)x^i$, а краевые условия для x^i тоже произвольные, лишь бы они обеспечивали линейную независимость совокупности векторов $x^i(t)$, $i = 1, 2, \ldots, p$, при всех t.

Как известно, достаточно проверить линейную независимость при каком-то одном значении t. Что касается (скалярных) коэффициентов α_i, то они произвольны, и этот произвол «тратится» на выполнение p заданных краевых условий (2). То, что конструкция (3) при любых α_i удовлетворяет уравнению (1), очевидно. Подставим ее в краевые условия:

$$C\left[x^0(0) + \sum_{i=1}^{p} \alpha_i\, x^i(0)\right] + D\left[x^0(T) + \sum_{i=1}^{p} \alpha_i\, x^i(T)\right] = f,$$

или

$$\sum_{i=1}^{p} \alpha_i [C x^i(0) + D x^i(T)] = f - C x^0(0) - D x^0(T). \qquad (4)$$

Получена система p линейных алгебраических уравнений с матрицей, i-й столбец которой есть $Cx^i(0) + Dx^i(T)$. Если система (4) имеет единственное решение ($\det \neq 0$), краевая задача имеет единственное решение. Но это не есть обязательный факт, хотя его можно считать типичным. Отсутствие решения (или неединственность при подходящей правой части) следует считать вырождением задачи.

Все, что было сказано выше, полностью взято из курса обыкновенных дифференциальных уравнений. Специалист по вычислительной математике должен добавить только четкое указание, откуда взять функции $x^i(t)$, $i = 0, 1, \ldots, p$. Ответ почти очевиден: раз мы научились численно интегрировать задачу Коши, то просто нужно сконструировать такие задачи Коши, которые дадут то, что нужно.

Решение $x^0(t)$ можно получить, взяв задачу Коши с начальными данными $x^0(0) = 0$. Само решение находим каким-либо численным методом, хотя бы по схеме Эйлера. Обозначая $x_k^0 \approx x^0(t_k)$, где $\{t_k\}$ — сетка, покрывающая интервал $[0, T]$, используем простейшую схему

$$x_{k+1}^0 = x_k^0 + \tau_k A(t_k) x_k^0 + a(t_k),$$

$$\tau_k = t_{k+1} - t_k, \quad x_0^0 = 0, \quad k = 0, 1, \ldots, K.$$

Конечно, реально на практике используют более точные методы, Рунге—Кутты например, но сейчас важна принципиальная схема. При вычислении линейно-независимых решений $x^i(t)$ используем для них напрашивающиеся данные Коши:

$$x^i(0) = e^i, \quad i = 1, 2, \ldots, p,$$

где $e^i = \{0, \ldots, 0, 1_i, 0, \ldots, 0\}$, т.е. i-й орт p-мерного пространства. Итак, $x_0^i = e^i$, и далее

$$x_{k+1}^i = x_k^i + \tau_k A_k x_k^i, \quad k = 0, 1, \ldots, K-1.$$

Отметим, что такой способ решения краевой задачи «стоит» $(p+1)$-кратного решения задачи Коши. Однако часто объем работы можно сократить. Это относится к очень распространенному типу краевых задач: $r < p$ компонент x задано при $t = 0$ и $p - r$ компо-

нент — при $t = T$, т.е. краевые условия имеют вид

$$x_1(0) = f_1, \quad x_2(0) = f_2, \quad \ldots, \quad x_r(0) = f_r$$

(здесь нижний индекс — номер компоненты). Правые краевые условия произвольные; например,

$$Dx(T) = b \quad \text{или} \quad Bx(0) + Dx(T) = d,$$

где B, D — прямоугольные матрицы $p \to (p - r)$ (p столбцов, $p - r$ строк), b — $(p - r)$-вектор. В этом случае для решения $x^0(t)$ берем данные Коши:

$$x_i^0(0) = f_i, \quad i = 1, 2, \ldots, r; \qquad x_i^{(0)} = 0, \quad i = r + 1, r + 2, \ldots, p,$$

а решение краевой задачи ищем в виде

$$x(t) = x^0(t) + \sum_{i=r+1}^{p} \alpha_i x^i(t)$$

($x^i(t)$ находятся так же, как и раньше). Легко видеть, что такая конструкция при любых α_i удовлетворяет уравнению $\dot{x} = Ax + a$ и левым краевым условиям, а свободных параметров α_i как раз столько, чтобы за их счет выполнить $p - r$ условий на правом конце интервала времени.

Нелинейные краевые задачи. Метод «стрельбы». Перейдем теперь к нелинейным краевым задачам. Как всегда, в нелинейной ситуации лучше говорить о возможном подходе, чем о методе. Итак, пусть требуется найти решение уравнения

$$\frac{dx}{dt} = f(x, t), \qquad 0 \leqslant t \leqslant T,$$

при общих, например, краевых условиях $\Phi(x(0), x(T)) = 0$.

Используем умение достаточно надежно решать задачу Коши. Введем данные Коши $x(0)$ в качестве искомых неизвестных. Обозначая их через $\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_p\}$, определим траекторию $x(t, \alpha)$ задачи Коши:

$$\dot{x} = f(x, t), \qquad x(0) = \alpha.$$

Когда мы говорим «определим траекторию», это означает, что при каждом заданном значении вектора параметров α мы можем с какой-то точностью численно проинтегрировать задачу Коши.

Введем функцию

$$F(\alpha) \equiv \Phi(\alpha, x(T, \alpha)).$$

Решение краевой задачи свелось к решению системы нелинейных уравнений $F(\alpha) = 0$ (p уравнений с p неизвестными). Еще раз под-

черкнем, что функция F задана нам достаточно сложным алгоритмом, позволяющим для любого α вычислить вектор F; такое вычисление «стоит» одного численного интегрирования задачи Коши с начальными данными α.

Решение системы можно осуществить методом Ньютона и его модификациями. Конечно, в этом случае вычисление матрицы $F_\alpha(\alpha)$ проще всего выполнить численным дифференцированием, хотя есть и более аккуратные методы, используемые в вариационном исчислении (они предполагают использование так называемой системы уравнений в вариациях; см. § 27, 28).

Спектральная задача Штурма–Лиувилля. Специальный, но очень важный класс краевых задач связан с определением точек спектра для уравнения Штурма–Лиувилля. Рассмотрим простейший случай. Задано линейное однородное дифференциальное (самосопряженное) уравнение

$$\frac{d}{dt}\left[p(t)\frac{dx}{dt}\right] + q(t)\,x(t) = \lambda\, r(t)\,x(t), \qquad 0 \leqslant t \leqslant T,$$

содержащее параметр λ (функции $p(t) > 0$, $q(t)$, $r(t)$ заданы). Уравнение дополнено простыми краевыми условиями (тоже линейными однородными), например $x(0) = 0$, $x(T) = 0$. При почти всех λ краевая задача имеет тривиальное решение $x(t) \equiv 0$, но при некоторых специальных значениях λ, называемых точками спектра, появляются и нетривиальные решения. Они-то (и соответствующие им значения λ) представляют основной интерес в приложениях.

Соединим технику решения задачи Коши и решение нелинейных уравнений. Поставим для уравнения условия Коши $x(0) = 0$, $\dot{x}(0) = 1$. (Нетрудно видеть, что вместо $\dot{x}(0) = 1$ можно взять в краевом условии любое число.) После этого, если λ задано, определяется траектория $x(t, \lambda)$ — решение задачи Коши.

Разумеется, при произвольном λ эта траектория не удовлетворяет второму краевому условию, и теперь надо подобрать λ так, чтобы на траектории $x(t, \lambda)$ было выполнено второе условие. Другими словами, определим функцию $F(\lambda) \equiv x(T, \lambda)$ (опять-таки, напомним, что вычисление F при заданном λ требует численного интегрирования задачи Коши) и станем решать уравнение $F(\lambda) = 0$. Корни этого уравнения — суть точки спектра задачи Штурма–Лиувилля (разумеется, приближенные, коль скоро функцию F мы вычисляем лишь приближенно).

Самое грубое решение задачи можно представить себе так. Заменив функции p, q, $r(t)$ на постоянные, равные, например, средним значениям по интервалу $[0, T]$, вычислим спектр такой модельной задачи. Тем самым будет получена ориентировочная информация о

расположении точек спектра исходной задачи — о расстояниях между ними. Исходя из этого, выберем некоторый шаг Δ, заметно меньший расстояния между собственными числами, но не слишком малый, и вычислим значения $F(k\Delta)$ в точках сетки $k\Delta$, $k = 0, \pm 1, \pm 2, \ldots$ (Вспомним, что каждое значение $F(k\Delta)$ — это интегрирование задачи Коши.) Построим «график» $F(\lambda)$ по точкам $F(k\Delta)$. Он будет выглядеть примерно так, как показано на рис. 8. Проведя через полученные точки (хотя бы с помощью лекала) гладкую кривую, найдем приблизительные значения корней $F(\lambda) = 0$. Потом, если нужно, их можно уточнить. Заметим, что все это сравнительно просто в самосопряженной задаче, когда нам точно известно, что корни $F(\lambda) = 0$ находятся на вещественной оси. В общем случае они комплексны и ситуация заметно осложняется (см. § 15, 16).

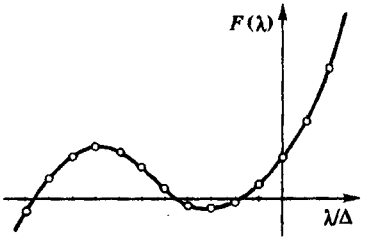


Рис. 8

Решение нелинейных краевых задач. Метод Ньютона. Закончим этот параграф описанием еще одной популярной алгоритмической конструкции, предназначенной для решения нелинейных краевых задач. Общую идею поясним на следующем примере.

Пусть требуется решить краевую задачу для системы уравнений

$$\dot{x} = f(x, t), \qquad 0 \leqslant t \leqslant T,$$

с краевыми условиями хотя бы общего вида $\Phi(x(0), x(T)) = 0$ (где Φ — p-вектор). Имеется некоторая функция $x^0(t)$, не удовлетворяющая ни краевым условиям, ни уравнению. Используем ее в качестве начального приближения и построим алгоритм типа метода Ньютона (в функциональном пространстве). Это обобщение и соответствующая теория разрабатывались Л. В. Канторовичем в начале сороковых годов.

В соответствии с общей схемой метода Ньютона следующее (первое) приближение ищем в виде

$$x^1(t) = x^0(t) + \delta x(t),$$

где $\delta x(t)$ — «малая» поправка. В результате получаем уравнение для δx:

$$\dot{x}^0 + \delta\dot{x} = f(x^0 + \delta x, t).$$

Линеаризуя его (отбрасывая малые второго порядка), имеем

$$\dot{x}^0 + \delta\dot{x} = f[x^0(t), t] + f_x[x^0(t), t]\,\delta x.$$

Чтобы функция $x^1(t)$ удовлетворяла краевым условиям, потребуем выполнения условий

$$\Phi[x^0(0) + \delta x(0),\ x^0(T) + \delta x(T)] = 0.$$

Они тоже линеаризуются:

$$0 = \Phi[x^0(0), x^0(T)] + \Phi_{x(0)}[x^0(0), x^0(T)]\ \delta x(0) + \Phi_{x(T)}\ \delta x(T).$$

Итак, $\delta x(t)$ находится решением краевой задачи для системы линейных неоднородных уравнений

$$\frac{d\,\delta x}{dt} - f_x[x^0(t), t]\ \delta x = f[x^0(t), t] - \dot{x}^0(t)$$

с известной правой частью. Далее процесс повторяется до получения нужной точности, если он сходится (что требует выбора не слишком случайного начального приближения). Ограничимся общим описанием и укажем, что в последнее время этот метод стали называть методом «квазилинеаризации» (по инициативе Р. Беллмана).

Пример решения краевой задачи. Покажем, как фактически реализуется алгоритм. Задача заимствована из американской литературы, где она характеризуется как «неустойчивая». Сложность решения этой задачи существенно зависит от длины интервала $[0, T]$. Поэтому ее решение, которое мы будем считать «точным», было получено методом продолжения по параметру. Сначала нашли решение при $T = 10$, затем, используя его как начальное приближение, нашли решение при $T = 11.6$, далее при $T = 13.2$, $T = 14.8$, $T = 16.4$, $T = 18.0$, $T = 19.6$ и, наконец, при $T = 20$.

Будем решать задачу модифицированным методом Ньютона в функциональном пространстве сразу на интервале $[0, 20]$. Сформулируем краевую задачу. На интервале $0 \leqslant t \leqslant 20$ ищется решение $x(t) = \{x^1, x^2, \ldots, x^5\}$, удовлетворяющее системе уравнений

$$\begin{aligned}
&\dot{x}^1 = x^2,\\
&\dot{x}^2 = x^3,\\
&\dot{x}^3 = -1.55\ x^1x^3 + 0.1\ (x^2)^2 - (x^4)^2 + 0.2\ x^2 + 1,\\
&\dot{x}^4 = x^5,\\
&\dot{x}^5 = 1.55\ x^1x^5 + 1.1\ x^2x^4 + 0.2\ (x^4 - 1).
\end{aligned}$$

Краевые условия:

$$x^1 = x^2 = x^4 = 0, \quad t = 0;$$

$$x^2 = 0, \quad x^4 = 1, \quad t = T = 20.$$

В качестве начального приближения берется функция, удовлетворяющая краевым условиям

$$x^1(t) = x^3(t) = 0, \qquad x^4(t) = t/T,$$

$$x^2(t) = \begin{cases} 0.5t/T, & t \leqslant 0.5T, \\ 0.5(T-t)/T, & t \geqslant 0.5T \end{cases}$$

$$x^5(t) = \begin{cases} t/T, & t \leqslant 0.1T, \\ 0.15 - 0.5t/T, & 0.1T \leqslant t \leqslant 0.5T, \\ 0.5t/T - 0.35, & 0.5T \leqslant t \leqslant 0.9T, \\ 1 - t/T, & t > 0.9T. \end{cases}$$

Это приближение очень грубое, оно выбрано без использования известного решения задачи.

Теперь поясним детали технической реализации алгоритма.

Сетка, сеточная функция. На интервале $[0, T]$ вводится равномерная сетка

$$\{t_n\}_{n=0}^N, \qquad t_n = n\tau, \quad \tau = T/N.$$

В узлах сетки определяется сеточная функция $\{x_n\}_{n=0}^N$, $x_n \in R^5$.

Невязка. Это очень важный объект, определяющий (наряду с числом узлов N) точность приближенного решения. Обозначая систему $\dot{x} = f(x)$, определяем невязку $r(t)$ как кусочно-постоянную на сетке функцию:

$$r(t) = \frac{x_{n+1} - x_n}{\tau} - f\left(\frac{x_n + x_{n+1}}{2}\right), \qquad t \in (t_n, t_{n+1}).$$

Итерационный процесс имеет целью свести норму невязки к нулю, т.е. приводит (при успешных вычислениях) к решению системы разностных уравнений, аппроксимирующих задачу со вторым порядком. Нормой невязки R считаем $\left(\int r^2\, dt\right)^{1/2}$, добавляя еще нормы невязок в краевых условиях.

Уравнение в вариациях. Напомним, что процесс итераций по схеме модифицированного метода Ньютона состоит в вычислении поправки $\delta x(t)$ и образовании однопараметрического семейства функций

$$y(t, h) \equiv x(t) + h\, \delta x(t),$$

где h — подлежащий определению шаг. Для определения $\delta x(t)$ решается краевая задача, получающаяся линеаризацией исходной нелинейной задачи на имеющемся уже приближении $x(t)$. Эта краевая задача имеет вид

$$\delta \dot{x} = f_x[x(t)]\, \delta x(t) - r(t).$$

Линеаризация краевых условий строится очевидным образом. Пояснения заслуживает способ вычисления матрицы $f_x[t] \equiv f_x(x(t))$. В расчетной схеме в качестве этой матрицы использовалась кусочно-постоянная матрица

$$f_x[t] \equiv f_x\left(\frac{x_n + x_{n+1}}{2}\right), \qquad t \in (t_n, t_{n+1}).$$

Решение линейной краевой задачи осуществляется методом ортогональной прогонки, описанным в § 18. Не следует думать, что шаг численного интегрирования совпадает с шагом сетки τ, он в несколько раз меньше. В процессе интегрирования запоминается ограничение функции $\delta x(t)$ на сетку $\{t_n\}$ — сеточная функция $\{\delta x_n\}$.

Определение шага. Определяется функция $R(h)$ — норма невязки сеточной функции $\{x_n + h\,\delta x_n\}_{n=0}^{N}$, находится (не очень точно) $\min R(h)$ по h. После определения h новое приближение вычисляется по формуле

$$x_n := x_n + h\,\delta x_n, \qquad n = 0, 1, \ldots, N.$$

Выше описан стандартный шаг итерационного процесса. Итерации продолжаются до получения достаточно малой невязки. Отметим, что во многих прикладных задачах одной из наиболее трудоемких операций является линеаризация, т.е. вычисление f_x. Конечно, такая операция трудна при достаточно сложной форме правой части. В рассматриваемом примере это не так. Трудоемкой операцией, вообще говоря, является и вычисление шага h, требующее нескольких вычислений невязки. Естественно, число операций пропорционально числу интервалов сетки N, и следующее ниже усовершенствование имеет целью повысить точность, не увеличивая существенно N или даже совсем его не меняя. Это достигается изменением определения невязки.

Уточненная формула вычисления невязки. Имея сеточную функцию $\{x_n\}$, восполняем ее до непрерывной функции $\tilde{x}(t)$ с помощью кусочно-гладкого интерполяционного аппарата (с помощью сплайна, например). Теперь, в принципе, можно говорить о непрерывной функции $r(t) = \dot{\tilde{x}} - f(\tilde{x}(t))$ и вычислять ее норму $\left(\int r^2\,dt\right)^{1/2}$. Разумеется, норма вычисляется по какой-нибудь хорошей квадратурной формуле, так что на самом деле такая «непрерывная» невязка вычисляется в дискретном наборе узлов квадратуры (в описываемых ниже расчетах на каждом интервале (t_n, t_{n+1}) интеграл вычислялся по квадратуре Гаусса с тремя узлами). Остальные элементы вычислительной схемы остаются без изменений.

Приведем результаты, характеризующие эффективность алгоритма и достигнутую точность.

В табл. 7 представлены следующие данные: i — номер итерации, R — норма невязки, h — шаг на i-й итерации. Начальное приближение выбрано таким, что краевые условия выполнены; они линейны, поэтому не нарушаются в процессе итераций. Невязка содержит только компоненту $\dot{x} - f$; невязки в краевых условиях остаются нулевыми. Расчет проводился при шаге основной сетки $\tau = T/50 = 0.4$.

Таблица 7

i	0	1	2	3	4	5
R	15.60	14.15	13.73	12.40	11.36	10.63
h	–	0.155	0.053	0.153	0.130	0.115
i	6	7	8	9	10	11
R	9.31	6.84	1.43	0.20	0.007	0.0002
h	0.197	0.413	1.04	1.03	1.0	1.0

Таблица 8 характеризует точность расчета. В ней приведены значения компонент вектора x в нескольких характерных точках времени. Каждая величина представлена тремя числами: а) результат, полученный за одиннадцать итераций при грубом вычислении невязки; б) результат, полученный после одной дополнительной итерации с более точным вычислением невязки, в) «точное» решение.

Таблица 8

	$x^1(10.4)$	$x^1(20)$	$x^2(5.6)$	$x^2(12)$	$x^3(0)$	$x^3(10.4)$
а)	−1.0808	−1.1808	0.1244	−0.0455	−0.9621	−0.0438
б)	−1.091	−1.190	0.1214	−0.0444	−0.9863	−0.04308
в)	−1.091	−1.190	0.1213	−0.0443	−0.9863	−0.04312
	$x^4(5.6)$	$x^4(12)$	$x^5(0)$	$x^5(5.6)$	$x^5(12)$	$x^5(20)$
а)	1.1838	0.9867	0.6442	−0.09101	0.02496	0.00915
б)	1.1817	0.9867	0.6529	−0.08878	0.02434	0.008806
в)	1.1820	0.9866	0.6529	−0.08874	0.02433	0.008622

§ 9. Метод дифференциальной прогонки

Начнем изучение самого, вероятно, популярного в вычислительной математике алгоритма. С ним связаны существенные успехи в решении сложных задач, и его изобретение считается по праву ярким событием в истории развития современной вычислительной математики. Любопытно, что метод прогонки предназначен для решения задачи, на первый взгляд не содержащей никаких проблем.

Пусть требуется решить краевую задачу для системы двух уравнений

$$\dot{x} = ay + f, \qquad \dot{y} = bx + \varphi, \qquad 0 \leqslant t \leqslant 1, \quad a > 0, \quad b > 0,$$

с краевыми условиями $x(0) = \mathscr{X}_0$, $y(1) = 0$. Коэффициенты a и b ради простоты считаем постоянными. Очень важны их числовые значения. Пусть $a \approx b \approx 40$. Эти значения взяты не случайно, в них суть дела. Как мы увидим дальше, a и b — «большие параметры». Именно то, что они большие, определяет специфику задачи и требует разработки принципиально новых алгоритмов.

Объясним причины, по которым описанный в § 8 метод сведения краевой задачи к задачам Коши не работает. Проанализируем известный («школьный») метод. Ради простоты положим $f = \varphi = 0$ (не в них дело). Найдя решение двух задач Коши:

$$\text{а)}\ \begin{aligned} \dot{x}^1 &= ay^1, \quad x^1(0) = 1, \\ \dot{y}^1 &= bx^1, \quad y^1(0) = 0; \end{aligned} \qquad \text{б)}\ \begin{aligned} \dot{x}^2 &= ay^2, \quad x^2(0) = 0, \\ \dot{y}^2 &= bx^2, \quad y^2(0) = 1, \end{aligned}$$

ищем решение в виде линейной комбинации:

$$\begin{Bmatrix} x(t) \\ y(t) \end{Bmatrix} = \alpha_1 \begin{Bmatrix} x^1(t) \\ y^1(t) \end{Bmatrix} + \alpha_2 \begin{Bmatrix} x^2(t) \\ y^2(t) \end{Bmatrix},$$

а коэффициенты α_1 и α_2 определим, подставив его в краевые условия. Что же из этого получится? Посмотрим, какие последствия имеют большие значения коэффициентов a и b (и в самом ли деле они большие).

Как известно, точное общее решение рассматриваемой системы уравнений с постоянными коэффициентами имеет вид

$$\begin{Bmatrix} x(t) \\ y(t) \end{Bmatrix} = C_1 e^{\lambda_1 t} \begin{Bmatrix} X_2 \\ Y_2 \end{Bmatrix} + C_2 e^{\lambda_2 t} \begin{Bmatrix} X_2 \\ Y_2 \end{Bmatrix},$$

где X_i, Y_i — некоторые числа (которые легко вычисляются, но они нам не нужны), C_1, C_2 — произвольные постоянные, а λ_1, λ_2 — корни характеристического уравнения

$$\det \begin{pmatrix} -\lambda & a \\ b & -\lambda \end{pmatrix} = 0, \quad \text{т.е.} \quad \lambda^2 = ab, \quad \lambda_{1,2} = \pm\sqrt{ab} \approx \pm 40.$$

Таким образом, почти любое частное решение (в том числе построенные нами два линейно независимых решения) есть сумма (с примерно равными коэффициентами C_1, C_2) двух экспонент: одной — сильно растущей (типа e^{40t}), второй — сильно убывающей (типа e^{-40t}).

Теперь обратимся к исходной задаче. Прежде всего подчеркнем, что выбор больших значений коэффициентов a и b не был произвольным, он продиктован практикой. Наиболее близким содержательным примером задачи, качественные черты которой хорошо передает разбираемая модельная, является прохождение излучения через слой большой оптической толщины, например прохождение потока нейтронов, источником которого является ядерный реактор, через слой защиты. В этом случае одно краевое условие $x_0 = \mathscr{X}_0$ задает поток нейтронов, падающий на внутреннюю поверхность защиты, второе условие $y(1) = 0$ означает отсутствие потока нейтронов, падающих на внешнюю границу защиты. Интересующая же нас величина $x(1)$ имеет смысл потока нейтронов, выходящего со стороны внешней границы защиты.

Искомое решение есть функция типа $e^{-40t}\mathscr{X}_0$ (в этом, собственно, и состоит назначение защиты: ослабить поток нейтронов примерно в $e^{40} \approx 10^{16}$ раз). Мы же пытаемся получить его в виде линейной комбинации двух линейно независимых решений, в каждом из которых решающую роль играет именно растущая экспонента. Получить функцию типа e^{-40t} в виде линейной комбинации решений, в которых главную роль играют компоненты типа e^{40t} (они должны взаимно погаситься), — очень трудная вычислительная задача, сопровождающаяся резким падением точности.

Следует принять во внимание и накопление погрешностей вычислений. В рассматриваемом случае, допустим, при интегрировании задач Коши методом k-го порядка погрешность аппроксимации у левого конца траектории имеет величину порядка $\tau^k\mathscr{X}_0$ (τ — шаг численного интегрирования). Ее последствия у правого конца траектории достигнут величины $e^{40}\tau^k\mathscr{X}_0$, и нужно, чтобы они были заметно меньше искомого решения, т.е. величины порядка $e^{-40}\mathscr{X}_0$.

Итак, имеем ориентировочное соотношение для шага численного интегрирования:

$$e^{40}\tau^k \ll e^{-40}, \quad \text{т.е.} \quad \tau^k \ll 10^{-30}.$$

Даже при $k = 5$ получаем $\tau < 10^{-6}$, т.е. нужно брать сетку с миллионом узлов. Это чудовищно. Забегая вперед, укажем, что

фактически такие задачи решаются методом прогонки при условиях $|\lambda_{1,2}\tau| \ll 1$. В нашем случае $\tau \ll 1/40$, т.е. вполне приемлем, например, расчет при $\tau \approx 1/200$ и даже при $\tau \approx 1/100$.

Убедившись в провале «школьного» метода решения простой краевой задачи, приступим к описанию метода прогонки. Отметим только в качестве «морали»: в вычислительной математике важна не столько внешняя форма задачи, сколько качественные свойства искомых решений. Абсолютно одинаковые внешне задачи часто требуют существенно разных методов. Та же самая задача при $a \approx b \approx 5$ (или на интервале $0 \leqslant t \leqslant 0.1$) без всяких затруднений может быть решена только что скомпрометированным методом фундаментальных решений.

Рассмотрим детально метод дифференциальной прогонки. Будем искать связь между компонентами решения вида

$$x(t) = \alpha(t)y(t) + \beta(t),$$

где $\alpha(t)$, $\beta(t)$ — неизвестные пока «прогоночные коэффициенты». Получим уравнения для них. Дифференцируя прогоночное соотношение:

$$\dot{x} = \dot{\alpha}y + \alpha\dot{y} + \dot{\beta},$$

и учитывая, что $\dot{x} = ay + f$, $\dot{y} = bx + \varphi$, имеем

$$ay + f = \dot{\alpha}y + \alpha(bx + \varphi) + \dot{\beta}.$$

Заменяя x на $\alpha y + \beta$:

$$ay + f = \dot{\alpha} + b\alpha^2 y + b\beta\alpha + \alpha\varphi + \dot{\beta},$$

и приводя подобные члены (коэффициенты при y и единице), получаем

$$y[\dot{\alpha} + b\alpha^2 - a] + [\dot{\beta} + b\beta\alpha - f] = 0.$$

Приравнивая нулю коэффициенты при y и единице, приходим к уравнениям для α и β:

$$\dot{\alpha} + b\alpha^2 - a = 0, \qquad \dot{\beta} + \alpha\beta b + \alpha\varphi - f = 0.$$

Эти уравнения дополним начальными данными, используя стандартный прием метода прогонки. Левое краевое условие $x(0) = \mathscr{X}_0$ запишем в виде того же самого прогоночного соотношения: $x(0) = \alpha(0)y(0) + \beta(0)$. Очевидно, следует положить $\alpha(0) = 0$, $\beta(0) = \mathscr{X}_0$. Итак, получены задачи Коши для $\alpha(t)$ и $\beta(t)$. Они могут быть проинтегрированы (например, численно), и можно считать, что функции $\alpha(t)$ и $\beta(t)$ у нас уже есть.

Перейдем к следующему характерному элементу прогонки — разрешению правого краевого условия. Имея условие $y(1) = 0$ и

прогоночное соотношение при $t = 1$: $x(1) = \alpha(1)y(1) + \beta(1)$, легко находим значение $x(1) = \beta(1)$.

Наконец, рассмотрим заключительный этап прогонки. Опять-таки отклоним напрашивающийся рецепт: раз мы знаем $x(1)$ и $y(1)$, можно (формально) интегрировать задачу Коши справа налево. Но эта задача так же неустойчива, как и задача Коши, решаемая слева направо. Мы воспользуемся уравнением $\dot{y} = bx + \varphi$. Заменяя x из прогоночного соотношения $x = \alpha y + \beta$, получаем уравнение для y:

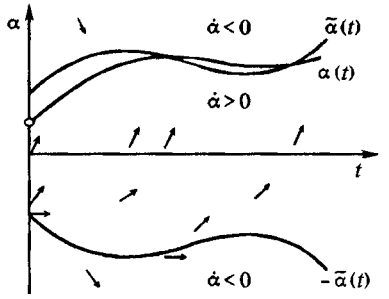


Рис. 9

$$\dot{y} = b\alpha y + \beta b + \varphi, \quad y(1) = 0.$$

Проинтегрируем задачу справа налево, попутно определяя $x(t) = \alpha(t)y(t) + \beta(t)$.

Перейдем к анализу метода прогонки, рассматривая для общности краевые условия вида $Ax(0) + By(0) = C$. В этом случае $\alpha(0) = -B/A$, $\beta(0) = C/A$. Разберемся в том, действительно ли «большой» параметр (который так и остался в задаче) уже не страшен и процесс вычислений устойчив. Нам нужны некоторые оценки для $\alpha(t)$. Ограничимся физически наиболее естественными условиями при $t = 0$:

$$A > 0, \quad B \leqslant 0, \quad \text{т.е.} \quad \alpha(0) > 0.$$

Рассмотрим поле направлений $\dot{\alpha}$. На плоскости (t, α) введем кривую (рис. 9)

$$\dot{\alpha} = 0, \quad \text{т.е.} \quad \tilde{\alpha}(t) = \sqrt{b(t)/a(t)} = O(1).$$

При $\alpha = 0$, очевидно, $\dot{\alpha} > 0$. Выделим области $\dot{\alpha} > 0$ (ниже кривой $\tilde{\alpha}(t)$) и $\dot{\alpha} < 0$ (выше кривой $\tilde{\alpha}(t)$). Несложный анализ показывает, что

$$0 \leqslant \alpha(t) \leqslant \max_t \tilde{\alpha}(t) = O(1).$$

Этого нам достаточно для дальнейшего.

Посмотрим, что дает теория численного интегрирования, примененная к уравнению $\dot{\alpha} = -b\alpha^2 + a$. Если бы мы оценивали устойчивость численного интегрирования для этой системы по самой общей теореме, оперирующей с оценкой погрешности типа εe^{Ct} (где C — константа Липшица правой части, ε — локальная погрешность), то картина была бы пессимистической. В самом деле,

$$C \approx \left| \frac{\partial}{\partial \alpha} (b\alpha^2) \right| \approx 2b\alpha \approx 80,$$

и все трудности были бы такими же, как и при методе фундаментальных решений. Но

$$\frac{\partial}{\partial \alpha}(-b\alpha^2 + a) = -2b\alpha < 0,$$

т.е. мы получили устойчивое решение, для которого специальная теорема о точности численного интегрирования не содержит экспоненциального множителя, и для шага τ достаточно иметь только соотношение

$$\tau \frac{\partial}{\partial \alpha}(b\alpha^2) \ll 1, \quad \text{т.е. } 80\,\tau \ll 1.$$

Итак, нас выручает устойчивость искомого решения $\alpha(t)$. В задаче для β та же самая ситуация:

$$\frac{\partial}{\partial \beta}(-b\alpha\beta + \alpha\varphi + f) = -b\alpha \approx -40,$$

т.е. мы имеем дело с интегрированием устойчивой задачи.

Наконец, обратная прогонка. Ее уравнение имеет вид

$$\dot{y} = b\alpha y + (\beta b + \varphi), \qquad (b\alpha)_y = b\alpha \approx 40 > 0.$$

Эта задача неустойчива вправо и, соответственно, устойчива влево. Но ведь нам нужно интегрировать ее именно справа-налево! И здесь все в порядке, несмотря на присутствие большого параметра.

Заметим, что прогонку можно осуществить в обратном направлении — решая уравнение для α справа-налево. В этом случае (см. рис. 9) траектория $\alpha(t)$ «притягивается» к кривой $\alpha = -\sqrt{b(t)/a(t)}$ $(\alpha(t) < 0)$ и интегрируется устойчивая влево задача.

§ 10. Прогонка в разностной задаче Штурма–Лиувилля

Рассмотрим классическую краевую задачу Штурма—Лиувилля:

$$\frac{d}{dt}\left[p(t)\frac{dx}{dt}\right] + q(t)\frac{dx}{dt} + r(t)x(t) = f(t),$$

с краевыми условиями общего вида:

$$\alpha\dot{x} + \beta x = \gamma, \quad t = 0, \qquad \alpha_1\dot{x} + \beta_1 x = \gamma_1, \quad t = T.$$

Начнем с построения разностной схемы, т.е. разностной аппроксимации задачи. Введем сетку, для простоты равномерную:

$$\{t_n\}_{n=0}^{N}, \quad t_n = n\tau, \quad \tau = T/N, \quad N \gg 1,$$

и счетные величины x_n, $n = 0, 1, \ldots, N$.

Построим разностное уравнение, аппроксимирующее дифференциальное:

$$\frac{1}{\tau}\left[p_{n+1/2}\frac{x_{n+1}-x_n}{\tau}-p_{n-1/2}\frac{x_n-x_{n-1}}{\tau}\right]+q_n\frac{x_{n+1}-x_{n-1}}{2\tau}+r_nx_n=f_n,$$

где $p_{n+1/2}=p(t_n+\tau/2)$, $q_n=q(t_n)$ и т.д. Это уравнение можно написать только при $n=1, 2, \ldots, N-1$ (во внутренних узлах сетки). Для дальнейшего уравнениям удобно придать стандартную форму:

$$a_nx_{n-1}-b_nx_n+c_nx_{n+1}=d_n,$$

где a_n, b_n, c_n, d_n — так называемые локальные коэффициенты схемы. Имеем для них выражения

$$a_n=\frac{1}{\tau^2}p_{n-1/2}-\frac{1}{\tau}q_n,\qquad c_n=\frac{1}{\tau^2}p_{n+1/2}+\frac{1}{\tau}q_n,$$

$$b_n=a_n+c_n-r_n,\qquad d_n=f_n.$$

Примем довольно естественные с физической точки зрения условия $p>0$, $r\leqslant 0$ и отметим важное соотношение $b_n\geqslant a_n+c_n$.

Аппроксимируем левое краевое условие:

$$\alpha\frac{x_1-x_0}{\tau}+\beta x_0=\gamma,$$

и запишем его в стандартной форме:

$$-b_0x_0+c_0x_1=d_0,\quad b_0=\frac{\alpha}{\tau}-\beta,\quad c_0=\frac{\alpha}{\tau},\quad d_0=\gamma.$$

Ради простоты ограничимся физически наиболее естественными условиями $\alpha>0$, $\beta<0$; следовательно $b_0>c_0$.

Аппроксимируем правое краевое условие:

$$\alpha_1\frac{x_N-x_{N-1}}{\tau}+\beta_1x_N=\gamma_1,$$

и запишем его в стандартной форме:

$$a_Nx_{N-1}-b_Nx_N=d_N.$$

Итак, мы получили специальную, но очень распространенную в приложениях систему линейных алгебраических уравнений:

$$-b_0x_0+c_0x_1=d_0,$$

$$a_nx_{n-1}-b_nx_n+c_nx_{n+1}=d_n,\quad n=1, 2, \ldots, N-1,$$

$$a_Nx_{N-1}-b_Nx_N=d_N.$$

Матрица системы имеет так называемую трехдиагональную (якобиеву) форму:

$$\begin{pmatrix} -b_0 & c_0 & & & & & 0 \\ a_1 & -b_1 & c_1 & & & & \\ & a_2 & -b_2 & c_2 & & & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ & & a_n & -b_n & c_n & & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ & & & a_{N-1} & -b_N & c_N & \\ 0 & & & & a_N & -b_N & \end{pmatrix}.$$

Такие матрицы часто появляются при аппроксимации дифференциальных уравнений разностными. Их специфика — большой порядок ($N = T/\tau$) и огромное число нулей (так как операторы дифференцирования являются операторами локального типа: значение производной функции в какой-то точке зависит только от значений функции в сколь угодно малой окрестности этой точки).

Большую роль в вычислительной математике играют так называемые *экономные* методы решения подобных систем уравнений. Это такие методы, в которых количество операций пропорционально первой степени числа неизвестных, т.е. в данном случае $O(N)$. Напомним, что если бы мы просто сослались на то, что получена система линейных алгебраических уравнений, которую можно решать любой стандартной программой, дело было бы довольно скверным. В общем случае решение системы N алгебраических уравнений с N неизвестными требует $O(N^3)$ операций и $O(N^2)$ ячеек памяти.

Для систем уравнений с якобиевой матрицей был разработан специальный *метод прогонки*, требующий $O(N)$ операций и $O(N)$ ячеек памяти. Этот метод был разработан почти одновременно в нескольких местах учеными, занимавшимися в сущности одной и той же проблемой. Она была связана с закрытыми работами, поэтому публикации последовали спустя годы после того, как была придумана и весьма эффективно применена прогонка.

Алгоритм прогонки. Решение ищется в форме *прогоночного соотношения*:

$$x_{n-1} = P_n x_n + Q_n, \qquad n = 1, 2, \ldots, N,$$

где P_n, Q_n — неизвестные пока *прогоночные коэффициенты*. Нетрудно видеть, что, определив P_n, Q_n, мы в сущности приведем систему с трехдиагональной матрицей к системе с двухдиагональной матрицей.

Алгоритм начинается с того, что левое краевое условие записывается в форме прогоночного соотношения:

$$x_0 = (c_0/b_0)\, x_1 - d_0/b_0,$$

т.е. $P_1 = c_0/b_0$, $Q_1 = -d_0/b_0$ (отметим, что $P_1 < 1$). Далее следует процесс *прямой прогонки*: последовательно (по рекуррентным формулам) вычисляются P_2, Q_2, затем P_3, Q_3, и т.д. вплоть до P_N, Q_N.

Выведем эти рекуррентные формулы. Пусть прогоночные коэффициенты P_n, Q_n уже вычислены. Подставляя $x_{n-1} = P_n x_n + Q_n$ в n-е уравнение $a_n x_{n-1} - b_n x_n + c_n x_{n+1} = d_n$, получаем

$$a_n(P_n x_n + Q_n) - b_n x_n + c_n x_{n+1} = d_n.$$

Соотношение между x_n и x_{n+1} представим в форме прогоночного соотношения, т.е. разрешим его относительно x_n:

$$x_n = \frac{c_n}{b_n - a_n P_n}\, x_{n+1} + \frac{a_n Q_n - d_n}{b_n - a_n P_n}.$$

Оно примет стандартный вид $x_n = P_{n+1} x_{n+1} + Q_{n+1}$, если положить

$$P_{n+1} = \frac{c_n}{b_n - a_n P_n}, \qquad Q_{n+1} = \frac{a_n Q_n - d_n}{b_n - a_n P_n}.$$

Это и есть рекуррентные формулы прямой прогонки. По ним вычисляем P_n, Q_n вплоть до P_N, Q_N. (Для этой цели мы последний раз можем использовать стандартное трехчленное уравнение с номером $N-1$.)

Имея неиспользованное пока правое краевое условие (N-е уравнение) $a_N x_{N-1} - b_N x_N = d_N$ и прогоночное соотношение $x_{N-1} = P_N x_N + Q_N$, можно найти величину x_N (разрешение правого краевого условия).

Неизвестные x_n последовательно определяются справа-налево по формулам (обратная прогонка)

$$x_{n-1} = P_n x_n + Q_n, \qquad n = N, N-1, \ldots, 1.$$

Исследование устойчивости прогонки. Проведем исследование вычислительной устойчивости прогонки, т.е. покажем, что погрешности вычислений, связанные с конечной разрядностью машинных чисел (погрешности округления), и погрешности, связанные с машинной реализацией арифметических операций, накапливаются в такой мере, что это не приводит к существенным погрешностям в результате.

Рассмотрим сначала процесс прямой прогонки коэффициента P_n. Введем величины P_n, вычисленные по формулам прогонки $P_n = c_n/(b_n - a_n P_n)$ в идеальной арифметике, т.е. без погрешностей округления и погрешностей в выполнении операций. Реальные расчеты на ЭВМ дают $P_n^{\text{м}} = P_n + \delta_n$, где δ_n — погрешность предшествующих получению P_n вычислений.

Предположим, что δ_n — малая погрешность (такова она, во всяком случае, на начальном этапе прогонки) и проанализируем процесс ее эволюции, записывая соотношение между δ_n и δ_{n+1} в виде

$$P_{n+1} + \delta_{n+1} = \frac{c_n}{b_n - a_n(P_n + \delta_n)} + \varepsilon_n.$$

Здесь ε_n — суммарная погрешность, связанная с выполнением операций в правой части формулы. В нее включаются также погрешности машинного представления коэффициентов c_n, b_n, a_n, погрешность выполнения машинных операций и погрешность округления, связанная с записью P_{n+1} в памяти в виде $P_{n+1}^{\text{м}}$. Последняя погрешность очень мала и зависит от разрядности представления чисел в ЭВМ.

Таким образом, погрешность δ_{n+1} есть следствие двух погрешностей: наследственной погрешности δ_n, в которой суммируются погрешности предшествующих вычислений, и локальной погрешности ε_n, для которой нетрудно указать хорошую оценку через погрешности c_n, b_n и т.д.

Воспользуемся тем, что $|\delta_n| \ll P_n$, и применим линейный анализ, пренебрегая величинами $O(\delta^2)$. Используя обычное исчисление дифференциалов, получаем

$$P_{n+1} + \delta_{n+1} = \frac{c_n}{b_n - a_n P_n} + \frac{c_n a_n}{(b_n - a_n P_n)^2}\,\delta_n + \varepsilon_n.$$

Из этого соотношения следует формула

$$\delta_{n+1} = \frac{a_n}{c_n}(P_{n+1})^2\,\delta_n + \varepsilon_n, \quad n = 0, 1, \ldots, N-1.$$

Заметим, что число шагов, выполняемых по рекуррентной формуле, достаточно велико (порядка T/τ), и нас будет интересовать асимптотическое поведение погрешности (при $N = T/\tau \to \infty$).

Сначала получим важные для анализа свойства прогоночных коэффициентов.

Лемма 1. При значениях $b_n > a_n + c_n$ и $P_1 < 1$ для всех n имеем $0 \leqslant P_n \leqslant 1$.

В самом деле, если $0 \leqslant P_n \leqslant 1$, то

$$\text{а)}\quad P_{n+1} = \frac{c_n}{b_n - a_n P_n} \geqslant \frac{c_n}{b_n - a_n} > 0.$$

$$\text{б)}\quad P_{n+1} = \frac{c_n}{b_n - a_n P_n} \leqslant \frac{c_n}{a_n + c_n - a_n P_n} = \frac{c_n}{c_n + a_n(1 - P_n)} \leqslant 1.$$

Разумеется была использована положительность локальных коэффициентов a_n, b_n, c_n.

Лемма 2. Пусть $p(t)$ удовлетворяет условию Липшица. Тогда $|a_n/c_n| \leqslant 1 + C\tau$, где постоянная C не зависит от τ.

Действительно,

$$\frac{a_n}{c_n} = \frac{p(t_n - \tau/2) - \tau q_n}{p(t_n + \tau/2) + \tau q_n} = 1 + O(\tau).$$

С учетом приведенных оценок имеем

$$|\delta_{n+1}| \leqslant (1 + C\tau)|\delta_n| + \varepsilon, \quad |\varepsilon_n| \leqslant \varepsilon.$$

Используя это соотношение и стандартные рассуждения (см. § 5, 8), получаем

$$|\delta_n| \leqslant (1 + C\tau)^n |\delta_0| + \frac{(1 + C\tau)^n}{C\tau}\varepsilon.$$

При $n \leqslant T/\tau$ и достаточно малых τ, таких, что $C\tau < 1$, мы приходим к оценке

$$|\delta_n| \leqslant e^{CT}|\delta_0| + \frac{\varepsilon}{C\tau} e^{CT}.$$

Обсудим этот результат. Прежде всего есть величина $\varepsilon/C\tau$, из которой следует, что при слишком малом шаге τ погрешность может стать недопустимо большой. Напомним, что ε зависит от разрядности чисел ЭВМ. То, что при разностной аппроксимации дифференциальных уравнений конечная разрядность чисел ограничивает снизу разумный малый шаг τ, нам уже известно; здесь это обстоятельство проявилось еще раз. Однако в большинстве расчетов N равны 100, 1000, так что отношение ε/τ очень мало (на БЭСМ-6, во всяком случае; на ЕС уже нужно быть осторожнее).

В полученной оценке есть неприятный множитель e^{CT}. При решении задач на больших интервалах времени, при $CT \gg 1$, возможны серьезные затруднения. Однако дело не только в величине T, но и в гладкости функции $p(t)$, т.е. чем меньше ее константа Липшица, тем благоприятнее ситуация. Нужно, однако, иметь в виду некоторый неиспользованный нами в грубой оценке резерв: если, как это часто бывает в прикладных задачах, $b_n > a_n + c_n + s^2$ $(s > 0)$, то $P_n \leqslant q < 1$, оценка может быть существенно улучшена и в неко-

торых случаях можно исключить множитель e^{CT}. Но мы этим заниматься не будем.

Неприятным является предположение о гладкости $p(t)$. В приложениях часто встречаются задачи с кусочно-гладкими $p(t)$, т.е. $p(t)$ имеет небольшое число точек разрыва, а между ними гладкая. Такие изолированные разрывы не имеют катастрофических последствий. Несложный анализ, являющийся простым обобщением проведенного выше, показывает, что в этом случае в оценке множитель e^{CT} заменится на

$$e^{CT} \prod_{k=1}^{K} \left| \frac{p(t_k^* - 0)}{p(t_k^* + 0)} \right|,$$

где t_k^* $(k = 1, 2, \ldots, K)$ — точки разрыва $p(t)$.

Мы не будем проводить подробно анализ остальных частей алгоритма прогонки. Ограничимся лишь самыми простыми соотношениями. Обозначим $Q_n^{\text{м}} = Q_n + \xi_n$, где ξ_n — погрешность вычисления Q_n. Запишем соотношение для Q_{n+1}:

$$Q_{n+1} = \frac{a_n}{c_n} P_{n+1} Q_n - \frac{d_n}{c_n} P_n.$$

С учетом погрешностей вычислений имеем

$$Q_{n+1}^{\text{м}} = Q_{n+1} + \xi_{n+1} =$$
$$= \frac{a_n}{c_n} (P_{n+1} + \delta_{n+1})(Q_n + \xi_n) - \frac{d_n}{c_n} (P_{n+1} + \delta_{n+1}) + \varepsilon_{n+1}.$$

Отсюда

$$\xi_{n+1} = \frac{a_n}{c_n} P_n \xi_n + \varepsilon'_{n+1},$$

где

$$\varepsilon'_{n+1} = \frac{a_n}{c_n} Q_n \delta_{n+1} - \frac{d_n}{c_n} \delta_{n+1} + \varepsilon_{n+1}.$$

Далее погрешность ξ_n оценивается так же, как это было сделано выше при оценке δ_n.

Аналогично анализируется обратная прогонка $x_{n-1} = P_n x_n + Q_n$. С учетом погрешностей вычислений, обозначая через η_n погрешность в x_n, имеем

$$x_{n-1}^{\text{м}} = x_{n-1} + \eta_{n-1} = (P_n + \delta_n)(x_n + \eta_n) + (Q_n + \xi_n) + \varepsilon_n,$$

откуда

$$\eta_{n-1} = P_n \eta_n + (x_n \delta_n + \xi_n + \varepsilon_n).$$

И здесь ключевой факт: $|P_n| \leqslant 1$. Основную роль, как было видно из предшествующего анализа, играет «коэффициент усиления наследственной погрешности». Если он меньше единицы или хотя бы не более $1 + O(1/N)$, накопление погрешностей не имеет катастрофического характера. Крайне неприятной является ситуация, в которой этот коэффициент превосходит некоторое не зависящее от τ число $q > 1$. Тогда погрешность накапливается, усиливаясь за каждый шаг в q раз, и при достаточно малом шаге τ величина $q^N = q^{T/\tau}$ может стать катастрофически большой.

§ 11. Численное интегрирование задачи Коши для уравнений с частными производными

Изучение одного из важнейших разделов современной вычислительной математики начнем с простой задачи, которая даст повод ввести основные идеи метода конечных разностей. Это классическая задача для уравнения теплопроводности.

Итак, в области $0 \leqslant x \leqslant X$, $0 \leqslant t \leqslant T$ нужно найти функцию $u(t, x)$, удовлетворяющую:

а) уравнению $\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(t, x)$ (всюду в области);

б) левому краевому условию $-\alpha_1 \frac{\partial u}{\partial x} + \beta_1 u = \psi_1(t)$ при $x = 0$;

в) правому краевому условию $\alpha_2 \frac{\partial u}{\partial x} + \beta_2 u = \psi_2(t)$ при $x = X$;

г) начальным условиям Коши $u(0, x) = u_0(x)$, $x \in [0, X]$, при $t = 0$.

Заметим сразу же, что формально метод без существенных изменений можно применить и для решения более сложной задачи с нелинейным уравнением, например

$$c(t, x, u) \frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left[\varkappa(t, x, u) \frac{\partial u}{\partial x}\right] + f(t, x, u).$$

Возможны и другие усложнения задачи, и они на первый взгляд легко вписываются в метод конечных разностей. Рассмотрим вначале линейное уравнение, затем проведем формальное обобщение метода на более сложные задачи и обсудим, так ли все это просто на самом деле. Введем основные элементы метода сеток.

Сетка. Область определения функции u покрывается дискретным множеством точек

$$\{x_m, t_n\}, \quad m = 0, 1, \ldots, M, \quad n = 0, 1, \ldots, N.$$

Ради простоты изложения будем считать сетку равномерной, т.е. $x_m = mh$, где $h = X/M$ — шаг сетки по x, M — число узлов по x; $t_n = n\tau$, где $\tau = T/N$ — шаг сетки по t, N — число узлов по t.

Сеточная функция. Приближенное решение задачи ищем в виде сеточной функции, т.е. функции, определенной в каждом узле сетки. Эту функцию обозначим $\{u_m^n\}$. Значение u_m^n будем трактовать как приближенное значение функции $u(t, x)$ в узле (t_n, x_m), т.е. $u_m^n \approx u(t_n, x_m)$.

Разностная аппроксимация уравнения. Сеточную функцию получим как решение некоторого уравнения, аппроксимирующего дифференциальное. Существует много технических приемов построения таких уравнений, мы начнем с самого простого и наглядного. Он состоит в том, что входящие в уравнение производные заменяются подходящими разностными отношениями. Это можно сделать тоже неоднозначно.

Приведем достаточно популярные разностные уравнения.

Явная схема:

$$\frac{u_m^{n+1} - u_m^n}{\tau} = \frac{u_{m-1}^n - 2u_m^n + u_{m+1}^n}{h^2} + f_m^n. \qquad (1)$$

Принятый способ разностной аппроксимации называют *схемой*. Обычно структуру схемы поясняют ее шаблоном. Шаблон — это совокупность узлов сетки, в которых берутся значения функции, участвующие в аппроксимации уравнения в данном узле (n, m).

Неявная схема:

$$\frac{u_m^{n+1} - u_m^n}{\tau} = \frac{u_{m-1}^{n+1} - 2u_m^{n+1} + u_{m+1}^{n+1}}{h^2} + f_m^n. \qquad (2)$$

Шаблоны явной и неявной схем показаны на рис. 10.

Обычно каждое разностное уравнение относят к некоторому узлу счетной сетки. Удобно считать уравнения (1), (2) отнесенными к точке $(n+1, m)$. Нетрудно видеть, что эти уравнения могут быть составлены не во всех узлах сетки, а только во *внутренних*, т.е. в тех узлах, в которых шаблон не выходит за пределы сетки, в данном случае для $m = 1, 2, \ldots, M-1$, $n = 0, 1, \ldots, N-1$. Тем самым мы имеем уравнения для точек $[m = 1, 2, \ldots, M-1] \times [n = 1, 2, \ldots, N]$.

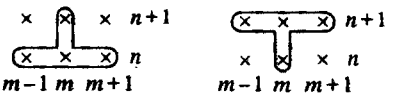


Рис. 10

Таким образом, без уравнений остались пока самый нижний ряд узлов и крайние левый и правый ряды. В этих узлах следует составить уравнения, аппроксимирующие начальные данные и краевые условия:

аппроксимация начальных данных:

$$u_m^0 = u_0(x_m), \quad m = 0, 1, \ldots, M;$$

аппроксимация левого краевого условия:

$$-\alpha_1 \frac{u_1^n - u_0^n}{h} + \beta_1 u_0^n = \psi_1(t_n), \quad n = 1, 2, \ldots, N; \tag{3}$$

аппроксимация правого краевого условия:

$$\alpha_2 \frac{u_M^n - u_{M-1}^n}{h} + \beta_2 u_M^n = \psi_2(t_n), \quad n = 1, 2, \ldots, N, \tag{4}$$

где $\alpha_i \geqslant 0$, $\beta_i \geqslant 0$, $\alpha_i + \beta_i > 0$, $i = 1, 2$. Теперь мы имеем столько неизвестных, сколько точек, и столько же уравнений.

Решение разностных уравнений. (Или, как принято говорить, *реализация* разностной схемы.) Чтобы завершить описание схемы, нужно дать алгоритм вычисления u_m^n, подсчитать количество операций и требуемые ресурсы памяти.

Общей чертой реализации разностных схем для так называемых эволюционных задач (т.е. задач, в которых одна из независимых переменных играет особую роль времени) является *счет по слоям*. Слоем мы называем совокупность неизвестных, определенных в узлах одного горизонтального ряда; n-й слой будем обозначать u^n, имея в виду величины $\{u_m^n\}_{m=0}^M$.

Схема счета по слоям очень проста. Пусть n-й слой уже сосчитан, т.е. переменные, входящие в u^n, и переменные всех предшествующих слоев $u^0, u^1, \ldots, u^{n-1}$ уже известны. Имеется алгоритм, который по значениям u^n вычисляет u^{n+1}, используя, быть может, и другие нижние слои $u^{n-1}, \ldots$ Этот алгоритм называют «реализацией шага», обозначим его S.

Заметим, что разностные уравнения (1), (2) связывают неизвестные только на двух соседних слоях. В этом случае реализация шага может быть записана в виде $u^{n+1} = S(u^n)$. Такие схемы называют «двуслойными». В трехслойных схемах реализация шага имеет вид $u^{n+1} = S(u^n, u^{n-1})$. Так как слой u^0 известен из начальных условий, можно находить последовательно слой за слоем: $u^1 = S(u^0)$, $u^2 = S(u^1)$ и т.д.

Реализация явной схемы. Она совсем проста. Итак, пусть u^n (n-й слой) известен. Запишем (1) в форме

$$u_m^{n+1} = u_m^n + \frac{\tau}{h^2}(u_{m-1}^n - 2u_m^n + u_{m+1}^n) + \tau f_m^n. \qquad (5)$$

Тем самым мы имеем явную формулу вычисления u_m^{n+1}, но только для $m = 1, 2, \ldots, M-1$.

Для завершения шага нужно вычислить еще u_0^{n+1} и u_M^{n+1}. Из левого краевого условия (3) находим

$$u_0^{n+1} = \frac{\alpha_1}{\alpha_1 + h\beta_1} u_1^{n+1} + \frac{h\psi_1(t_{n+1})}{\alpha_1 + h\beta_1}. \qquad (6)$$

Аналогично вычисляется u_M^{n+1} из правого краевого условия (4):

$$u_M^{n+1} = \frac{\alpha_2}{\alpha_2 + h\beta_2} u_{M-1}^{n+1} + \frac{h\psi_2(t_n)}{\alpha_2 + h\beta_2}. \qquad (7)$$

Вычисление u_0^{n+1}, u_M^{n+1} производится после расчета по формуле (5), так что значения u_1^{n+1}, u_{M+1}^{n+1} уже известны. Легко подсчитать, что реализация шага требует $O(M)$ операций и, следовательно, вся задача решается за $O(MN)$ операций.

Оценим ресурсы памяти. На первый взгляд кажется, что требуется $(N+1)(M+1)$ ячеек памяти. Но нетрудно видеть, что можно обойтись и $2(M+1)$-й ячейкой, если заметить, что предшествующие слои больше не понадобятся и могут быть «забыты».

Приведем схему счета, в которой используются только два одномерных массива $u0(0{:}M)$, $u1(0{:}M)$. Можно обойтись и одним, используя на языке FORTRAN оператор EQUIVALENCE ($u0(0)$, $u1(2)$). Разумеется, перенос массива $u1$ на место $u0$ (см. схему) в этом случае делается обратным циклом. Обратим внимание на то, что печатаются и просматриваются не все полученные в расчете числа, а только некоторые слои u^n, соответствующие времени p, $2p$, $3p$ и т.д.

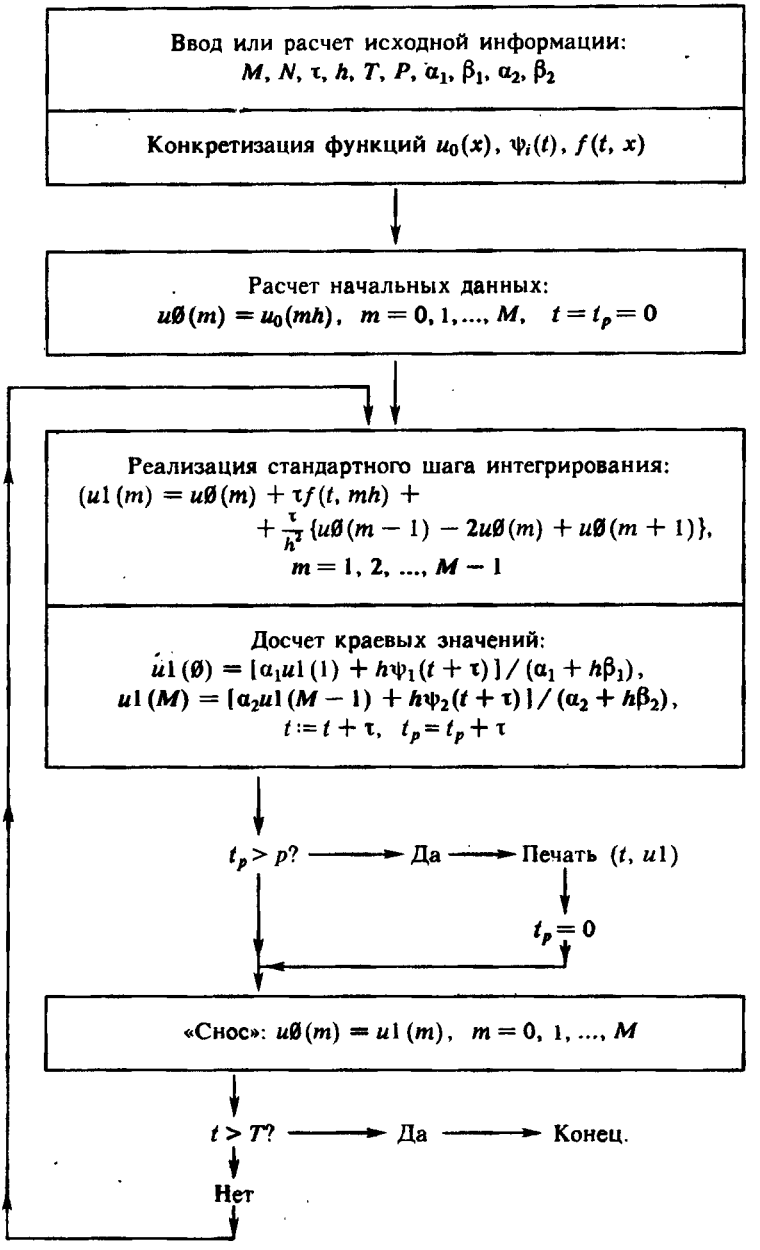
Реализация неявной схемы. Здесь мы сталкиваемся с характерной для всех неявных схем проблемой — необходимостью решения так называемых *уравнений на верхнем слое.* В самом деле, пусть слой u^n известен. Выписывая уравнения для точек $(n+1)$-го слоя и перенося неизвестные в левую часть, из (2)—(4) получаем

$$\frac{\tau}{h^2} u_{m-1}^{n+1} - \left(1 + 2\frac{\tau}{h^2}\right) u_m^{n+1} + \frac{\tau}{h^2} u_{m+1}^{n+1} = -u_m^n + \tau f_m^n,$$

$$(\alpha_1 + h\beta_1) u_0^{n+1} - \alpha_1 u_1^{n+1} = h\psi_1^{n+1},$$

$$(\alpha_2 + h\beta_2) u_M^{n+1} - \alpha_2 u_{M-1}^{n+1} = h\psi_2^{n+1}$$

Реализация явной схемы. Счет по слоям



($m = 1, 2, \ldots, M$). Это уже знакомая нам система уравнений с трехдиагональной матрицей (матрицей Якоби). Она может быть решена методом прогонки с затратой $O(M)$ операций.

Реализация неявной разностной схемы отличается одним новым моментом: стандартный шаг, переход от известного n-го слоя к $(n+1)$-му, требует решения системы (обычно высокого порядка $M = X/h$) линейных алгебраических уравнений со специфической матрицей. Для решения используется специальный алгоритм. Это характерная черта современных методов решения уравнений с частными производными. В них большую роль играют именно неявные схемы (аппроксимация пространственных производных «на верхнем слое»), и в связи с этим возникают проблемы решения «уравнений на верхнем слое». Причины, побуждающие выносить аппроксимацию пространственных производных «на верхний слой» будут обсуждены чуть позже.

Обобщения на нелинейные задачи. Характерной особенностью метода сеток является легкость перехода к гораздо более сложным задачам. Правда, речь идет о формальной операции, и не надо эту легкость понимать слишком уж буквально. Пусть нужно решать гораздо более сложную нелинейную задачу:

$$c(t, x, u)\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left[\varkappa(t, x, u)\frac{\partial u}{\partial x}\right] + f(t, x, u)$$

с теми же самыми начальными и краевыми условиями.

Есть простой способ справляться с нелинейным характером уравнений, аппроксимируя нелинейные зависимости на нижнем слое:

$$c_m^n \frac{u_m^{n+1} - u_m^n}{\tau} = \frac{1}{h}\left[\varkappa_{m+1/2}^n \frac{u_{m+1}^n - u_m^n}{h} - \varkappa_{m-1/2}^n \frac{u_m^n - u_{m-1}^n}{h}\right] + f_m^n,$$

где

$$c_m^n = c(t_n, x_m, u_m^n), \qquad f_m^n = f(t_n, x_m, u_m^n),$$

$$\varkappa_{m+1/2}^n = \varkappa\left(t_n, \frac{x_m + x_{m+1}}{2}, \frac{u_m^n + u_{m+1}^n}{2}\right).$$

Все функции, в которые u входит нелинейно, вычисляются по нижнему n-му слою, т.е. при реализации шага они вычисляются через уже известные значения u_m^n. В этом случае схема решения нелинейного уравнения теплопроводности ничем в сущности не отличается от явной схемы решения линейного уравнения.

Точно так же можно записать и неявную схему с нелинейностью на нижнем слое:

$$c_m^n \frac{u_m^{n+1} - u_m^n}{\tau} = \frac{1}{h}\left[\varkappa_{m+1/2}^n \frac{u_{m+1}^{n+1} - u_m^{n+1}}{h} - \varkappa_{m-1/2}^n \frac{u_m^{n+1} - u_{m-1}^{n+1}}{h}\right] + f_m^n,$$

и счет реализуется так же, как и в линейном случае.

Итак, переход к нелинейным задачам почти ничего не стоит. Так ли это? Действительно ли все так просто или есть какие-то скрытые от поверхностного взгляда сложности? Ответ, который можно дать уже здесь, такой: если нелинейный характер уравнений не порождает в решении каких-то особенностей, сложного характера функций, больших градиентов, высокочастотных осцилляций и т.п., то, как правило, решение нелинейных уравнений методом конечных разностей немногим сложнее решения линейных.

Таким образом, не нелинейность сама по себе, а связанные с ней возможные нарушения гладкости искомого решения $u(t, x)$ (их может и не быть) осложняют фактическое решение нелинейных задач. Этим разъяснением мы пока и ограничимся.

Нелинейные уравнения на верхнем слое. В некоторых ситуациях приходится часть нелинейных зависимостей «выносить на верхний слой». Тогда возникают проблемы решения нелинейных уравнений (с большим числом неизвестных) на верхнем слое. В этом случае методы типа прогонки комбинируются с итерационными методами решения нелинейных уравнений. Как это делается, покажем на самом простом уравнении, в котором нелинейность входит только в правую часть.

Пусть решается стандартная краевая задача для уравнения

$$u_t = u_{xx} + f(u)$$

и по каким-то причинам используется неявная нелинейная схема

$$\frac{u_m^{n+1} - u_m^n}{\tau} = \frac{u_{m+1}^{n+1} - 2u_m^{n+1} + u_{m-1}^{n+1}}{h^2} + f(u_m^{n+1}).$$

Здесь, как видим, система уравнений для неизвестных u_m^{n+1} нелинейная. Решается она методом итераций с линеаризацией по Ньютону. Обозначим i-е приближение к искомому u_m^{n+1} через $u_m^{(i)}$.

Рассмотрим стандартную ситуацию: известны u^n и $u_m^{(i)}$, т.е. i итераций проделано. Надо найти $u_m^{(i+1)}$ $(m = 0, 1, \ldots, M)$. Линеаризуем выражение $f(u_i^{(i+1)})$, т.е. заменим его на

$$f(u_m^{(i)} + (u_m^{(i+1)} - u_m^{(i)})) \approx f(u_m^{(i)}) + f_u(u_m^{(i)})\,(u_m^{(i+1)} - u_m^{(i)}).$$

Теперь для определения $u^{(i+1)}$ мы имеем уже линейную систему:

$$\frac{u_m^{(i+1)} - u_m^n}{\tau} = \frac{u_{m-1}^{(i+1)} - 2u_m^{(i+1)} + u_{m+1}^{(i+1)}}{h^2} + f(u_m^{(i)}) + f_u(u_m^{(i)})u_m^{(i+1)} - f_u(u_m^{(i)})u_m^{(i)}.$$

Иногда используют более простую (но менее эффективную и надежную) схему простых итераций (см. § 1), основанную на принципе «нелинейность с предыдущей итерацией»:

$$\frac{u_m^{(i+1)}-u_m^n}{\tau}=\frac{u_{m-1}^{(i+1)}-2u_m^{(i+1)}+u_{m+1}^{(i+1)}}{h^2}+f(u_m^{(i)}).$$

Эти уравнения по своей структуре не отличаются от обычных уравнений на верхнем слое в неявной схеме. Существенным фактором, облегчающим решение нелинейных уравнений, является наличие хорошего начального приближения. В качестве такового естественно взять $u_m^{(0)}=u_m^n$ $(m=0, 1, \ldots, M)$: за малое время τ решение $u(t, x)$ меняется мало, значит, и u^n мало отличается от u^{n+1}. Для таких итерационных процессов часто доказываются теоремы о сходимости при достаточно малом шаге τ.

Обоснование метода сеток. Как обычно, нам нужно установить факт сходимости численного решения к точному, т.е. сравнить u_m^n с $u(t_n, x_m)$ и получить оценку

$$|u_m^n-u(t_n, x_m)|\leqslant C_1(\tau^k+h^p),\quad \forall\ n\leqslant T/\tau,\ \ m\leqslant X/h,$$

где C_1 не зависит от τ, h и оценка относится к семейству решений, зависящему от шагов τ, h сетки. Если такая оценка будет получена, будем говорить, что «разностная схема имеет порядок точности k по τ и p по h».

В большинстве случаев в практике решения сложных задач такую оценку получить не удается. Но в любом случае мы должны иметь если не полное обоснование метода, то хотя бы какие-то соображения, грубые оценки. И здесь появляется первое необходимое требование к разностной схеме: она должна аппроксимировать решаемую дифференциальную задачу. Это минимальное требование.

Напомним общую схему исследования аппроксимации. Дифференциальную задачу записываем в операторной форме:

$$LU=F,$$

где L — оператор, U — искомая функция, F — заданная правая часть. Аппроксимирующую задачу (точнее, семейство задач) представим в виде

$$L_s u_s=F_s.$$

Здесь s — символ сетки (т.е. параметров h и τ в нашем случае), u_s — сеточное решение, F_s — сеточная правая часть, L_s — оператор, действующий в пространстве сеточных функций.

Обозначим через U_s ограничение на сетку точного решения U дифференциальной задачи. Можно подставить U_s в разностные уравнения и вычислить невязку:

$$\eta_s = L_s U_s - F_s.$$

Если она стремится к нулю (при $h \to 0$, $\tau \to 0$), то говорят, что разностная задача аппроксимирует дифференциальную. Если установлена оценка $\|\eta_s\| \leqslant C(\tau^q + h^p)$, то говорят, что разностная задача (схема) имеет порядок аппроксимации q по τ и p по h.

Обычно устанавливается так называемая *формальная аппроксимация*, основанная на предположении такой гладкости решения, какая понадобится при оценке $\|\eta\|$. Это операция несложная, но нужна некоторая аккуратность и педантичность в оформлении задач в операторном виде.

Итак, сначала надо описать оператор L, отображающий функцию U, определенную на $[0, X] \times [0, T]$, в аналогичную функцию, притом так, чтобы запись $LU = F$ включала все, что есть в задаче. Положим

$$(LU)(t, x) = \begin{cases} U(0, x), & t = \emptyset, \quad x \in [\emptyset, X], \\ [U_t - U_{xx}](t, x), & t \in (\emptyset, T], \quad x \in (\emptyset, X), \\ [-\alpha_1 U_x + \beta_1 U](t, 0), & x = \emptyset, \quad t \in (\emptyset, T], \\ [\alpha_2 U_x + \beta_1 U](t, X), & x = X, \quad t \in (\emptyset, T]. \end{cases}$$

(Оператор L отображает U в комплекс из четырех разных функций.) Функцию F определим, используя ту или иную компоненту правой части в соответствующем диапазоне изменения t, x. Они те же, что и при определении LU:

$$F(t, x) = \{u_0(x);\ f(t, x);\ \psi_1(t);\ \psi_2(t)\}.$$

Очевидно, наша цель достигнута, и $LU = F$ есть компактная запись всей задачи.

То же самое надо проделать с разностной задачей: нужно определить L_s — отображение сеточной функции в сеточную. Для явной схемы имеем

$$(L_s u_s)_m^n = \begin{cases} u_m^0, & n = 0, \ m = 0, 1, \ldots, M, \\ \dfrac{u_m^n - u_m^{n-1}}{\tau} - \dfrac{u_{m-1}^{n-1} - 2u_m^{n-1} + u_{m+1}^{n-1}}{h^2}, & n = 1, 2, \ldots, N, \\ & m = 1, 2, \ldots, M-1, \\ -\alpha_1 \dfrac{u_1^n - u_0^n}{h} + \beta_1 u_0^n, & m = 0, \ n = 1, 2, \ldots, N, \\ \alpha_2 \dfrac{u_M^n - u_{M-1}^n}{h} + \beta_2 u_M^n, & m = M, \ n = 1, 2, \ldots, N. \end{cases}$$

Определим и сеточную правую часть:

$$(F_s)_m^n = \left\{u_0(x_m);\ f_m^{n-1};\ \psi_1^n;\ \psi_2^n\right\},$$

где

$$f_m^{n-1} = f(t_{n-1},\ x_m), \qquad \psi_1^n = \psi_1(t_n), \qquad \psi_2^n = \psi_2(t_n).$$

При вычислении $(F_s)_m^n$ используется тот или иной вариант правой части в зависимости от диапазона изменения индексов m, n.

Теперь можно оценить невязку η_s. Оценки эти тривиальны, они связаны с известными оценками погрешностей аппроксимации при замене производных теми или иными разностями. Они, разумеется, носят формальный характер и в данном случае основаны на предположениях о существовании у решения $U(t,x)$ ограниченных вторых производных по t и четвертых производных по x. Вычислим невязку:

$$(\eta_s)_m^n = (L_s U_s)_m^n - F_m^n.$$

Используя разложение в ряд Тейлора функции $U(n\tau, mh)$ по τ и h, получаем

$$(\eta_s)_m^n = \{\emptyset;\ O(\tau + h^2);\ O(h);\ O(h)\}.$$

Каждая компонента правой части относится к своему диапазону изменения индексов m, n. В данном случае схема имеет первый порядок аппроксимации по τ и h.

Итак, мы получаем знакомую ситуацию. Приближенное решение находится решением задачи $L_s u_s = F_s$, а точное U_s можно было бы найти решением почти такой же задачи $L_s U_s = F_s + \eta_s$. Но мы не знаем η_s, знаем только оценку для нее и то, что это есть величина сколь угодно малая, если шаги сетки достаточно малы. Мы не имеем права утверждать, что из совпадения (с точностью до погрешности аппроксимации) уравнений следует, что и их решения совпадают с точностью до величин порядка погрешности аппроксимации. Это право мы получим, если докажем *устойчивость разностной задачи.*

Разностная задача называется устойчивой, если из

$$L_s u_s = F_s + \varepsilon_s', \qquad L_s v_s = F_s + \varepsilon_s''$$

следует, что $\|u_s - v_s\| \leqslant C(\|\varepsilon'\| + \|\varepsilon''\|)$, причем постоянная C не зависит от сетки s (т.е. от τ, h). Установить устойчивость обычно бывает очень трудно. Но для линейного уравнения теплопроводности это можно сделать.

Устойчивость явной схемы. Рассмотрим уравнение с теми же краевыми условиями и начальными данными:

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left[\varkappa(t, x)\frac{\partial u}{\partial x}\right] + f(t, x).$$

Это есть линейное уравнение с переменными коэффициентами. Прежде всего заметим, что нужно конкретизировать вид норм. Устойчивость можно устанавливать в разных нормах. Здесь мы используем самую наглядную и надежную:

$$\|f\| = \max_{t,\,x} |f(t, x)|, \qquad \|u\| = \max_{t,\,x} |u(t, x)| \qquad \text{и т.д.}$$

Введем аналогичные нормы для сеточных функций:

$$\|u_s\| = \max_{m,\,n} |u_m^n|, \qquad \|u^n\| = \max_{m} |u_m^n| \qquad \text{и т.д.}$$

Основой для установления устойчивости является следующая лемма.

Лемма 1. Нормы сеточных функций u^n и u^{n+1} связаны между собой неравенством

$$\|u^n\| \leqslant \max\{\|u^n\| + \tau\|f\|, \|\psi\|\}, \quad \|\psi\| \equiv \max\{\|\psi_1/\beta_1\|, \|\psi_2/\beta_2\|\},$$

если выполнено *условие Куранта* $\|\varkappa\|\tau/h^2 \leqslant 1/2$ (мы ограничимся случаем $\beta_i > 0$; в случае $\beta = 0$, $\alpha = 1$ оценки проще).

Доказательство. Имеем соотношение

$$\frac{u_m^{n+1} - u_m^n}{\tau} = \frac{1}{h^2}\left[\varkappa_{m+1/2}^n(u_{m+1}^n - u_m^n) - \varkappa_{m-1/2}^n(u_m^n - u_{m-1}^n)\right] + f_m^n,$$

или

$$u_m^{n+1} = \frac{\tau}{h^2}\varkappa_{m-1/2}^n u_{m-1}^n + \left(1 - \frac{\tau}{h^2}\varkappa_{m-1/2}^n - \frac{\tau}{h^2}\varkappa_{m+1/2}^n\right)u_m^n + \\ + \frac{\tau}{h^2}\varkappa_{m+1/2}^n u_{m+1}^n + \tau f_m^n, \quad m = 1, 2, \ldots, M-1.$$

Отсюда следует (так как $|u_m^n| \leqslant \|u^n\|$, $\varkappa > 0$)

$$|u_m^{n+1}| \leqslant \frac{\tau}{h^2}\varkappa_{m-1/2}^n\|u^n\| + \frac{\tau}{h^2}\varkappa_{m+1/2}^n\|u^n\| + \\ + \left|1 - \frac{\tau}{h^2}(\varkappa_{m-1/2}^n + \varkappa_{m+1/2}^n)\right|\|u^n\| + \tau\|f\|.$$

Условие Куранта дает важное соотношение

$$\left|1 - \frac{\tau}{h^2}(\varkappa_{m-1/2}^n + \varkappa_{m+1/2}^n)\right| = 1 - \frac{\tau}{h^2}(\varkappa_{m-1/2}^n + \varkappa_{m+1/2}^n),$$

и мы получаем

$$|u_m^{n+1}| \leqslant \|u^n\| + \tau\|f\|, \quad m = 1, 2, \ldots, M-1.$$

Заметим, что

$$\|u^{n+1}\| = \{\text{либо } \max_{m=1,\ldots,M-1} |u_m^{n+1}|, \text{ либо } |u_0^{n+1}|, \text{ либо } |u_M^{n+1}|\}.$$

В первом случае имеем

$$\|u^{n+1}\| \leqslant \|u^n\| + \tau\|f\|.$$

Во втором случае используем краевое условие:

$$(\alpha_1 + h\beta_1)u_0^{n+1} = \alpha u_1^{n+1} + h\psi_1^{n+1}, \qquad \alpha_1, \beta_1 > 0.$$

Если $\|u^{n+1}\| = |u_0^{n+1}|$, то

$$|u_1^{n+1}| \leqslant |u_0^{n+1}|, \qquad (\alpha_1 + h\beta_1)|u_0^{n+1}| \leqslant \alpha_1|u_0^{n+1}| + h|\psi_1^{n+1}|,$$

т.е.

$$\|u^{n+1}\| = |u_0^{n+1}| \leqslant |\psi_1^{n+1}/\beta| \leqslant \|\psi\|.$$

Такую же оценку мы получим и в случае $\|u^{n+1}\| = |u_M^{n+1}|$. Итак, либо $\|u^{n+1}\| \leqslant \|\psi\|$, либо $\|u^{n+1}\| \leqslant \|u^n\| + \tau\|f\|$. В любом случае $\|u^{n+1}\| \leqslant \max\{\|u^n\| + \tau\|f\|, \|\psi\|\}$.

Теперь докажем следующую теорему.

Теорема 1. Явная линейная разностная схема при выполнении условия Куранта $\|\varkappa\|\tau/h^2 \leqslant 1/2$ устойчива по начальным данным, краевым условиям и правым частям.

Доказательство. Используем лемму 1 рекуррентно, обозначая ради простоты $u^n \equiv \|u^n\|$, $f \equiv \|f\|$ и т.д.:

$$u^n \leqslant \max\{u^{n-1} + \tau f, \psi\} \leqslant \quad .$$

$$\leqslant \max\{\max[u^{n-2} + \tau f, \psi] + \tau f, \psi\} =$$

$$= \max\{u^{n-2} + 2\tau f, \psi + \tau f\} \leqslant$$

$$\leqslant \max\{\max[u^{n-3} + \tau f, \psi] + 2\tau f, \psi + \tau f\} =$$

$$= \max\{u^{n-3} + 3\tau f, \psi + 2\tau f\} \leqslant$$

$$\cdots\cdots\cdots\cdots\cdots$$

$$\leqslant \max\{u^0 + n\tau f, \psi + (n-1)\tau f\}.$$

Так как $n\tau \leqslant T$, получаем результат

$$\|u^n\| \leqslant \|u^0\| + T\|f\| + \|\psi\|.$$

Обозначая $\|F_s\| = \|u^0\| + T\|f\| + \|\psi\|$, запишем оценку в форме $\|u_s\| \leqslant \|F_s\|$, т.е. $\|L_s^{-1}\| \leqslant 1$.

Мы установили оценку нормы решения разностной задачи через нормы начальных данных, правых частей и краевых условий. Это еще

не совсем то, что нужно. Нам нужно установить, что при малых возмущениях начальных данных, правых частей и краевых условий решение изменится соответственно мало. Но это следует из линейности задачи (как известно, ограниченность и непрерывность для линейных операторов — это одно и то же).

Воспроизведем это рассуждение. Если $L_s u_s = F_s$, $L_s U_s = F_s + \eta_s$, то в силу линейности $L_s(U_s - u_s) = \eta_s$ из ограниченности L_s^{-1} получаем $\|U_s - u_s\| \leqslant \|\eta_s\|$. Таким образом, для линейных разностных задач устойчивость есть равномерная (по всем сеткам) ограниченность обратного оператора.

Устойчивость неявной схемы. Покажем, что неявная схема дает разностную задачу *безусловно-устойчивую*, т.е. для ее устойчивости не требуется выполнения условия Куранта. Ограничимся доказательством следующей леммы.

Лемма 2. При любых шагах h, τ, нормы сеточных функций u^n и u^{n+1} связаны неравенством

$$\|u^{n+1}\| \leqslant \max\{\|u^n\| + \tau\|f\|, \|\psi\|\}.$$

Доказательство. Имеем альтернативу:

$$\|u^{n+1}\| = \{\text{либо } |u_0^{n+1}|, \text{ либо } |u_M^{n+1}|, \text{ либо } \max_{m=1,\ldots,M-1} |u_m^{n+1}|\}.$$

В двух первых случаях, как было установлено выше, $\|u^{n+1}\| \leqslant \|\psi\|$. Нужно исследовать третий случай. Для неявной схемы имеем

$$u_m^{n+1}\left[1 + \frac{\tau}{h^2}(\varkappa_{m-1/2}^n + \varkappa_{m+1/2}^n)\right] = \tau f_m^n + u_m^n + \frac{\tau}{h^2}\varkappa_{m-1/2}^n u_{m-1}^{n+1} + \frac{\tau}{h^2}\varkappa_{m+1/2}^n u_{m+1}^{n+1}.$$

Пусть m — внутренняя точка, для которой $\|u^{n+1}\| = |u_m^{n+1}|$. Тогда

$$\left[1 + \frac{\tau}{h^2}(\varkappa_{m-1/2}^n + \varkappa_{m+1/2}^n)\right]\|u^{n+1}\| \leqslant \|u^n\| + \tau\|f\| + \frac{\tau}{h^2}(\varkappa_{m-1/2}^n\|u^{n+1}\| + \varkappa_{m+1/2}^n\|u^{n+1}\|),$$

или, после сокращения,

$$\|u^{n+1}\| \leqslant \|u^n\| + \tau\|f\|.$$

На этом мы закончим исследование устойчивости неявной схемы. Еще раз подчеркнем, что она носит безусловный характер: схема всегда устойчива. В этом ее отличие и решающее преимущество перед явной схемой, счет по которой возможен лишь при

$\tau \leqslant 0.5h^2/\|\varkappa\|$. В других задачах, как мы увидим, тоже появляется это характерное условие: явные схемы устойчивы лишь при некоторых ограничениях на шаг по времени τ: он должен быть достаточно малым относительно шага по пространству h. Переход к неявным схемам, как правило, либо снимает условие устойчивости, либо существенно его ослабляет.

Возникает естественный вопрос: действительно ли условие Куранта существенно для явной схемы (ведь оно было необходимо для проведения достаточно простых оценок) или, может быть, оно связано с грубостью оценок, а не с существом дела? Оказывается, условие Куранта носит принципиальный характер, его нарушение делает результаты расчета совершенно бессмысленными.

Покажем это на простом примере (который, кстати, иллюстрирует возможный экспериментальный прием исследования устойчивости разностной схемы: он состоит в фактическом вычислении последствий «единичной погрешности в начальных данных»). Если в качестве начальных данных взять $u_m^0 = 0$, то решение будет нулевым (мы не учитываем здесь краевых условий, считая, что задача решается на бесконечном интервале: $-\infty \leqslant m \leqslant \infty$).

Таблица 9

−6	21	−50	90	−126	141	−126	90	−50	21	−6
1	−5	15	−30	45	−51	45	−30	15	−5	1
0	1	−4	10	−16	19	−16	10	−4	1	0
0	0	1	−3	6	−7	6	−3	1	0	0
0	0	0	1	−2	3	−2	1	0	0	0
0	0	0	0	1	−1	1	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0

Теперь возьмем начальные данные с изолированной погрешностью: $u_0^0 = 1$, остальные $u_m^0 = 0$, и станем решать задачу по явной схеме, нарушив условие Куранта. Для иллюстрации удобно взять $\tau = h^2$. Тогда $u_m^{n+1} = u_{m-1}^n - u_m^n + u_{m+1}^n$. В табл. 9 представлены результаты, полученные для $n = 1, 2, \ldots, 6$.

Из таблицы видно, что решение возрастает почти в три раза за шаг и примерно через 20 шагов достигает катастрофического значения (порядка 10^9). А ведь в расчетах делаются сотни шагов по времени! Обратите внимание на характерный (по m) профиль функции u_m^n. Он носит «пилообразный» характер: $u_m^n \approx (-3)^n(-1)^m v_m^n$, где v_m^n — «гладкая» сеточная функция. Это характерный признак вычислительной неустойчивости. Обычно вычислители просматривают

полученные результаты, строят графики сеточных функций. Часто уже внешний вид таких функций содержит «намек» на какое-то неблагополучие, на сомнительность результата.

На рис. 11 показаны два примерных графика u_m^n, $m = 1, 2, 3, \ldots$ Первый, естественно, воспринимается как сеточная проекция «хорошей» функции, второй — типичный пример «подозрительного» решения. Общее качественное соображение носит простой характер: в методе конечных разностей каждое «событие» должно быть разрешено несколькими точками. «Событием» мы называем колебание функции, переход с одного уровня на другой и т.п. Если такое «событие» происходит на одном счетном интервале — это явно подозрительно, настораживает, делает результаты сомнительными.

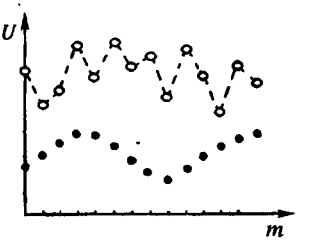


Рис. 11

Вычислители очень не любят «пилообразных» графиков. Однако не следует все абсолютизировать. Не следует думать, что если получены точки u_m^n, легко укладывающиеся на гладкую функцию, то имеется гарантия правильности расчета. «Пила» на решении — тоже не 100 %-ная гарантия ошибочности расчета, хотя ничего хорошего в этом нет.

Появление «пилы» на графике сеточной функции часто является признаком вычислительной неустойчивости разностной схемы. Но настоящая вычислительная неустойчивость сопровождается еще и очень быстрым нарастанием амплитуды «пилы», настолько быстрым, что за несколько шагов решение может вообще выйти за пределы машинной бесконечности.

Сходимость разностных схем. (Точнее, следует говорить о сходимости приближенного решения к точному при $\tau, h \to 0$.) Установив устойчивость схемы и оценив погрешность аппроксимации, воспользуемся теоремой Рябенького—Филиппова и получим оценку

$$\|U_s - u_s\| = O(\tau + h).$$

В явной схеме $\tau = O(h^2)$. Такое же соотношение во многих случаях приходится выдерживать по соображениям точности расчета и в неявных схемах (см. § 21). Было бы желательно иметь в оценке $O(\tau + h^2)$, тем более что почти во всех узлах сетки невязка есть $O(h^2)$; мешает только аппроксимация краевых условий с погрешностью $O(h)$. Улучшим ее, используя характерный прием.

Выпишем погрешность аппроксимации (2) более аккуратно, используя ряд Тейлора:

$$U_1^n = U(t_n, h) = U(t_n, 0) + hU_x + \frac{h^2}{2} U_{xx} + O(h^3).$$

Тогда

$$-\alpha_1 \frac{U_1^n - U_0^n}{h} + \beta_1 U_0^n - \psi_1^n = -\alpha_1 U_x + \beta U - \psi_1 - \frac{1}{2}\alpha_1 h U_{xx} + O(h^3).$$

Перенося главную часть погрешности $\frac{1}{2}\alpha_1 h U_{xx}$ из правой части в левую и заменяя U_{xx} разностной аппроксимацией, получаем аппроксимацию краевого условия второго порядка:

$$-\alpha_1 \frac{U_1^n - U_0^n}{h} + \beta_1 U_0^n + \frac{1}{2h}\alpha_1 (U_0^n - 2U_1^n + U_2^n) - \psi_1 = \\ = -\alpha_1 U_x + \beta_1 U - \psi_1 + O(h^2).$$

В реализации явной схемы никаких осложнений не возникает. В неявной схеме это приводит к нарушению трехдиагональной структуры уравнений на верхнем слое. Предоставим читателю внести необходимые дополнения в алгоритм решения уравнений на верхнем слое прогонкой.

§ 12. Спектральный признак устойчивости

Рассмотрим основной аналитический аппарат исследования устойчивости разностных схем, который имеет дело не с реальной вычислительной схемой, а с некоторой ее моделью. Он связан с более или менее обозримой и выполнимой аналитической работой, благодаря чему и получил самое широкое распространение. Хотя этот метод исследования не дает точного ответа на вопрос об устойчивости, он позволяет отбраковать подавляющее большинство заведомо неустойчивых схем, а схемы, признанные на основе спектрального признака устойчивыми, как правило, на самом деле являются таковыми.

Начнем с двух упрощений, которые приходится произвести, чтобы можно было применять аппарат спектральной устойчивости. Имеются в виду:

а) линейные, однородные с постоянными коэффициентами схемы;

б) задача Коши на всем пространстве, без краевых условий (их место занимают условия типа «ограниченности на бесконечности»).

Итак, если нас интересует разностная схема для общего уравнения теплопроводности

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left[\varkappa(t, x, u)\frac{\partial u}{\partial x}\right] + f(t, x, u),$$

то исследование проводится для уравнения

$$\frac{\partial u}{\partial t} = \varkappa \frac{\partial^2 u}{\partial x^2} + au, \qquad \varkappa, a = \text{const}.$$

Рассмотрим явную разностную схему (1.10):

$$\frac{u_m^{n+1}-u_m^n}{\tau}=\varkappa\frac{u_{m-1}^n-2u_m^n+u_{m+1}^n}{h^2}+au_m^n,$$

$$m=0,\ \pm 1,\ \ldots,\ \pm\infty,\qquad n=0,\ 1,\ \ldots,\ N=T/\tau.$$

Исследование основано на следующем общем факте: все линейные однородные разностные уравнения с постоянными коэффициентами, заданные на всем пространстве (с условием ограниченности на бесконечности), имеют универсальное полное семейство частных решений

$$u_m^n=\lambda^n e^{im\varphi},\quad 0\leqslant\varphi\leqslant 2\pi. \tag{1}$$

Здесь φ — параметр семейства, $\lambda(\tau, h,$ «схема»; $\varphi)$ — функция, зависящая от шагов τ, h, параметра φ и вида схемы.

Каждая схема характеризуется своей функцией $\lambda(\varphi)$ (остальные аргументы (τ, h, «схема») мы будем всегда иметь в виду, не выписывая их явно). Функция $\lambda(\varphi)$ называется *спектральной функцией* схемы. Совокупность значений, пробегаемых точкой $\lambda(\varphi)$ (в комплексной плоскости), когда φ пробегает $[0, 2\pi]$, называют *спектром разностной схемы.*

Введем формальные определения.

Разностную схему называют *спектрально-устойчивой*, если

$$|\lambda(\varphi)|\leqslant 1+C\tau,\quad\forall\ \varphi\in[0, 2\pi], \tag{2}$$

где C — не зависящая от τ постоянная. Другими словами, спектр устойчивой (по спектральному признаку) схемы должен лежать в $C\tau$-расширении единичного круга.

Разностную схему называют *спектрально-неустойчивой*, если существуют $q>1$ (q не зависит от τ) и $\varphi_0\in[0, 2\pi]$, такие, что

$$|\lambda(\varphi_0)|\geqslant q>1. \tag{3}$$

Это пока чисто формальные определения. Сейчас мы научимся вычислять спектр разностных схем, а затем выясним содержательный смысл введенных понятий. Он будет простым: спектрально-неустойчивые схемы не годятся для вычислений, расчеты по таким схемам сопровождаются катастрофическим нарастанием последствий погрешностей вычислений (т.е. погрешностей машинного представления чисел, округления и т.п.).

Примеры вычисления спектра. Рассмотрим примеры вычисления спектра для различных разностных схем с учетом вышеприведенных определений.

Явная схема. Вычисление $\lambda(\varphi)$ проводится просто: нужно решение $u_m^n = \lambda^n e^{im\varphi}$ подставить в разностные уравнения. Для явной схемы имеем

$$\frac{\lambda^{n+1}e^{im\varphi} - \lambda^n e^{im\varphi}}{\tau} = \frac{\lambda^n e^{i(m-1)\varphi} - 2\lambda^n e^{im\varphi} + \lambda^n e^{i(m+1)\varphi}}{h^2}.$$

Сокращая на $\lambda^n e^{im\varphi}$, получаем

$$\frac{\lambda - 1}{\tau} = \frac{e^{-i\varphi} - 2 + e^{i\varphi}}{h^2}.$$

Используем соотношение $e^{-i\varphi} - 2 + e^{i\varphi} = -4\sin^2(\varphi/2)$. В результате спектральная функция явной схемы принимает вид

$$\lambda(\varphi) = 1 - 4\frac{\tau}{h^2}\sin^2\frac{\varphi}{2}.$$

Легко видеть, что $\lambda(\varphi)$ вещественна и

$$\lambda(\pi) = 1 - 4\tau/h^2 \leqslant \lambda(\varphi) \leqslant 1 = \lambda(0).$$

Итак, спектр есть отрезок $[1 - 4\tau/h^2, 1]$. Условие устойчивости:

$$1 - 4\tau/h^2 \geqslant -1, \qquad \text{или} \quad \tau \leqslant h^2/2.$$

Это есть условие Куранта, которое нам уже знакомо. Таким образом, явная схема для уравнения теплопроводности устойчива при выполнении условия Куранта (условно-устойчива).

Неявная схема. Проведем те же вычисления:

$$\frac{\lambda^n e^{im\varphi}(\lambda - 1)}{\tau} = \lambda^n e^{im\varphi}\lambda\frac{e^{i\varphi} - 2 + e^{-i\varphi}}{h^2}.$$

После очевидных преобразований получаем

$$\lambda(\varphi) = \left(1 + 4\frac{\tau}{h^2}\sin^2\frac{\varphi}{2}\right)^{-1}.$$

Очевидно, $\lambda(\varphi) \in [0, 1]$, $\forall\ \tau, h$. Неявная схема безусловно-устойчива (по спектральному признаку). Этот факт (правда, с другим пока смыслом термина «устойчивость») нам уже известен.

Схема «крест» для волнового уравнения $u_{tt} = u_{xx}$. Схема имеет вид

$$\frac{u_m^{n+1} - 2u_m^n + u_m^{n-1}}{\tau^2} = \frac{u_{m-1}^n - 2u_m^n + u_{m+1}^n}{h^2}.$$

Подставляя $u_m^n = \lambda^n e^{im\varphi}$ и сокращая на $\lambda^{n-1}e^{im\varphi}$, получаем

$$\frac{\lambda^2-2\lambda+1}{\tau^2} = -4\frac{\lambda}{h^2}\sin^2\frac{\varphi}{2},$$

т.е. λ есть решение квадратного характеристического уравнения

$$\lambda^2 - 2\left(1 - 2\frac{\tau^2}{h^2}\sin^2\frac{\varphi}{2}\right)\lambda + 1 = 0.$$

Исследовать спектр можно, не решая уравнения. Заметим, что свободный член есть единица, т.е. $\lambda_1\lambda_2 = 1$. Здесь имеются две возможности:

а) если корни вещественны, то один корень меньше единицы, второй больше единицы, т.е. схема неустойчива;

б) если корни комплексно-сопряженные, то $|\lambda_1| = |\lambda_2| = 1$, т.е. схема устойчива.

Итак, схема устойчива, если корни комплексные, т.е. если отрицателен (при всех φ) дискриминант

$$\left(1 - 2\frac{\tau^2}{h^2}\sin^2\frac{\varphi}{2}\right)^2 - 1 = 1 - 4\frac{\tau^2}{h^2}\sin^2\frac{\varphi}{2} + 4\frac{\tau^4}{h^4}\sin^4\frac{\varphi}{2} - 1 =$$
$$= 4\frac{\tau^2}{h^2}\sin^2\frac{\varphi}{2}\left(\frac{\tau^2}{h^2}\sin^2\frac{\varphi}{2} - 1\right).$$

Очевидно, что при всех $\varphi \in [0, 2\pi]$ это выражение отрицательно только для $\tau/h \leqslant 1$. Это и есть условие Куранта для схемы «крест».

Шахматная схема. Рассмотрим систему уравнений, описывающих распространение звука

$$u_t + v_x = 0, \qquad v_t + u_x = 0.$$

Поясним некоторые новые объекты. Прежде всего удобно ввести так называемую шахматную сетку, т.е. определить сеточные функции u и v в разных точках.

Итак, введем «целые» точки, или u-точки: $t_n = n\tau$, $x_m = mh$. В этих точках определим u_m^n. Введем «полуцелые» точки, или v-точки: $t_{n+1/2} = (n+1/2)\tau$, $x_{m+1/2} = (m+1/2)h$. В этих точках определим $v_{m+1/2}^{n+1/2}$. На такой сетке удобно аппроксимировать систему:

$$\frac{u_m^{n+1}-u_m^n}{\tau} + \frac{v_{m+1/2}^{n+1/2}-v_{m-1/2}^{n+1/2}}{h} = 0, \qquad \frac{v_{m+1/2}^{n+1/2}-v_{m+1/2}^{n-1/2}}{\tau} + \frac{u_{m+1}^n-u_m^n}{h} = 0.$$

Обобщим конструкцию стандартного решения:

$$\begin{Bmatrix} u_m^n \\ v_{m+1/2}^{n+1/2} \end{Bmatrix} = \lambda^n(\varphi)\begin{Bmatrix} Ue^{im\varphi} \\ Ve^{i(m+1/2)\varphi} \end{Bmatrix},$$

где U, V — некоторые постоянные. Подставляя это решение в разностные уравнения, после сокращения на $\lambda^n e^{im\varphi}$ и $\lambda^{n-1}e^{i(m+1/2)\varphi}$ получаем

$$U\frac{\lambda-1}{\tau}+V\frac{e^{i\varphi/2}-e^{-i\varphi/2}}{h}=0, \qquad V\frac{\lambda-1}{\tau}+U\lambda\frac{e^{i\varphi/2}-e^{-i\varphi/2}}{h}=0.$$

Система (относительно U, V) имеет нетривиальное решение при

$$\det\begin{pmatrix}\frac{\lambda-1}{\tau} & 2\frac{i}{h}\sin\frac{\varphi}{2}\\ 2\lambda\frac{i}{h}\sin\frac{\varphi}{2} & \frac{\lambda-1}{\tau}\end{pmatrix}=0.$$

Это и есть уравнение, определяющее $\lambda(\varphi)$:

$$\left(\frac{\lambda-1}{\tau}\right)^2+4\frac{\lambda}{h^2}\sin^2\frac{\varphi}{2}=0,$$

или

$$\lambda^2-2\lambda\left(1-2\frac{\tau^2}{h^2}\sin^2\frac{\varphi}{2}\right)+1=0.$$

Такое уравнение мы уже исследовали в связи со схемой «крест» для волнового уравнения, и ответ нам известен: схема устойчива при условии Куранта $\tau\leqslant h$.

Схема «ромб» для уравнения теплопроводности. Этот пример интересен тем, что он связан с поиском явных безусловно-устойчивых схем. Схема имеет вид

$$\frac{u_m^{n+1}-u_m^{n-1}}{2\tau}=\frac{1}{h^2}\left(u_{m-1}^n-2\frac{u_m^{n+1}+u_m^{n-1}}{2}+u_{m+1}^n\right).$$

Эта схема трехслойная, расчет требует задания двух начальных слоев u^0 и u^1 (u^1 можно вычислить, например, по явной двухслойной схеме). В приведенном уравнении предполагаются известными значения u^{n-1} и u^n, u_m^{n+1} явно выражается через известные величины на двух предыдущих слоях.

Стандартное исследование устойчивости приводит к характеристическому уравнению

$$\frac{\lambda^2-1}{2\tau}=\frac{2\lambda\cos\varphi-(\lambda^2+1)}{h^2}.$$

Обозначая $r=h^2/2\tau$, представим уравнение в другой форме:

$$\lambda^2-2\frac{\cos\varphi}{1+r}\lambda+\frac{1-r}{1+r}=0.$$

Решение выписывается просто:

$$\lambda_{1,2}=\frac{\cos\varphi\pm\sqrt{\cos^2\varphi-1+r^2}}{1+r}.$$

Рассмотрим два случая.

а) $\cos^2\varphi - 1 + r^2 < 0$. Корни комплексно-сопряженные, их произведение $|\lambda_1\lambda_2| = |(1-r)/(1+r)| < 1$, т.е. $|\lambda_1| = |\lambda_2| < 1$.

б) $p^2 \equiv r^2 - (1-\cos^2\varphi) > 0$. Очевидно, что $p < r$ и корни $\lambda_{1,2} = (\cos\varphi \pm p)/(1+r)$. Несложный анализ, основанный на том, что $|\cos\varphi| \leqslant 1$, $|\pm p| < r$, показывает, что в этом случае корни $|\lambda_1| \leqslant 1$, $|\lambda_2| < 1$.

Итак, схема «ромб» безусловно-устойчива. К сожалению, она не годится для решения уравнения теплопроводности, так как не аппроксимирует его. Причина этого состоит в замене значения u_m^n на $0.5(u_m^{n+1} + u_m^{n-1})$. Погрешность такой замены есть $O(\tau^2)$ и была бы допустимой в других ситуациях, но это среднее используется при аппроксимации второй производной, т.е. в выражении, делящемся на h^2. В результате в погрешности аппроксимации появляется член $O(\tau^2/h^2)$, что делает такую схему допустимой лишь при очень малых τ, например при $\tau = O(h^2)$, т.е. мы не получаем серьезных преимуществ от безусловной устойчивости. Однако она представляет определенный интерес, особенно в задаче «на установление», когда решение уравнения теплопроводности (с не зависящими от времени правыми частями и краевыми условиями) в пределе при $t \to \infty$ переходит в решение уравнения Пуассона. В этом случае детали процесса выхода решения на предел игнорируются.

Схема «квадрат». Для уравнения переноса $u_t + u_x = f$ часто используется схема «квадрат» (в теории переноса излучения эта схема получила название «алмазная»):

$$\frac{1}{\tau}\left(\frac{u_{m+1}^{n+1}+u_m^{n+1}}{2} - \frac{u_{m+1}^{n}+u_m^{n}}{2}\right) + \frac{1}{h}\left(\frac{u_{m+1}^{n+1}+u_{m+1}^{n}}{2} - \frac{u_m^{n+1}+u_m^{n}}{2}\right) = f_{m+1/2}^{n+1/2}.$$

Опуская несложные выкладки, приведем выражение для спектральной функции:

$$\lambda(\varphi) = \left(1 + i\frac{\tau}{h}\operatorname{tg}\frac{\varphi}{2}\right) \Big/ \left(1 - i\frac{\tau}{h}\operatorname{tg}\frac{\varphi}{2}\right).$$

Очевидно, $|\lambda(\varphi)| = 1$. Таким образом, эта схема безусловно-устойчива. На первый взгляд она неявная, так как в каждое разностное уравнение входят две величины с верхнего слоя. Однако решение уравнений на верхнем слое в данном случае столь просто выписывается в «явном» виде, что подобные схемы относят к явным.

В самом деле для, уравнения переноса $u_t + u_x = f$ математически корректной является задача с начальными данными и одним краевым условием на левой границе. Пусть для простоты задано

значение u на левой границе: $u(t, 0) = \psi(t)$, т.е. при переходе со слоя n на слой $n+1$ известны величины u^n и значение u_0^{n+1}. В этом случае уравнения верхнего слоя разрешаются явно слева-направо (такие алгоритмы получили название «маршевых»). Из разностного уравнения легко выразить неизвестное u_{m+1}^{n+1} через известные уже u_m^n, u_{m+1}^n и u_m^{n+1}:

$$u_{m+1}^{n+1} = u_m^n + \frac{\tau - h}{\tau + h}(u_m^{n+1} - u_{m+1}^n) + \tau f_{m+1/2}^{n+1/2}.$$

Заметим, что «маршевый» алгоритм (последовательного вычисления u_1^{n+1}, u_2^{n+1}, ...) вычислительно-устойчив, так как модуль «коэффициента усиления» накопившейся погрешности при переходе u_m^{n+1} к u_{m+1}^{n+1} есть $|(\tau - h)/(\tau + h)| < 1$. Кстати, при попытке решать по этой схеме «неправильную» краевую задачу, когда заданы значения u на правой границе области (т.е. известны u_M^n), мы легко получим формулу «маршевого» алгоритма, действующего справа-налево, однако в этом случае коэффициент усиления погрешности есть $(h + \tau)/(h - \tau)$ и такой «марш» вычислительно-неустойчив.

Содержательный смысл спектральной устойчивости. Выясним, что, собственно, следует из спектральной устойчивости или неустойчивости разностной схемы. Покажем, что, если схема спектрально-неустойчива, она непригодна для решения задач, так как погрешности в начальных данных катастрофически нарастают и портят решение до такой степени, что оно становится полностью бессмысленным. Если схема спектрально-устойчива, этого не происходит.

Пусть проведен расчет по какой-то разностной схеме, начиная с начальных данных u_m^0, $m \in (-\infty, \infty)$, и этот расчет дает решение u_m^n, $n = 0, \ldots, N$. Расчет, начинающийся с начальных данных с малой погрешностью $\tilde{u}_m^0 = u_m^0 + \delta_m^0$, $\|\delta^0\| \leqslant \varepsilon$, дает возмущенное решение $\tilde{u}_m^n$. В силу линейности задачи $\tilde{u}_m^n = u_m^n + \delta_m^n$, где δ_m^n — расчет по той же схеме, начинающийся с δ_m^0. Нас интересует величина $|\tilde{u}_m^n - u_m^n|$. Если она мала, то все в порядке: погрешности в начальных данных приводят к малым последствиям.

Рассмотрим погрешности, малые в норме l_2:

$$\|\delta^0\| = \left[\sum_m (\delta_m^0)^2\right]^{1/2} \leqslant \varepsilon.$$

Используя теорию дискретного преобразования Фурье, разложим сеточную функцию в интеграл Фурье:

$$\delta_m^0 = \int_0^{2\pi} c(\varphi)\, e^{im\varphi}\, d\varphi.$$

Здесь $c(\varphi)$ — фурье-образ сеточной функции δ_m^0. Функция $c(\varphi)$ вычисляется по формуле $c(\varphi) = \nu \sum_m \delta_m^0\, e^{-im\varphi}$, где ν — нормирующий множитель, для дальнейшего несущественный ($\nu = O(1)$).

Важным является равенство Парсеваля

$$\|\delta^0\| = \left[\sum_m (\delta_m^0)^2\right]^{1/2} = \left[\int_0^{2\pi} |c(\varphi)|^2\, d\varphi\right]^{1/2}.$$

Предположим, что задача решается по разностной схеме со спектральной функцией $\lambda(\tau, h; \varphi)$. Поскольку функции $\lambda^n(\varphi)\, e^{im\varphi}$ удовлетворяют разностному уравнению, можно сразу же выписать решение:

$$\delta_m^n = \int_0^{2\pi} \lambda^n(\varphi)\, c(\varphi)\, e^{im\varphi}\, d\varphi.$$

Проанализируем эту основную формулу.

Пусть схема спектрально-устойчива, т.е. имеет место равномерная по τ оценка $|\lambda(\varphi)| \leqslant 1 + C\tau$. Тогда

$$\|\delta^n\| = \left[\int_0^{2\pi} |\lambda^n(\varphi)\, c(\varphi)|^2\, d\varphi\right]^{1/2}.$$

Далее,

$$|\lambda^n(\varphi)| \leqslant (1 + C\tau)^n \leqslant (1 + C\tau)^{T/\tau} \approx e^{CT}$$

(при $\tau \to 0$). И наконец,

$$\|\delta^n\| \leqslant e^{CT}\left[\int_0^{2\pi} |c(\varphi)|^2\, d\varphi\right]^{1/2} \leqslant e^{CT}\|\delta^0\|.$$

Таким образом, погрешность в начальных данных в процессе решения может увеличиться не более чем в e^{CT} раз; эта оценка остается справедливой, когда $\tau \to 0$, а число шагов $N \to \infty$, как T/τ. Итак, спектральная устойчивость схемы означает непрерывную зависимость решения разностной задачи по начальным данным с оценкой, равномерной по $\tau \to 0$.

Пусть схема спектрально-неустойчива, т.е. существует $q > 1$, не зависящее от τ, и $|\lambda(\varphi_0)| \geqslant q$ при некотором $\varphi_0 \in [0, 2\pi]$. В силу непрерывности $\lambda(\varphi)$ существует малая окрестность Δ, в которой $|\lambda(\varphi)| \geqslant q' > 1$, $\varphi \in \Delta$. Рассмотрим возмущение δ_m^0, порожденное фурье-образом $c(\varphi) = \{0$ при $\varphi \notin \Delta$; $\varepsilon/\text{mes}\,\Delta$ при $\varphi \in \Delta\}$:

$$\delta_m^0 = \int_0^{2\pi} c(\varphi)\, e^{im\varphi}\, d\varphi.$$

Очевидно, $\|\delta^0\| = \varepsilon$.

Оценим последствия такого возмущения. Как уже отмечалось, они имеют вид

$$\delta_m^n = \int_\Delta \lambda^n(\varphi)\, c(\varphi)\, e^{im\varphi}\, d\varphi.$$

Вычислим норму

$$\|\delta^n\|^2 = \int_\Delta |\lambda^n(\varphi)|^2\, |c(\varphi)|^2\, d\varphi \geqslant (q')^{2n}\, \|\delta^0\|^2.$$

При достаточно малом τ число шагов N становится сколь угодно большим и множитель $(q')^{T/\tau} \to \infty$ при $\tau \to \emptyset$.

Итак, при расчете по спектрально-неустойчивой схеме сколь угодно малая погрешность в начальных данных приводит (при достаточно малом τ) к сколь угодно большим погрешностям в решении. Мы рассмотрели последствия специально сконструированного возмущения начальных данных. Более или менее очевидно, что почти любое начальное возмущение имеет фурье-образ $c(\varphi) \neq 0$ в Δ и такое возмущение тоже будет катастрофически нарастать: чем меньше шаг τ, тем сильнее будут сказываться последствия неустойчивости.

Перейдем к обсуждению спектрального признака устойчивости и практики его применения в реальных ситуациях. Рассмотрим два вопроса.

1. Мы уже знаем, что устойчивость метода приближенного решения — это, грубо говоря, непрерывная зависимость решения от исходной информации, которой являются функции, входящие в начальные данные, краевые условия и в правую часть уравнения. Спектральный признак оценивает только устойчивость по начальным данным. В более или менее общем случае из такой устойчивости следует устойчивость по правой части (дело в том, что начальные данные можно трактовать, как правую часть, имеющую характер δ-функции). Устойчивость по краевым условиям — свойство совсем иного характера, она не связана однозначно с устойчивостью по начальным данным. Краевые условия требуют отдельного, самостоятельного исследования. Теоретические основы такого анализа

были разработаны И. М. Гельфандом, К. И. Бабенко. Технически это более сложные исследования.

2. Почему при исследовании устойчивости мы ограничились функциями $\lambda^n e^{im\varphi}$ для $\varphi \in [0, 2\pi]$? Ведь эта функция будет решением линейного однородного разностного уравнения с постоянными коэффициентами при любом φ, в том числе и комплексном, и спектр будет совсем другим. Мы ограничились вещественными φ потому, что при комплексных φ такая функция для $n = 0$, т.е. $e^{im\varphi}$, уже содержит бесконечные (при $m \to \pm\infty$) значения. И если такие начальные данные приводят к очень большим решениям $\lambda^n e^{im\varphi}$, тут нет ничего удивительного и этот факт не компрометирует схему.

Другое дело, когда при вещественном φ из ограниченных всюду начальных данных получается бесконечно большое решение — это уже дефект разностной схемы. В функции $e^{im\varphi}$ параметр φ определен с точностью до 2π, поэтому ограничимся только интервалом $[0, 2\pi]$. Кстати, упоминавшийся выше анализ устойчивости по краевым условиям приводит к изучению, например, полуограниченной части оси x ($m > 0$). В этом случае ограниченные начальные данные дают все φ, для которых $\operatorname{Re} i\varphi \leqslant 0$. Но среди таких φ нужно отобрать те, для которых функция $e^{im\varphi}$ удовлетворяет рассматриваемым разностным краевым условиям (однородным).

Устойчивость нелинейных разностных схем. Спектральный признак устойчивости используется для анализа самых сложных задач. При этом руководствуются правилом, получившим несколько высокопарное название «принцип замороженных коэффициентов». Имеется в виду следующий рецепт. Все входящие в уравнение коэффициенты, зависящие от t, x и самой искомой функции, полагаются постоянными, и разностная схема становится линейной с постоянными коэффициентами. Правые части игнорируются, краевые условия переносятся в бесконечность (в форме требования ограниченности решения) и получается схема, допускающая исследование спектральным методом.

Найдем условие устойчивости, в которое входят «замороженные» коэффициенты. Используя «принцип замороженных коэффициентов» и явную схему, например, для уравнения теплопроводности

$$c(t, x, u)\, u_t = [\varkappa(t, x, u)\, u_x]_x + f(t, x, u)$$

получаем как объект исследования разностную схему

$$c\,\frac{u_m^{n+1} - u_m^n}{\tau} = \frac{\varkappa}{h^2}\,(u_{m-1}^n - 2u_m^n + u_{m+1}^n)$$

и условие устойчивости Куранта $\varkappa\tau \leqslant 0.5ch^2$. Возвращаясь к реальной схеме, нужно решить вопрос: какие же значения c и $\varkappa$ следует брать

при выборе τ? Ответ прост: шаг τ должен быть таким, чтобы условие Куранта выполнялось при всех значениях $\varkappa$ и c, встречающихся в данном расчете.

Счет с автоматическим выбором шага. Как выбрать τ, когда коэффициенты c и $\varkappa$ зависят от u, а эта функция с самого начала нам неизвестна? И здесь рецепт прост и очень полезен. Рассмотрим ситуацию стандартного шага: $\{u_m^n\}_{m=0}^M$ известны, надо вычислить u^{n+1}. Расчет начинается с того, что находится

$$b = \max_m \frac{\varkappa(t_n, x_m, u_m^n)}{c(t_n, x_m, u_m^n)},$$

затем вычисляется шаг $\tau_{n+1/2}$, определяющий переход от t_n к $t_{n+1} = t_n + \tau_{n+1/2}$:

$$\tau_{n+1/2} = h^2/2b;$$

далее все делается стандартно.

В схемах с «нелинейностью с верхнего слоя» поступают так же, так как u^{n+1} мало отличается от u^n. В некоторых задачах может оказаться, что одно узкое место определяет слишком малый шаг τ, хотя в остальной части условие Куранта допускает гораздо больший. Это неприятно, и понятен интерес к безусловно-устойчивым схемам, в которых шаг τ может выбираться без учета требования вычислительной устойчивости. К сожалению, такими являются лишь неявные схемы.

Практика использования спектрального признака. Практика показала, что в большинстве случаев ситуация такая:

а) если схема спектрально-неустойчива, она для расчетов заведомо непригодна; нелинейность, переменность коэффициентов и прочие факторы, которые не учитывались при спектральном анализе, только усугубляют неустойчивость;

б) если схема устойчива по спектральному признаку, то это в реальной схеме, конечно, не гарантия, но очень серьезный довод в пользу ее устойчивости; наиболее серьезные коррективы вносят краевые условия.

В целом исследование спектрального признака позволяет отбрасывать подавляющее большинство неустойчивых схем, остальные исследуются, в частности, и экспериментально. Наиболее типичной причиной фактической неустойчивости схемы, устойчивой по спектральному признаку, является неустойчивость разностной реализации краевых условий. Внешне она проявляется в том, что численное решение оказывается испорченным большими пилообразными возмущениями (в первую очередь около соответствующей границы об-

ласти). Особенно хорошо это видно на начальной стадии расчета, при больших n это возмущение распространяется на всю область.

Построены примеры разностных схем, устойчивых при исследовании по «принципу замороженных коэффициентов», но неустойчивых фактически. Это противоречие связано не с краевыми условиями, а с переменностью коэффициентов уравнений по t. Однако существенным для таких примеров является сильное изменение коэффициентов за один шаг по времени. Такая ситуация не является характерной для схем, используемых для решения дифференциальных уравнений: в них шаг τ должен быть настолько малым, чтобы за один шаг по n ситуация (т.е. коэффициенты уравнения, решение и т.п.) менялась незначительно. Поэтому такие примеры не опровергают указанной выше практической точки зрения на спектральную устойчивость.

Устойчивость и структура пространства сеточных функций. Рассмотрим сетку по пространству $\{x_m\}_{m=0}^{M}$ $(x_m = mh)$ и пространство сеточных функций $\{u_m\}_{m=0}^{M}$. В этом пространстве функции $e^{im\varphi}$ образуют базис, причем вещественные φ принимают дискретные значения $\varphi_k = k\pi/M$ $(k = 0, 1, \ldots, M-1)$. Совокупность таких сеточных функций (назовем их $v^{(k)} = \{v_m^{(k)}\}_{m=0}^{M}$) образует в пространстве всех сеточных функций полный линейно-независимый базис. Остальные φ можно не рассматривать, так как они не вносят в пространство новых функций.

Среди функций базиса можно (достаточно условно) выделить две качественно разных части:

а) гладкие сеточные функции, соответствующие малым номерам $k = 0, 1, 2, \ldots$;

б) негладкие сеточные функции, соответствующие большим номерам $k = M-1, M-2, \ldots$

Основанием для такого разделения служит следующий фундаментальный факт: действие разностного оператора, например, $(u_{m+1} - u_m)/h$ дает результат, близкий к результату действия аппроксимируемого им дифференциального оператора d/dx, если функция $\{u_m\}$ — гладкая сеточная функция (т.е. если в ее разложении по базису $\{v^{(k)}\}$ определяющую роль играют первые члены с $k = 0, 1, 2, \ldots$). Если же функция $\{u_m\}$ — негладкая (например, совпадает с одной из базисных функций $v^{(k)}$, $k = M-1, M-2, \ldots$), результаты действия этих операторов не имеют между собой ничего общего.

Условная граница между гладкими и негладкими функциями базиса зависит, очевидно, от требований к точности. Численное решение какой-то задачи, если оно претендует на точность аппро-

ксимации решения исходной задачи (сформулированной, например, в терминах дифференциальных уравнений) должно быть гладким. С этой точки зрения бо́льшая часть пространства сеточных функций является в некотором смысле лишней. Высокая размерность пространства сеточных функций ($M = 10^2 \div 10^3$) — своеобразная плата за обеспечение точности разностных аппроксимаций на существенно более бедном подпространстве гладких сеточных функций; только оно и представляет интерес с точки зрения вычислителя.

При решении разностного уравнения мы сталкиваемся со следующей ситуацией. Если решение является гладкой функцией, оно аппроксимирует решение того дифференциального уравнения, которое аппроксимирует разностное. Если решение является негладкой функцией, оно не имеет никакого отношения к решению дифференциальной задачи. Это общий факт как для устойчивых, так и для неустойчивых схем. Разница между ними в том, что такие бессмысленные решения в неустойчивых схемах растут с катастрофической скоростью, а в устойчивых остаются ограниченными.

Конечно, в реальном расчете присутствуют обе компоненты, но если в начальных данных, правых частях разностных уравнений (к ним в процессе решения «добавляются» и погрешности машинной арифметики) негладкие компоненты малы, то их наличие не оказывает существенного влияния на решение устойчивой схемы. Решение же, полученное по неустойчивой схеме, через несколько шагов по n оказывается полностью бессмысленным. Приведенные качественные рассуждения объясняют, в частности, тот важный для практики факт, что внешним проявлением неустойчивости является появление в численном решении высокочастотной (негладкой) компоненты с быстро нарастающей с ростом n амплитудой.

Другое практическое следствие — приближенный, ориентировочный анализ спектра. Обычно при исследовании спектра схемы легко выписывается характеристическое уравнение для $\lambda(\varphi)$, это чисто техническая выкладка. Но если порядок такого уравнения высок (выше второго), оно не имеет удобного явного решения: «нарисовать» и проанализировать весь спектр $\lambda(\varphi)$ оказывается не так-то просто. Иногда начинают с частных вопросов, например с исследования функции $\lambda(\tau, h; \varphi)$ при «критических» значениях $\varphi = \pi, \pi/2, \ldots$, при которых особенно сильно проявляется неустойчивость (если она есть), а характеристическое уравнение часто упрощается. Если такой анализ обнаружил спектральную неустойчивость, остальные значения φ можно не рассматривать. В особо сложных случаях используют численное решение характеристического уравнения при разных характерных значениях τ, h, φ и других параметров (например, коэффициентов уравнения).

Отметим, наконец, и такой практический рецепт: можно проводить расчет по спектрально-неустойчивой схеме, но при этом пери-

одически (после каждой серии из небольшого числа шагов по n) проводить сглаживание полученного численного решения u^n, т.е. «отфильтровывать» паразитическую, негладкую компоненту. Этот прием организации расчетов используется на практике, хотя большинство вычислителей предпочитают все-таки производить фильтрацию неявно, за счет спектральной устойчивости используемой разностной схемы.

Устойчивость — асимптотическое свойство. Каждый конкретный расчет производится обычно на конкретной сетке, т.е. при фиксированных значениях шагов τ и h. Процесс $\tau \to 0$, $h \to 0$ подразумевается, но фактически, конечно, не осуществляется. Иногда (для контроля) расчеты проводятся на нескольких разных сетках (допустим, для h^0, $h^0/2$, $h^0/4$). Но на конкретной, фиксированной сетке формальные критерии спектральной устойчивости и неустойчивости схемы — неравенства $\{|\lambda(\varphi)| \leqslant 1 + C\tau, \ \forall \varphi\}$ и $\{\exists \varphi_0: |\lambda(\varphi_0)| \geqslant q > 1\}$ могут оказаться совместными при подходящем выборе C. Однако при $\tau \to 0$ и при условии, что C и q от τ не зависят, эти неравенства оказываются несовместными и происходит однозначная характеристика схемы как устойчивой или неустойчивой.

Таким образом, спектральная устойчивость — это асимптотическое свойство «схемы», т.е. семейства систем уравнений, построенных по определенному общему закону. Бо́льшая часть современных прикладных расчетов проводится при столь малых шагах сетки, при которых уже вступают в действие асимптотические (при $\tau \to 0$, $h \to 0$) свойства схемы, т.е. ее спектральная устойчивость.

Общая теория устойчивости разностных схем. Спектральный признак устойчивости имеет дело с сильно упрощенной моделью вычислительного алгоритма. Естественно, возникает потребность в более полной и адекватной теории. Ограничимся классом линейных разностных задач эволюционного типа, т.е. задач со временем, в которых счет реализуется по слоям. Такие схемы можно записать в виде

$$A_s^n u_s^n + B_s^n u_s^{n-1} = 0, \qquad n = 1, 2, \ldots, N_s. \tag{4}$$

Здесь s — набор малых параметров (например, τ, h); u_s^n — n-й слой (вектор размерности M_s); A_s^n, B_s^n — матрицы $M_s \to M_s$; A_s^n предполагается обратимой. Ради простоты мы опускаем в (4) «правую часть».

Таким образом, будем учитывать устойчивость по начальным данным и краевым условиям. Очевидно, схему (4) можно переписать в явной форме: $u_s^n = R_s^n u_s^{n-1}$, где $R_s^n = -(A_s^n)^{-1} B_s^n$ — оператор

перехода с $(n-1)$-го на n-й слой. В схему (4) укладываются аппроксимации линейных краевых задач для уравнений с частными производными с переменными (по t и x) коэффициентами при однородных краевых условиях. Вопрос об устойчивости такой схемы сводится к оценке

$$\left\| \prod_{n=1}^{N_s} R_s^n \right\| \leqslant C.$$

Если такая оценка получена, постоянная C не зависит от s и не является неприемлемо большой с практической точки зрения, то схему естественно считать устойчивой.

Обычно исследуют более простые схемы, в которых матрицы A, B не зависят от n (коэффициенты аппроксимируемого уравнения не зависят от t). Тогда оператор перехода есть R_s, и оценке подлежит $\|R_s^{N_s}\|$. Оценка $\|R_s\| \leqslant 1 + C\tau$ достаточна для вывода об устойчивости, однако общие критерии устойчивости предпочитают формулировать в терминах матриц A, B, так как именно они получаются в явном виде при конструировании разностной схемы.

Существуют и другие «канонические» формы записи разностных схем. В частности, в трудах А. А. Самарского и его учеников, активно развивавших общую теорию устойчивости разностных схем, принята следующая форма записи двухслойных схем:

$$A_s \frac{u_s^n - u_s^{n-1}}{\tau} = B_s u_s^{n-1}. \tag{5}$$

Легко перейти от (4) к (5) и наоборот. В их теории введены и исследованы трехслойные схемы в канонической форме и т.п. На этом пути получены необходимые и достаточные условия устойчивости в форме некоторых матричных неравенств. К сожалению, проверка таких неравенств возможна лишь в очень простых случаях, аналогичных схеме, аппроксимирующей уравнение теплопроводности, для которой было проведено полное исследование устойчивости, например в теореме 11.1.

Устойчивость краевых условий. Опишем в общих чертах алгоритм исследования устойчивости краевых условий, предложенный К. И. Бабенко и И. М. Гельфандом. Он относится к той же упрощенной схеме, которая была использована для исследования спектральной устойчивости. Однако учитывается то, что схема имеет дело с сеточной функцией, определенной при $m = 0, 1, \ldots, M$, и стандартные уравнения во внутренних узлах сетки дополнены краевыми условиями. Практический рецепт таков: нужно исследовать спектральную устойчивость трех задач, вычислить три спектра. Если все

три задачи устойчивы, схема оказывается устойчивой (по начальным данным и краевым условиям).

Первая задача — это стандартное исследование спектральной устойчивости. Перейдем ко второй задаче — к анализу разностной схемы на правой полупрямой, т.е. при $m = 0, 1, 2, \ldots$ Устойчивость исследуется с помощью той же конструкции общего решения $\lambda^n e^{im\varphi}$, но теперь, как было указано, кроме $\varphi \in [0, 2\pi]$, необходимо учесть φ, которые совместимы с левыми краевыми условиями и для которых функции $e^{im\varphi}$ не возрастают вправо. Точнее, следует учесть, что m ограничено величиной $O(1/h)$, поэтому допустимы значения $|e^{i\varphi}| \leqslant 1 + Ch$, где C, естественно, не зависит от h, т.е. $\mathrm{Re}(i\varphi) \leqslant Ch$.

Поясним сказанное простым примером. Рассмотрим явную схему для уравнения теплопроводности с краевым условием

$$\frac{u_1^n - u_0^n}{h} - \beta u_0^n = 0, \quad \text{или} \quad (1 + h\beta)u_0^n - u_1^n = 0.$$

Подставляя в него $\lambda^n e^{im\varphi}$, получаем уравнение для дополнительных значений φ:

$$e^{i\varphi} = 1 + \beta h, \qquad \varphi = \frac{1}{i}\ln(1 + \beta h) \approx -i\beta h, \quad (1 + \beta h)^{1/h} = O(1).$$

Вычислим для φ точку спектра по стандартной формуле:

$$\lambda(\varphi) = 1 + \frac{\tau}{h^2}(e^{i\varphi} - 2 + e^{-i\varphi}) = 1 + \frac{\tau}{h^2}\left(1 + \beta h - 2 + \frac{1}{1 + \beta h}\right) \approx 1 + \frac{\tau\beta^2}{2}.$$

Третья задача аналогична второй, только рассматривается разностная задача на левой полупрямой $m = 0, -1, -2, \ldots$ Итак, схема оказалась устойчивой по краевым условиям, согласно критерию Бабенко—Гельфанда. Связь этой формальной устойчивости с содержательной, т.е. с оценкой роста вычислительных погрешностей, составляет существо этой весьма нетривиальной теории, развитие которой (С. К. Годунов, В. С. Рябенький) привело к выделению тонких и нестандартных в классической спектральной теории понятий (спектр семейства разностных операторов и др.).

Заметим, что вычислительная устойчивость схемы имеет место как при $\beta > 0$, когда исходная дифференциальная задача действительно устойчива, так и при $\beta < 0$, когда она неустойчива. В последнем случае, если β меньше некоторого $\beta_0 < 0$, решение дифференциальной задачи растет, должно, соответственно, расти и решение разностной задачи, но этот рост не имеет катастрофического характера, его темп от τ практически не зависит.

Таким образом, следует отличать неустойчивость решения разностного уравнения, являющуюся аппроксимацией неустойчивости решения дифференциальной задачи, от вычислительной неустойчиво-

сти разностной схемы, которая является неприемлемым недостатком данной разностной схемы и к дифференциальной задаче отношения не имеет. Рекомендуем читателю провести численное решение задачи с разными β, а также проверить, что краевое условие $u_0 - 2u_1 = 0$ вычислительно-неустойчиво. Однако это лишь методический пример, так как он соответствует аппроксимации физического краевого условия с ненормальным значением $\beta = -2/h$.

Устойчивость и погрешности расчетов. Итак, мы выделили два основных свойства разностных схем: аппроксимацию и устойчивость, наличие которых по теореме Рябенького—Филиппова обеспечивает точность расчета. Однако не секрет, что часто расчеты дают неверные результаты. Более того, практически каждый достаточно сложный расчет не имеет никаких гарантированных (в математическом смысле) оценок точности. Даже в тех относительно простых ситуациях, когда имеются оценки точности, ими лучше не пользоваться: они настолько завышены, что могут привести к незаслуженной дискредитации полученных результатов.

В сущности, любая теорема о сходимости содержит оценку погрешности вида, например, $O(\tau^q + h^p)$, и при желании ее можно превратить в реальную оценку типа $C_1\tau^q + C_2h^p$, однако постоянные C_1, C_2 столь велики, что авторы подобных теорем о сходимости предусмотрительно не вычисляют их. (Во всяком случае, автор ни разу не видел, чтобы подобные оценки доводились до числа в том или ином расчете. Замечательным исключением являются работы К. И. Бабенко и его последователей по так называемым «доказательным вычислениям». Но это тема отдельного обсуждения.)

Выше было указано, что установление факта аппроксимации — стандартная элементарная выкладка, ошибиться в ней трудно. Использование для расчетов схемы, не аппроксимирующей исходную задачу, маловероятно. Напротив, установление устойчивости очень сложно, и, строго говоря, в большинстве случаев прикладные расчеты проводятся по схеме, теоретическая устойчивость которой не установлена. Можно ли из этого сделать вывод, что причиной ошибочных численных результатов является фактическая неустойчивость алгоритма? Нет. Дело обстоит как раз наоборот: неустойчивость схемы практически никогда не приводит к ошибкам, так как ее последствия носят столь катастрофический характер, что не заметить их невозможно. Часто их «замечает» ЭВМ, сигнализируя об этом «авостом» из-за выхода чисел в область машинной бесконечности. Нелепость таких результатов столь очевидна, что они не рассматриваются как содержательно ценные.

Реальным источником погрешностей, иногда полностью обесценивающих расчет, является именно погрешность аппроксимации.

Легко устанавливается наличие «формальной аппроксимации», т.е. оценка погрешности аппроксимации величиной типа $O(\tau^q + h^p)$ в предположении существования у искомого решения такого числа ограниченных производных, которое понадобилось для этой элементарной оценки. В теорему же Рябенького—Филиппова входит «фактическая погрешность аппроксимации», о точном значении которой a priori мало что можно сказать, так как для этого часто не хватает данных о точной характеристике гладкости искомого решения. Поэтому заранее, на основе теоретических оценок, трудно сказать, достаточно ли мал используемый в расчетах шаг сетки.

Доверие к результатам расчетов обычно основано на других неформальных соображениях. Такими средствами контроля являются, например: сопоставление результатов на разных сетках или результатов, полученных разными методами; сравнение с известными, иногда точными решениями, качественно близкими к найденным численно; сопоставление с данными экспериментов или с результатами расчетов, прошедших тщательный контроль и считающихся «эталонными». Проблема контроля численных результатов сложна, большую роль имеют опыт и неформальные знания в той области естествознания, к которой относится расчет.

Линеаризация схемы и исследование устойчивости. Исследование спектральной устойчивости схемы предполагает переход к некоторой ее модели — к линейному однородному разностному уравнению с постоянными коэффициентами. Построение такой модели требует некоторой аккуратности, иначе можно получить модель другой схемы, а не той, которая нас интересует.

Наиболее апробированный путь построения модели — это линеаризация разностной схемы. Речь идет о достаточно простой формальной операции. Пусть $L_s(u_s) = F_s$ — разностная схема на сетке «s». Рассмотрим задачу для малого возмущения F_s. Другими словами, рассмотрим решение возмущенной задачи, мало отличающейся от исходной. Это возмущение вызвано, например, погрешностями вычисления F_s, т.е. заменой его на $F_s + \delta F_s$. В δF_s можно включить и последствия погрешностей машинной арифметики. Такое возмущенное решение определяется уравнением для $\tilde{u}_s = u_s + \delta u_s$:

$$L_s(u_s + \delta u_s) = F_s + \delta F_s.$$

Линеаризуя его (т.е. разлагая входящие в него выражения в ряд Тейлора с точностью до первого члена), получаем

$$L_s(u_s) + R_s\,\delta u_s = F_s + \delta F_s,$$

где R_s — вычисленная на решении u_s производная от L_s.

Итак, для δu_s имеем линейную разностную схему

$$R_s\,\delta u_s = \delta F_s.$$

Производя в ней «замораживание» коэффициентов, перенося краевые условия «в бесконечность» и игнорируя в δF_s все, кроме возмущений начальных данных, получаем схему, поддающуюся спектральному анализу. Поясним это на простом примере.

Пусть решается уравнение теплопроводности с нелинейным источником

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left[\varkappa(u)\frac{\partial u}{\partial x}\right] + Q(u).$$

Используя явную схему с источником «на верхнем слое»:

$$\frac{u_m^{n+1}-u_m^n}{\tau} = \frac{1}{h}\left[\varkappa_{m+1/2}^n \frac{u_{m+1}^n-u_m^n}{h} - \varkappa_{m-1/2}^n \frac{u_m^n-u_{m-1}^n}{h}\right] + Q(u_m^{n+1}),$$

$$\varkappa_{m+1/2}^n = \varkappa\left(\frac{u_m^n+u_{m+1}^n}{2}\right),$$

получаем для δu схему

$$\frac{\delta u_m^{n+1}-\delta u_m^n}{\tau} = \frac{1}{h}\left[\varkappa_{m+1/2}\frac{\delta u_{m+1}^n-\delta u_m^n}{h} - \varkappa_{m-1/2}\frac{\delta u_m-\delta u_{m-1}}{h}\right] +$$
$$+ Q_u(u_m^n)\,\delta u_m^{n+1} + \varkappa'_{m+1/2}\frac{\delta u_{m+1}^n+\delta u_m^n}{2} - \varkappa'_{m-1/2}\frac{\delta u_m^n+\delta u_{m-1}^n}{2}.$$

Полагая $\varkappa_{m+1/2}$, $\varkappa'_{m+1/2}$, $Q_u(u_m^n)$ равными постоянным a, b, c соответственно («замораживая» коэффициенты), приходим к следующей схеме для δu:

$$\frac{\delta u_m^{n+1}-\delta u_m^n}{\tau} = a\frac{\delta u_{m-1}^n - 2\delta u_m^n + \delta u_{m+1}^n}{h^2} + b\frac{\delta u_{m+1}^n - \delta u_{m-1}^n}{2h} + c\,\delta u_m^{n+1}.$$

Применяя технику вычисления спектра схемы, получаем для $\lambda(\varphi)$ выражение

$$\lambda(\varphi) = \frac{1}{1-c\tau}\left(1 - 4\frac{\tau}{h^2}a\sin^2\frac{\varphi}{2} + i\frac{\tau}{h}b\sin\varphi\right).$$

Несложный анализ, который мы предоставим провести читателю, показывает, что мы, кажется, зря старались с аккуратной линеаризацией: появляющиеся дополнительные члены $(i\tau/h)b\sin\varphi$ и $c\tau$ определяют малые (порядка $O(\tau)$) поправки к величине $1-(4\tau/h^2)\sin^2(\varphi/2)$, которая получается в более простой модели. А такие поправки на выводы об устойчивости не влияют. Это наблюдается не только в рассмотренном простом примере, но и в общем случае: спектральная устойчивость как асимптотическое свойство определяется аппроксимацией «главных» дифференциальных членов решаемого уравнения, младшие члены вносят в характеристическое уравнение лишь малые (порядка $O(\tau)$, $O(h)$) возмущения.

Все это так, если бы не следующее обстоятельство. Иногда приходится решать задачи, в которых в отдельных узких частях области определения решения коэффициенты при младших дифференциальных членах становятся очень большими (например, порядка $O(1/h)$). В этом случае младшие члены в характеристическом уравнении уже оказывают на $\lambda(\varphi)$ такое же влияние, как и главные, и их надо учитывать. Пример такой ситуации — расчет ударной волны методом искусственной вязкости, который в дальнейшем будет описан подробно. Здесь укажем только, что речь идет о расчете разрыва в решении, который в вычислениях «размазывается», т.е. заменяется узкой зоной (порядка $4h$) непрерывного решения с большим градиентом $O(1/h)$, причем при $\tau \to 0$, $h \to 0$ длина зоны размазывания тоже стремится к нулю, а градиент решения всегда имеет величину порядка $O(1/h)$. Наличие таких зон может привести к своеобразной неустойчивости за счет младших членов.

§ 13. Метод переменных направлений

Рассмотрим простейшую двумерную задачу для уравнения теплопроводности. В области $0 \leqslant x \leqslant X$, $0 \leqslant y \leqslant Y$, $0 \leqslant t \leqslant T$ ищется функция $u(t, x, y)$, удовлетворяющая уравнению теплопроводности

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + f(t, x, y)$$

с начальными данными $u(0, x, y) = u_0(x, y)$ и краевыми условиями на боковых стенках области

$$u(t, 0, y) = \varphi_1(t, y), \qquad u(t, X, y) = \varphi_2(t, y),$$

$$u(t, x, 0) = \varphi_3(t, x), \qquad u(t, x, Y) = \varphi_4(t, x).$$

(Можно рассматривать и другие условия, свои на разных частях границы.)

Метод сеток строится, как обычно, из стандартных элементов.

1. Сетка — множество точек (n, k, m) с геометрическими координатами t_n, x_k, y_m. Ради простоты рассмотрим равномерную сетку: $t_n = n\tau$, $x_k = kh$, $y_m = mh$ (можно брать разные шаги h_x, h_y).

2. Приближенное решение ищется в виде сеточной функции

$$\{u^n_{k,m}\}, \qquad n = 0, 1, \ldots, N, \quad k = 0, 1, \ldots, K, \quad m = 0, 1, \ldots, M.$$

3. Разностное уравнение строится так же, как в § 11.

Явная схема:

$$\frac{u^{n+1}_{k,m} - u^n_{k,m}}{\tau} = \left(\frac{\partial^2 u}{\partial x^2}\right)^n_{k,m} + \left(\frac{\partial^2 u}{\partial y^2}\right)^n_{k,m} + f^n_{k,m}.$$

Здесь, ради краткости, мы используем компактные обозначения типа

$$\left(\frac{\partial^2 u}{\partial x^2}\right)^n_{k,m} \equiv \frac{u^n_{k-1,m} - 2u^n_{k,m} + u^n_{k+1,m}}{h^2}.$$

Несложное исследование спектральной устойчивости с универсальной конструкцией

$$u^n_{k,m} = \lambda^n e^{ik\varphi + im\psi}, \qquad \varphi, \psi \in [0, 2\pi],$$

приводит к спектральной функции

$$\lambda(\varphi, \psi; \tau, h_x, h_y, \text{«схема»}) = 1 - 4\frac{\tau}{h_x^2}\sin^2\frac{\varphi}{2} - 4\frac{\tau}{h_y^2}\sin^2\frac{\psi}{2},$$

откуда получаем условие Куранта $\tau(1/h_x^2 + 1/h_y^2) \leqslant 1/2$.

Неявная схема:

$$\frac{u^{n+1}_{k,m} - u^n_{k,m}}{\tau} = \left(\frac{\partial^2 u}{\partial x^2}\right)^{n+1}_{k,m} + \left(\frac{\partial^2 u}{\partial y^2}\right)^{n+1}_{k,m} + f^n_{k,m},$$

$$\lambda(\varphi, \psi) = \left(1 + 4\frac{\tau}{h_x^2}\sin^2\frac{\varphi}{2} + 4\frac{\tau}{h_y^2}\sin^2\frac{\psi}{2}\right)^{-1}.$$

$$|\lambda(\varphi, \psi)| \leqslant 1, \quad \forall\, \varphi, \psi \in [0, 2\pi].$$

Эта схема безусловно устойчивая, но уравнение на верхнем слое очень сложное: каждое уравнение связывает пять неизвестных. Картина «связности» имеет вид, показанный на рис. 12, т.е. получается система KM уравнений с матрицей, в каждой строке которой всего пять ненулевых элементов.

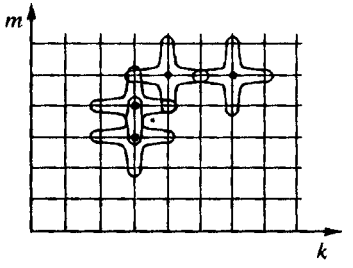


Рис. 12

Мы не останавливаемся специально на аппроксимации начальных данных и краевых условий, так как здесь ничего нового по сравнению с одномерной задачей не появляется. Эти уравнения замыкают систему уравнений на верхнем слое. Матрица системы имеет специальную структуру: все ненулевые элементы расположены на пяти диагоналях.

Матрицы подобного рода часто появляются при аппроксимации краевых задач методом сеток. Они получили специальное название «ленточные матрицы». Этот термин связан с тем, что в такой матрице можно выделить «ленту» около главной диагонали, в которой расположены все ненулевые элементы, и «площадь» ленты (в нашем случае $2KMK$) существенно меньше «площади» матрицы K^2M^2.

В настоящее время созданы специальные методы решения систем уравнений с такими матрицами. Их основная особенность состоит в том, что в процессе решения ненулевые элементы появляются только в области исходной ленты, т.е. можно проводить вычисления с объемом памяти, существенно меньшим объема полной матрицы; соответственно уменьшается и число операций.

Однако в нашем случае есть другой путь: построение такой аппроксимации уравнения теплопроводности, которая совмещает безусловную устойчивость с возможностью построения чрезвычайно эффективного алгоритма решения уравнений на верхнем слое. Эта конструкция (так называемый *метод переменных направлений*) — одно из важных изобретений в современных численных методах решения задач математической физики. Существенным ее элементом является метод прогонки. Шаги по времени не одинаковы. Они выполняются по чередующимся формулам (четные по одной схеме, нечетные по другой).

Рассмотрим пару шагов: u^n известно. Вычислим сначала u^{n+1}, затем u^{n+2}.

1. Первый шаг: $u^n \to u^{n+1}$. Используем схему

$$\frac{u^{n+1}_{k,m} - u^n_{k,m}}{\tau} = \left(\frac{\partial^2 u}{\partial x^2}\right)^{n+1}_{k,m} + \left(\frac{\partial^2 u}{\partial y^2}\right)^{n}_{k,m} + f^n_{k,m},$$

в которой производная по x аппроксимируется на верхнем слое, производная по y — на нижнем.

Система уравнений на верхнем слое расщепляется на независимые системы. Каждая такая система объединяет неизвестные, лежащие на одной горизонтальной линии, и каждая группа переменных $\{u^n_m\} \equiv \{u^n_{k,m}\}$ $(k = 0, 1, \ldots, K)$ может быть найдена независимо от всех остальных. Более того, эта система является системой с трехдиагональной матрицей и может быть решена прогонкой по горизонтальной линии ценой $O(K)$ операций. Всего таких линий M; следовательно, весь массив u^{n+1} может быть найден ценой $O(KM)$ операций, т.е. число операций пропорционально числу неизвестных.

2. Второй шаг: $u^{n+1} \to u^{n+2}$. Он осуществляется по аналогичной схеме, но с переменой ролей x и y:

$$\frac{u^{n+2}_{k,m} - u^{n+1}_{k,m}}{\tau} = \left(\frac{\partial^2 u}{\partial x^2}\right)^{n+1}_{k,m} + \left(\frac{\partial^2 u}{\partial y^2}\right)^{n+2}_{k,m} + f^n_{k,m}.$$

Здесь та же ситуация, только система расщепляется на независимые подсистемы, объединяющие переменные на одной вертикали:

$$\{u^{n+2}_k\} \equiv \{u^{n+2}_{k,m}\}, \qquad m = 0, 1, \ldots, M.$$

Таким образом, алгоритм метода переменных направлений «экономичен», т.е. число операций пропорционально числу неизвестных.

Спектральная устойчивость схемы переменных направлений. Рассмотрим эволюцию универсальной функции $e^{ik\varphi+im\psi}$ за два стандартных шага. Схема эволюции такая:

$$u^n \equiv e^{ik\varphi+im\psi} \to u^{n+1} = \lambda_1(\varphi, \psi)\, e^{ik\varphi+im\psi} \quad \text{(первый шаг)};$$

$$u^{n+1} \equiv e^{ik\varphi+im\psi} \to u^{n+2} = \lambda_2(\varphi, \psi)\, e^{ik\varphi+im\psi} \quad \text{(второй шаг)}.$$

Сдвоенный шаг дает

$$u^n \equiv e^{ik\varphi+im\psi} \to u^{n+1} = \lambda_1 u^n \to \lambda_2 u^{n+1} =$$

$$= \lambda_2\lambda_1 u^n = \lambda_1(\varphi, \psi)\lambda_2(\varphi, \psi)e^{ik\varphi+im\psi}.$$

Здесь $\lambda(\varphi, \psi) = \lambda_1\lambda_2$ отвечает за устойчивость схемы.

Вычисляем λ_1:

$$\frac{\lambda_1 - 1}{\tau} = -4\frac{\lambda_1}{h_x^2}\sin^2\frac{\varphi}{2} - 4\frac{1}{h_y^2}\sin^2\frac{\psi}{2},$$

т.е.

$$\lambda_1 = \left(1 - 4\frac{\tau}{h_y^2}\sin^2\frac{\psi}{2}\right)\Big/\left(1 + 4\frac{\tau}{h_x^2}\sin^2\frac{\varphi}{2}\right).$$

Аналогично (с переменой ролей x и y) вычисляем λ_2:

$$\lambda_2 = \left(1 - 4\frac{\tau}{h_x^2}\sin^2\frac{\varphi}{2}\right)\Big/\left(1 + 4\frac{\tau}{h_y^2}\sin^2\frac{\psi}{2}\right).$$

Окончательно:

$$\lambda(\varphi, \psi) = \lambda_1\lambda_2 = \frac{1 - 4\,(\tau/h_x^2)\sin^2(\varphi/2)}{1 + 4\,(\tau/h_x^2)\sin^2(\varphi/2)} \cdot \frac{1 - 4\,(\tau/h_y^2)\sin^2(\psi/2)}{1 + 4\,(\tau/h_y^2)\sin^2(\psi/2)}.$$

Очевидно, $|\lambda(\varphi, \psi)| \leqslant 1$ и схема безусловно устойчива.

О краевых условиях. При осуществлении прогонки «по линии» система разностных уравнений замыкается соответствующими краевыми условиями. В случае задания значений u на границе области дело совсем просто: правые и левые значения u^{n+1} на данной линии уже известны. Не возникает трудностей и в том случае, когда заданы общие краевые условия третьего рода: $-\alpha\frac{\partial u}{\partial \nu} + \beta u = \psi$, где ν — направление внешней нормали к грани-

це. Аппроксимация, например, на правой границе (при $x = X$) имеет очевидную форму:

$$\alpha \frac{u^{n+1}_{K,m} - u^{n+1}_{K-1,m}}{\tau} + \beta u^{n+1}_{K,m} = \psi^{n+1}_m, \qquad m = 1, 2, \ldots, M-1.$$

Она использует величины u^{n+1} на одной m-й горизонтали сетки и не препятствует расщеплению системы на изолированные подсистемы. Однако если заданы условия с косой производной (это нечасто встречающийся в приложениях случай), их аппроксимация уже не может быть осуществлена по величинам на одной линии.

Аналогичные препятствия к непосредственному обобщению схемы метода переменных направлений возникают и при краевых условиях с нормальной производной на границе непрямоугольной области, если ее граница не проходит по координатной линии сетки и направление нормали к такой границе является «косым» по отношению к линии сетки.

Основная конструктивная идея метода переменных направлений оказалась очень плодотворной и была в дальнейшем обобщена. Ниже кратко описываются два обобщения, часто применяемые в современной практике конструирования разностных схем для краевых задач математической физики.

Метод дробных шагов. Суммарная аппроксимация. Рассмотрим эволюционную систему дифференциальных уравнений с частными производными в общей форме:

$$\frac{\partial u}{\partial t} = L_1 u + L_2 u + f.$$

Здесь L_1, L_2 — операторы дифференцирования по x и y соответственно; u в принципе может быть вектор-функцией; f — заданная правая часть. (В такой системе отсутствуют члены со смешанными производными.)

В методе дробных шагов переход от слоя u^n к слою u^{n+1} совершается с использованием дробного промежуточного шага $u^{n+1/2}$, который условно можно отнести к моменту времени $t_{n+1/2} = t_n + \tau/2$ (τ — шаг сетки по t). Схема стандартной группы дробных шагов такова. Пусть u^n известно.

1. Вычисляем $u^{n+1/2}$, используя, например, неявную схему

$$\frac{u^{n+1/2} - u^n}{\tau} = L_1 u^{n+1/2} + f_1.$$

Ради простоты мы не вводим особых обозначений для сеточных функций и разностных аппроксимаций дифференциальных операторов. Читатель по тексту без труда догадается, где речь идет о

дифференциальном уравнении, а где — о его конечно-разностной аппроксимации.

Уравнение первого дробного шага на верхнем слое распадается на серии независимых уравнений, связывающих неизвестные на одной линии сетки. Такие разностные уравнения называются *локально-одномерными*. Второе пространственное измерение в этих уравнениях присутствует в качестве параметра, определяющего совокупность «одномерных» задач.

2. Второй дробный шаг строится аналогично:

$$\frac{u^{n+1}-u^{n+1/2}}{\tau} = L_2 u^{n+1} + f_2.$$

Он имеет такую же локально-одномерную структуру. Уравнения на верхнем слое в обоих случаях обычно решаются алгоритмами типа прогонки, как и в методе переменных направлений. В этих общих терминах метод переменных направлений записывается в форме

$$\frac{u^{n+1}-u^{n}}{\tau} = L_1 u^{n+1} + L_2 u^{n} + f,$$

$$\frac{u^{n+2}-u^{n+1}}{\tau} = L_1 u^{n+1} + L_2 u^{n+2} + f.$$

Для метода дробных шагов нам надо еще уточнить два вопроса. Как «разбить» правую часть f на f_1 и f_2? Какому времени соответствует функция u^{n+1}: $t_n + \tau$ или $t_n + 2\tau$? На оба вопроса мы получим ответ, используя формальную процедуру «исключения промежуточного слоя», которая приведет к сравнительно обычной форме аппроксимации.

Запишем уравнения дробных шагов в виде

$$(E - \tau L_1) u^{n+1/2} = u^n + \tau f_1, \qquad (E - \tau L_2) u^{n+1} = u^{n+1/2} + \tau f_2.$$

Подействуем на второе из них оператором $E - \tau L_1$:

$$(E - \tau L_1)(E - \tau L_2) u^{n+1} = (E - \tau L_1) u^{n+1/2} + \tau (E - \tau L_1) f_2.$$

Используя уравнение первого дробного шага, исключаем $u^{n+1/2}$:

$$(E - \tau L_1)(E - \tau L_2) u^{n+1} = u^n + \tau f_1 + \tau (E - \tau L_1) f_2.$$

Это уравнение можно привести к сравнительно стандартной форме (предварительно раскроем скобки):

$$\frac{u^{n+1}-u^{n}}{\tau} = L_1 u^{n+1} + L_2 u^{n+1} + (f_1 + f_2) + \tau (L_1 f_2 + L_1 L_2 u^{n+1}).$$

Последний член имеет величину порядка $O(\tau)$ и относится к погрешности аппроксимации.

На два поставленных вопроса следуют очевидные ответы: два дробных шага осуществляют продвижение решения по времени на τ, а не на 2τ, как во внешне похожем методе переменных направлений. Правые части f_1 и f_2 вводят так, чтобы $f_1 + f_2 = f$, т.е., например, $f_1 = f_2 = f/2$. Разумеется, равенство $f_1 + f_2 = f$ можно трактовать с точностью, например, до $O(\tau)$.

Заметим, что мы провели выкладки формально, упуская при этом из вида существенное обстоятельство — краевые условия. Дело в том, что проделывая выкладки достаточно аккуратно, мы должны в общую операторную форму включить краевые условия примерно в том стиле, как это делалось в § 11 при записи конкретной задачи в абстрактной форме. Эта аккуратность имеет определенные практические последствия: в методе дробных шагов нужно достаточно ответственно подходить к аппроксимации краевых условий. На это впервые, видимо, обратил внимание Е. Г. Дьяконов, он же разработал формализм соответствующего анализа. Здесь мы ограничимся тем, что обратим внимание читателя на этот момент конструирования схемы.

Полезно следующее рассуждение, приводящее к понятию о суммарной (аддитивной) аппроксимации. Будем трактовать u^n, $u^{n+1/2}$, u^{n+1} как величины, относящиеся к моментам t_n, $t_{n+1/2}$, t_{n+1}, и вычислим погрешность аппроксимации каждого полушага. После очевидных преобразований получаем (полагая для простоты $f_1 = f_2 = f/2$)

$$\frac{u^{n+1} - u^n}{\tau/2} = (L_1 + L_2)\, u^{n+1/2} + f + \{(L_1 - L_2)\, u^{n+1/2}\},$$

$$\frac{u^{n+1} - u^{n+1/2}}{\tau/2} = (L_1 + L_2) u^{n+1} + f - \{(L_1 - L_2) u^{n+1}\}.$$

Члены, стоящие вне фигурных скобок аппроксимируют исходное уравнение в обычном смысле слова, члены же в фигурных скобках следует отнести к погрешностям аппроксимации. Их особенность в том, что они имеют недопустимую (согласно общим представлениям) величину порядка $O(1)$, но они почти равны друг другу по модулю (с точностью до $O(\tau)$ из-за того, что в них участвуют u с разных слоев) и противоположны по знаку.

Влияние больших альтернирующих погрешностей в среднем компенсируется, и при определенных условиях они не препятствуют сходимости приближенного решения к точному. Грубо говоря, дело в том, что на решение дифференциального уравнения существенное влияние оказывают не мгновенные значения правых частей, а их средние значения по малым интервалам времени. В связи с этим можно сказать,

что разностные формулы дробных шагов не аппроксимируют дифференциальных уравнений, если оценивать погрешность аппроксимации в какой-нибудь «сильной» норме (типа C или L_2), но аппроксимирует в «слабой» норме, в которой такая альтернирующая по знаку функция в среднем близка к нулю (к «слабому» нулю).

Разумеется, введение «слабой» аппроксимации может быть оправдано существенным усилением теоремы типа «аппроксимация + устойчивость = сходимость». Нужно доказать, что возмущение правой части разностного уравнения погрешностью, малой в слабом смысле, но большой в обычном смысле слова, должно привести к малому (в обычном смысле слова) отличию решений. Конечно, это более тонкий факт, чем устойчивость при действительно малом возмущении, и устанавливать его в конкретных случаях гораздо труднее. Тем не менее схемы с суммарной аппроксимацией часто оказываются практически очень удобными и в последние годы все смелее вводятся в расчетные методики сложных задач (конечно, без строгого теоретического обоснования; впрочем, такого обоснования не имеют и схемы с полной аппроксимацией). Теория схем с суммарной аппроксимацией разрабатывалась Н. Н. Яненко, А. А. Самарским и их учениками.

Схемы расщепления. В сущности то, что называют схемами расщепления, формально не отличается от схем дробных шагов. Пусть решаемое дифференциальное уравнение имеет вид

$$\frac{\partial u}{\partial t} = L_1 u + L_2 u + L_3 u + f,$$

где L_1, L_2, L_3 — операторы, описывающие разные физические процессы (например, L_1 — перенос, L_2 — диффузию, L_3 — еще что-нибудь; можно разделять процессы и по направлениям, т.е. L_1 описывает диффузию по x, L_2 — диффузию по y и т.д.).

Предположим, что каждое из «частичных» уравнений

$$\frac{\partial u}{\partial t} = L_i u + f, \quad i = 1, 2, 3,$$

уже хорошо освоено в вычислительной практике, для них построены апробированные схемы, удовлетворяющие, кроме формального требования аппроксимации, еще каким-то дополнительным требованиям (в дальнейшем мы познакомимся с ними при описании некоторых методов решения задач газовой динамики), а для всего уравнения в целом таких схем построить не удается. Тогда можно использовать расчет по схеме, формально совпадающей со схемой дробных шагов. Этот общий подход получил название «расщепление по физическим процессам» (его более ранний вариант — «расщепление по направлениям»).

§ 14. Решение эллиптических задач методом сеток

В различных задачах математической физики в качестве важной составляющей части входят краевые задачи для эллиптического уравнения, особенно часто для уравнения Пуассона. Например, в уравнения, описывающие движение плазмы (см. § 24), входит уравнение для потенциала электрических сил u:

$$\Delta u = -4\pi\rho,$$

где ρ — плотность заряда. Такое же уравнение входит в систему уравнений, описывающих эволюцию совокупности гравитационно взаимодействующих тел, в систему уравнений Навье—Стокса (динамика вязкой несжимаемой жидкости) и т.д. При расчете описываемых этими уравнениями явлений уравнение Пуассона приходится решать много раз: на каждом шаге по времени. Часто трудоемкость расчета определяется именно временем, затрачиваемым на решение этого уравнения.

Рассмотрим вопросы, связанные с быстрыми методами решения уравнения Пуассона. В общем случае нас интересует эллиптическое уравнение

$$\sum_i \sum_j \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right) = f$$

в произвольной области Ω с краевыми условиями, для определенности, первого рода:

$$u|_{\partial\Omega} = \varphi.$$

Здесь f, φ — заданные функции, $a_{ij}(x)$ — известные функции, удовлетворяющие следующим естественным условиям:

а) $a_{ij} = a_{ji}$ (симметричность);

б) $\sum_i \sum_j a_{ij}\xi_i\xi_j \geqslant \alpha \sum_i \xi_i^2$, $\forall\,\xi$ (эллиптичность уравнения, ξ — вещественные).

Индексы i, j меняются от 1 до 2 или 3 (в зависимости от размерности пространства, в котором решается задача).

Имея в виду именно такую общую эллиптическую краевую задачу, мы начнем анализ с самого ее простого варианта, так как многие факты уже здесь могут быть обнаружены. Рассмотрим уравнение Пуассона

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y), \quad 0 \leqslant x,\ y \leqslant 1,$$

с краевыми условиями первого рода: $u(x, y) = \varphi(x, y)$ на границе квадрата, или, подробнее:

$$u(x, 0) = \varphi_1(x), \quad u(x, 1) = \varphi_3(x),$$
$$u(0, y) = \varphi_2(y), \quad u(1, y) = \varphi_4(y).$$

Введем основные объекты метода сеток.

Сетка. Область покрываем сеткой из точек (k, m) с координатами $x_k = kh$, $y_m = mh$ $(k, m = 0, 1, \ldots, N)$, $h = 1/N$. (Ради простоты, считаем сетку равномерной, с одинаковыми шагами, хотя это, конечно, совсем необязательно.)

Сеточная функция. Приближенное решение ищем в виде сеточной функции $u_{k,m}$, которую, как обычно, трактуем как приближенное значение $u(x_k, y_m)$. Функция $u_{k,m}$ определена во всех узлах сетки: $\{u_{k,m}\}$, $(k, m \in 0, N)$.

Аппроксимация уравнения. Сеточную функцию будем искать как решение системы уравнений, полученных простейшим способом, — прямой заменой входящих в уравнение производных на соответствующие разностные отношения:

$$\frac{u_{k-1,m} - 2u_{k,m} + u_{k+1,m}}{h^2} + \frac{u_{k,m+1} - 2u_{k,m} + u_{k,m-1}}{h^2} = f_{k,m}.$$

Это уравнение имеет крестообразный шаблон и называется простейшей *пятиточечной аппроксимацией уравнения* Пуассона. Оно определено только в так называемых внутренних узлах сетки, т.е. при $k, m = 1, 2, \ldots, N-1$. В дальнейшем мы будем использовать более компактные формы записи этого уравнения:

$$(\Delta u)_{k,m} = f_{k,m}$$

и даже $\Delta u = f$ (из контекста будет ясно, о каком, дифференциальном или разностном, уравнении идет речь). Используем и такую форму:

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{k,m} + \left(\frac{\partial^2 u}{\partial y^2}\right)_{k,m} = f_{k,m}.$$

В сущности в этом параграфе всюду в дальнейшем производные обозначают соответствующие разностные аппроксимации.

Аппроксимация краевых условий. Это вопрос совсем простой: $u_{k,0} = \varphi_1(x_k)$ и т.д. Если бы на границе $x = 1$ краевое условие имело более сложный вид: $\alpha u_x + \beta u = \varphi(y)$, его можно было бы аппроксимировать так:

$$\alpha \frac{u_{N,m} - u_{N-1,m}}{h} + \beta u_{N,m} = \varphi(y_m), \quad m = 1, 2, \ldots, N-1.$$

(Обратим внимание на то, что в угловых точках условий нет, но там функция в сущности и не нужна. Впрочем, можно было бы просто краевые условия определять для $k, m = 0, 1, \ldots, N$, потребовав согласования функций φ_i в угловых точках.)

Итак, построение аппроксимирующей разностной задачи закончено. Мы получили систему линейных алгебраических уравнений высокого порядка. (В современных расчетах $N \approx 10^2$, стало быть, система имеет порядка 10^4 неизвестных.) Эта система имеет специальную структуру: каждое уравнение связывает значения только пяти неизвестных.

Как всегда, возникают два вопроса.

1. Теоретический вопрос: как обосновать метод сеток, т.е. доказать (при тех или иных предположениях), что

$$|u_{k,m} - u(x_k, y_m)| \leqslant Ch^p?$$

Такое обоснование распадается на установление аппроксимации и устойчивости разностной схемы.

2. Практический вопрос: как фактически решить систему уравнений, т.е. вычислить $u_{k,m}$?

Обоснование метода. Рассмотрим свойства аппроксимации и устойчивости системы разностных уравнений в несколько более общей ситуации. Пусть задача решается в произвольной области Ω с гладкой границей. В этом случае сначала надо уточнить построение сетки и аппроксимирующих задачу уравнений. Покроем плоскость (x, y) квадратной, для простоты, сеткой с шагом h. Множество узлов (k, m), для которых точки (x_k, y_m) попадают строго внутрь Ω, назовем *внутренними*. В каждой такой внутренней точке поместим шаблон используемой схемы и отметим узлы сетки, входящие в шаблон. В случае простейшей схемы шаблон в точке (k, m) «отмечает» еще четыре узла: $(k-1, m)$, $(m-1, k)$, $(k+1, m)$ и $(k, m+1)$.

Множество отмеченных узлов назовем *счетными* узлами; именно в них будет определена сеточная функция $u_{k,m}$. Разумеется, внутренние узлы являются счетными; все остальные счетные узлы образуют множество *граничных* узлов. В каждом внутреннем узле может быть записано стандартное разностное уравнение; в граничных узлах следует использовать краевое условие. Простейший вариант: для граничного узла (k, m) можно найти на контуре области $\partial\Omega$ ближайшую точку $(\tilde{x}_{k,m}, \tilde{y}_{k,m})$ и реализовать краевое условие сносом: $u_{k,m} = \varphi(\tilde{x}_{k,m}, \tilde{y}_{k,m})$. Очевидно, расстояние между точками (x_k, y_k) и $(\tilde{x}_{k,m}, \tilde{y}_{k,m})$ есть $O(h)$. (Этого достаточно, чтобы считать точку на контуре «ближайшей».)

Пусть $U(x, y)$ — решение исходной дифференциальной задачи, $U_{k,m}$ — ограничение решения на сетку. Подставляя $U_{k,m}$ в разностное уравнение, вычисляем (формально) погрешность аппроксимации для внутренних и граничных узлов (k, m) соответственно:

$$r_{k,m} = \begin{cases} (\Delta U)_{k,m} - f_{k,m} = \frac{h^2}{24}(U_{xxxx} + U_{yyyy}) + O(h^4), \\ U(x_k, y_m) - \varphi(\tilde{x}_{k,m}, \tilde{y}_{k,m}) = O(h). \end{cases}$$

Итак, если решение имеет четыре непрерывных производных (это обеспечивается двумя производными f и гладкостью φ на контуре), разностная задача имеет первый порядок аппроксимации. Если граничные узлы сетки точно попадают на границу области Ω (как для задачи в прямоугольной области), порядок аппроксимации равен двум.

Перейдем к устойчивости. Эллиптические линейные задачи в этом отношении достаточно благополучны. Часто можно установить их устойчивость, используя специфическое свойство — «принцип максимума».

Лемма 1. Пусть сеточная функция $u_{k,m}$ удовлетворяет условию $(\Delta u)_{k,m} > 0$ во всех внутренних узлах. Тогда $\max\limits_{k,m} u_{k,m}$ достигается в граничном узле. (Здесь, конечно, $(\Delta u)_{k,m}$ — аппроксимация оператора Лапласа на пятиточечном шаблоне.)

Доказательство. Предположим, что максимум $u_{k,m}$ не достигается на границе. Тогда он достигается в какой-то внутренней точке (i, j). В этой точке определена положительная по условию величина $(\Delta u)_{i,j}$. Распишем ее в полных обозначениях:

$$\frac{1}{h^2}(u_{i-1,j} + u_{i,j-1} - 4u_{i,j} + u_{i+1,j} + u_{i,j+1}) > 0.$$

Отсюда

$$u_{i,j} < \frac{1}{4}(u_{i-1,j} + u_{i,j-1} + u_{i+1,j} + u_{i,j+1}).$$

Это явно противоречит тому, что $u_{i,j}$ — максимум, т.е. не меньше каждого из четырех входящих в правую часть неравенства значений u. Таким же образом можно установить, что из условия $(\Delta u)_{k,m} < 0$ следует, что минимум $u_{k,m}$ достигается на границе.

Построим специальную функцию сравнения — разностную мажоранту Гершгорина. (Читатель, знакомый с начальными фактами теории уравнения Пуассона, легко поймет, что нижеследующее есть простое ее обобщение для разностного уравнения Пуассона.) Предположим, что точка (0, 0) находится внутри Ω близко к ее «центру». Введем мажоранту — сеточную функцию

$$w_{k,m} = \frac{1}{4}\|f\|[(x_k^2 + y_m^2) - R^2],$$

где $\|f\| = \max\limits_{x,y} |f(x, y)|$, R — пока произвольная постоянная.

Утверждение 1. Функция $w_{k,m}$ удовлетворяет разностному уравнению

$$(\Delta w)_{k,m} = \|f\|, \quad (k, m) \text{ — внутренний узел.}$$

Оно непосредственно следует из того, что для функции $x^2 + y^2$ значения вторых производных и вторых разностных производных совпадают (при любом шаге сетки). Выберем значение R таким, чтобы область Ω помещалась в круг радиусом R и функция $w_{k,m}$, таким образом, была отрицательной в Ω. Введем нормы

$$\|u\| = \max_{k,m} |u_{k,m}|, \quad \|f\| = \max_{(x,y)\in\Omega} |f(x, y)|, \quad \|\varphi\| = \max_{(x,y)\in\partial\Omega} |\varphi(x, y)|.$$

Теперь мы имеем все для доказательства устойчивости разностного уравнения Пуассона.

Теорема 1. Пусть функция $u_{k,m}$ удовлетворяет разностному уравнению Пуассона. Тогда имеет место общая для всех задач (т.е. равномерная по шагу сетки h) оценка

$$\|u\| \leqslant \|\varphi\| + \tfrac{1}{4} R^2 \|f\|.$$

Доказательство. Рассмотрим функцию $v_{k,m} = u_{k,m} + w_{k,m}$. Во внутренних узлах сетки она удовлетворяет разностному уравнению:

$$(\Delta v)_{k,m} = (\Delta u)_{k,m} + (\Delta w)_{k,m} = f_{k,m} + \|f\| \geqslant 0.$$

Следовательно, максимум $v_{k,m}$ достигается на границе и для любого счетного узла (i, j) можно записать соотношение (условимся, что (k, m) пробегают лишь граничные значения)

$$v_{i,j} \leqslant \max_{k,m} v_{k,m} = \max_{k,m} (u_{k,m} + w_{k,m}) \leqslant \max_{k,m} u_{k,m} = \max_{k,m} \varphi_{k,m} \leqslant \|\varphi\|$$

(здесь мы использовали отрицательность $w_{k,m}$).

Далее, из $v_{i,j} < \|\varphi\|$ следует

$$\|\varphi\| \geqslant u_{i,j} + w_{i,j} = u_{i,j} + \tfrac{1}{4}\|f\|\,(x_i^2 + y_j^2) - \tfrac{1}{4}\|f\|R^2.$$

Отсюда $u_{i,j} \leqslant \|\varphi\| + R^2\|f\|/4$. Так как функция $-u_{k,m}$ удовлетворяет тому же разностному уравнению Пуассона, но с изменением знаков f и φ, имеем второе неравенство: $-u_{k,m} \leqslant \|\varphi\| + R^2\|f\|/4$. Из двух полученных неравенств следует утверждение теоремы.

Таким образом, для всего семейства разностных задач Пуассона (с параметром $h \to 0$) установлена равномерная оценка решения через нормы функций, входящих в «правую часть» уравнения. Как уже разъяснялось, для линейной задачи она означает

непрерывную зависимость решения от правой части задачи, а отсюда (в силу теоремы Рябенького—Филиппова) следует сходимость разностного решения к точному с порядком, равным порядку аппроксимации.

Можно придать этому утверждению более определенную формулировку. Пусть $u^h_{k,m}$ — решение разностной задачи на сетке с шагом h, $U^h_{k,m}$ — ограничение точного решения на ту же сетку. Тогда

$$\|u^h_{k,m} - U^h_{k,m}\| \leqslant C_1 h + C_2 R^2 h^2/12,$$

где C_1 — оценка первой нормальной производной $U(x, y)$ на границе, C_2 — оценка четвертых производных $U(x, y)$. Тем самым мы установили устойчивость и сходимость в метрике C. Хотя доказанная теорема имеет внешне вполне законченный характер, ее не следует переоценивать: она основана на слишком сильных априорных предположениях о гладкости решения $U(x, y)$.

Дальнейшее развитие теории разностного уравнения Пуассона связано со стремлением существенно ослабить эти предположения, доведя их до «естественных». Например, если f только ограничена, $U(x, y)$ имеет лишь две производные. Реальные задачи обычно связаны с функциями, существенно лучшими по гладкости, чем просто ограниченные. Так, часто f — кусочно-гладкая функция, имеющая небольшое число линий разрыва самой функции или ее производных. Это приводит к тому, что погрешность аппроксимации устроена «неравномерно»: она мала (порядка $O(h^2)$) почти всюду в области. Только в окрестности линий нарушения гладкости $U(x, y)$ она существенно больше и может достигать даже величины $O(1)$. Следовательно, невязка может оказаться малой лишь в какой-то интегральной норме, более слабой, чем норма в C.

Соответственно, нужны и более тонкие теоремы об устойчивости. Такие теоремы тем ценнее, чем более слабая норма используется для невязки и чем более сильная — для погрешности. Все это, конечно, выходит за рамки принятого в книге теоретического уровня. Однако надо понимать, что дело не только в качестве теорем, но и в существе самой проблемы. Более слабые требования к невязке приводят к более слабым утверждениям о погрешности численного решения потому, что ухудшение гладкости искомого решения ведет к росту погрешности. Бороться с этим можно, просто тупо увеличивая число узлов сетки. Но иногда удаются более остроумные и квалифицированные способы повышения точности расчета, не требующие существенного увеличения объема вычислений. Один из таких приемов — *выделение особенности* (*регуляризация*). Поясним идею на простой задаче.

В квадрате $0 \leqslant x, y \leqslant 1$ решается задача Лапласа: $\Delta u = 0$, однако краевые условия $u(x, y) = \varphi(x, y)$ на границе содержат разрыв. Для

определенности, пусть $\varphi(x, 0) \to a$ при $x \to 0$, $\varphi(0, y) \to b$ при $y \to 0$, $a \neq b$. Решение такой задачи существует, но около точки (0, 0) оно теряет гладкость (в точке (0, 0) нет непрерывных первых производных). Это обстоятельство приводит к снижению точности расчета в окрестности точки (0, 0). Правда, влияние такого локального нарушения гладкости носит локальный характер и погрешность расчета быстро убывает при удалении от точки (0, 0).

Метод регуляризации состоит в следующем. Решение ищется в форме $u(x, y) = v(x, y) + w(x, y)$, где $v(x, y)$ — некоторая известная гармоническая функция, имеющая в граничных условиях тот же разрыв, который имеет заданная граничная функция φ, а в остальном — гладкая. Второе слагаемое $w(x, y)$ подлежит расчету. Для w ставится, очевидно, следующая краевая задача:

$$\Delta w = 0 \text{ внутри}, \qquad w(x, y) = \varphi(x, y) - v(x, y) \text{ на границе}.$$

Граничные значения функции w непрерывны, и при ее расчете методом сеток не происходит потери точности.

Для реализации такого выделения особенности нужно иметь функцию $v(x, y)$. В данном случае такая гармоническая функция известна. В теории функций комплексного переменного устанавливается, что, например, $\arg(x + iy) = \operatorname{arctg}(y/x)$ является гармонической функцией в положительном квадранте и она имеет разрыв в краевых условиях именно в той точке, где нам нужно:

$$\operatorname{arctg}\frac{y}{x} \to 0 \text{ при } y = 0,\ x \to 0, \qquad \operatorname{arctg}\frac{y}{x} = \frac{\pi}{2} \text{ при } x = 0,\ y \to 0.$$

Следовательно, в качестве v можно взять гармоническую функцию

$$v(x, y) = a + 2\frac{b-a}{\pi}\operatorname{arctg}\frac{y}{x}.$$

Диагональное преобладание. Устойчивость системы разностных уравнений удалось установить благодаря важному их свойству, называемому «диагональным преобладанием». В разностном уравнении $(\Delta u)_{k,m} = f_{k,m}$ «вес» центрального члена (коэффициент при $u_{k,m}$, равный $-4/h^2$) по модулю не меньше суммы модулей остальных «весов». Это есть следствие как математической структуры оператора Лапласа Δ (значение гармонической функции в какой-то точке равно среднему значению этой же функции по окружности с центром в рассматриваемой точке), так и использованной нами простейшей его аппроксимации.

Свойство диагонального преобладания настолько полезно, что его стараются обеспечить и при построении аппроксимации в более сложных ситуациях. Не всегда оно получается автоматически, иногда нужно потрудиться над конструкцией схемы. Для эллиптического

уравнения $u_{xx} + u_{xy} + u_{yy} = f$ можно предложить несколько схем, например:

$$(\Delta u)_{k,m} + \frac{1}{4h^2}(u_{k+1,m+1} - u_{k+1,m-1} - u_{k-1,m+1} + u_{k-1,m-1}) + f_{k,m}.$$

На рис. 13*а* показан шаблон схемы, около узлов которого проставлены коэффициенты схемы (умноженные на h^2). Видно, что диагонального преобладания нет.

Смешанную производную можно аппроксимировать иначе:

$$\frac{\partial^2 u}{\partial x \partial y} \approx \frac{1}{h^2}(u_{k+1,m+1} - u_{k,m+1} - u_{k+1,m} + u_{k,m}).$$

Коэффициенты схемы показаны на рис. 13*б*. Возможно, читателя смутит сокращение числа узлов в схеме, хотя диагональное преобладание здесь есть. Автора это обстоятельство тоже смущает, хотя ничего определенно компрометирующего эту схему мы сказать не

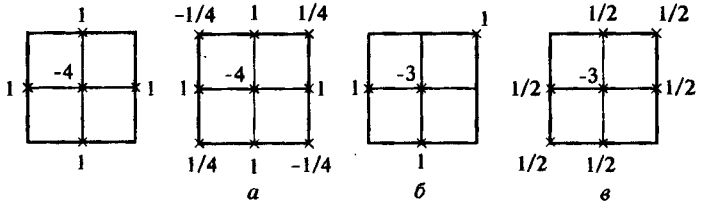


Рис. 13

можем. Попробуем аппроксимировать смешанную производную еще одним способом:

$$\frac{1}{2h^2}(u_{k+1,m+1} - u_{k,m+1} - u_{k+1,m} + u_{k,m}) + \frac{1}{2h^2}(u_{k,m} - u_{k,m-1} - u_{k-1,m} + u_{k-1,m-1}).$$

Шаблон схемы показан на рис. 13*в*.

Если бы смешанная производная входила в уравнение со знаком минус, следовало бы (для сохранения диагонального преобладания) ориентировать шаблон по другой диагонали. Что касается рассматриваемого уравнения, то это то же самое уравнение Пуассона, только в косоугольной системе координат. Для его решения, конечно, справедлив принцип максимума и теорема о среднем в соответствующей редакции. Надо, однако, предупредить, что отсутствие диагонального преобладания не является фатальным недостатком схемы, делающим ее непригодной для использования.

Перейдем к практическим проблемам: как найти решение системы линейных алгебраических уравнений очень высокого порядка и очень специальной структуры? Нас интересует не принципиальная возможность решить систему (это тривиально). Суть проблемы — в числе необходимых для решения операций. Предварительно отметим лишь, что из доказанных оценок (устойчивости) следует, что однородная система ($f \equiv 0, \varphi \equiv 0$) имеет только тривиальное решение $u \equiv 0$. Следовательно, разностное уравнение Пуассона однозначно разрешимо при любых правых частях.

Метод простой итерации. Основным средством решения больших систем линейных алгебраических уравнений, возникающих при разностной аппроксимации краевых эллиптических задач, являются методы итераций (последовательных приближений). В этих методах, начиная с какой-то сеточной функции $u^0_{k,m}$ (верхний индекс означает номер итерации), по тем или иным правилам находят $u^1_{k,m}, u^2_{k,m}, \ldots$ Если при этом $u^i_{k,m} \to u_{k,m}$ $(i \to \infty)$, то метод называют сходящимся.

Однако в этих вопросах одного факта сходимости мало, нам нужна еще и оценка скорости сходимости. Обычно она имеет вид

$$|u^i_{k,m} - u_{k,m}| \leqslant Cq^i \quad (\forall\, k, m;\ q < 1).$$

Числа $q < 1$ — свои для разных методов: чем меньше q, тем лучше метод, тем быстрее он сходится. Поскольку $u_{k,m}$ совпадает с решением $U_{k,m}$ с точностью до $\varepsilon = O(h^p)$, то итерации следует проводить до тех пор, пока $u^i_{k,m}$ не совпадет с $u_{k,m}$ с той же точностью $\varepsilon = O(h^p)$.

Дальнейшие итерации особого смысла не имеют. Поэтому обычно назначается некоторое $\varepsilon = O(h^p)$ и делается такое число $i(\varepsilon)$ итераций, которое обеспечивает оценку

$$|u^i_{k,m} - u_{k,m}| \leqslant \varepsilon, \quad \text{т.е. } Cq^i = \varepsilon, \qquad \text{откуда } i(\varepsilon) = \frac{\ln(\varepsilon/C)}{\ln q}.$$

Кроме числа итераций, мы должны учитывать и число операций, которых требует выполнение одной итерации. Обозначим его через T. Можно считать T временем выполнения одной итерации, так как в конечном счете нас интересует именно машинное время, необходимое для получения ε-решения. Очевидно,

$$t(\varepsilon) = i(\varepsilon)T = T\,\frac{\ln(\varepsilon/C)}{\ln q}.$$

(Если T — число операций, то мы получаем характеристику метода как такового. Если T — машинное время, то мы получаем характеристику, учитывающую быстродействие ЭВМ, приспособленность данного алгоритма к ее архитектуре и качество программирования.)

Метод простой итерации несложен. Опишем стандартную итерацию. Пусть i-е приближение $\{u^i_{k,m}\}$ известно (будем обозначать его просто u^i). Тогда для внутренних и граничных узлов имеем соответственно

$$u^{i+1}_{k,m} = \begin{cases} u^i_{k,m} + \tau(\Delta u^i - f)_{k,m}, & (k, m) \text{ внутри,} \\ u^i_{k,m} = \varphi, & (k, m) \text{ на границе.} \end{cases}$$

Число операций $T = K^* N^2$ ($K^* \approx 10$), так как число узлов (k, m) есть N^2, а вычисление $(\Delta u^i - f)_{k,m}$ требует около шести операций. (Считаем, что массив $f_{k,m}$ хранится в памяти.) Разумеется, в памяти не хранятся массивы u^i для каждого i. Как и при решении уравнения теплопроводности, заведомо достаточно двух массивов, а при незначительном усложнении программы и одного (плюс N буферных ячеек памяти). Особую роль играет τ — итерационный параметр.

Анализ сходимости. Оценка скорости сходимости. Выбор оптимального значения τ. Для анализа сходимости введем фундаментальные объекты — погрешность $v_{k,m}$ и невязку $r^i_{k,m}$:

$$v^i_{k,m} \equiv u^i_{k,m} - u_{k,m}, \qquad r^i_{k,m} = \begin{cases} (\Delta u^i - f)_{k,m}, & (k, m) \text{ внутри,} \\ (u^i - \varphi)_{k,m}, & (k, m) \text{ на границе.} \end{cases}$$

Равенство нулю невязки или погрешности означает, что найдено точное решение. Погрешность имеет очевидный смысл, но, как правило, она нам неизвестна. Невязка удобна тем, что ее всегда можно вычислить. Поэтому обычно итерации обрывают, когда невязка достигает достаточно малой величины. В дальнейшем мы увидим, что между погрешностью и невязкой есть простая связь: $\|v\| \leqslant \mu \|r\|$. (Число μ будет указано; нормы здесь и в дальнейшем гильбертовы.)

Выведем формулу итераций погрешности. Из формулы итераций и тождества $u = u + \tau(\Delta u - f)$ имеем для внутренних и граничных узлов (k, m) соответственно

$$\begin{cases} u^{i+1}_{k,m} = u^i_{k,m} + \tau(\Delta u^i - f)_{k,m}, \\ u_{k,m} = u_{k,m} + \tau(\Delta u - f)_{k,m}, \end{cases} \qquad \begin{cases} u^{i+1}_{k,m} = u^i_{k,m}, \\ u_{k,m} = u_{k,m}. \end{cases}$$

Вычитая, получаем уравнение эволюции погрешности:

$$v^{i+1}_{k,m} = \begin{cases} v^i_{k,m} + \tau(\Delta v^i)_{k,m}, & (k, m) \text{ внутри,} \\ 0, & (k, m) \text{ на границе.} \end{cases}$$

Введем операторную запись. Определим оператор D, переводящий сеточную функцию в такую же функцию:

$$(Du)_{k,m} = \begin{cases} (\Delta u)_{k,m}, & (k, m) \text{ внутри,} \\ u_{k,m}, & (k, m) \text{ на границе.} \end{cases}$$

Тогда для сеточных функций v, равных нулю на границе, имеем

$$v^{i+1} = (E + \tau D)v^i.$$

Тем самым проблема свелась к оценке нормы итерационного оператора $E + \tau D$, так как $\|v^{i+1}\| \leqslant \|E + \tau D\| \|v^i\|$ и, следовательно, $\|v^i\| \leqslant \|E + \tau D\|^i \|v^0\|$. Этой оценкой мы и займемся. Наиболее эффективным аппаратом подобных оценок является метод Фурье или, что то же самое, спектральный анализ оператора. Задача простая и известны аналитические выражения для собственных векторов и чисел.

Лемма 2. Сеточные функции $\varphi^p_k = \sin(kp\pi/N)$ суть собственные векторы разностного оператора $\partial^2/\partial x^2$, которым соответствуют собственные значения $\lambda^p = (4/h^2)\sin^2(p\pi/2N)$. Здесь p — номер собственной функции ($p = 1, 2, \ldots, N-1$), k — номер узла.

Доказательство состоит в простой проверке, которая опускается. Заметим только, что нам удобно ввести спектр формулой

$$\left(\frac{\partial^2 \varphi^p}{\partial x^2}\right)_k \equiv \frac{\varphi^p_{k-1} - 2\varphi^p_k + \varphi^p_{k+1}}{h^2} = -\lambda^p \varphi^p_k.$$

Введем собственные векторы оператора $\partial^2/\partial y^2$: $\psi^q_m \equiv \sin(mq\pi/N)$ ($q = 1, 2, \ldots, N-1$). Очевидно, $(\partial^2 \psi^q/\partial y^2)_m = -\lambda^q \psi^q_m$. Отметим, что на границах ($m = k = 0$, $m = k = N$) эти функции обращаются в нуль, что нам и нужно.

Лемма 3. Собственными функциями оператора D являются функции $z^{p,q}_{k,m} = \varphi^p_k \psi^q_m$. Им соответствуют собственные значения $\lambda^{p,q} = \lambda^p + \lambda^q$.

Доказательство состоит в прямом вычислении $(Dz^{p,q})_{k,m}$. В дальнейшем особую роль будут играть границы спектров:

$$0 \leqslant l' \leqslant \lambda^p \leqslant L', \quad p = 1, 2, \ldots, N-1,$$

где

$$l' = \lambda^1 = \frac{4}{h^2}\sin^2\frac{\pi}{2N} \approx \pi^2 \quad (\text{так как } N \gg 1),$$

$$L' = \lambda^{N-1} = \frac{4}{h^2}\sin^2\frac{(N-1)\pi}{2N} \approx \frac{4}{h^2}\sin^2\frac{\pi}{2} = \frac{4}{h^2} = 4N^2.$$

Очевидно, для $\lambda^{p,q}$ имеем

$$\lambda^{p,q} \in [l, L], \quad \text{где } l \approx 2\pi^2, \ L = 8/h^2 = 8N^2.$$

Будем использовать следующие факты из теории дискретных рядов Фурье.

Всякая функция $v_{k,m}$, равная нулю на границе, имеет представление

$$v_{k,m} = \sum_{p,q} c_{p,q}\, z^{p,q}_{k,m}.$$

При этом

$$\|v\| = \left\{\sum_{k,m} (v_{k,m})^2\right\}^{1/2} = \left\{\sum_{p,q} (c_{p,q})^2\right\}^{1/2}.$$

Теперь все готово для анализа сходимости итераций.

Разлагая в ряд Фурье начальную погрешность v^0, имеем

$$v^1 = (E + \tau D)v^0 = (E + \tau D)\sum c_{p,q}\, z^{p,q} =$$

$$= \sum c_{p,q}(E + \tau D)z^{p,q} = \sum c_{p,q}(1 - \tau\,\lambda^{p,q})z^{p,q}.$$

Обозначим $g(\tau) = \max\limits_{l \leqslant \lambda \leqslant L} |1 - \tau\lambda|$. Тогда

$$\|v^1\| = \left\{\sum (c_{p,q})^2(1 - \tau\,\lambda^{p,q})^2\right\}^{1/2} \leqslant g(\tau)\left\{\sum (c_{p,q})^2\right\}^{1/2} = g(\tau)\|v^0\|.$$

Точно так же:

$$\|v^i\| \leqslant g^i(\tau)\|v^0\|.$$

Выберем итерационный параметр τ так, чтобы сходимость была максимально быстрой. Получаем типичную задачу: найти

$$\min_{\tau}\ \{\max_{l \leqslant \lambda \leqslant L} |1 - \tau\lambda|\}.$$

Начнем ее решение с внутренней операции:

$$\max_{l \leqslant \lambda \leqslant L} |1 - \tau\lambda| = \max\{|1 - \tau l|,\ |1 - \tau L|\}.$$

Почти очевидно, что $\max |1-\tau\lambda|$ достигается на правой или левой границе интервала $[l, L]$. Таким образом, нужно найти

$$\min_{\tau} \{\max \{|1-\tau l|,\ |1-\tau L|\}\}.$$

Из простого анализа графиков функций $|1-\tau L|$, $|1-\tau l|$ и $\max\{|1-\tau L|, |1-\tau l|\}$ видно (рис. 14), что оптимальное значение $\tau_{\text{опт}}$ находится из соотношения

$$1-\tau_{\text{опт}} l = -(1-\tau_{\text{опт}} L), \quad \text{т.е.} \quad \tau_{\text{опт}} = 2/(L+l).$$

Оценим скорость сходимости:

$$g_{\text{опт}} = 1-\tau_{\text{опт}} l = 1-\frac{2l}{L+l} = \frac{L-l}{L+l} \approx 1-\frac{2l}{L}$$

(так как $l \ll L$). При $l = 2\pi^2$, $L = 8N^2$ получаем характеристику оптимальной сходимости метода простой итерации:

$$g_{\text{опт}} = 1-2(\pi/N)^2.$$

Для $N = 100$ показатель сходимости $g_{\text{опт}} = 0.9995$. Полезна будет и формула для числа итераций, необходимых для уменьшения погрешности начального приближения в ε^{-1} раз. При этом $i(\varepsilon)$ находится из соотношения $\|v^i\| \approx \varepsilon \|v^0\|$, т.е.

$$i(\varepsilon) = \frac{\ln \varepsilon}{\ln g_{\text{опт}}} = \frac{\ln \varepsilon}{\ln (1-2l/L)} \approx \frac{L}{2l} \ln \frac{1}{\varepsilon}.$$

Число $\eta = L/l$ — важная характеристика матрицы системы разностных уравнений (так называемое *число обусловленности*). Чем больше η, тем «хуже» матрица, тем труднее проводить вычисления. В данном случае обусловленность существенно повлияла на число итераций. В примере с $N = 100$ при $\varepsilon \approx 10^{-5}$ число итераций порядка 10^4. Учитывая, что каждая итерация «стоит» порядка $10N^2$ операций, оцениваем необходимое для решения задачи число операций в 10^9; для БЭСМ-6 это около часа работы. Ясно, что с таким методом нельзя браться за серьезную вычислительную работу. Большие усилия были затрачены на разработку методов ускорения итерационных процессов, на создание новых, существенно более эффективных. Наиболее быстрые современные методы решения разностного уравне-

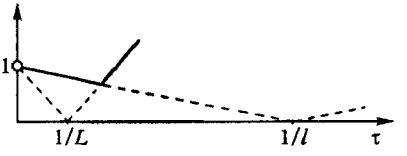


Рис. 14

ния Пуассона на сетке 100×100 требуют $1 \div 10$ секунд работы БЭСМ-6. Ниже мы опишем некоторые из таких методов. Метод же простой итерации послужил нам удобным поводом ввести читателя в суть проблемы и определить основные объекты и термины.

Чебышевское ускорение простых итераций. В методе простой итерации мы получили такое выражение для погрешности:

$$v^i = \sum_{p,\,q} c_{p,\,q}\,(1 - \tau_{\text{опт}}\,\lambda^{p,\,q})^i\,z^{p,\,q}.$$

Из него видно, что погашение фурье-компонент погрешности происходит неравномерно: в средней части спектра (при $\lambda \approx L/2$) существенно быстрее, чем на его краях, и результат определяется именно скоростью погашения на краях спектра.

Возникает идея сделать погашение более или менее равномерным по всему спектру. Это достигается простым видоизменением итераций. Параметр τ выбирается своим на каждой итерации:

$$u^{i+1} = u^i + \tau_{i+1}\,(\Delta u^i - f).$$

Для погрешности имеем соотношение $v^{i+1} = (E + \tau_{i+1}D)v^i$; следовательно, $v^i = \prod_{j=1}^{i} (E + \tau_j D)v^0$. В фурье-представлении:

$$v^i = \sum c_{p,\,q} \prod_{j=1}^{i} (1 - \tau_j \lambda^{p,\,q})\,z^{p,\,q}.$$

Стремясь получить наиболее эффективный процесс, приходим к следующей характерной задаче: найти

$$g = \min_{\tau_1, \ldots, \tau_i} \left\{ \max_{l \leqslant \lambda \leqslant L} \prod_{j=1}^{i} |1 - \tau_j \lambda| \right\}.$$

Это есть классическая задача о полиноме, наименее уклоняющемся от нуля на интервале $[l, L]$ (нормировка: полином равен единице при $\lambda = 0$). Она нам уже известна (см. § 3), известно и ее решение. Параметры τ_j^i — величины, обратные значениям корней полинома Чебышева степени i:

$$\tau_j^i = \left[\frac{L+l}{2} + \frac{L-l}{2} \cos\left(\frac{2j-1}{2i}\,\pi \right) \right]^{-1}, \quad j = 1, 2, \ldots, i.$$

Полиномы Чебышева хорошо изучены, поэтому можно оценить среднюю эффективность одной итерации. Опуская выкладки, полу-

чаем следующую оценку (напомним, что $\eta = L/l$):

$$\|v^i\| \approx (1 - 2/\sqrt{\eta})^i \|v^0\|, \quad \text{т.е. } g \approx 1 - 2\sqrt{l/L},$$

что существенно лучше метода простой итерации. Для числа итераций имеем $i(\varepsilon) \approx 0.5\sqrt{\eta}\ln(1/\varepsilon)$. При $N = 100$, $\varepsilon = 10^{-5}$ значения $g \approx 0.968$, $i(\varepsilon) \approx 360$. Однако попытки применения метода чебышевского ускорения привели к парадоксальному результату: не было не только ускорения сходимости, пропала даже та медленная сходимость, которая была в методе простой итерации! Метод стал расходящимся: даже при умеренных $i \approx 20 \div 30$ величины u^i выходили в «машинную бесконечность». Ниже мы проведем анализ причин этого неприятного явления. Он позволил разработать алгоритмически несложную модификацию метода, работающую так, как предсказывает теория.

Анализ вычислительной устойчивости. Устойчивая форма алгоритма. Причиной неустойчивости является наличие погрешностей округления в расчетах и некоторые свойства полинома Чебышева. Посмотрим, какие последствия имеют погрешности округления. Дело в том, что в действительности итерация имеет вид

$$v^{j+1} = (E + \tau_{j+1}D)v^j + \varepsilon^{j+1},$$

где $\varepsilon^{j+1}_{k,m}$ — погрешность округления, «случайная» сеточная функция, которая имеет порядок $\varepsilon\,|u^{j+1}_{k,m}|$. Это есть очевидное следствие того, что любая величина a в машинном представлении становится величиной $a^{\text{м}} = a(1 + \varepsilon)$, где ε — «случайная» величина порядка 10^{-12} или 10^{-7} в зависимости от длины мантиссы в машинном представлении чисел.

Ради простоты предположим, что все вычисления в итерационном процессе делаются точно и только на k-й итерации вносится погрешность:

$$v^k = \prod_{j=1}^{k} (E + \tau_j^i D)v^0 + \varepsilon, \quad \|\varepsilon\| \approx 10^{-p}\|u^k\|.$$

После полного цикла из i итераций

$$v^i = \prod_{j=k+1}^{i} (E + \tau_j^i D)v^k = \prod_{j=1}^{i} (E + \tau_j^i D)v^0 + \prod_{j=k+1}^{i} (E + \tau_j^i D)\varepsilon.$$

Первое слагаемое оценивается так, как это делалось выше, а вот второе, порожденное малой погрешностью, оказывается очень неприятным.

Чтобы понять, в чем тут дело, рассмотрим представление полинома Чебышева в виде произведения двух «частичных» полиномов:

$$T_i(\lambda) = P_k(\lambda)\, Q_k(\lambda),$$

$$P_k(\lambda) = \prod_{j=1}^{k} (1 - \tau_j^i \lambda), \qquad Q_k(\lambda) = \prod_{j=k+1}^{i} (1 - \tau_j^i \lambda).$$

На рис. 15 изображены графики этих полиномов на интервале $[l, L]$, содержащем спектр оператора D. Полином $P_k(\lambda)$ — очень малая величина в левой части спектра и очень большая в правой, причем не просто очень большая, а очень-очень большая, и

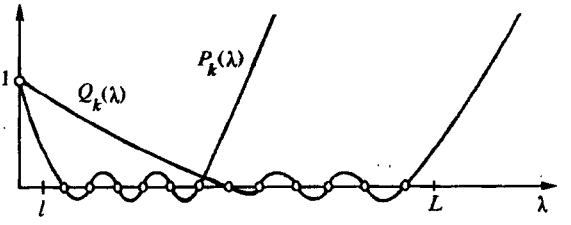


Рис. 15

в этом все дело. Полином $Q_k(\lambda)$ — величина порядка единицы в левой части спектра и очень-очень маленькая в правой. В целом произведение $T_i = P_k\, Q_k(\lambda)$ — малая величина на всем интервале $[l, L]$.

Что же такое v^k (k-е приближение в чебышевском итерационном процессе)? Это есть

$$v^k = \sum\nolimits_1 c_{p,q} P_k(\lambda^{p,q}) z^{p,q} + \sum\nolimits_2 c_{p,q} P_k(\lambda^{p,q}) z^{p,q} + \varepsilon,$$

где первая сумма (по $\lambda^{p,q}$ в левой части $[l, L]$) очень мала, вторая же сумма (по $\lambda^{p,q}$ в правой части $[l, L]$), напротив, очень велика. Таким образом, $\|u^k\| \approx \|v^k\| \gg \|v^0\|$, но функция v^k имеет специальный спектральный состав: последующие итерации $k+1$, $k+2$, ..., i ее погасят.

Теперь оценим погрешность ε. Она имеет порядок $10^{-p}\|u^k\|$ и оказывается по модулю достаточно большой. Однако погрешность округления — величина практически случайная, она более или менее равномерно распределена по всему спектру. Другими словами, $\varepsilon = \sum e_{p,q} z^{p,q}$, и все коэффициенты Фурье $e_{p,q}$ — величины одного порядка.

Проделаем оставшиеся итерации:

$$v^i = \prod_{j=k+1}^{i} (E + \tau_j^i D) v^k = Q_k(D) v^k.$$

В пространстве коэффициентов Фурье получаем

$$v^i = \sum\nolimits_1 c_{p,q} Q_k(\lambda^{p,q}) P_k(\lambda^{p,q}) z^{p,q} + \sum\nolimits_2 c_{p,q} Q_k(\lambda^{p,q}) P_k(\lambda^{p,q}) z^{p,q} + \\ + \sum\nolimits_1 e_{p,q} Q_k(\lambda^{p,q}) z^{p,q} + \sum\nolimits_2 e_{p,q} Q_k(\lambda^{p,q}) z^{p,q}.$$

Первые два слагаемых малы, так как $Q_k(\lambda)P_k(\lambda) = T_i(\lambda)$ — малая величина на всем интервале $[l, L]$, четвертое слагаемое тоже мало, так как полином $Q_k(\lambda)$ мал на правой части $[l, L]$, а вот третье слагаемое не мало, так как полином $Q_k(\lambda) \approx 1$ на левой части $[l, L]$.

Таков качественный механизм неустойчивости метода чебышевских итераций. Поняв его, можно понять и то, что́ нужно изменить, чтобы метод стал устойчивым. Достаточно очевидно, что нужно перебирать параметры $\tau_1 < \tau_2 < \ldots < \tau_i$ не в их естественном порядке (и не в обратном), а как-то «в разбивку», с тем чтобы частичные произведения

$$P_k(\lambda) = \prod_{j=1}^{k} (1 - \tau_{n(j)}\lambda), \qquad Q_k(\lambda) = \prod_{j=k+1}^{i} (1 - \tau_{n(j)}\lambda)$$

были при любом k более или менее равномерно ограниченными на всем интервале $[l, L]$. Здесь последовательность параметров $\tau_{n(j)}$ — та же самая, но как-то переставленная.

Но это наводящие соображения, а точная постановка задачи и ее решение достаточно сложны. Тем не менее задача решена: рациональные перестановки итерационных параметров, приводящие к устойчивому итерационному процессу, были получены почти одновременно В. И. Лебедевым и В. Н. Финогеновым, а также А. А. Самарским и Е. С. Николаевым. Рецепт построения этих устойчивых перестановок достаточно прост в случае $i = 2^r$. Нужная перестановка получается рекуррентно.

Пусть имеется перестановка для $i = 2^r$:

$$\{n(j)\}_{j=1, 2, \ldots, 2^r}.$$

Тогда перестановка для $i' = 2^{r+1}$ получается заменой каждого $n(j)$ парой $n(j)$, $2^{r+1} + 1 - n(j)$. Этот рецепт дает

$$r = 1, \ i = 2: \quad n = \{1, 2\}$$
$$r = 2, \ i = 4: \quad n = \{1, 4, 2, 3\},$$
$$r = 3, \ i = 8: \quad n = \{1, 8, 4, 5, 2, 7, 3, 6\}$$

и т.д. Если теперь $1/\tau_j^i$ $(j=1, 2, \ldots, i=2^r)$ — корни полинома Чебышева на $[l, L]$ степени i, расположенные в естественном порядке: $\tau_1 > \tau_2 > \ldots$, а $\{n(j)\}$ $(j=1, 2, \ldots, i)$ — «устойчивая» перестановка, то итерационный процесс

$$u^{j+1} = u^j + \tau^i_{n(j+1)}(Du^j - f), \quad j=0, 1, \ldots, i-1,$$

сходится в соответствии с теорией, не учитывающей погрешностей округления. Процесс устойчив, и погрешности округления не оказывают существенного влияния.

При вышеизложенном порядке использования итерационных параметров получаем процесс со следующими характеристиками:

$$\|v^i\| \approx \|v^0\| \exp(-2i\sqrt{l/L}).$$

Такая же формула имеет место и для невязки:

$$\|r^i\| \approx \|r^0\| \exp(-2i\sqrt{l/L}).$$

В нашем случае имеем $L=8N^2$, $l=2\pi^2$, т.е.

$$\exp(-2i\sqrt{l/L}) = \exp(-i\pi/N)$$

и, например, за N итераций погрешность убывает в 20 раз. Это не так уж быстро, но в методе простой итерации за то же время погрешность умножится лишь на 0.95 (при $N=100$).

Отметим, что метод простой итерации и чебышевское ускорение имеют широкую сферу применения при решении систем линейных уравнений вида $Au=f$. Изложенное выше основано на таких свойствах оператора A:

а) самосопряженность: $A^*=A$ (отсюда следует вещественность спектра A и ортонормированность базиса собственных векторов);

б) положительность спектра A: $0 < l \leqslant \lambda_k \leqslant L$;

в) для организации расчетов нужно иметь только оценки для границ спектра l, L (снизу для l, сверху для L).

В заключение приведем два полезных замечания.

Замечание 1. Обычно ход итерационного процесса контролируют, выводя на экран PC, например, норму невязки (она вычисляется по ходу работы алгоритма, так как основная расчетная формула имеет вид $u^{j+1}=u^j+\tau r^j$). В методе простой итерации $\|r^j\|$ монотонно убывает. При чебышевском ускорении $\|r^j\|$ меняется немонотонно: она убывает только за полный цикл из i итераций. На промежуточных итерациях возможен сильный рост $\|r^j\|$. Так как $r=D^{-1}v$, то $\|r\| \leqslant \|v\|/\lambda^{1,1}$.

Замечание 2. Выбор длины цикла i (степени полинома Чебышева) не совсем прост. Можно, задав погрешность ε, рассчитать i,

используя приведенные выше оценки. Анализ показывает, что более эффективным является другой способ: использовать полиномы степени $i \approx \sqrt{L/l} \sim N$, повторяя цикл итераций длиной $\sqrt{L/l}$ до получения нужной точности.

Метод переменных направлений. Эффект чебышевского ускорения недостаточен. Имеются и более быстрые итерационные методы. В частности, большие успехи были получены на основе метода переменных направлений, в котором используется так называемый *принцип установления*. Решение стационарного уравнения $\Delta u = f$, $u|_{\partial\Omega} = \varphi$ является пределом при $t \to \infty$ решения $u(t, x)$ уравнения теплопроводности $u_t = \Delta u - f$. Метод простой итерации, как нетрудно заметить, есть просто решение этого уравнения по явной схеме, а условие $\tau_0 = 2/(L + l) \approx 2/8N^2 \approx 0.25h^2$ похоже на условие Куранта. Таким малым шагом трудно получить достаточно большие t, отсюда и большое число итераций.

Как известно, метод переменных направлений, допускает счет с произвольным шагом τ. Кажется, можно получить сколь угодно большие t за один шаг! К сожалению, дело не так просто, так как при этом теряется аппроксимация. Тем не менее метод переменных направлений при достаточно аккуратном его оформлении действительно приводит к существенно более эффективным итерационным процессам. (Полезно оценить оптимальный эффект в процессе, использующем схему «ромб»; см. § 12.)

Одна стандартная итерация (переход u^i в u^{i+1}) метода переменных направлений состоит из двух «полуитераций».

1. Сначала по известной u^i находится промежуточная функция u^* из уравнения

$$\frac{u^* - u^i}{\tau} = \frac{\partial^2 u^*}{\partial x^2} + \frac{\partial^2 u^i}{\partial y^2} - f, \quad (k, m) \text{ внутри}$$

$$u^* = u^i = \varphi, \quad (k, m) \text{ на границе}$$

Функция u^* находится серией раздельных прогонок по горизонтальным линиям сетки, и это требует $O(N^2)$ операций.

2. Затем по известной u^* находится функция u^{i+1} из уравнения

$$\frac{u^{i+1} - u^*}{\tau} = \frac{\partial^2 u^*}{\partial x^2} + \frac{\partial^2 u^{i+1}}{\partial y^2} - f, \quad (k, m) \text{ внутри},$$

$$u^{i+1} = u^* = \varphi, \quad (k, m) \text{ на границе}$$

(серией прогонок по вертикальным линиям сетки за $O(N^2)$ операций).

Анализ сходимости. Здесь точный результат дает спектральный метод исследования. Уравнения для погрешности получаются известным способом — вычитанием из формул итераций тождества для решения $u_{k,m}$ (системы разностных уравнений):

$$\frac{u-u}{\tau}=\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}-f.$$

Очевидно,

$$\frac{v^*-v^i}{\tau}=\frac{\partial^2 v^*}{\partial x^2}+\frac{\partial^2 v^i}{\partial y^2}, \qquad \frac{v^{i+1}-v^*}{\tau}=\frac{\partial^2 v^*}{\partial x^2}+\frac{\partial^2 v^{i+1}}{\partial y^2}.$$

Рассмотрим эффект одной итерации в терминах коэффициентов Фурье. Разложим v^i в сумму:

$$v^i=\sum_{p,q} c^i_{p,q}\, z^{p,q}.$$

Легко проверить, что $z^{p,q}$ суть собственные векторы разностных операторов $\partial^2/\partial x^2$, $\partial^2/\partial y^2$, которые действуют в пространстве двумерных сеточных функций, обращающихся в нуль на границе квадрата:

$$\left(\frac{\partial^2 z^{p,q}}{\partial x^2}\right)_{k,m}=-\lambda'_p z^{p,q}_{k,m}, \qquad \left(\frac{\partial^2 z^{p,q}}{\partial y^2}\right)_{k,m}=-\lambda''_q z^{p,q}_{k,m}$$

($0<l\leqslant\lambda'_p,\ \lambda''_q\leqslant L$; здесь $l\approx\pi^2$, $L=4N^2$). Разложим в сумму и v^*:

$$v^*=\sum_{p,q} c^*_{p,q}\, z^{p,q}.$$

Представим связь между v^* и v^i в виде

$$\left(E-\tau\frac{\partial^2}{\partial x^2}\right)v^*=\left(E+\tau\frac{\partial^2}{\partial y^2}\right)v^i.$$

В терминах рядов Фурье имеем

$$\left(E-\tau\frac{\partial^2}{\partial x^2}\right)\sum_{p,q} c^*_{p,q}\, z^{p,q}=\left(E+\tau\frac{\partial^2}{\partial y^2}\right)\sum_{p,q} c^i_{p,q}\, z^{p,q}.$$

Вводя операторы под знак суммы, получаем

$$\sum_{p,q} c^*_{p,q}\,(1+\tau\,\lambda'_p)z^{p,q}=\sum_{p,q} c^i_{p,q}\,(1-\tau\,\lambda''_q)z^{p,q}.$$

В силу единственности разложения функций по базису $\{z^{p,q}\}$ приходим к соотношению для коэффициентов $c^*_{p,q}$:

$$c^*_{p,q}=\frac{1-\tau\,\lambda''_q}{1+\tau\,\lambda'_p}\,c^i_{p,q}.$$

Точно так же анализируется вторая «полуитерация», и для коэффициентов $c^{i+1}_{p,q}$ получается соотношение

$$c^{i+1}_{p,q} = \frac{1-\tau\lambda'_p}{1+\tau\lambda''_q} c^*_{p,q} = \frac{1-\tau\lambda'_p}{1+\tau\lambda''_q}\frac{1-\tau\lambda''_q}{1+\tau\lambda'_p} c^i_{p,q}.$$

Введем функцию

$$g(\tau) = \max_{l\leqslant\lambda\leqslant L}\left|\frac{1-\tau\lambda}{1+\tau\lambda}\right|.$$

Очевидно, $|c^{i+1}_{p,q}| \leqslant g^2(\tau)\,|c^i_{p,q}|$, $\forall\, p, q$; следовательно, $\|v^{i+1}\| \leqslant g^2(\tau)\|v^i\|$, и наконец, $\|v^i\| \leqslant g^{2i}\|v^0\|$. Теперь осталось найти наилучшее значение для параметра τ, т.е. $\min\limits_{\tau} g(\tau)$.

Итак, нужно решить задачу: найти

$$\min_{\tau}\left\{\max_{l\leqslant\lambda\leqslant L}\left|\frac{1-\tau\lambda}{1+\tau\lambda}\right|\right\}.$$

Задача решается по знакомой схеме. Соотношение

$$\max_{l\leqslant\lambda\leqslant L}\left|\frac{1-\tau\lambda}{1+\tau\lambda}\right| = \max\left\{\left|\frac{1-\tau l}{1+\tau l}\right|, \left|\frac{1-\tau L}{1+\tau L}\right|\right\}$$

проверяется простым анализом графика при разных τ. Построим график функции

$$g(\tau) = \max\left\{\left|\frac{1-\tau l}{1+\tau l}\right|, \left|\frac{1-\tau L}{1+\tau L}\right|\right\}.$$

Минимум $g(\tau)$ достигается в ситуации

$$\frac{1-\tau l}{1+\tau l} = -\frac{1-\tau L}{1+\tau L} \Rightarrow 1-\tau l+\tau L-\tau^2 lL =$$

$$= \tau^2 lL - 1 - \tau l + \tau L \Rightarrow 2\tau^2 lL = 2.$$

Оптимальное значение итерационного параметра $\tau_{\text{опт}} = 1/\sqrt{Ll}$.

Вычислим теперь $g_{\text{опт}}$:

$$g_{\text{опт}} = g(\tau_{\text{опт}}) = \frac{1-l/\sqrt{Ll}}{1+l/\sqrt{Ll}} \approx 1 - 2\sqrt{l/L},$$

т.е. за одну сдвоенную итерацию погрешность уменьшается в $g^2_{\text{опт}} \approx 1 - 4\sqrt{l/L}$ раз.

Таким образом, эффективность метода переменных направлений при едином оптимальном значении параметра оказывается примерно такой же, как при чебышевском ускорении: для получения $\|v^i\| \leqslant \varepsilon\|v^0\|$ требуется $i(\varepsilon) \approx \frac{1}{4\sqrt{l/L}} \ln \varepsilon^{-1}$ итераций.

Метод переменных направлений с серией параметров. Естественно, возникает идея перейти от одного параметра $\tau_{\text{опт}}$ к серии итераций со своим значением τ_j на каждой. Действительно, эта идея оказывается весьма плодотворной. Очевидным образом обобщая проделанные выше выкладки, получаем соотношение между коэффициентами Фурье v^i и v^0:

$$c^i_{p,q} = \prod_{j=1}^{i} \frac{1-\tau_j\lambda'_p}{1+\tau_j\lambda'_p}\,\frac{1-\tau_j\lambda''_q}{1+\tau_j\lambda''_q}\,c^0_{p,q}.$$

Выбор «оптимальной» серии $\tau^i_1, \tau^i_2, \ldots, \tau^i_i$ приводит к минимаксной задаче

$$\min_{\tau}\left\{\max_{l\le\lambda\le L}\prod_{j=1}^{i}\left|\frac{1-\tau_j\lambda}{1+\tau_j\lambda}\right|^2\right\}.$$

Это уже достаточно сложная задача. Она была решена в 1960 г. Е. Вашпрессом, однако в дальнейшем было обнаружено, что еще в начале века решение было получено по другому поводу Е. И. Золотаревым. Тем не менее оптимальные итерационные параметры называют параметрами Вашпресса.

Мы не будем здесь излагать точное решение задачи и следующий из него алгоритм расчета оптимальной серии параметров. Существует достаточно простой рецепт выбора параметров, дающий эффект, близкий к оптимальному. Эта конструкция хорошо иллюстрирует характерную в таких вопросах идею «равномерного подавления компонент погрешности». Имеется в виду, что каждая итерация со своим значением τ эффективно гасит свою часть фурье-разложения погрешности; итерация эффективна на своей части спектра. А в совокупности полный набор параметров обеспечивает погашение всей погрешности. Та же идея, очевидно, лежит и в основе метода чебышевского ускорения простых итераций.

Введем функцию $g(\xi) = (1-\xi)^2/(1+\xi)^2$ и переформулируем минимаксную задачу:

$$\min_{\tau_1,\ldots,\tau_i}\left\{\max_{l\le\lambda\le L}\prod_{j=1}^{i} g(\tau_j\lambda)\right\}.$$

Если приближенное решение задачи даст оценку

$$\prod_{j=1}^{i} g(\tau_j\lambda) \le g_i \quad \text{для всех } \lambda\in[l, L],$$

мы получим $\|v^i\| \le g_i\|v^0\|$, а «средняя эффективность» одной итерации будет, очевидно, $(g_i)^{1/i}$.

Выберем некоторое $\theta < 1$ и выделим интервал, на котором $g(\xi) \leqslant \theta$. Его границы обозначим через $\Lambda(\theta)$ и $\Pi(\theta)$. Итак, $g(\xi) \leqslant \theta$ при $\xi \in [\Lambda(\theta), \Pi(\theta)]$ и $g(\xi) < 1$ на остальной части положительной полуоси. Параметр τ_1 выберем так, чтобы левая граница θ-интервала функции $g(\tau_1\lambda)$ совпала с l: $\tau_1 l = \Lambda(\theta)$, или $\tau_1 = \Lambda(\theta)/l$. Тогда правая граница θ-интервала функции $g(\tau_1\lambda)$ определяется соотношением $\tau_1\lambda = \Pi(\theta)$, т.е. $\lambda = \Pi(\theta)/\tau_1 = l\,\Pi(\theta)/\Lambda(\theta)$.

Утверждение 2. Пусть τ_1 выбрано так, как указано выше, а остальные $\tau_2, \ldots, \tau_i$ произвольные (положительные). Тогда

$$\prod_{j=1}^{i} g(\tau_j\lambda) \leqslant \theta \quad \text{при} \quad l \leqslant \lambda \leqslant lr, \quad \text{где} \quad r = \Pi(\theta)/\Lambda(\theta) > 1.$$

Для доказательства достаточно заметить, что все множители $g(\tau_2\lambda)$, ... не превосходят единицы. Итак, выделен тот «участок спектра», за который отвечает параметр τ_1. Выберем τ_2 так, чтобы левая граница θ-интервала для $g(\tau_2\lambda)$ совпала с правой для $g(\tau_1\lambda)$: $\tau_2 lr = \Lambda(\theta)$, или $\tau_2 = \Lambda(\theta)/lr = \tau_1/r$. Тогда правая граница θ-интервала функции $g(\tau_2\lambda)$ определится соотношением $\tau_2\lambda = \Pi(\theta)$, т.е. $\lambda = \Pi(\theta)/\tau_2 = lr^2$.

Утверждение 3. Пусть τ_1 и τ_2 выбраны так, как указано выше, остальные $\tau_3, \ldots, \tau_i$ произвольные (положительные). Тогда

$$\prod_{j=1}^{i} g(\tau_j\lambda) \leqslant \theta \quad \text{при} \quad l \leqslant \lambda \leqslant lr^2.$$

Продолжая строить последовательность примыкающих друг к другу θ-интервалов для функций $g(\tau_3\lambda)$, ..., получаем, очевидно, последовательность параметров:

$$\tau_1 = l/r, \quad \tau_2 = \tau_1/r, \quad \ldots, \quad \tau_{j+1} = \tau_j/r = \tau_1/r^j.$$

Так продолжаем до тех пор, пока очередная правая граница θ-интервала не выйдет за пределы правой границы спектра L, т.е. в качестве i следует взять наименьшее целое, при котором $lr^i \geqslant L$, т.е.

$$i(\theta) = \frac{\ln (L/l)}{\ln r} + 1.$$

Теорема 2. Проделав $i(\theta)$ итераций метода переменных направлений с указанным выше выбором параметров $\tau_1, \tau_2, \ldots, \tau_i$, получим оценку для погрешности

$$\|v^i\| \leqslant \theta \|v^0\|.$$

Для достижения нужной погрешности ε мы имеем два пути: либо сразу назначить $\theta = \varepsilon$, либо использовать найденную последовательность циклически и за $ki(\theta)$ итераций получить в оценке множитель $\theta^k \approx \varepsilon$. Выясним, что же выгоднее, т.е. оптимизируем процесс за счет рационального выбора θ. Задачу решаем, используя «среднюю эффективность» одной итерации, т.е. вводя характеристику

$$\gamma(\theta) = -\ln \theta^{1/i(\theta)} = \frac{1}{i(\theta)} \ln \theta^{-1}.$$

В этих терминах имеем оценку

$$\|v^i\| \leqslant \|v^0\| \, e^{-i\gamma(\theta)}.$$

Конечно, это соотношение не следует понимать буквально, оно выполняется только после каждой серии из $i(\theta)$ итераций. Очевидно, $\gamma(\theta)$ и есть та характеристика итерационного процесса, которую следует сделать максимальной.

Итак, для выбора θ получаем задачу: найти

$$\max_{\theta} \gamma(\theta) = \max_{\theta} \frac{\ln \theta^{-1} \ln [\Pi(\theta)/\Lambda(\theta)]}{\ln (L/l)} =$$

$$= \frac{1}{\ln (L/l)} \max_{\theta} \{\ln \theta^{-1} \ln [\Pi(\theta)/\Lambda(\theta)]\}.$$

Обратим внимание на то, что наилучшее значение θ определяется независимо от границ спектра l, L, т.е. один раз для всех задач. Вычислив $\theta_{\text{опт}}$, найдем универсальные характеристики:

$$r_0 = \Lambda_0/\Pi_0, \quad \gamma_0 = \ln \theta_0^{-1} \ln r_0^{-1}, \quad i_0 = 1/\ln r_0.$$

В конкретной задаче, имея оценки границ спектра l, L, рассчитываем длину итерационного цикла $i = i_0 \ln (L/l) + 1$, среднюю эффективность итерации $\gamma = \gamma_0/\ln (L/l)$ и набор итерационных параметров $\tau_1 = \Lambda_0/l$, $\tau_2 = \tau_1/r_0$, $\tau_3 = \tau_2/r_0$, ... Для уменьшения нормы погрешности в ε^{-1} раз в конкретной задаче потребуется

$$i(\varepsilon, l/L) \approx \gamma_0^{-1} \ln (L/l) \ln \varepsilon^{-1}, \qquad \gamma_0 \approx 3.2,$$

итераций. Значение $\theta_{\text{опт}}$ находится по табл. 10, из которой видно, что $\theta_{\text{опт}} \approx 0.16 \div 0.2$ (бóльшая точность, очевидно, здесь не нужна) и $i_0 = 5 \div 4$ соответственно. Значения Π_0, Λ_0, r_0 предоставим вычислить читателю. Для задачи на сетке 100×100 ($l = \pi^2$, $L = 4N^2$) получаем убывание норм погрешности и невязки в процессе итераций со скоростью

$$\|v^i\| \approx \|v^0\| e^{-0.38i}, \qquad \|r^i\| = \|r^0\| e^{-0.38i}.$$

Таким образом, для того чтобы уменьшить погрешность начального приближения в 10^5 раз, потребуется около 30 итераций метода переменных направлений. Важно отметить, что, хотя эти формулы

Таблица 10

θ	0.01	0.04	0.09	0.125	0.160	0.200	0.250	0.360
$i(\theta)$	22	9	7	6	5	4	4	3
$\gamma(\theta)$	0.21	0.35	0.36	0.375	0.384	0.382	0.375	0.340

нужно понимать «в среднем», в данном случае убывание норм погрешности и невязки происходит монотонно: $\|v^{i+1}\| < \|v^i\|$ при любом порядке использования итерационных параметров. Проблемы устойчивости здесь, в отличие от метода чебышевского ускорения, не возникает.

Величины $\|v^0\|$ и $\|r^0\|$, фигурирующие в формулах, легко оцениваются при самом простом выборе начального приближения:

$$u^0_{k,m} = 0 \quad \text{внутри}, \qquad u^0_{k,m} = \varphi_{k,m} \quad \text{на границе}.$$

В этом случае

$$\|v^0\| \approx \|u\|, \qquad \|r^0\| \approx \|f\| + \|\varphi\|/h^{3/2}$$

(где u — решение разностной задачи, $\|\varphi\| \approx (\oint \varphi^2\, ds)^{1/2}$), и содержательный смысл достигнутой в расчете точности $\|v^\nu\| = \|u^\nu - u\| \leqslant \leqslant 10^{-5}\|v^0\| \approx 10^{-5}\|u\|$ очевиден. Если же в этом примере использовать оптимальные параметры Вашпресса, результат будет такой: за 16 итераций получается оценка $\|v^{16}\| \leqslant 0.3\cdot 10^{-6}\|v^0\|$. В формуле эффективного убывания погрешности множитель 3.2 меняется на 9, т.е. оптимальный вариант примерно в 2.8 раза эффективнее упрощенного.

Обсудим вопрос о возможности применения метода переменных направлений в более общей ситуации. Проанализировав весь ход рассуждений, легко убедимся в том, что были использованы следующие факторы.

1. Уравнение $Du = f$ имеет форму

$$D_1 u + D_2 u = f.$$

2. Уравнения «на верхнем слое» для схем суть

$$\frac{u - u^*}{\tau} = D_i u + q,$$

где u^*, q известны; они легко (т.е. «экономно») разрешаются относительно u.

3. Операторы D_i самосопряженные с положительным (отрицательным) спектром. Для границ спектра есть достаточно эффективные оценки l_i, L_i.

4. Операторы (разностные) D_1, D_2 имеют общую систему собственных функций, что, как известно, эквивалентно их перестановочности: $D_1D_2 = D_2D_1$. Этот важный факт существенно определяет возможность обобщения приведенной выше теории.

Если перевести приведенные формальные признаки на содержательный язык применительно к эллиптическим уравнениям второго порядка, то мы получим следующий класс уравнений, простейшие разностные аппроксимации которых имеют нужные свойства:

а) уравнение

$$\frac{\partial}{\partial x}\left[a_1(x)\frac{\partial u}{\partial x}\right] + b_1(x)\,u + \frac{\partial}{\partial y}\left[a_2(y)\frac{\partial u}{\partial y}\right] + b_2(y)\,u = f(x, y),$$

б) область — прямоугольник,

в) достаточно общие краевые условия: $\alpha\frac{\partial u}{\partial n} + \beta u = \varphi$; здесь n — внешняя нормаль, α, β — постоянные, свои на каждой стороне прямоугольника.

В сущности это — случай, когда работает разделение переменных. Мы не обсуждаем известных в теории эллиптических задач условий на a_i, b_i, обеспечивающих знакоопределенность операторов

$$D_1 = \frac{\partial}{\partial x}\left(a_1\frac{\partial}{\partial x}\right) + b_1, \qquad D_2 = \frac{\partial}{\partial y}\left(a_2\frac{\partial}{\partial y}\right) + b_2.$$

Мы применили грубую теорию, в которой границы спектров D_1, D_2 взяты в виде $l = \min(l_1, l_2)$, $L = \max(L_1, L_2)$. В точной теории Золотарева—Вашпресса используются свои границы спектров для D_1, D_2.

Первое собственное число разностного оператора часто уже при не очень больших N почти совпадает с соответствующим собственным числом аппроксимируемого дифференциального оператора. Для его приближенного вычисления можно привлечь весь арсенал аналитических оценок. В частности, можно заменить переменные коэффициенты дифференциального оператора на средние значения и вычислить аналитически первое собственное число оператора с постоянными коэффициентами.

При оценке верхней границы L используется другое соображение. Известно, что для любой матрицы $\{a_{i,j}\}$ оценкой любого собственного числа сверху является $\max\limits_i \sum\limits_j |a_{i,j}|$. В нашем случае в каждом узле схемы следует просуммировать модули коэффициентов разностной схемы и взять наибольшее (по узлам сетки) значение

такой суммы. Для оператора Лапласа, аппроксимированного по пятиточечной схеме, получим значение $L = 8/h^2$, почти не отличающееся от точного: $L = (8/h^2)\cos[\pi(N-1)/2N]$.

Величину L можно оценить и снизу, используя известное соотношение Рэлея для самосопряженных матриц D (операторов). Для любого собственного числа λ имеем

$$\min_u \frac{(Du, u)}{(u, u)} \leqslant \lambda \leqslant \max_u \frac{(Du, u)}{(u, u)}.$$

Эти соотношения часто применяют для оценок границ спектра. Они эффективно работают в том случае, когда имеется априорная информация о том, какую форму имеют собственные функции, соответствующие крайним точкам спектра. В частности, максимальному (по модулю) собственному числу разностной аппроксимации оператора Δ (и других аналогичных операторов) соответствует сеточная функция типа $u_{k,m} = (-1)^{k+m}$. Для оценки можно взять даже функцию, равную -1 в одном узле сетки и $+1$ в четырех соседних узлах, в соответствии с шаблоном пятиточечной схемы. Предоставим читателю проверить, насколько близко к верхней оценке $L = 8/h^2$ будет отношение Рэлея на такой пробной функции.

Как было указано выше, теория выбора итерационных параметров и оценка эффективности работают лишь в случае разделения переменных. Однако сама схема формально применима и в более общей ситуации, например при уравнении вида

$$\frac{\partial}{\partial x}\left[a_1(x, y)\frac{\partial u}{\partial x}\right] + \frac{\partial}{\partial y}\left[a_2(x, y)\frac{\partial u}{\partial y}\right] + c(x, y)\,u = f(x, y).$$

Для краевых условий первого рода (задано u на границе области) форма области более или менее безразлична: простейшие разностные аппроксимации операторов D_1, D_2 включают точки только одной горизонтали (вертикали) и краевые условия этого обстоятельства не нарушают. Если область — прямоугольник (или объединение прямоугольников), допустимы краевые условия третьего рода: в них входит нормальная производная, а при ее аппроксимации используются точки той же горизонтали (вертикали) сетки (если опустить тонкую проблему аппроксимации в угловой точке границы).

Если граница области не проходит по координатной линии сетки, нормальная производная аппроксимируется по шаблону, захватывающему как минимум две соседние горизонтали (вертикали) сетки. Такие краевые условия препятствуют непосредственному расщеплению уравнений «на верхнем слое» метода переменных направлений и прямое обобщение вычислительной схемы не проходит. Что касается выбора параметров, то они, естественно, рассчитываются для системы с «замороженными» коэффициентами (в качестве таковых берут, например, средние значения) и для минимального прямо-

угольника, содержащего данную область. Опыт показывает, что такой способ часто приводит к успеху (к быстрой сходимости итераций), особенно когда коэффициенты уравнения изменяются не очень сильно. Однако при появлении в уравнении смешанной производной u_{xy} отказывает не только теория сходимости, но и алгоритмическая схема. Что делать в этом случае?

Общая схема итерационных процессов. Пусть D — разностная аппроксимация общего эллиптического самосопряженного дифференциального оператора. И пусть разностная аппроксимация выбрана так, что $D = D^*$ (сохраняется самосопряженность). Нужно решить уравнение $Du = f$.

Многие итерационные методы решения этого уравнения укладываются в общую схему:

$$B\frac{u^{i+1}-u^i}{\tau} = Du^i - f,$$

где B — некоторый разностный оператор, называемый (по терминологии А. А. Самарского) *регуляризатором*. Он должен обладать следующими свойствами:

1) $B = B^*$ (самосопряженность);

2) $B > 0$, т.е. $(Bu, u) \geqslant \gamma(u, u)$ для всех u;

3) (очень важное свойство) оператор B должен быть легко обратимым, т.е. задача $Bv = z$ легко решается; мы имеем алгоритм, позволяющий сравнительно «дешево» определить v из этого уравнения («дешево» по сравнению с «ценой» исходной задачи $Du = f$);

4) (очень важное свойство) оператор B должен быть «энергетически эквивалентен» оператору $-D$ в смысле неравенств

$$\gamma_1(Bu, u) \leqslant (-Du, u) \leqslant \gamma_2(Bu, u), \quad \forall\, u.$$

Положительные постоянные $0 < \gamma_1 < \gamma_2$ называются *константами эквивалентности* B и $-D$; будем считать их известными.

Итерационный процесс фактически реализуется так:

1) вычисляем $r^i = Du^i - f$ (u^i — известно);

2) решаем уравнение $Bv^i = r^i$;

3) вычисляем $u^{i+1} = u^i + \tau v^i$.

Формально процесс можно записать в виде

$$u^{i+1} = u^i + \tau B^{-1}Du^i - \tau B^{-1}f.$$

Алгоритм напоминает метод простой итерации с оператором $B^{-1}D$. Если B и D неперестановочны, из самосопряженности B и D не следует самосопряженность $B^{-1}D$. Однако это легко исправить. Так как

$B = B^*$ и $B > 0$, существуют операторы $B^{1/2}$ и $B^{-1/2}$. Сделаем замену переменных $w = B^{1/2}u$ и умножим формулу итерации на $B^{1/2}$:

$$B^{1/2}u^{i+1} = B^{1/2}u^i + \tau B^{1/2}B^{-1}DB^{-1/2}B^{1/2}u^i - \tau B^{1/2}B^{-1}f,$$

т.е. $w^{i+1} = w^i + \tau B^{-1/2}DB^{-1/2}w^i - \tau B^{-1/2}f$.

Обозначим $S = B^{-1/2}DB^{-1/2}$. Легко видеть, что $S^* = S$ и итерация может изучаться в виде $w^{i+1} = w^i + \tau S w^i + \tilde{f}$. Из предположения об энергетической эквивалентности можно вывести важный факт: спектр оператора S (в смысле $S\varphi = -\lambda\varphi$) положительный и $\gamma_1 \leqslant \lambda \leqslant \gamma_2$. В самом деле,

$$\gamma_1(Bu, u) \leqslant (-Du, u) \Rightarrow \gamma_1(B^{1/2}u, B^{1/2}u) \leqslant (-Du, u), \quad \forall\, u$$

(так как $B = B^{1/2}B^{1/2}$, $(B^{1/2})^* = B^{1/2}$). Полагая $B^{1/2}u = w$, имеем

$$\gamma_1(w, w) \leqslant (-DB^{-1/2}w, B^{-1/2}w) = (-B^{-1/2}DB^{-1/2}w, w),$$

т.е. $\gamma_1(w, w) \leqslant (-Sw, w)$, $\forall\, w$. Аналогично, из $(-Du, u) \leqslant \gamma_2(Bu, u)$, $\forall\, u$ следует $(-Sw, w) \leqslant \gamma_2(w, w)$. Если $Sw = -\lambda w$, то $(-Sw, w) = \lambda(w, w)$, т.е. $\lambda \in [\gamma_1, \gamma_2]$.

Сведя исследование итерационного процесса с регуляризатором B к исследованию простой итерации с оператором S, можно воспользоваться уже знакомой нам теорией. В частности, если границы γ_1 и γ_2 близки друг к другу (т.е. оператор $B^{-1}D$ «хорошо обусловлен»), то в качестве оптимального итерационного параметра можно взять $\tau = 2/(\gamma_1 + \gamma_2)$, что приведет к сходимости с множителем $(\gamma_2 - \gamma_1)/(\gamma_1 + \gamma_2)$ за одну итерацию. Правда, не следует забывать, что «цена» такой итерации зависит от затрат вычислительной работы на решение уравнения $Bv = z$. Если оператор $B^{-1}D$ «плохо обусловлен» ($\gamma_2/\gamma_1 \gg 1$), то можно использовать чебышевское ускорение и получить процесс, в котором средняя за итерацию эффективность соответствует множителю $(1 - 2\sqrt{\gamma_1/\gamma_2})$.

Применение общей схемы. Конкретные итерационные алгоритмы получаются при конкретном выборе регуляризатора B. Приведем некоторые примеры.

Метод простой итерации. Он получается при $B = E$.

Метод переменных направлений (точнее, некоторые его обобщения). Он получается при выборе в качестве регуляризатора «факторизованной» конструкции

$$B \equiv \left(E - \sigma_1 \frac{\partial^2}{\partial x^2}\right)\left(E - \sigma_2 \frac{\partial^2}{\partial y^2}\right).$$

(Напомним, что мы договорились производные понимать в смысле их простейших разностных аппроксимаций.) Оператор B легко «обращается». Решение уравнения $Bv = z$ сводится к решению последовательности задач:

$$\text{а)}\ \left(E - \sigma_1 \frac{\partial^2}{\partial x^2}\right) v^* = z, \qquad \text{б)}\ \left(E - \sigma_2 \frac{\partial^2}{\partial y^2}\right) v = v^*.$$

Каждая из этих задач расщепляется на серии «одномерных», легко решаемых прогонкой систем уравнений.

Здесь мы сталкиваемся с характерной ситуацией: конструкция оператора B содержит некоторые параметры (σ_1, σ_2 в данном случае). Вместе с параметром τ они должны быть найдены таким образом, чтобы получить возможно более высокую скорость сходимости. Точный анализ сходимости в общем случае провести не удается. Поэтому задача «оптимизации параметров» обычно решается раздельно: сначала за счет выбора параметров регуляризатора стремятся уменьшить обусловленность S — величину γ_2/γ_1, т.е. сблизить оценки в неравенствах энергетической эквивалентности операторов B и $-D$.

Теоретической предпосылкой для существования хороших оценок такого типа является известный факт из теории эллиптических операторов: два любых эллиптических дифференциальных оператора одного порядка энергетически эквивалентны друг другу. Следствием этого является и энергетическая эквивалентность их разностных аппроксимаций с постоянными γ_1, γ_2, не зависящими по существу от шага сетки h. Рассматриваемый здесь факторизованный оператор B является, как легко заметить, аппроксимацией дифференциального оператора четвертого порядка (правда, вырожденного). Дифференциальные операторы разных порядков не могут быть энергетически эквивалентными. Это приводит к существенной зависимости констант γ_i от h: $\gamma_2/\gamma_1 = O(h^{-1})$ при «оптимальном» выборе σ_1, σ_2.

Попеременно-треугольный метод. В качестве регуляризатора используется факторизованный оператор

$$B = R_1 R_2 = \left[E + \sigma\left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\right)\right]\left[E - \sigma\left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\right)\right].$$

Очевидно, $B = E - \sigma^2\left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\right)^2$. Оператор $\left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\right)^2$ имеет второй порядок, но, к сожалению, не является строго эллиптическим. Слагаемое E при подходящем выборе σ придает разностной аппроксимации B «эллиптический» характер.

Оператор B легко обратим: решение уравнения $Bv = z$ требует числа операций, пропорционального числу неизвестных, т.е. числу

узлов сетки. Чтобы убедиться в этом, выпишем подробно разностную аппроксимацию R_1 и R_2:

$$(R_1 v)_{k,m} = v_{k,m} + \sigma\left(\frac{v_{k,m} - v_{k-1,m}}{h} + \frac{v_{k,m} - v_{k,m-1}}{h}\right),$$

$$(R_2 v)_{k,m} = v_{k,m} - \sigma\left(\frac{v_{k+1,m} - v_{k,m}}{h} + \frac{v_{k,m+1} - v_{k,m}}{h}\right).$$

На рис. 16 показана сетка и расположение в ней шаблонов операторов R_1, R_2; такая аппроксимация обеспечивает важное свойство $R_2 = R_1^*$ Из этого рисунка очевиден алгоритм «обращений» R_1, R_2 при известных значениях v на границе. Решение, например, уравнения $R_1 v = z$ осуществляется «маршевым» алгоритмом вычисления слева-направо и снизу-вверх, начиная с левого нижнего угла области. На каждом шаге такого алгоритма в выражении $(R_1 v)_{k,m}$ из трех значений v два ($v_{k-1,m}$ и $v_{k,m-1}$) уже известны и можно вычислить

$$v_{k,m} = \frac{1}{1+2\sigma/h}\left[z_{k,m} + \frac{\sigma}{h}(v_{k-1,m} + v_{k,m-1})\right].$$

Устойчивость этого «марша» легко устанавливается (при $\sigma > 0$).

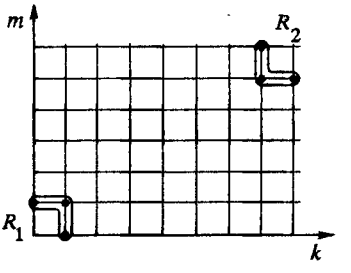


Рис. 16

Обращение R_2 осуществляется аналогично, но в обратном направлении (начиная с правого верхнего угла области). Оптимизация оценок эквивалентности за счет σ приводит к $\gamma_2/\gamma_1 = O(h^{-1})$. Выбирая далее последовательность итерационных параметров τ_i в соответствии с теорией чебышевского ускорения, получаем алгоритм, в котором число итераций, необходимых для уменьшения погрешности начального приближения в ε^{-1} раз, есть (для сетки $N \times N$) $i(\varepsilon) = O(\sqrt{N} \ln \varepsilon^{-1})$. Ограничимся здесь этими общими сведениями, отправляя интересующихся деталями (как оценивать γ_1, γ_2, как выбирать σ и т.п.) к специальной литературе.

«Двухступенчатые» итерационные методы. Почти очевидно, в каком направлении следует искать операторы B, наилучшие с точки зрения оценок энергетической эквивалентности: оператор B должен быть возможно более похожим на $-D$ (идеальный случай: $B = -D$, $\gamma_1 = \gamma_2 = 1$, достаточно одной итерации; к сожалению, она просто эквивалентна решению исходной задачи). Однако по мере сближения B с $-D$ возрастают трудности решения уравнения $Bv = z$.

Удачным компромиссом является, например, выбор

$$B \equiv -\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) = -\Delta.$$

В этом случае γ_2/γ_1 не зависит от h, а для решения разностного уравнения $-\Delta v = z$ можно использовать построенные в последние годы эффективные алгоритмы для решения в прямоугольной области уравнения с постоянными коэффициентами. Ограничимся здесь только названиями методов, тем более что многие из них уже оформлены как стандартные быстро работающие программы математического обеспечения современных ЭВМ: это методы циклической редукции, быстрого преобразования Фурье (см. § 24), маршевый и некоторые другие. Кстати, включение некоторых из них в арсенал средств практических вычислений было связано с анализом вычислительной неустойчивости и разработкой вычислительно устойчивых модификаций (как это было с чебышевским ускорением). Простейшие формы алгоритмов неустойчивы.

Можно сближать B с $-D$, решая уравнение $Bv = z$ подходящим итерационным методом, например методом переменных направлений. При этом, естественно, нет необходимости в очень точном решении уравнения, достаточно ограничиться каким-то числом «внутренних» итераций. Так мы приходим к семейству «двухступенчатых» итерационных алгоритмов. Их оптимизация, в частности выбор наилучшего (с точки зрения эффективности процесса в целом) числа внутренних итераций, связана с достаточно сложным анализом. Такие итерационные процессы и теория их оптимизации построены Е. Г. Дьяконовым.

Многосеточный метод. Опишем конструкцию своеобразного итерационного алгоритма, получившего в последние годы широкое применение по причинам, которые удобно объяснить несколько позже. Метод достаточно сложен алгоритмически, сложен он и для теоретического анализа даже в самом простом случае. Поэтому мы ограничимся самым общим описанием и качественным объяснением механизма, обеспечивающего высокую скорость сходимости итераций. Исходной идеей этой конструкции является следующее замечание. Все собственные функции оператора Δ (разностного) $z_{k,m}^{p,q} = \sin\frac{kp\pi}{N}\sin\frac{qm\pi}{N}$ условно разделим на две части: гладкие ($p < N/2$ и $q < N/2$) и негладкие ($p \geqslant N/2$ или $q \geqslant N/2$) функции, т.е. разделим низкие и высокие гармоники и частоты λ некоторой условной границей.

Легко построить итерационный метод, эффективно гасящий негладкую компоненту погрешности (и невязки). В самом деле, метод простой итерации $u^i = u^i + \tau(\Delta u^i - f)$, как было показано, гасит (p, q)-компоненту погрешности, умножая ее за один шаг на

$1 - \tau\lambda_{p,q}$. Высокие частоты расположены, очевидно, между $\lambda_{1,N/2} = (4/h^2)\sin^2(\pi N/4N) \approx 2/h^2$ и $\lambda_{N-1,N-1} \approx 8/h^2$. Выбирая τ оптимальным для этой части спектра, т.е. $\tau = 2h^2/(2+8) = 0.2h^2$, получаем убывание негладких компонент погрешности (невязки) в процессе итераций со скоростью $(0.6)^l$, т.е. достаточно быстрое.

На остальной части спектра сходимость, конечно, очень медленная. Так, компонента (1, 1) погрешности убывает с показателем $1 - 2\pi^2\tau = 1 - 0.4(\pi/N)^2$ за шаг. В целом итерационный процесс оказывается неэффективным. Однако небольшое число таких итераций «сглаживает» невязку: высокие гармоники в ней гасятся, основную роль играет гладкая компонента.

Таким образом, после i итераций имеем приближение u^i, удовлетворяющее уравнениям

$$(\Delta u^i)_{k,m} - f_{k,m} = r^i_{k,m} \quad \text{(внутри)}$$

$$u^i_{k,m} - \varphi_{k,m} = 0 \qquad \text{(на границе)}$$

(это просто определение невязки r). Если бы мы могли решить уравнения

$$(\Delta w)_{k,m} = r^i_{k,m} \quad \text{(внутри)}$$

$$w_{k,m} = 0 \qquad \text{(на границе)}$$

то функция w была бы поправкой в том смысле, что точное решение $u_{k,m} = u^i_{k,m} - w_{k,m}$. На первый взгляд, найти поправку w — задача такой же степени трудности, как и исходная. Но после i итераций с $\tau = 0.2h^2$ ситуация изменилась: невязка r^i стала гладкой функцией и уравнение для поправки можно решать на другой, более грубой сетке.

Предположим для простоты изложения, что $N = 2^l$, и наряду с основной сеткой с шагом $h = 1/N$ введем вспомогательную сетку с шагом $H = 2h$. Узлы этой сетки совпадают с четными узлами основной сетки (т.е. с теми узлами (k, m), для которых четны оба индекса k и m). На этих сетках мы будем рассматривать близкие по смыслу функции — в этих случаях будем использовать одинаковые буквы для H-сетки и h-сетки (большие и малые соответственно).

Итак, имеем i-е приближение $u^i_{k,m}$ и его невязку $r^i_{k,m}$. Возьмем ограничение невязки на H-сетку, т.е., проще говоря, $R_{k,m} = r_{2k,2m}$ $(k, m = 0, 1, \ldots, N/2)$, и решим на H-сетке уравнение

$$(DW)_{k,m} = R_{k,m}$$

(с нулевыми значениями на границе). Здесь D — аппроксимация оператора Лапласа на H-сетке. Эта задача заметно проще исходной хотя бы потому, что в ней в четыре раза меньше неизвестных и число обусловленности для системы меньше (тоже в четыре раза). Тем не менее решение такой системы не настолько проще решения исходной задачи на h-сетке, чтобы можно было пренебречь проблемой решения вспомогательной задачи. Пока будем считать, что вспомогательная задача так или иначе решена (приближенно, вообще говоря). Интерполируя (линейно по x и y, например) функцию W на узлы основной h-сетки, получаем функцию $w_{k,m}$. Вычитая ее из u^i, приходим к новому приближению $\tilde{u} = u^i - w$.

Что можно сказать о невязке этой функции? Вычислим ее во внутренних узлах:

$$\tilde{r}_{k,m} = (\Delta(u^i - w))_{k,m} - f_{k,m} = r_{k,m} - (\Delta w)_{k,m}.$$

Теперь вопрос в следующем: из того, что $(DW)_{k,m} = r_{2k,2m}$, следует ли (хотя бы приближенно), что $(\Delta w)_{k,m} \approx r^i_{k,m}$? Если да, то проведенная коррекция явно целесообразна. Вычисления (простые, но довольно

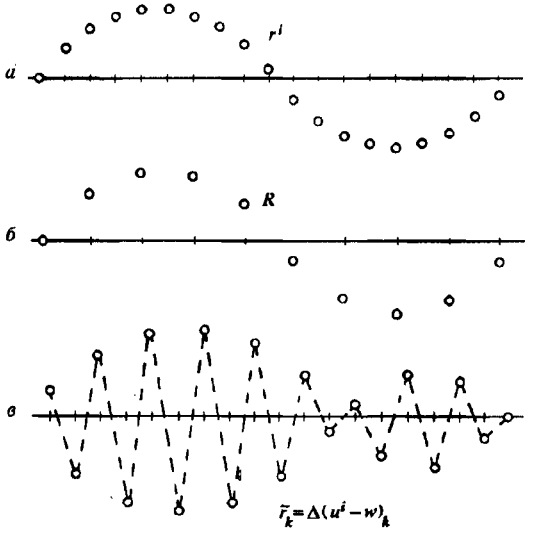


Рис. 17

громоздкие) показывают, что ответ неоднозначен. Точнее, соотношение $\Delta w \approx r^i$ выполняется «в среднем», в слабом смысле слова.

Поясним это на одномерном примере. На рис. 17 показаны этапы процесса: $а$ — гладкая на h-сетке невязка r^i; $б$ — ее ограничение

на H-сетку R; $\tilde{r}$ — функция $\tilde{r}_k = \Delta(u^i - w)_k$. Обратим внимание на характер функции $\tilde{r}$: она в среднем близка к нулю и состоит, в основном, из негладких собственных функций. После этого очередная серия простых итераций с $\tau = 0.2h^2$ эффективно гасит невязку. Причина именно такой структуры $\tilde{r}$ разъясняется ниже.

Итак, основная идея коррекции с помощью вспомогательной сетки состоит в том, что невязка из подпространства гладких сеточных функций «перегоняется» в подпространство негладких сеточных функций, где эффективно работает метод простой итерации.

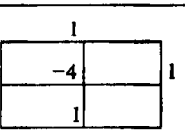
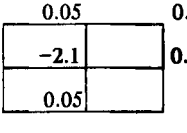
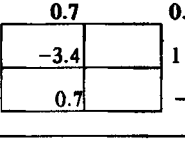
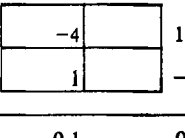
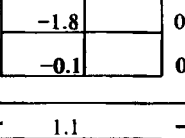
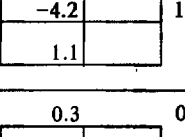
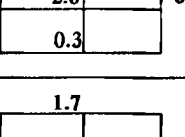
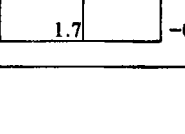
Вернемся к вопросу о том, как решать задачу на вспомогательной H-сетке, ведь она немногим проще исходной. Ответ очевиден: для этого надо использовать свою $2H$-сетку, для нее — $4H$-сетку, и т.д. до тех пор, пока число узлов на очередной вспомогательной сетке не станет совсем уж незначительным. Однако окупит ли эффект ускорения сходимости затраты на вспомогательные вычисления? Ответ на этот вопрос не очевиден, он требует проведения сложных и достаточно аккуратных оценок.

Такие оценки были проделаны и было доказано (сначала для задачи Пуассона в прямоугольнике, затем для гораздо более общих и сложных эллиптических задач), что норма невязки убывает со скоростью, не зависящей от шага сетки, т.е. необходимое для ее уменьшения в ε^{-1} раз число арифметических операций есть $CN^2 \ln \varepsilon^{-1}$ (C не зависит от N). Это асимптотически рекордный результат. При достаточно большом N описанные выше (и в сущности все остальные известные сейчас) методы уступают многосеточному. Однако постоянная C в этой оценке вычисляется настолько грубо, что лучше считать ее неизвестной.

Могло бы оказаться, что преимущество многосеточного метода по сравнению, например, с методом переменных направлений наступает при столь больших N, какие на практике не используются. Такие вопросы выясняются путем вычислительного эксперимента. Многочисленные реализации многосеточного метода и результаты вычислений показали, что метод оказывается эффективным уже на тех сравнительно скромных сетках (N порядка 50, 100), которые давно используются в практических расчетах.

Приведем для иллюстрации табл. 11, в которой показаны результаты вычислительного эксперимента по решению некоторых уравнений многосеточным методом. Основная сетка имела 108×108 узлов, первая вспомогательная ($H = 3h$) — 36×36, вторая — 12×12. В табл. 11 показаны вид аппроксимируемого оператора, шаблон аппроксимации, n — число итераций, затрачиваемых на сглаживание невязки, предшествующее обращению к вспомогательной задаче, и, наконец, средняя скорость убывания невязки с номером итерации i. При этом i — это число итераций на основной сетке 108×108, включая время работы, затрачиваемое на решение вспомогательных

Таблица 11

№	Оператор	Шаблон	n	$\|\|r^i\|\|$
1 2 3	$u_{xx}+u_{yy}$		4 8 10	$e^{-0.38i}$ $e^{-0.42i}$ $e^{-0.38i}$
4	$u_{xx}+1.9u_{xy}+u_{yy}$		10	$e^{-0.18i}$
5	$u_{xx}+1.6u_{xy}+0.7u_{yy}$		10	$e^{-0.12i}$
6	$u_{xx}+1.6u_{xy}+u_{yy}$		10	$e^{-0.18i}$
7 8	$u_{xx}+1.2u_{xy}+0.5u_{yy}$		8 10	$e^{-0.13}$ $e^{-0.125}$
9	$u_{xx}+1.2u_{xy}+0.5u_{yy}$		10	$e^{-0.13i}$
10	$u_{xx}+1.4u_{xy}+0.5u_{yy}$		8	$e^{-0.37i}$
11	$u_{xx}+1.4u_{xy}+u_{yy}$		8	$e^{-0.18i}$

задач (в единицах, равных времени на одну итерацию на основной сетке).

В дальнейшем многосеточный метод был усовершенствован. Усовершенствования касались таких деталей, как способ вычисления невязки R на H-сетке через невязку r на h-сетке, форма основных итераций, методы интерполяции с H- на h-сетку, и некоторых других. Все эти технические усовершенствования сделали алгоритм одним из наиболее эффективных, выдерживающих конкуренцию даже с некоторыми узкоспециализированными, применимыми только к задаче Пуассона (уравнение с постоянными коэффициентами $\Delta u = f$) в квадрате.

Замечательным оказался тот факт, что и алгоритм метода, и теорема о независимости скорости сходимости от шага сетки, выдержали обобщения при усложнении задачи за счет переменности коэффициентов, произвольного вида области и т.п. Однако наибольшая популярность и широта применения метода в настоящее время связаны с его приложением к такому мощному средству решения эллиптических задач, как метод конечных элементов. Основные идеи этого способа построения аппроксимирующей конечномерной задачи были изложены в § 3.

Напомним, что характерной особенностью системы линейных алгебраических уравнений, аппроксимирующих, например, задачу Пуассона в достаточно произвольной области, являются высокий порядок системы (достигающий в современных расчетах $10^4 \div 10^5$) и слабая заполненность матрицы. (Эти черты присущи и системам метода конечных разностей в прямоугольной области.)

Следующее свойство специфично для метода конечных элементов. Расположение ненулевых элементов в матрице системы не имеет такой простой и удобной структуры, с которой мы до сих пор имели дело, применяя метод сеток в простых областях. Это делает невозможным применение наиболее эффективных итерационных методов (переменных направлений, например). Пожалуй, единственный из знакомых нам методов, который в такой ситуации может быть использован, — это метод простой итерации с чебышевским ускорением. Но его эффективность недостаточна для решения сложных задач, поэтому в методе конечных элементов обычно используются «ленточные» варианты метода исключения Гаусса, что все-таки является довольно дорогой операцией, часто вынуждающей ограничиваться расчетами на относительно грубых сетках.

При описании основных идей метода конечных элементов специально было обращено внимание на процедуру автоматической триангуляции «произвольной» области, при которой возникает иерархическая структура вложенных друг в друга сеток. Она позволяет удобно реализовать алгоритм многосеточных итераций.

Комбинация техники метода конечных элементов с многосеточным итерационным алгоритмом привела к созданию мощных

средств вычислений. Надо отметить, что логическая структура метода заметно сложнее, чем структура методов, описанных выше. Это приводит к определенным трудностям в программной реализации. Поэтому в простых задачах обычно предпочитают более простые с точки зрения программирования методы, хотя они работают медленнее.

Формирование задач на вспомогательных сетках. Рассмотрим две сетки — основную и первую вспомогательную, которую назовем грубой. В современной практике приходится строить грубую сетку, учитывая геометрию области, разрывы коэффициентов и т.п. Все это приводит к тому, что грубая сетка не имеет такой простой связи с основной, как было описано выше. Например, грубая сетка может формироваться так: задается список номеров основной сетки k_i, i-й узел грубой сетки совпадает с k_i-м узлом основной. Имеется в виду сетка по переменной x. Аналогично, списком m_j определяется сетка по y.

Таким образом узел (i, j) грубой сетки совпадает с узлом (k_i, m_j) основной. Числа k_i, естественно, возрастают, и все разности $k_{i+1} - k_i$ достаточно малы, в остальном они произвольны. Возможны и более сложные способы построения грубой сетки. В таких ситуациях возникает вопрос: как строить аппроксимацию на грубой сетке? Он еще более обостряется, если коэффициенты уравнения достаточно сильно отличаются даже в близких узлах основной сетки, т.е. если решается уравнение с разрывными коэффициентами.

Пусть на основной сетке получено приближение u с гладкой невязкой $r = \Delta u - f$ (здесь Δ — оператор на основной сетке, аппроксимирующий произвольный эллиптический, а не обязательно оператор Лапласа). Определим грубую сетку и оператор I, интерполирующий функцию, заданную на грубой сетке, на основную сетку. Попытаемся найти на грубой сетке такую функцию W, чтобы получить

$$\Delta(u - IW) - f = 0.$$

Очевидно, это невозможно, так как уравнений здесь столько, сколько внутренних узлов на основной сетке, а неизвестных W столько, сколько внутренних узлов на грубой сетке (функция IW должна удовлетворять однородным краевым условиям исходной задачи, чтобы коррекция $u - IW$ не портила краевые условия). Однако это уравнение можно решить в «слабом», галеркинском, смысле:

$$0 = (\Delta(u - IW) - f, IV) = (I^*r - I^*\Delta IW, V), \quad \forall V$$

или в явной форме — в виде уравнения для W:

$$DW = R, \quad \text{где } D = I^*\Delta I, \ R = I^*r.$$

Таким образом все определяется только конструкцией оператора интерполяции с грубой сетки на основную I. Что представляет собой оператор I^*, сопряженный к оператору интерполяции? Он отображает функции, определенные на основной сетке, в функции, определенные на грубой сетке. Структура его достаточно проста. Предположим, что k — индекс (точнее, мультииндекс) некоторого узла грубой сетки. Вычислим $(I^*z)_k$, где z — некоторая функция на основной сетке. Пусть $j(1{:}N_k)$ — список номеров узлов основной сетки, при интерполяции в которые используется значение интерполируемой функции в k-м узле грубой сетки. Если N_k — число таких узлов, а $\sigma(1{:}N_k)$ — значения соответствующих коэффициентов интерполяции (т.е. при интерполяции в $j(n)$-й узел в сумму входит слагаемое $\sigma(n)Z_k$), то

$$(I^*z)_k = \sum_{n=1,\ldots,N_k} \sigma(n)\, z_{j(n)}.$$

Итак, I^* — это оператор «сбора» значений в узлах основной сетки в узел грубой. Он является оператором локального типа в том смысле, что значение $(Iz)_k$ зависит только от значений z в узлах основной сетки, примыкающих к k-му узлу грубой. Разумеется, это есть следствие локальности оператора интерполяции, в качестве которого обычно используют линейную по каждой переменной интерполяцию.

Найдя W и осуществив коррекцию, получим функцию $\tilde{u} = u - IW$. Об ее невязке $\tilde{r} = \Delta\tilde{u} - f$ известно, что $(\tilde{r}, IV) = 0$ $(\forall\ V)$. Взяв в качестве V функцию, равную единице в k-м узле грубой сетки и нулю в остальных, получим следующее свойство невязки $\tilde{r}$: взвешенная сумма значений $\tilde{r}_n$ в узлах основной сетки, примыкающих к k-му узлу грубой, равна нулю (см. рис. 17). Конечно, уравнение для W решается тоже приближенно, поэтому $DW = R + \varepsilon$, и вышеупомянутая взвешенная сумма не равна нулю, но она есть $O(\varepsilon)$. После коррекции невязка нового приближения $u - IW$ стала очень маленькой «в слабом смысле» функцией.

Поясним, почему такие невязки эффективно «подавляются» простыми итерационными методами. Рассмотрим одномерную модель задачи — систему уравнений

$$u_{n-1} - 2u_n + u_{n+1} = f_n, \quad n = 1, 2, \ldots, N-1, \quad u_0 = u_N = 0.$$

Простейший, так называемый релаксационный метод решения этой системы состоит из итераций типа

$$u_n = 0.5\,(u_{n-1} + u_{n+1} - f_n), \quad n = 1, 2, \ldots, N-1.$$

Здесь u_{n-1} берется уже с «верхней итерации». Проделав пересчет в n-м узле, мы, очевидно, обратим в нуль невязку именно в этом узле. Однако, как нетрудно проверить, невязки в соседних узлах изменятся следующим образом:

$$r_{n-1} := r_{n-1} + 0.5r_n, \qquad r_{n+1} := r_{n+1} + 0.5r_n.$$

Если знаки r_{n-1}, r_n, r_{n+1} совпадают, операция не меняет нормы невязки, определенной формулой $\|r\| = \sum |r_n|$. Она уменьшается в случаях, когда $n = 1$ или $n = N - 1$ и когда знак r_n противоположен знаку r_{n-1} или r_{n+1}. Именно такую ситуацию создает коррекция в узлах основной сетки, совпадающих с узлами грубой.

Коррекция u эффективна, если ее невязка r достаточно гладкая функция в том смысле, что при вычислении $R = I^* r$, как взвешенной суммы, не происходит сильного сокращения слагаемых с противоположными знаками. Что касается коэффициентов разностного оператора на грубой сетке $D = I^* \Delta I$, то они являются некоторой взвешенной суммой коэффициентов аппроксимации на основной сетке, вычисление которых однозначно определяется заданием оператора интерполяции.

ЧАСТЬ ВТОРАЯ

ПРИБЛИЖЕННЫЕ МЕТОДЫ ВЫЧИСЛИТЕЛЬНОЙ ФИЗИКИ

§ 15. Спектральная задача Штурма–Лиувилля

Рассмотрим некоторые приближенные методы вычисления собственных значений и функций линейных дифференциальных операторов. Важнейшим прикладным источником подобных задач является квантовая механика. В качестве характерного примера рассмотрим задачу вычисления волновой функции частицы, движущейся в центрально-симметричном поле с потенциалом $U(r)$. Определение волновой функции $\psi(r, \theta, \varphi)$ приводит к уравнению Шредингера

$$\Delta\psi + \frac{2\mu}{\hbar^2}(E - U(r))\psi = 0. \tag{1}$$

Здесь Δ — оператор Лапласа в сферических переменных r, θ, φ; μ, $\hbar$ — известные постоянные. Функция ψ определена во всем трехмерном пространстве; «граничным условием» для нее является ограниченность гильбертовой нормы. В уравнении (1) подлежат определению те дискретные вещественные числа E, при которых задача имеет нетривиальное решение (точки дискретного спектра оператора Шредингера).

Решение ищем в виде $\psi = Y_{l,m}(\theta, \varphi)R(r)/r$, где $Y_{l,m}$ есть известная сферическая функция (l, m — целые числа). Обозначая

$$\lambda = \frac{2\mu}{\hbar^2}E, \qquad V(r) = \frac{2\mu}{\hbar^2}U(r) + \frac{l(l+1)}{r^2},$$

получаем окончательную форму задачи:

$$\frac{d^2R}{dr^2} - (V(r) - \lambda)R = 0. \tag{2}$$

Уравнение (2) определено при $0 \leqslant r \leqslant \infty$. При $r = 0$ ставится условие $R(0) = 0$. Вторым условием является условие нормировки

$$\int_0^\infty R^2(r)\,dr = 1, \tag{3}$$

имеющее важное следствие: оно определяет знак точек спектра λ; из него следует $\lambda < 0$. В самом деле, при достаточно большом r потенциал V становится очень малым и решения уравнения (2) качественно совпадают с решениями уравнения $R'' + \lambda R = 0$, т.е.

$$R(r) \approx C_1 e^{r\sqrt{-\lambda}} + C_2 e^{-r\sqrt{-\lambda}}.$$

Если $\lambda > 0$, это — функция типа $\sin\sqrt{\lambda}\, r$, что, очевидно, несовместимо с нормировкой (3). (Однако эти решения ограничены, поэтому все $\lambda > 0$ образуют *сплошной спектр*, которым мы не интересуемся.) Если $\lambda < 0$, нормировка достигается при $C_1 = 0$.

Интервал $[0 \leqslant r \leqslant \infty]$ можно (грубо) представить в виде двух частей. При r, близких к нулю, знак множителя $V(r) - \lambda$ определяется потенциалом $V(r)$ и, так как $U(r) < 0$, может быть отрицательным. В этом случае решения уравнения $R'' = (V - \lambda)R$ имеют колебательный характер. При больших r знак множителя $V(r) - \lambda$ определяется величиной $(-\lambda) > 0$, т.е. функция $R(r)$ имеет экспоненциальный характер типа $e^{-r\sqrt{-\lambda}}$ (рис. 18), причем, как мы увидим в дальнейшем, в расчетах придется иметь дело с достаточно большими значениями $r\sqrt{|\lambda|}$.

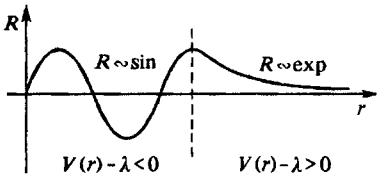


Рис. 18

Чтобы правильно оценить целесообразность метода, который будет изложен ниже, начнем с анализа трудностей, встречающихся при попытке решения задачи стандартными средствами. Используем метод «пристрелки». Значение $R(0) = 0$. Зададим $R'(0)$ произвольно, например $R'(0) = 1$, и решим (приближенно) задачу Коши для уравнения (2), считая λ тем или иным образом заданным.

Таким образом, мы имеем функцию $R(r, \lambda)$. Искомые собственные значения — это те дискретные величины λ, при которых функция $R(r, \lambda)$ при $r \to \infty$ имеет асимптотику $Ce^{-r\sqrt{-\lambda}}$, т.е. не содержит второй растущей компоненты общего решения. На практике просто назначают достаточно большое значение r^* и ставят условие $R(r^*, \lambda) = 0$. Это есть уравнение для собственных значений, которое решается, например, методом Ньютона или просто подбором: при λ_1 получаем $R(r^*, \lambda_1) > 0$, при λ_2 имеем $R(r^*, \lambda_2) < 0$, возьмем $\lambda_3 = 0.5(\lambda_1 + \lambda_2)$, и т.д.

Однако численная реализация этой процедуры наталкивается на серьезные затруднения. Дело в том, что на экспоненциальном участке решение задачи Коши неустойчиво. Малое отклонение численного решения от точного приведет к появлению в решении ком-

поненты $C_1 e^{r\sqrt{-\lambda}}$ (пусть даже с малым коэффициентом C_1), и при больших r эта компонента станет основной частью решения. Кроме того, погрешности численного интегрирования в этом случае также дают вклад в приближенные решения порядка $h^p e^{r\sqrt{-\lambda}}$ (h — шаг численного интегрирования, p — порядок точности используемого метода).

Трудности подобного рода преодолеваются методом прогонки. В этом случае решение ищется в виде «прогоночного соотношения» $R(r) = \alpha(r)R'(r)$. Для искомой функции $\alpha(r)$ легко выводится (см. § 9) уравнение

$$\alpha' = 1 - \alpha^2(V(r) - \lambda).$$

Левое краевое условие очевидно: $\alpha(0) = 0$. Условием на правом конце является условие выхода на асимптотику $R \sim e^{-r\sqrt{-\lambda}}$, которое можно аппроксимировать условием $\alpha(r^*) = -1/\sqrt{-\lambda}$. Реализация такого подхода связана с двумя затруднениями. На участке, где $V(r) - \lambda < 0$ и решение имеет колебательный характер, обязательно есть точки $\tilde{r}$, в которых $R' = 0$ и, следовательно, $\alpha = \infty$.

На участке, где $R(r)$ экспоненциально убывает, тоже возникают осложнения. Разберемся в существе дела, пренебрегая величиной $V(r)$ по сравнению с λ. Уравнение $\alpha' = 1 - |\lambda|\alpha^2$ имеет два положения равновесия: $\alpha = \pm 1/\sqrt{|\lambda|}$. Несложный анализ поля направлений показывает, что ветвь $\alpha = 1/\sqrt{|\lambda|}$ является асимптотически устойчивой, а ветвь $\alpha = -1/\sqrt{|\lambda|}$, наоборот, — неустойчивой. Нужно попасть именно на эту вторую ветвь.

Перейдем к описанию алгоритма, справляющегося с этими трудностями, — к алгоритму тригонометрической прогонки. Перед изложением существа дела сделаем несколько замечаний о характере вычислительной проблемы. Для приложений необходимо несколько первых собственных значений (нумерация собственных значений делается в порядке возрастания $|\lambda_k|$). Такие расчеты надо делать для нескольких l, т.е. задачу нужно решать многократно. Следовательно, алгоритм ее приближенного решения должен быть достаточно эффективным, экономичным. Оператор (2) самосопряжен, поэтому все λ_k действительны. Напомним еще осцилляционную теорему: естественная нумерация собственных значений связана с числом нулей функции $R(r)$.

«Тригонометрическая прогонка» основана на введении функции $\varphi(r)$, связанной с $R(r)$ соотношением

$$R(r)\sin\varphi(r) - R'(r)\cos\varphi(r) = 0. \qquad (4)$$

Поясним его происхождение, а заодно выясним «геометрический» смысл величины $\varphi(r)$ (он будет полезен). На рис. 19 (качественно) изображена фазовая плоскость (R, R') и траектория $\{R(r), R'(r)\}$.

Показан случай, когда траектория $R(r)$ имеет шесть нулей (точек $R(r) = 0$). Это означает, что показана пятая собственная функция (первая, например, имеет только два нуля: $r = 0$ и $r = r^*$).

Представим точку (R, R') в полярных координатах:

$$R(r) = A(r) \cos \varphi(r), \qquad R'(r) = A(r) \sin \varphi(r). \tag{5}$$

Соотношение (4) есть очевидное следствие (5). Угол $\varphi(r)$ — это угол точки (R, R'). Краевому условию $R(0) = 0$ соответствует угол $\varphi(0) = \pi/2$. Каждый нуль $R(r)$ соответствует изменению $\varphi(r)$ на $-\pi$. Отметим еще, что искомая функция $R(r)$ определена с точностью до множителя: здесь выбрана нормировка $R'(0) > 0$, при которой угол $\varphi(r)$ убывает; при нормировке $R'(0) < 0$ угол $\varphi(r)$ возрастает.

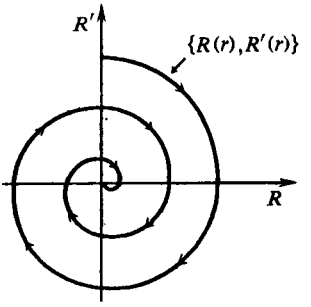


Рис. 19

Теперь можно сформулировать краевое условие для $\varphi(r^*)$: $\varphi(r^*) = \pi/2 - k\pi$, где k — номер собственного значения. Это очень удобный в практических расчетах факт: он позволяет явно задавать номер собственного значения. Получим дифференциальное уравнение для φ так, как это обычно делается в алгоритмах прогонки. Продифференцируем соотношение (4) по r:

$$R' \sin \varphi + R\varphi' \cos \varphi - R'' \cos \varphi + R'\varphi' \sin \varphi = 0.$$

Подставляя в это выражение $R'' = (V(r) - \lambda)R$, имеем

$$R'(1 + \varphi') \sin \varphi + R(\varphi + \lambda - V(r)) \cos \varphi = 0. \tag{6}$$

Соотношения (4) и (6) образуют систему линейных однородных уравнений для R и R'. Приравнивая нулю определитель, получаем уравнение для φ:

$$\varphi' + \sin^2 \varphi + (\lambda - V(r)) \cos^2 \varphi = 0, \quad \varphi(0) = \pi/2. \tag{7}$$

Численным интегрированием этой задачи Коши определена функция $\varphi(r^*, \lambda)$. Ее график напоминает «лестницу» с почти вертикальными участками в окрестности значений $\pi/2 - k\pi$. Это надо учитывать при решении уравнения $\varphi(r^*, \lambda) = \pi/2 - k\pi$.

Пусть найдено число λ_k, при котором решение (7) удовлетворяет обоим краевым условиям. Нужно найти саму собственную функцию $R(r)$, для чего достаточно определить амплитуду $A(r)$. Выведем дифференциальное уравнение для $A(r)$. Дифференцируем первое из соотношений (5): $R' = A' \cos \varphi - A\varphi' \sin \varphi$. Согласно второму из со-

отношений (5) $R' = A \sin \varphi$. С учетом этого получаем

$$A' = A(1 + \varphi') \sin \varphi / \cos \varphi.$$

Используя (7), вычисляем

$$1 + \varphi' = 1 - \sin^2 \varphi + (V(r) - \lambda) \cos^2 \varphi = (1 + V(r) - \lambda) \cos^2 \varphi.$$

В результате получаем

$$A' = A \cos \varphi \sin \varphi \, (1 + V(r) - \lambda).$$

Значение $A(0)$ произвольно, например $A(0) = 1$. В любом случае решение $A(r)\cos \varphi(r)$ нужно нормировать.

В заключение выясним качественную структуру траектории $\varphi(r)$. При r, близких к нулю, величина $\lambda - V(r) > 0$ и $R(r)$ осциллирует, $\varphi'(r) < 0$ и $\varphi(r)$ монотонно убывает. На правой части интервала $[0, r^*]$ функция $R(r) \approx e^{-r\sqrt{-\lambda}}$, поэтому $R'(r)/R(r) = \operatorname{ctg} \varphi(r) \approx \text{const}$. Итак, $\varphi(r)$ сначала монотонно убывает, затем скорость убывания замедляется, и график $\varphi(r)$ становится почти горизонтальным.

Столь простая структура решения $\varphi(r)$ может создать впечатление, что уравнение для φ можно интегрировать очень крупным шагом. В § 5 указывалось, что при выборе шага численного интегрирования следует ориентироваться на следующее простое соображение: каждый характерный участок решения нужно «покрыть» хотя бы пятью-десятью счетными точками. И кажется, что при интегрировании уравнения для φ достаточно взять пятьдесят точек на участке осцилляций и столько же точек на экспоненциальном участке, хотя, например, на участке осцилляций функция $R(r)$ имеет, допустим, пять полуволн. К сожалению, это не так: численный расчет «простой» траектории $\varphi(r)$ требует такого же числа точек (такого же шага интегрирования), какого требует расчет «сложной» траектории $R(r)$, если, конечно, используются стандартные методы, например методы Рунге—Кутты.

В самом деле, на участке осцилляций решение имеет (грубо, ориентировочно) вид $\varphi(r) \approx \pi/2 - Cr$. Постоянная C тем больше, чем больше номер собственного значения. Шаг численного интегрирования должен быть таким, чтобы за это время правая часть уравнения изменилась пренебрежимо мало. Но функции $\sin^2 \varphi$ и $\cos^2 \varphi$ за шаг h мало меняются лишь при условии $Ch \ll 1$, т.е. чем больше C (чем больше полуволн имеет решение $R(r)$ на участке осцилляций), тем меньше должен быть шаг h.

И наконец, несколько слов о погрешности, связанной с переносом граничного условия $R(\infty) = 0$ в точку r^*. Мы рассмотрим его опять-таки на модельном для экспоненциальной части траектории

уравнении $R'' = |\lambda| R$. Точное решение должно иметь вид

$$R(r) = C \exp(-r\sqrt{|\lambda|}).$$

Приближенное решение $\tilde{R}(r)$, обращающееся в нуль при $r = r^*$, очевидно, есть

$$\tilde{R} = C\,[\exp(-r\sqrt{|\lambda|}) - \exp(-2r^*\sqrt{|\lambda|})\exp(r\sqrt{|\lambda|})].$$

На участке $[0, r^*]$ вклад в $\tilde{R}(r)$ от «лишнего» слагаемого очень мал (порядка $\exp(-r^*\sqrt{|\lambda|})$).

Вычисление точек комплексного спектра. Спектральная задача существенно осложняется в случае несамосопряженного дифференциального оператора, когда собственные значения могут оказаться комплексными. Рассмотрим несложную задачу

$$-\frac{d^2x}{dt^2} + a\frac{dx}{dt} + bx = \lambda x, \quad 0 \leqslant t \leqslant 1, \tag{8}$$

с краевыми условиями $x(0) = x(1) = 0$. Пусть коэффициенты a, b — какие-то несложные функции или даже постоянные комплексные числа (или x — вектор, a, b — матрицы).

Исследование спектра (вообще говоря, комплексного) несколько облегчается тем, что обычно задачей численного анализа является определение точек спектра, расположенных в некоторой окрестности нуля. Размеры этой окрестности, конечно, зависят от модуля a. В области $|\lambda| \gg |a|$ часто можно воспользоваться подходящими асимптотическими методами, т.е. изучать спектр задачи (8) как слабое возмущение хорошо изученного спектра задачи при $a = 0$. Эти аналитические методы достаточно эффективны при больших $|\lambda|$, но теряют точность и иногда просто непригодны в области малых $|\lambda|$.

Однако именно точки спектра с малыми $|\lambda|$ нередко представляют наибольшую прикладную ценность, и численные методы в этой области удачно дополняют аналитические исследования. Таким образом, речь идет о приближенном вычислении относительно небольшого числа точек спектра. И здесь в зависимости от ситуации может быть использован метод «пристрелки», когда интегрируется задача Коши с начальными данными $x(0, \lambda) = 0$, $\dot{x}(0, \lambda) = 1$ и этим алгоритмом численного интегрирования определяется функция комплексного переменного $\Phi(\lambda) \equiv x(1, \lambda)$.

В другой ситуации может оказаться целесообразным применение метода прогонки, когда, например, решение ищется в форме $x(t) = \alpha(t)\,\dot{x}(t)$, а для «прогоночного коэффициента» $\alpha(t)$ обычным способом получается уравнение (типа уравнения Риккати), содержащее параметр λ. Интегрируя эту задачу (конечно, численно), вычисляем $\alpha(t, \lambda)$ и определяем $\Phi(\lambda) \equiv \alpha(1, \lambda)$. Если на правом конце за-

дано более общее краевое условие (например, $\dot{x}(1) + \beta x(1) = 0$), оно вместе с прогоночным соотношением $x(1) = \alpha(1)\,\dot{x}(1)$ образует систему линейных уравнений относительно $x(1)$ и $\dot{x}(1)$, а $\Phi(\lambda)$ определяется как детерминант этой системы. Так или иначе мы получаем функцию комплексного переменного, значения которой вычисляются процедурой интегрирования задачи Коши (в комплексных числах). Точки спектра исходной задачи суть нули этой функции.

При определенных условиях, которые хорошо изучены в теории обыкновенных дифференциальных уравнений, $\Phi(\lambda)$ — аналитическая функция (это следствие простой формы зависимости уравнения от λ), не имеющая полюсов при ограниченных λ. Поэтому для подсчета числа ее нулей в некоторой области G следует вычислить известный интеграл по контуру или, проще говоря, вращение векторного поля $\{\mathrm{Re}\,\Phi(\lambda), \mathrm{Im}\,\Phi(\lambda)\}$ при обходе контура ∂G. Конечно, это — громоздкая операция: надо «покрыть» контур сеткой точек $\{\lambda_i\}$ и в каждой точке λ_i вычислить $\Phi(\lambda_i)$, т.е. проинтегрировать задачу Коши.

Несколько облегчает работу то, что сетка $\{\lambda_i\}$ не должна быть особенно густой. Точнее, дело обстоит так. Расчет начинается с достаточно широкой области G, имеющей (для простоты и определенности) форму прямоугольника. Вычислив вращение вдоль контура (оно будет равно $2\pi n$, где n — число точек спектра в G), делят его пополам (пополам делится та сторона прямоугольника, которая на данном этапе процесса локализации корней $\Phi(\lambda)$ длиннее).

Вычислив вращение вдоль контура одного из полученных меньших прямоугольников, определяют число точек спектра в двух частях исходной области, и т.д. Когда прямоугольник велик (и его контур достаточно длинен), шаг сетки на контуре может быть взят достаточно большим. Например, в некоторых расчетах, проводившихся по этой схеме, было принято считать «нормальной» ситуацию, в которой при переходе от λ_i к λ_{i+1} значения $\arg \Phi(\lambda)$ изменялись в пределах интервала $[\pi/6, \pi/3]$. Если изменение было меньшим, шаг увеличивался, если бо́льшим, — происходил возврат в точку λ_i, шаг Δ изменения λ уменьшался (например, $\Delta := \Delta/2$) и делался переход в точку $\lambda_{i+1} = \lambda_i + \Delta$ на контуре.

Таким образом, сетка $\{\lambda_i\}$ не задавалась заранее, а «генерировалась» простым алгоритмом с адаптацией (с регулированием шага в зависимости от градиента функции $\arg \Phi(\lambda)$). Конечно, такая тактика сопряжена с некоторым риском: при вычислении $\Phi(\lambda_i)$ и $\Phi(\lambda_{i+1})$ приращение аргументов определено с точностью до $2k\pi$ (k — любое целое), причем число k вычислитель назначает сам.

Поясним сказанное подробнее. Вычислив комплексное число $x + iy$, обычно обращаются к подпрограмме, входящей в стандартную

библиотеку и имеющей на языке FORTRAN имя ATAN2(x, y). Результатом является главное значение $\operatorname{arctg}(y/x)$; к нему из каких-то дополнительных соображений нужно добавить $2k\pi$.

В рассматриваемом случае, используя предположение о том, что шаг $\Delta = \lambda_{i+1} - \lambda_i$ «достаточно мал», число k выбирают таким, чтобы изменение $\arg \Phi(\lambda_{i+1})$ по сравнению с $\arg \Phi(\lambda_i)$ было минимальным. Здесь, конечно, есть риск ошибиться на $2k\pi$. Можно получить достаточно надежное подтверждение правильности этого решения (или обнаружить его ошибочность), «пройдя» участок $(\lambda_i, \lambda_{i+1})$ с существенно меньшим шагом, но это слишком «дорого» и не делается без достаточных к тому оснований.

Вероятность ошибочности такого способа вычисления вращения векторного поля существенным образом зависит от расстояния контура до какой-то точки спектра. Если случайно оказалось, что точка спектра расположена очень близко от прямолинейного участка контура $(\lambda_i, \lambda_{i+1})$, описанная выше процедура определения аргумента может привести к принципиальной ошибке: в окрестности корня $\Phi(\lambda)$ (тем более, если близко к друг другу расположены несколько корней Φ или около контура находится кратный корень) изменение направления поля может быть большим на малом расстоянии.

В начале расчета, когда область G выбирается на основании грубых априорных соображений, вероятность столкнуться с такой ситуацией очень мала. По мере дробления области, когда происходит локализация корня в области все меньшей и меньшей, корень, конечно, приближается к контуру, но одновременно происходит и уменьшение шага, с которым обходится контур. В принципе, при благоприятном ходе вычислительного процесса обход каждого контура требует примерно одного и того же числа вычислений $\Phi(\lambda)$: ведь одновременно с уменьшением шага уменьшается и длина контура очередной области локализации. Если, например, в области G есть всего один корень и шаг $\lambda_{i+1} - \lambda_i$ регулируется так, что в среднем при переходе от λ_i к λ_{i+1}, значение $\arg \Phi(\lambda)$ изменяется на $\pi/4$, вычисление вращения «стоит» всего десяти вычислений $\Phi(\lambda)$.

Кроме того, имеется дополнительная возможность сокращения объема вычислений за счет использования уже имеющейся информации: производя деление очередного прямоугольника пополам, можно вычислить вращение поля вдоль каждого из двух новых контуров, вычисляя вращение лишь вдоль введенной на этом шаге линии раздела. Однако такой способ требует в общем случае достаточно хитроумного программирования и хранения полученной ранее информации.

Если читатель сочтет изложенное выше не слишком надежным, не гарантирующим правильного решения задачи вычисления всех

точек спектра в некоторой заданной области, он будет совершенно прав. Такую процедуру можно сделать сколь угодно надежной, уменьшая шаг обхода контура (т.е. увеличивая объем вычислений) и, наконец, просто «точной», если заменить описанную выше процедуру вычисления вращения на хорошо известный в теории функций комплексного переменного контурный интеграл. Однако эта точность обманчива: ведь интеграл нужно вычислять по какой-то квадратурной формуле, и на стадии ее реализации в расчет войдет какая-то сетка со всеми вытекающими отсюда последствиями.

Мы встречаемся здесь с достаточно типичной в вычислительной математике ситуацией: практически, доведенный до числа расчет редко дает полностью гарантированный ответ. Содержательная интерпретация такого численного результата содержит элемент риска, уменьшение которого связано с увеличением объема вычислительной работы.

Алгебраические методы. Аппроксимируя задачу конечномерной, мы получаем формально стандартную алгебраическую спектральную проблему. Для ее решения разработаны надежные алгоритмы, они включены в системы математического обеспечения ЭВМ, и можно просто воспользоваться одним из них. Этот путь возможен, но нужно внимательно отнестись к выбору средства аппроксимации. Общие алгебраические методы весьма чувствительны к такому фактору, как размерность пространства. Следует использовать методы дискретизации, которые при относительно невысокой размерности пространства позволяют получать достаточно высокую точность. Метод конечных разностей к таковым не относится, его достоинства в другом.

Проиллюстрируем сказанное, описав в общих чертах алгоритм, разработанный К. И. Бабенко. Эффективность этого метода основана на двух основных идеях:

а) обращение главной части дифференциального оператора;

б) выбор эффективного аппарата конечномерной аппроксимации.

Обращение главной части оператора состоит в решении краевой задачи

$$\frac{d^2x}{dt^2} = a\frac{dx}{dt} + bx - \lambda x, \qquad x(0) = x(1) = 0,$$

причем правая часть считается «известной». Решение имеет явное выражение: применение к обеим частям оператора $(d^2/dt^2)^{-1}$ состоит в интегрировании после умножения на функцию Грина $K(t, \xi)$. В результате получаем эквивалентное уравнение:

$$x(t) = \int_0^1 K(t, \xi)\,[a(\xi)\,\dot{x}(\xi) + b(\xi)\,x(\xi) - \lambda\,x(\xi)]\,d\xi. \qquad (9)$$

Явный вид ядра $K(t, \xi)$ считается, разумеется, известным. В рассматриваемом случае

$$K(t, \xi) = \{\xi(1-t) \text{ при } \xi \leqslant t, \ t(1-\xi) \text{ при } \xi \geqslant t\}.$$

Дальнейшее продвижение связано с заменой искомой функции $x(t)$ подходящей аппроксимацией. В частности, предлагается искать $x(t)$ в форме сеточной функции, определенной в чебышевских узлах и восполняемой до непрерывной с помощью интерполяционного полинома. Итак, приближенное решение ищется в виде

$$x(t) = \sum_{n=0}^{N} x_n \, l_N^n(t), \tag{10}$$

где $l_N^n(t)$ есть интерполяционный базис — полиномы степени N (см. § 3). Значения x_n пока неизвестны, для них будет получена система линейных уравнений с параметром λ.

Прямая подстановка конструкции (10) в (9) приведет, очевидно, к неразрешимой задаче, так как в нашем распоряжении имеется всего $N+1$ параметр, а (9) — это «континуум» уравнений. Поэтому вводится сетка так называемых точек коллокации $\{t_k^*\}_{k=0}^N$, и выполнение (9) для $x(t)$ в форме (10) требуется лишь в точках t_k^*. Таким способом получается система уравнений

$$\sum_{n=0}^{N} x_n \, l_N^n(t_k^*) = \sum_{n=0}^{N} b_{k,n} \, x_n - \lambda \sum_{n=0}^{N} c_{k,n} \, x_n = 0, \quad k = 0, 1, \ldots, N, \tag{11}$$

где

$$b_{k,n} = \int_0^1 K(t_k^*, \xi) \, [a(\xi) \, \dot{l}_N^n(\xi) + b(\xi) \, l_N^n(\xi)] \, d\xi,$$

$$c_{k,n} = \int_0^1 K(t_k^*, \xi) \, l_N^n(\xi) \, d\xi.$$

Теперь для определения $\{x_n\}$ и λ мы имеем $N+1$ уравнений (11), которым можно придать стандартную форму спектральной задачи $Ax = \lambda Cx$. Опыт показал, что сравнительно небольшие значения N, приводящие к не очень трудоемкой алгебраической проблеме собственных значений, при таком подходе позволяют получить с хорошей точностью сравнительно большое число точек спектра дифференциального оператора (примерно $N/2$ и даже больше). Например, для уравнения Ламе такой алгоритм уже при $N = 21$ дает для первых восьми собственных значений величины, совпадающие с

«каноническими табличными» в девяти десятичных знаках, следующие четыре собственных значения совпадают с табличными в восьми десятичных знаках, и т.д.

При оценке трудоемкости и точности алгоритма нужно иметь в виду, что она существенно зависит от того, как вычисляются интегралы, определяющие коэффициенты матриц A и C. Наиболее эффективные результаты получаются в том случае, когда интегралы $b_{k,n}$, $c_{k,n}$ «берутся» в конечном виде и вычисления проводятся по каким-то не очень сложным формулам с очень высокой точностью. Если же таких формул нет и приходится вычислять интегралы (а их не так уж мало, примерно $2N^2$) по каким-то квадратурным формулам, трудоемкость алгоритма заметно возрастает.

§ 16. Главная спектральная задача для краевых задач математической физики

В приложениях часто возникает задача, которую формально можно записать в простой форме. Пусть A — линейный дифференциальный оператор, соответствующий некоторой краевой задаче для уравнений с частными производными. Нужно найти главное собственное число и соответствующую ему собственную функцию:

$$Au = \lambda u. \tag{1}$$

Главным собственным числом называют обычно крайнюю точку спектра, например с наибольшим значением $\operatorname{Re}\lambda$. Поясним суть дела, рассмотрев важный в приложениях пример — математическую модель ядерного реактора. Разумеется, мы ограничимся сравнительно простой моделью. Ядерный реактор будем представлять себе в виде некоторого прямоугольного тела (например, в виде трехмерного куба).

Распределение нейтронов в реакторе описывается системой двухгрупповых уравнений диффузии:

$$\begin{aligned} \frac{1}{v_1}\frac{\partial \Phi_1}{\partial t} &= \operatorname{div} D_1 \operatorname{grad} \Phi_1 - A_{11}\Phi_1 + A_{12}\Phi_2, \\ \frac{1}{v_2}\frac{\partial \Phi_2}{\partial t} &= \operatorname{div} D_2 \operatorname{grad} \Phi_2 - A_{22}\Phi_2 + A_{21}\Phi_1, \end{aligned} \tag{2}$$

с краевыми условиями на границе Γ куба $\Phi_1 = \Phi_2 = 0$ и какими-то начальными данными. Здесь $\Phi_i(t, x, y, z)$ $(i = 1, 2)$ — функции, описывающие распределение быстрых (Φ_1) и медленных (Φ_2) нейтронов. Уравнения (2) описывают их эволюцию во времени с учетом следующих процессов:

1) диффузия (члены $\mathrm{div} D_i \operatorname{grad} \Phi_i$);

2) поглощение нейтронов (члены $-A_{11}\Phi_1$ и $-A_{22}\Phi_2$);

3) рождение быстрых нейтронов при поглощении медленных (член $A_{12}\Phi_2$) и наоборот (член $A_{21}\Phi_1$).

Коэффициенты системы D, A суть некоторые функции x, y, z, определяемые физическими константами материалов, из которых составлен реактор. Задача линейна относительно Φ_i. Ее можно записать в компактной форме:

$$\varphi_t = A\varphi, \tag{3}$$

где $\varphi = \{\Phi_1, \Phi_2\}$, A — линейный дифференциальный оператор эллиптического типа, главная дифференциальная часть которого $\mathrm{div} D \operatorname{grad}$ при постоянном D есть просто $D\Delta$ (свойства оператора A во многом похожи на свойства оператора Лапласа).

Чтобы представить себе характер процессов, протекающих в реакторе, воспользуемся методом Фурье для решения уравнения (3). Так как A не зависит от t, частные решения (3) можно искать в виде $\varphi(t) = e^{\lambda t}u$. Подставляя в (3), имеем

$$\lambda e^{\lambda t}u = Ae^{\lambda t}u = e^{\lambda t}Au, \qquad \text{или} \quad Au = \lambda u,$$

т.е. λ должно быть собственным числом, u — соответствующей собственной функцией оператора A. Из теории эллиптических уравнений мы знаем, что имеется дискретное множество собственных значений λ_k и соответствующих собственных функций ψ_k, образующих полную систему. Занумеруем собственные значения в порядке убывания $\mathrm{Re}\,\lambda_k$: $\mathrm{Re}\,\lambda_k \to -\infty$ при $k \to \infty$. Начальную функцию φ_0 разложим в ряд по ψ_k: $\varphi_0 = \sum_k c_k\psi_k$. Тогда сразу имеем решение:

$$\varphi(t) = \sum_k c_k e^{\lambda_k t}\psi_k. \tag{4}$$

Нетрудно убедиться, что при достаточно большом времени t в решении (4) выделяется главный член с наибольшим значением $\mathrm{Re}\,\lambda$: $\varphi(t) \approx c_1 e^{\lambda_1 t}\psi_1$. Кстати, нужно понимать, какое время является большим для процесса, описываемого системой (2). О нем естественно судить по величине $e^{(\lambda_2-\lambda_1)t}$; это следует из выражения

$$\varphi(t) = e^{\lambda_1 t}(c_1\psi_1 + c_2 e^{(\lambda_2-\lambda_1)t} + \ldots).$$

Например, в некоторых реакторах $\lambda_2 - \lambda_1 \approx -(50 \div 100)$, т.е. время 0.1 с уже является очень большим. Что же происходит с реактором? Все зависит от значения λ_1: если $\lambda_1 > 0$, реактор «взрывается», если

$\lambda_1 < 0$, реактор «тухнет». Рабочий режим реактора — это ситуация $\lambda_1 = 0$. Разумеется, значение λ_1 зависит от коэффициентов системы, т.е. от физического состава реактора. Он поддается регулированию с помощью стержней. Цель этого регулирования — обеспечить значение $\lambda_1 = 0$.

Теперь понятно, почему в практике расчетов ядерных реакторов одной из главных вычислительных задач является вычисление крайней точки спектра линейного дифференциального оператора (1). На практике A — это не дифференциальный оператор, а конечно-разностный, если мы решаем задачу (2) методом сеток. Для нас важно следующее обстоятельство: размерность конечномерного пространства u очень велика (порядка $10^3 \div 10^5$). Поэтому матрицы A мы обычно в явном виде не имеем.

Что же реально мы имеем? Рассмотрим для простоты двумерный случай (функции зависят от x и y). Введем в зоне реактора сетку с шагом h, узлами $x_k = kh$, $y_m = mh$ $(k, m = 0, 1, \ldots, N)$ и сеточные функции $\Phi 1_{k,m}$, $\Phi 2_{k,m}$. Тогда вместо дифференциального уравнения можно рассматривать аппроксимирующее конечно-разностное уравнение:

$$\frac{1}{h}\left\{D1_{k+1/2,m}\frac{\Phi 1_{k+1,m}-\Phi 1_{k,m}}{h} - D1_{k-1/2,m}\frac{\Phi 1_{k,m}-\Phi 1_{k-1,m}}{h}\right\} +$$

$$+\frac{1}{h}\left\{D1_{k,m+1/2}\frac{\Phi 1_{k,m+1}-\Phi 1_{k,m}}{h} - D1_{k,m-1/2}\frac{\Phi 1_{k,m}-\Phi 1_{k,m-1}}{h}\right\} -$$

$$- A11_{k,m}\Phi 1_{k,m} + A12_{k,m}\Phi 2_{k,m} = \frac{\lambda}{v_1}\Phi 1_{k,m}$$

(второе уравнение запишется точно так же).

Обычно в памяти ЭВМ мы имеем вектор $\{\Phi 1_{k,m}, \Phi 2_{k,m}\}$ и коэффициенты $D1_{k+1/2,m}, \ldots, A21_{k,m}$. Иногда для них нет места в памяти и приходится использовать подпрограммы, которые по индексам k, m и какой-то относительно небольшой информации о структуре реактора позволяют получить $D1_{k+1/2,m}$ и остальные коэффициенты схемы. Таким образом, имея точку u, можно вычислить точку (той же размерности) Au. Это сравнительно «дешевая» операция: она требует $O(N^2)$ арифметических действий.

Степенной метод определения границ спектра матрицы. Ограничимся сравнительно простым, но важным в приложениях случаем, когда оператор A самосопряженный: $A = A^*$. В этом случае все собственные значения λ_k вещественны. Следующий алгоритм позволяет вычислять максимальное по модулю собственное значение и соответствующий собственный вектор. Выбирается некоторый более

или менее произвольный вектор u^0 (начальное приближение). Затем производятся итерации (i — номер итерации):

$$u^{i+1} = Au^i, \quad i = 0, 1, 2, \ldots$$

Нетрудно убедиться, что при $i \to \infty$ вектор u^i стремится к собственной функции, соответствующей $\max |\lambda_k|$. В самом деле, $u^i = A^i u^0$. Пусть ψ_k — собственные векторы матрицы A. Разложим u^0 в ряд по базису ψ: $u^0 = \sum_k c_k \psi_k$. Тогда

$$u^i = A^i \sum_k c_k \psi_k = \sum_k c_k A^i \psi_k = \sum_k c_k (\lambda_k)^i \psi_k.$$

Обозначая $L = \max |\lambda_k|$, имеем

$$u^i = L^i \sum_k \left(\frac{\lambda_k}{L}\right)^i c_k \psi_k.$$

Очевидно, что компоненты суммы, у которых $|\lambda_k| < L$, стремятся к нулю и в конце концов остается только тот член, у которого $|\lambda_k| = L$ (для простоты считаем, что такой член только один).

Оценим скорость сходимости. Пусть K — номер максимального (по модулю) λ_k, $K-1$ — номер следующего за ним (по модулю) собственного значения. Тогда

$$u^i = (\lambda_K)^i \left\{ c_K \psi_K + \sum_{k \neq K} c_k \left(\frac{\lambda_k}{\lambda_K}\right)^i \psi_k \right\}.$$

Таким образом, u^i состоит из слагаемого, пропорционального ψ_K, и убывающей при $i \to \infty$ погрешности. Ее можно оценить по норме

$$\left\| \sum_{k \neq K} c_k \left(\frac{\lambda_k}{\lambda_K}\right)^i \psi_k \right\| \leqslant \left| \frac{\lambda_{K-1}}{\lambda_K} \right| \left\{ \sum_k c_k^2 \right\}^{1/2}.$$

Итак, погрешность убывает, как $|\lambda_K/\lambda_{K-1}|^i$.

Разумеется, мы неявно предполагаем, что коэффициент c_K не слишком мал, т.е. что начальное приближение не слишком плохое. Плохое оно при $c_K = 0$, однако и в этом случае метод дает правильный результат: за счет погрешностей округления в каком-то приближении обязательно появится ненулевой коэффициент c_K. Но если он слишком мал, процесс придется доводить до слишком боль-

ших чисел i, растущих, впрочем, при убывании c_K со скоростью логарифма. Вычислители на это не очень надеются и стараются выбрать разумное приближение, возможно более близкое к ψ_K. На этот счет есть достаточно надежные рецепты. Если A — разностная аппроксимация дифференциального оператора, то ψ_K близка к функции $u_{k,m} = (-1)^{k+m}$.

Теперь добавим некоторые важные детали. Так как $|\lambda_K| \neq 1$, то при достаточно большом i величина $A^i u^0$ обращается либо в машинный нуль, либо в машинную бесконечность. Чтобы этого избежать, в процесс добавляется нормировка i-го приближения, после чего он выглядит так:

$$\tilde{u}^{i+1} = Au^i, \qquad u^{i+1} = \tilde{u}^{i+1}/\|\tilde{u}^{i+1}\|.$$

При этом i-е приближение к собственному значению

$$\lambda^{(i)} = (Au^i, u^i)/(u^i, u^i)$$

(если $u^i = \Psi_K$, то формула дает $\lambda^{(i)} = \lambda_K$). Критерием достигнутой точности служит соотношение

$$\|Au^i - \lambda^{(i)} u^i\| \leqslant \varepsilon,$$

где ε — заданная погрешность.

Полезно представить себе характерное значение скорости сходимости. Возьмем в качестве оператора A разностную аппроксимацию оператора Лапласа на сетке $N \times N$. (Легко проверить, что суть дела не в шаге сетки h, а в числе узлов, приходящихся на линейный размер области; поэтому можно рассматривать задачу в квадрате 1×1.) Тогда

$$\max |\lambda_k| = 8N^2 \sin^2 \frac{N-1}{2N} \pi \approx 8N^2 \left(1 - \frac{\pi^2}{4N^2}\right).$$

Следующее за ним по модулю собственное значение есть

$$4N^2 \left(\sin^2 \frac{N-1}{2N} \pi + \sin^2 \frac{N-2}{2N} \pi\right) \approx$$

$$\approx 4N^2 \left(1 - \frac{\pi^2}{4N^2} + 1 - \frac{\pi^2}{N^2}\right) = 8N^2 \left(1 - \frac{5}{8} \frac{\pi^2}{N^2}\right).$$

Итак,

$$\frac{\lambda_{K-1}}{\lambda_K} \approx 1 - \frac{1}{8} \frac{\pi^2}{N^2}.$$

Видно, что при больших N скорость сходимости степенного метода невелика.

Обратим внимание на то, что при исследовании спектра разностной аппроксимации дифференциального оператора вычисление

$\max |\lambda_k|$ не очень интересно. При $N \to \infty$ эта величина стремится к бесконечности и особого смысла не имеет, хотя ее значение (или хотя бы ее оценка) нам понадобится. Интересной величиной является не $\max |\lambda_k|$, а $\max \lambda_k$. Спектры эллиптического дифференциального и разностного операторов устроены примерно так:

$$\lambda_1 > \lambda_2 > \ldots > \lambda_k \ldots, \qquad \lambda_k \to -\infty \text{ при } k \to \infty.$$

Нас интересует именно крайняя правая точка спектра. Итак, спектр A расположен на $[-L, \lambda_1]$, причем $L \gg |\lambda_1|$, где λ_1 может иметь любой знак.

Построим простой оператор с теми же собственными векторами, что и A, но с другим спектром: $B = E + \tau A$. Очевидно, его собственные векторы — те же ψ_k, а собственные числа суть

$$\beta_k = 1 + \tau\lambda_k, \qquad B\psi_k = \psi_k + \tau A\psi_k = \psi_k + \tau\lambda_k\psi_k.$$

Подберем τ таким, чтобы $\max |\beta_k| = |\beta_1|$. Очевидно, это достигается при $\tau = 1/L$. В самом деле,

$$\beta_k \geqslant 1 - |\tau\lambda_k| \geqslant 1 - L/L = 0,$$

$$\max |\beta_k| = \max \beta_k = \max (1 + \tau\lambda_k) = 1 + \tau \max \lambda_k = 1 + \tau\lambda_1.$$

Спектр B устроен так: β_1 может быть больше или меньше единицы в зависимости от знака λ_1, $\beta_k \in [0, \beta_1]$.

Теперь можно применить степенной метод к оператору B:

$$\tilde{u}^{i+1} = Bu^i, \qquad u^{i+1} = \tilde{u}^{i+1}/\|\tilde{u}^{i+1}\|, \qquad \beta_1^{(i+1)} = (\tilde{u}^{i+1}, u^i).$$

Как было выяснено, $u^i \to C\psi_1$, $\beta_1^{(i)} \to \beta_1$. Можно оценить и скорость сходимости: погрешность убывает, как

$$\left|\frac{\beta_2}{\beta_1}\right|^i = \left|\frac{1+\tau\lambda_2}{1+\tau\lambda_1}\right|^i \approx |1 - \tau(\lambda_1 - \lambda_2)|^i \approx \left|1 - \frac{\lambda_1 - \lambda_2}{L}\right|^i,$$

так как обычно $|\lambda_1| \ll L$, $|\lambda_2| \ll L$.

Для того чтобы составить себе представление о скорости сходимости, обратимся опять-таки к разностному оператору Лапласа на сетке $N \times N$. В этом случае, как показывают простые вычисления, $(\lambda_1 - \lambda_2)/L \approx \pi^2/2N^2$. Скорость сходимости невысока. Легко понять, что ее можно повысить примерно вдвое, взяв $\tau \approx 2/L$.

Метод обратной итерации. Запишем уравнение (1) в форме $Bu = \beta u$ (где $B = A^{-1}$, $\beta = \lambda^{-1}$) и применим степенной метод к оператору B, определяя его максимальное по модулю собственное зна-

чение. Пусть, как это часто бывает, $|\lambda_1| = \min |\lambda_k|$, тогда $B^i u^0$ выделяет именно то, что нас интересует. Используя для иллюстрации оператор Лапласа, находим, что спектр $B = \Delta^{-1}$ расположен на интервале $[1/\lambda_1, 0]$, $\lambda_1 < 0$, а скорость сходимости степенного метода для B определяется величиной β_2/β_1. Для лапласиана это дает (в терминах § 14) $\lambda_{1,1}/\lambda_{1,2} \approx 0.4$, т.е. скорость сходимости практически на зависит от числа узлов сетки N.

Итак, метод обратной итерации очень эффективен (погрешность убывает, как $(0.4)^i$), но сама стандартная итерация весьма громоздка. Поскольку явного выражения для оператора B мы не имеем, метод реализуется в форме $Au^{i+1} = u^i$ и каждая итерация требует приближенного решения системы линейных уравнений с матрицей A. Если $\min |\lambda_k| \neq |\lambda_1|$, то предварительно следует сдвинуть спектр, т.е. перейти к оператору $A' = A - \alpha E$, подобрав соответствующее значение α. Легко понять, что итерации будут тем эффективнее, чем ближе α к λ_1. Однако при этом заметно осложняется решение уравнения $(A - \alpha E)u^{i+1} = u^i$, так как оператор $A - \alpha E$ приближается к вырожденному оператору $A - \lambda_1 E$.

Вычисление второй собственной функции. В некоторых задачах представляет интерес вторая собственная функция (соответствующая точке спектра λ_2), а иногда и последующие. Степенной метод позволяет вычислить ее, правда, после того, как уже вычислена первая собственная функция. Используется тот же алгоритм, но в подпространстве, ортогональном найденной собственной функции ψ_1. В выражении $u^i = B^i u^0 = \sum_k c_k(\beta_k)^i \psi_k$ слагаемое $c_2(\beta_2)^i \psi_2$ будет играть главную роль в случае, если начальное приближение u^0 выбрано так, что $c_1 = 0$. Другими словами, начальное приближение должно быть ортогонально первой собственной функции ψ_1 (считаем, что она уже найдена), так как $c_1 = (u^0, \psi_1)$.

Процесс итераций организуется так. Возьмем какую-то функцию u^0. Спроецируем ее на подпространство, ортогональное функции ψ_1: $u^0 := u^0 - (u^0, \psi_1)\psi_1$. Выполняем стандартные итерации степенного метода: $\tilde{u}^{i+1} = Bu^i$. В результате погрешностей округления (и неточности определения ψ_1) функция u^i содержит пусть маленькую, но все же ненулевую компоненту $c_1\psi_1$, поэтому нужно проводить систематическую ортогонализацию и нормировку:

$$\tilde{u}^{i+1} := \tilde{u}^{i+1} - (\tilde{u}^{i+1}, \psi_1)\psi_1, \qquad u^{i+1} = \tilde{u}^{i+1}/\|\tilde{u}^{i+1}\|.$$

(Иногда ортогонализацию и нормировку проводят не на каждой итерации, а периодически, например через пять-десять итераций.) Собственное значение $\lambda_2^{(i)}$ (i-е приближение) вычисляется по той же формуле: $\lambda_2^{(i)} = (Au^i, u^i)/(u^i, u^i)$.

Ускорение степенного метода. Один из наиболее часто употребляемых приемов ускорения степенного метода состоит в варьировании итерационного параметра τ. Вычисления ведутся практически по той же самой схеме, но с переменным τ:

$$u^{i+1} = (E + \tau_{i+1}A)u^i.$$

Разлагая u^0 в ряд Фурье по собственным векторам, получаем для u^i формулу

$$u^i = \sum_k c_k \prod_{j=1}^{i} (1 + \tau_j \lambda_k)\psi_k.$$

Здесь, конечно, тоже следует использовать идею чебышевского метода, но с некоторой модификацией. Мы заинтересованы в том, чтобы полином $\prod (1 + \tau_j\lambda_k)$ был как можно меньше на всем спектре, за исключением точки λ_1.

Для реализации этой идеи необходимо располагать некоторой информацией. Прежде всего нужно иметь оценку (сверху) для максимального (по модулю) собственного числа L и оценку для второго собственного числа. Достаточно знать число l, разделяющее λ_1 и λ_2, т.е. должно быть $\lambda_k \in [-L, l]$ $(k = 2, 3, 4, \ldots)$, $\lambda_1 > l$. Параметры τ_j следует брать обратными к корням полинома Чебышева (некоторой назначенной степени i) на интервале $[-L, l]$. Эффект ускорения тем выше, чем ближе l к λ_2.

Есть еще один простой прием, позволяющий ускорить сходимость: сдвиг спектра. Пусть имеется некоторая оценка первого собственного числа λ^*. Она может быть получена на первом этапе решения задачи, т.е. является не очень точной. Уточнение осуществляется теми же итерациями, но с небольшим изменением: используется оператор со сдвинутым спектром $A - \lambda^* E$, т.е.

$$u^{i+1} = u^i + \tau_{i+1}(A - \lambda^* E)u^i.$$

Оператор $A - \lambda^* E$ имеет те же собственные векторы, что и A, а спектр его получается из спектра A сдвигом на λ^*.

В идеальном случае ($\lambda^* = \lambda_1$ и для λ_2 известна достаточно хорошая оценка) чебышевское ускорение действует только на компоненты $c_k\psi_k$ при $k = 2, 3, 4, \ldots$; компонента $c_1\psi_1$ не меняется. Случай

$\lambda^* < \lambda_1$, когда первое сдвинутое число $\lambda_1 - \lambda^* > 0$, тоже достаточно благоприятен при условии, конечно, что $\lambda_2 - \lambda^* < 0$. В этом случае значение чебышевского полинома в точке $\lambda = \lambda_1$ велико и это способствует более быстрому выделению в функции u^i главной компоненты. (Полиномы Чебышева обладают многими экстремальными свойствами, в том числе и самым быстрым среди полиномов той же степени ростом за пределами интервала, на котором расположены корни.)

Применяется и часто дает хорошие результаты метод, иногда называемый методом регуляризации. В этом случае итерации проводятся по формуле

$$B\frac{u^{i+1}-u^i}{\tau} = Au^i,$$

где B — «легко-обратимый» оператор (не в том смысле, что легко найти B^{-1}, а в том, что легко решить уравнение $Bv = r$). Очевидно, такие итерации можно изучать в форме $u^{i+1} = (E + \tau B^{-1}A)u^i$.

Таким образом, можно говорить о методе простой итерации для оператора $B^{-1}A$. Эффект ускорения достигается в том случае, если число обусловленности (или, другими словами, отношение минимального собственного значения к максимальному) у матрицы $B^{-1}A$ близко к единице, т.е. $B^{-1}A \approx E$.

Однако применение этого метода связано с двумя неприятностями. Первая состоит в том, что если матрицы B и A не перестановочны (а это типичный в приложениях случай), то собственные векторы $B^{-1}A$ (именно один из таких векторов выделяется рассматриваемым итерационным процессом), вообще говоря, не совпадают с собственными векторами A. Для борьбы с этим недостатком используется метод сдвига спектра. Имея u^i, можно вычислить оценку для λ_1 (считая, что в u^i искомая компонента ψ_1 является уже доминирующей): $\lambda^* = (Au^i, u^i)/(u^i, u^i)$. Следующая итерация проводится по формуле $B(u^{i+1} - u^i)/\tau = (A - \lambda^* E)u^i$. Смысл этого приема в том, что при $\lambda^* = \lambda_1$ собственный вектор ψ_1 матрицы A, соответствующий точке спектра λ_1, является и собственным вектором матрицы $B^{-1}(A - \lambda^* E)$ при любом B.

Укажем и на вторую опасность, которую тоже надо иметь в виду. Пусть для простоты A и B перестановочны, а собственный вектор ψ_1 соответствует точкам спектра λ_1 и β_1 операторов A и B соответственно. При этом для A точка λ_1 является крайней правой в спектре, и именно поэтому она нас интересует. Однако собственное число λ_1/β_1 оператора $B^{-1}A$ может оказаться не крайней, а внутренней

точкой спектра, и степенной метод с таким регуляризатором выделит не ψ_1, а какой-то другой собственный вектор, соответствующий крайней точке множества λ_k/β_k ($k = 1, 2, 3, \ldots$).

Оператор Шредингера с периодическим потенциалом. Рассмотрим характерную спектральную задачу квантовой теории твердого тела. Ниже описан опыт ее решения в ИПМ им. М. В. Келдыша. Физическая задача связана с исследованиями возможности существования так называемого «металлического водорода». Для нас наиболее интересными будут вопросы вычислительной технологии; они характерны для задач квантовой механики.

Итак, рассматривается задача определения собственных чисел и функций оператора Шредингера $-\Delta + U$, где $U(x, y)$ — потенциал, являющийся периодической функцией переменных x, y:

$$U(x, y) = \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} V(x + ka_1, y + la_2), \tag{5}$$

$$V(x, y) = e^{-\sqrt{x^2+y^2}}/\sqrt{x^2 + y^2}. \tag{6}$$

Такой потенциал возникает в периодической прямоугольной решетке (a_1, a_2 — периоды по x, y соответственно). В центре каждой ячейки находится полюс потенциала.

По теореме Блоха, используя замену $u(x, y) = e^{i(k_1 x + k_2 y)}\varphi(x, y)$, можно свести проблему к спектральной задаче $H\varphi = \lambda\varphi$, где

$$H = -\Delta + 2i\left(k_1 \frac{\partial}{\partial x} + k_2 \frac{\partial}{\partial y}\right) + (k_1^2 + k_2^2) + U. \tag{7}$$

Функция $\varphi(x, y)$, определенная в ячейке $|x| \leqslant a_1/2$, $|y| \leqslant a_2/2$, является периодической по x, y. Задачу надо решать многократно для значений k_1, k_2 из некоторой сетки, покрывающей так называемую зону Бриллюэна.

Задача решается методом конечных разностей. Ячейка покрывается сеткой $N \times N$, оператор (7) аппроксимируется стандартным конечно-разностным. Первая неприятность состоит в том, что функция $U(x, y)$ может обратиться в бесконечность в каком-то узле сетки. Ее легко избежать, немного сдвигая ячейку, что допустимо в периодической задаче. Но это плохой выход: опыт показывает, что при тех небольших значениях N, которыми приходилось ограничиваться в практической работе, результат сильно зависит от сдвига и от числа узлов сетки N.

Вторая проблема связана с вычислением потенциала U по формуле (5). Вычисление таких сумм осложняется очень медленной сходимостью, необходима разработка методов ее ускорения. Один из часто

применяемых приемов состоит в том, что члены медленно сходящегося ряда $\sum a_k$ разбивают на две части:

$$\sum_k a_k = \sum_k a_k' + \sum_k a_k'', \quad a_k = a_k' + a_k'',$$

таким образом, чтобы ряд $\sum a_k'$ суммировался аналитически, а остаток $\sum a_k''$ был быстро сходящимся. Подобрать такое разбиение — это уже искусство. Поэтому в данном случае говорят не о методе, а о приеме. В следующем ниже алгоритме потенциал $U(x, y)$ нужно вычислять в каждом узле сетки, что вынуждает использовать относительно грубые сетки.

В расчетах с успехом был использован характерный для задач с особенностями прием регуляризации, или выделения особенности. Он основан на том, что часто бывает известно решение близкой задачи, содержащее такую же особенность. Решение ищется в виде $\varphi(x, y) = S(x, y)\psi(x, y)$, где $S = e^{-\sqrt{x^2+y^2}}$. Смысл замены переменных заключается в том, что функция S удовлетворяет уравнению

$$-\Delta S = (1 - 1/\sqrt{x^2 + y^2})S,$$

отличающемуся от (7) младшими дифференциальными членами. Другими словами, особенность в потенциале $1/\sqrt{x^2 + y^2}$ порождает в решении слабую особенность типа $S(x, y)$. Поэтому ищем решение, уже содержащее такую особенность, полагая, что после замены переменных искомая функция ψ должна быть гладкой.

В результате простых преобразований, которые мы опустим, получаем для ψ уравнение

$$L_k\psi = \lambda\psi,$$

где дифференциальный оператор

$$L_k \equiv -\Delta + 2\left(\frac{x}{\sqrt{x^2+y^2}}\frac{\partial}{\partial x} + \frac{y}{\sqrt{x^2+y^2}}\frac{\partial}{\partial y}\right) - \left(\frac{1}{\sqrt{x^2+y^2}} - U(x, y) - 1\right) -$$

$$- 2i\left(k_1\frac{\partial}{\partial x} + k_2\frac{\partial}{\partial y}\right) + 2i\frac{k_1x + k_2y}{\sqrt{x^2+y^2}} + (k_1^2 + k_2^2).$$

«Потенциал» $1/\sqrt{x^2 + y^2} - U(x, y) - 1$ особенности при $x = y = 0$ уже не имеет. Коэффициенты типа $x/\sqrt{x^2 + y^2}$, $y/\sqrt{x^2 + y^2}$ также остаются ограниченными всюду. Заметим, что для функции $\psi(x, y)$ краевые условия не являются условиями периодичности, как это было для $\varphi(x, y)$.

Чтобы иметь дело с хорошо освоенной в вычислениях периодической задачей, будем трактовать замену переменных следующим образом. Мы имеем уравнение $H\varphi = \lambda\varphi$. Используя замену $\varphi = S\psi$, получаем $L\psi = \lambda\psi$. Обратная замена $\psi = e^{\sqrt{x^2+y^2}}\varphi$ дает

$$e^{-\sqrt{x^2+y^2}} L\, e^{\sqrt{x^2+y^2}}\varphi = \lambda\varphi, \quad \text{т.е.} \quad H = e^{-\sqrt{x^2+y^2}} L\, e^{\sqrt{x^2+y^2}}.$$

Практически это следует понимать так. Расчет ведется в терминах $\varphi_{m,n}$ — периодической, но «негладкой» функции, которая имеет «особенность» типа $e^{-\sqrt{x^2+y^2}}$. Действие на нее оператора H состоит из следующих операций. Функция φ умножается на $e^{\sqrt{x^2+y^2}}$ и превращается в гладкую функцию. На эту гладкую функцию действует разностный оператор, аппроксимирующий дифференциальный оператор L. Затем результат умножается на $e^{-\sqrt{x^2+y^2}}$. Относительно оператора H известно, что он самосопряженный.

Для вычисления главной собственной функции и собственного числа оператора H применяется степенной метод с регуляризатором B и параметром τ:

$$B\frac{\varphi^{i+1}-\varphi^i}{\tau} = (H - \lambda^i E)\,\varphi^i, \quad \lambda^{i+1} = (H\varphi^{i+1}, \varphi^{i+1})/(\varphi^{i+1}, \varphi^{i+1}),$$

или

$$\varphi^{i+1} = \varphi^i + \tau B^{-1}(H - \lambda^i E)\varphi^i.$$

Здесь i — номер итерации. В качестве оператора B использовались

$$B_1 \equiv E, \qquad B_2 \equiv \left(E - \sigma\frac{\partial^2}{\partial x^2}\right)\left(E - \sigma\frac{\partial^2}{\partial y^2}\right), \qquad B_3 \equiv (E - \sigma\Delta).$$

Разумеется, имеется в виду разностная аппроксимация дифференциального оператора. Оператор B_2 легко обращается (сначала прогонкой по x, затем по y). При этом используется периодическая прогонка А. А. Абрамова (см. § 18). Параметр σ подбирался экспериментально. Обращение оператора B_3 выполнялось с помощью быстрого преобразования Фурье (см. § 24).

Таблица 12 показывает сравнительную эффективность разных методов (i — число итераций для достижения заданной точности, t — машинное время, расходуемое на выполнение одной итерации; во всех расчетах погрешность $\varepsilon = 10^{-3}$). Расчеты, проведенные для разных сеток, показали, что собственные числа практически не зависят от N даже при сравнительно грубых сетках (типа 10×10). Это, конечно, — эффект удачной «регуляризации», т.е. аналитического выделения особенности.

Отметим, что особенность вводится в решение «мультипликативно», а не «аддитивно» (т.е. в решении используется замена типа

$\varphi = e^{-\sqrt{x^2+y^2}}\psi$, а не замена типа $\varphi = \varphi_0 + \psi$, где φ_0 — известное решение с особенностью). Это связано с характером вхождения особенности в уравнение: особенность входит в потенциал, умножающийся на φ. Если бы особенность входила в правую часть или в краевые условия («аддитивно»), то выделение ее в решении носило бы аддитивный характер (см. § 14).

Таблица 12

B	B_1	B_1	B_1	B_2	B_2	B_2	B_2	B_2	B_3	B_3
σ	–	–	–	0.05	0.025	0.0167	0.01	0.005	1	1
N	11	15	20	11	11	11	11	11	16	32
i	125	232	412	60	41	35	32	37	3	3
t	0.1	0.2	0.35	0.5	0.5	0.5	0.5	0.5	1.5	6

Исследование устойчивости стационарного состояния. Другим источником главных спектральных задач является важная в приложениях проблема устойчивости некоторых состояний среды. В общих чертах возникновение такой задачи можно представить в следующем виде. Временная эволюция некоторой системы описывается дифференциальным уравнением

$$\frac{\partial u}{\partial t} = L(u), \quad t \geqslant 0, \quad u(0) = u_0,$$

где L — нелинейный дифференциальный оператор. (Уравнение дополнено краевыми условиями, которые мы явно в наше достаточно поверхностное изложение не вводим.) Пусть состояние u_0 стационарное, т.е. $L(u_0) = 0$, и функция $u(t) \equiv u_0$ (на самом деле от t не зависящая) является решением уравнения. Предположим, что по каким-то причинам мы заинтересованы в длительном существовании этого состояния.

Возникает вопрос: возможно ли оно? Ведь система подвергается различным возмущениям, т.е. более точно ее поведение описывается уравнением

$$\frac{\partial u}{\partial t} = L(u) + \varepsilon f, \quad u(0) = u_0 + \varepsilon v_0,$$

где εf и εv_0 — малые возмущения. Судьба стационарного состояния существенно зависит от того, приведет ли наличие возмущений к столь же малому возмущению решения или следствием будет уход системы из состояния u_0. В последнем случае представляет интерес

и темп ухода, т.е. оценка времени, на котором разница между u_0 и $u(t)$ будет достаточно мала.

Исследование таких вопросов начинается обычно в линейном приближении. Уравнение линеаризуется, и для возмущения $\varepsilon v(t)$ получаем линейную краевую задачу:

$$\frac{\partial v}{\partial t} = L_u(u_0)\, v + f, \quad v(0) = v_0.$$

Здесь $L_u(u_0)$ — производная оператора $L(u)$ в точке u_0. Она вычисляется формальным дифференцированием по u входящих в $L(u)$ членов. Отметим, что коэффициенты $L_u(u_0)$ не зависят от времени. Следовательно, при $f \equiv 0$ решение можно найти методом Фурье:

$$v(t) = \sum_k c_k e^{\lambda_k t} \psi_k,$$

где ψ_k — собственные функции оператора $L_u(u_0)$, λ_k — соответствующие собственные значения, c_k — коэффициенты Фурье функции v_0.

Суждение об устойчивости состояния u_0 зависит от крайней правой точки спектра. (Наличие f не вносит существенных корректив, так как в этом случае решение ищется методом вариации произвольных постоянных и имеет тот же качественный характер, что и решение при $f = 0$.) Подчеркнем, что спектр задачи существенно зависит от исследуемой стационарной точки u_0. Исследования подобного рода в настоящее время активно проводятся в таких областях, как гидро- и газодинамика, физика плазмы. В последнем случае особенно важными являются исследования некоторых состояний плазмы в установках типа токамак, стелларатор и других, в которых физики надеются получить управляемую термоядерную реакцию.

Обычно исследования подобного рода составляют лишь начальный этап. Обнаружив неустойчивость исследуемого состояния, переходят к следующему этапу. Неустойчивость приводит к быстрому росту малого возмущения, и через короткое время уже нельзя пользоваться линеаризованной теорией: нужно переходить к решению полных эволюционных уравнений. Линейная теория дает в этом случае достаточно разумные начальные данные. На линейной стадии развития процесса неустойчивости из очень малого случайного возмущения εv_0 естественно выделяется наиболее быстро растущая компонента (именно ее определяет решение главной спектральной задачи). В первую очередь нужно рассмотреть последствия конечного возмущения именно той формы, которая соответствует главной собственной функции.

Конечно, возможности численных методов решения эволюционного нелинейного уравнения ограничены. Для их успешного применения необходима достаточная гладкость начальных данных v_0. В противном случае требуется слишком мелкий шаг сетки, и прове-

дение расчетов может оказаться даже невозможным. К счастью, в большинстве случаев ситуация благоприятная: главная собственная функция оказывается достаточно гладкой, имеющей небольшое число нулей. (В многомерном случае вместо числа нулей следует говорить о числе подобластей, в которых функция сохраняет знак.)

Расчет нестационарных процессов в ядерном реакторе. В настоящее время наиболее освоенной (с вычислительной точки зрения) задачей математической теории ядерных реакторов является расчет стационарного состояния, т.е. решение главной спектральной задачи. Однако все более актуальным становится расчет динамических процессов, происходящих в реакторе при изменении внешних условий его работы, например при изменении положений регулирующих стержней.

По существу речь идет о расчете процессов перехода реактора в новое стационарное состояние, хотя, конечно, имеется в виду и математическое моделирование аварийных ситуаций. Задачи подобного рода обычно решаются для упрощенных моделей реактора. В настоящее время разрабатываются методы решения нестационарных задач для столь же развитых и подробных моделей реактора, какие используются для расчета стационарных состояний. Обсудим некоторые вычислительные проблемы, возникающие в таких задачах, и возможные пути их преодоления.

Уравнения нестационарного процесса запишем в форме

$$\mathscr{E}\frac{\partial \Phi}{\partial t} = L(\alpha)\Phi, \qquad \frac{\partial \alpha}{\partial t} = A(\alpha)\Phi, \qquad \Phi(0) = \Phi^0, \quad t \geqslant 0. \tag{8}$$

Первое уравнение есть компактная запись системы (2), матрица $\mathscr{E}$ — диагональная. Наряду с полями нейтронов $\Phi = \{\Phi_1, \Phi_2\}$ (если используется модель с большим числом групп, размерность вектора Φ, соответственно, увеличивается) учитываются еще и поля, описывающие другие физические характеристики состояния реактора.

Для определенности будем считать, что α — скалярная функция, описывающая температуру (именно такая модель изучалась в расчетах, результаты которых иллюстрируют излагаемые здесь подходы), $L(\alpha)$ — дифференциальный оператор, коэффициенты которого D_i, $A_{i,j}$ в системе (8) зависят от α. Таким образом, коэффициенты системы (8) зависят от пространственных координат и t явно и неявно — через зависимость от α. В коэффициенты может входить явная зависимость от t, если рассматриваются, например, процессы регулирования положения стержней.

Сложности решения системы (8) связаны с тем, что $\mathscr{E}$ — это малый «векторный» параметр, т.е. система является сингулярно-возмущенной, описывающей взаимодействие процессов с существенно разными характерными временами. Если, например, для переменных Φ время $0.1 \div 1$ с является большим, то для температуры α вре-

мя $0.1 \div 1$ с является малым. (Во избежание недоразумений отметим, что все конкретные цифры относятся к тем расчетам, результаты которых будут приведены ниже.)

Рассчитываемый процесс был связан с проработкой проекта научно-исследовательского реактора, в котором при «холодном» состоянии ($\alpha^0 \approx 0$ °C, $\Phi^0 \approx 0$) внезапно (поднятием регулирующих стержней) создается надкритическая ситуация ($\lambda_1 \approx 10 \div 200$). Начальные данные Φ^0 определяются по существу «флуктуационным» фоном. Первый этап процесса (сравнительно длительный) происходит практически при постоянной температуре α^0, потоки нейтронов экспоненциально нарастают ($\Phi(t) \approx c_1 e^{\lambda_1 t}\psi_1$), но пока еще слишком малы, чтобы привести к заметному изменению температуры. На этом этапе из всех компонент ряда (4), входящих в начальный фон Φ^0, выделяется первая.

Постепенно $\Phi(t)$ достигает такого значения, что начинает изменяться и температура $\alpha(t)$, сначала медленно, потом все быстрее и быстрее. Но одновременно начинает изменяться и λ_1. Характер зависимости коэффициентов оператора $L(\alpha)$ таков, что λ_1 смещается при росте α в сторону отрицательных значений (отрицательная обратная связь между α и λ_1). Темп роста Φ (и, следовательно, α) замедляется. Наконец, λ_1 становится отрицательной, Φ начинает экспоненциально убывать, а $\alpha(t)$ выходит на некоторое стационарное состояние (около $1000 \div 2000$ °C).

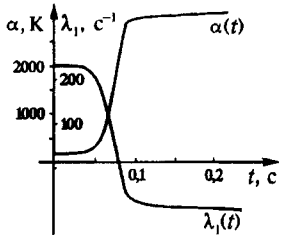


Рис. 20

Рисунок 20 иллюстрирует сказанное выше. На нем представлены $\lambda_1(t)$, и $\alpha(t)$ в точке максимума по пространственным переменным. Период начального разгона (от фона до значений Φ, вызывающих заметное изменение α) не рассчитывается. Вместо этого находится ψ_1 (главная спектральная задача!) и в качестве начальных данных Φ^0 берется $N\psi_1$, где N назначается на основании простых оценок $A(\alpha^0)$. Такое «волевое решение» приводит к некоторому неопределенному сдвигу математического времени t относительно физического. Этот сдвиг допускает несложную оценку и для приложений не очень важен. Расчету подлежит только переходной процесс, длительность которого порядка 1 с.

Опишем в общих чертах метод, который использовался в расчетах и позволил проинтегрировать уравнение (8) за $20 \div 30$ шагов по времени. Правда, шаг (переход от состояния в момент t к состоянию в момент $t + \tau$) предполагает выполнение достаточно сложной операции — решения главной спектральной задачи. В целом этот под-

ход оказался эффективным именно благодаря тому, что был использован очень быстрый способ нахождения главной собственной функции (многосеточный метод; см. § 14).

Основу метода составляет определение (разумеется, приближенное) искомого решения системы уравнений (11) из «асимптотического уравнения»

$$\frac{d\alpha}{dt} = A(\alpha)N_0 \exp\left(\int_0^t \lambda_1(t')\,dt'\right)\psi_1(t), \quad \alpha(0) = \alpha^0,$$

где $\psi_1(t)$ — главная собственная функция спектральной задачи

$$L[\alpha(t)]\,\psi = \lambda\mathscr{E}\psi;$$

λ_1 — соответствующая точка спектра; N_0 — нормирующий множитель, определяемый «начальными данными» (при $t = 0$ множитель N_0 выбирается таким, чтобы уже началось незначительное, но заметное изменение температуры α). Опуская некоторые технические детали, стандартный шаг численного интегрирования можно представить следующими операциями.

Пусть в некоторый момент времени t уже получены значения $\alpha(t)$ и $\int_0^t \lambda_1(t')\,dt'$. Определяются коэффициенты оператора $L(\alpha)$, решается главная спектральная задача, находятся $\psi_1(t)$ и $\lambda_1(t)$. После этого для вычисления $\alpha(t+\tau)$ делается один шаг по явной схеме Эйлера:

$$\alpha(t+\tau) = \alpha(t) + \tau A[\alpha(t)]N_0 \exp\left(\int_0^t \lambda_1\,dt'\right)\psi_1(t),$$

пересчитывается интеграл $\int_0^{t+\tau} = \int_0^t + \tau\lambda_1(t)$, и т.д.

Естественно возникает вопрос о правомочности такого приближения. Некоторые физически достаточно убедительные аргументы дает следующее рассуждение. Подставим используемую конструкцию в исходные уравнения задачи. Разумеется, они не будут выполнены: возникает некоторая «невязка». Если она очень мала относительно входящих в уравнение членов, то это в известной мере оправдает вышеизложенный подход. Обозначая для удобства $\Lambda(t) = \int_0^t \lambda_1\,dt'$, вычисляем

$$\left[\mathscr{E}\frac{\partial}{\partial t} - L(\alpha)\right]N_0 e^{\Lambda(t)}\,\psi_1(t) =$$

$$= N_0 e^{\Lambda(t)}\left[\lambda_1\mathscr{E}\psi_1 + \mathscr{E}\frac{\partial\psi_1}{\partial t} - L(\alpha)\psi_1\right] = N_0 e^{\Lambda(t)}\,\mathscr{E}\frac{\partial\psi_1}{\partial t}.$$

Итак, погрешность метода зависит от соотношения $\mathscr{E}\partial\psi_1/\partial t$ и членов, входящих в $L(\alpha)\psi_1$. (В эти выражения входят коэффициенты, погрешности измерения которых часто не так уж малы; кроме того, сама формулировка исходной задачи, которую мы принимаем за истину, в действительности основана на пренебрежении некоторыми относительно малыми величинами.) Заметим, что собственная функция ψ_1 определена с точностью до нормировки и этим фактором тоже можно разумно распорядиться с целью уменьшения погрешности. Можно обыграть два обстоятельства: величины v_1 и v_2 существенно разные (скорости быстрых и медленных нейтронов связаны соотношением $v_1 \approx 10^2 v_2$). Следовательно, важнее получить медленное изменение второй компоненты ψ_1.

Другое обстоятельство, которое можно было предвидеть на основе опыта расчетов в этой области: компоненты главной собственной функции ψ_1 (в зависимости от пространственных координат, которые в принятых обозначениях опущены) — просто устроенные функции «колоколообразной» формы. С течением времени (при изменении $\alpha(t)$) их форма меняется не очень существенно. Заметно меняется лишь отношение их амплитуд (т.е. соотношение между потоками быстрых и медленных нейтронов в холодном и горячем реакторе). Из сказанного выше следует рецепт нормировки: постоянной полагается гильбертова норма второй компоненты ψ_1. При численном решении задачи оценивалась и величина $\mathscr{E}\partial\psi_1/\partial t$: она фактически оказалась достаточно малой по сравнению с другими членами (около $1 \div 2$ %).

§ 17. Жесткие системы обыкновенных дифференциальных уравнений

В пятидесятых годах при решении задач Коши для систем, описывающих кинетику реагирующих друг с другом химических веществ, вычислители столкнулись с крайне неприятным явлением. Расчеты проводились с помощью хорошо отработанных программ, в которых использовались методы Рунге—Кутты и надежные алгоритмы автоматического выбора шага. Эти алгоритмы очень быстро вырабатывали шаг численного интегрирования, столь малый, что часто не было никакой возможности рассчитать процесс на требуемом для приложений отрезке времени, даже используя наиболее мощные ЭВМ той эпохи. Визуальный анализ правых частей, казалось, не давал оснований для каких-то опасений.

Типичная система уравнений химической кинетики выглядела (технические подробности опускаем) примерно так:

$$\dot{x}^i = \sum A^i_j x^j + \sum A^i_{jk} x^j x^k, \quad i, j, k = 1, 2, \ldots, N, \qquad (1)$$

где A_j^i, A_{jk}^i — константы, характеризующие скорости протекания тех или иных реакций. Бросалась в глаза, правда, существенная разница в их значениях: они отличались друг от друга часто на много порядков. В то же время исключить какие-то «малые» члены из (1), оставив только самые большие, было нельзя, ориентируясь лишь на значения A. Существенными были и концентрации разных веществ x^i: они могли очень сильно изменяться с течением времени.

Затрачивая значительное машинное время, удавалось получать начальные отрезки траекторий и провести анализ ситуации. Он выявил следующую характерную картину. В начале процесса происходит сильное изменение $x(t)$ и выбираемый программой шаг численного интегрирования вполне разумен: он очень мал, но так и должно быть для аккуратного интегрирования столь быстро меняющихся функций. Через небольшое время t характер траектории резко меняется, она становится гладкой, медленно меняющейся, но программа этого «не замечает» и выбирает такой же малый шаг. Попытки «подсказать» программе выбор существенно большего шага, согласованного с гладкостью решения, немедленно приводили к вычислительной катастрофе. В § 5, 7 специально указывалось, что при оценке вычислительной сложности задачи Коши для системы обыкновенных дифференциальных уравнений $\dot{x} = f(x)$ существенны два фактора: строение поля траекторий в окрестности интегрируемой траектории и свойства матрицы $f_x(x)$.

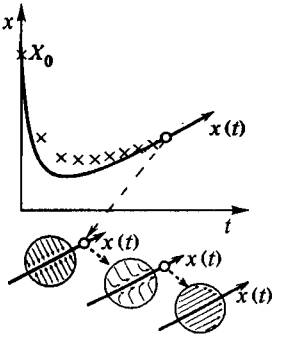


Рис. 21

Анализ поля направлений таких систем, получивших название «жестких», дал характерную картину, качественно представленную рис. 21. Траектория $x(t)$ состоит из короткого участка быстрого ее изменения (так называемого «пограничного слоя») и длительного участка очень медленной ее эволюции (иногда его называют «квазистационарным режимом»). Основные трудности связаны именно с расчетом последнего. Пограничный слой интегрируется очень малым шагом, но он настолько краток, что число шагов интегрирования вполне приемлемо. На рис. 21 при помощи «микроскопа» с последовательно увеличивающимся разрешением показана структура поля направлений в окрестности $x(t)$. Сначала видны траектории, отвесно падающие на $x(t)$. При следующем увеличении видно, что, приближаясь к $x(t)$, они поворачивают, стремясь двигаться «параллельно» $x(t)$. И лишь при еще большем увеличении видна стандартная картина практически параллельных линий.

Если из точки $x(t)$ траектории сдвинуться по касательной в точку $x^* = x(t) + \tau\dot{x}(t)$, то, хотя расстояние $x(t+\tau) - x^* = O(\tau^2)$ ничтожно, $f(x^*)$ не имеет ничего общего с $\dot{x}(t+\tau)$, направление $f(x^*)$ скорее напоминает перпендикуляр к траектории $x(t)$. То же самое получается и при определении x^* отрезком ряда Тейлора из трех-четырех и более членов при том значении τ, которое хотелось бы использовать для численного интегрирования квазистационарного режима.

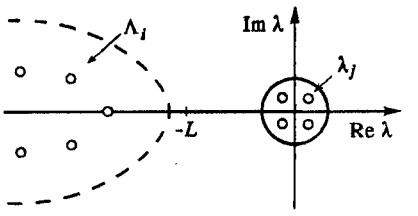


Рис. 22

Анализ матрицы $f_x(x)$ в окрестности траектории также привел к специфической картине, которую мы примем за основу при следующем формальном определении жесткой системы.

Определение 1. Задачу Коши

$$\dot{x} = f(x), \quad x(0) = X_0, \quad 0 \leqslant t \leqslant T, \quad x \in R^N \tag{2}$$

будем называть жесткой, если спектр матрицы $f_x(x)$ достаточно четко делится на две части (рис. 22).

Жесткий спектр. Собственные значения и векторы обозначим $\Lambda_i(x)$ и $\Phi_i(x)$ $(i = 1, 2, \ldots, I)$. Для жесткого спектра выполняются условия

$$\operatorname{Re}\Lambda_i(x) \leqslant -L, \qquad |\operatorname{Im}\Lambda_i(x)| < |\operatorname{Re}\Lambda_i(x)| \tag{3}$$

Мягкий спектр. Собственные значения и векторы обозначим $\lambda_j(x)$, и $\varphi_j(x)$ $(j = 1, 2, \ldots, J)$. Для мягкого спектра выполняются условия

$$|\lambda_j(x)| \leqslant l \ll L. \tag{4}$$

Время интегрирования T является средним относительно l и очень большим относительно L: lT равно, например, 10, 20, 30, а LT может быть порядка 10^3, 10^6 и больше. Отношение L/l называют показателем жесткости системы. В приложениях встречаются ситуации, когда L/l равно 10^6, 10^9, 10^{15}. Будем считать, что $l = 1$.

Стандартные методы интегрирования требуют шага τ^*, малого в том смысле, что $\tau^*\|f_x\| \ll 1$. Так как $\|f_x\| \sim |\Lambda_i| \approx L$, то $\tau^* \ll 1/L$ и расчет требует LT шагов численного интегрирования, что иногда просто неприемлемо. С точки зрения гладкости квазистационарной части траектории приемлем большой шаг, например $\tau \approx 10^{-3}T$. Основной проблемой в теории жестких систем является разработка ал-

горитмов численного интегрирования с таким большим шагом. Она была решена на основе неявных схем. Ниже мы объясним, почему оказалось достаточно такого относительно несложного вычислительного аппарата.

Заметим еще, что нелинейная система может быть жесткой в одной части фазового пространства и не быть таковой — в другой. Числа I, J следовало бы обозначать $I(x)$, $J(x)$ $(I + J = N)$.

Линейные жесткие системы. Этот удобный объект позволяет оценить возможности неявных схем, сравнивая точное решение с приближенным. Рассмотрим систему $\dot{x} = Ax$, $x(0) = X_0$, считая постоянную матрицу A жесткой (если $f(x) = Ax$, то $f_x = A$). Точное решение дается явной формулой

$$x(t) = \sum_i C_i e^{\Lambda_i t} \Phi_i + \sum_j c_j e^{\lambda_j t} \varphi_j. \tag{5}$$

Первое слагаемое убывает, как e^{-Lt}, и становится пренебрежимо малым вне пограничного слоя — интервала времени $\left[0, O\left(\frac{\ln L}{l}\right)\right]$; второе слагаемое представляет «квазистационарное» движение $x(t)$ (см. рис. 21).

Попытаемся интегрировать систему по явной схеме Эйлера (см. § 5):

$$(x_{n+1} - x_n)/\tau = Ax_n, \quad \text{или} \quad x_{n+1} = (E + \tau A)x_n. \tag{6}$$

Решение (6) легко получить в терминах спектра:

$$x_n = \sum C_i(1 + \tau\Lambda_i)^n \Phi_i + \sum c_j(1 + \tau\lambda_j)^n \varphi_j. \tag{7}$$

Для мягкой компоненты при шаге $\tau l \ll 1$ (на практике это означает, например, $\tau l \approx 0.1$) имеем $1 + \tau\lambda_j = e^{\tau\lambda_j}(1 + O(\tau^2 l^2))$ и

$$(1 + \tau\lambda_j)^n = e^{n\tau\lambda_j}(1 + O(\tau l)) \quad \text{при} \quad n \approx T/\tau.$$

Таким образом, мягкая компонента (7) аппроксимирует соответствующую компоненту точного решения (5) в обычном смысле слова. Для жесткой компоненты $(1 + \tau\Lambda)^n \approx (-\tau L)^n$, и при $\tau L \gg 1$ (а на практике это величины порядка $10^3 \div 10^6$ и т.д.) за несколько шагов числа x_n просто выходят из разрядной сетки ЭВМ.

Рассмотрим неявную схему:

$$(x_{n+1} - x_n)/\tau = Ax_{n+1}, \quad \text{или} \quad x_{n+1} = (E - \tau A)^{-1} x_n. \tag{8}$$

Точное решение (8) имеет вид

$$x_n = \sum C_i(1 - \tau\Lambda_i)^{-n} \Phi_i + \sum c_j(1 - \tau\lambda_j)^{-n} \varphi_j. \tag{9}$$

Мягкая компонента так же аппроксимирует соответствующую часть (5), а жесткая быстро стремится к нулю (как $(1/\tau L)^n$) и, таким образом, качественно правильно описывает поведение «погранслойного» слагаемого (5). Обычно в прикладных задачах подобного рода не представляют особого интереса ни структура пограничного слоя, ни его длительность (лишь бы она была много меньше T).

Наиболее интересным содержательным результатом является квазистационарный режим. Если это так, решение (9) нас устраивает. На рис. 21 крестиками показано удовлетворительное приближенное решение: толщина пограничного слоя (τ или 2τ, 3τ, ...) существенно больше его реальной толщины $O\left(\frac{\ln L}{L}\right)$. Структура слоя совершенно не описывается, но квазистационарный режим описан достаточно аккуратно. Кстати, если по каким-то причинам нас интересует и пограничный слой, его расчет малым шагом $\tau^* \ll 1/L$ не представляет труда и может быть выполнен по любой явной схеме.

При численном интегрировании линейной системы $\dot{x} = A(t)x$ с медленно меняющейся матрицей (в том смысле, что $\|\dot{A}\|\tau \ll \|A\|$) иногда используют квазианалитические методы численного интегрирования:

$$x_{n+1} = e^{A_n \tau} x_n. \tag{10}$$

Для их эффективной реализации нужно вычислять матричную экспоненту. В жесткой системе (при $\|A\|\tau \gg 1$) это не просто. Использовать ряд Тейлора практически нельзя как по «экономическим» причинам (предоставим читателю оценить необходимое число членов ряда, оно порядка τL), так и по причине вычислительной неустойчивости. В силу (3), (4) имеем $e^{A\tau} = O(1)$, но эта величина получается суммированием тех же величин, что и $e^{-A\tau} = O(e^{L\tau})$.

Метод удвоения аргумента. Для вычисления матричной экспоненты разработан достаточно эффективный алгоритм. Выберем некоторое целое p, такое, что $\|A\|\tau/2^p \ll 1$. Тогда $e^{A\tau/2^p} \approx B = E + (\tau/2^p)A$, а $e^{A\tau} \approx B^{2^p}$, последняя же матрица может быть вычислена за p умножений матриц (вычисляются значения $B_1 = B \times B$, $B_2 = B_1 \times B_1$ и т.д.). Очевидно, что $p = O(\ln(\tau L))$, и с точки зрения числа операций этот метод вполне эффективен. Однако его применение требует известной осторожности. В самом деле, мягкое собственное значение B есть $1 + (\tau/2^p)\lambda_j + \varepsilon$, где ε — относительная погрешность представления чисел в ЭВМ. Пусть, например, $\tau L/2^p = 1$. Тогда в B информация о λ_j передается с относительной погрешностью, не меньшей $|\varepsilon \tau L/\lambda_j|$. При некоторых вполне реальных соотношениях между жесткостью и длиной машинного слова величина $\beta_j^{2^p} = (1 + (\tau/2^p)\lambda_j)^{2^p}$ может не иметь ничего общего с

$e^{\lambda_i \tau}$ и квазистационарный режим будет рассчитан неверно. Плохо и то, что погрешность такого рода не имеет явных признаков: численное решение будет гладким и будет выглядеть внешне таким, каким оно могло бы быть. Поэтому численное интегрирование жестких систем на машинах типа IBM (четырехбайтное слово) ведется как минимум с двойной точностью.

Системные методы численного интегрирования. Вычисление матричной экспоненты используется в некоторых алгоритмах. В качестве характерного примера опишем метод, разработанный в Институте химической физики. Стандартный шаг численного интегрирования состоит из следующих операций.

Пусть точка x_n уже известна. Вычисляется матрица $A = f_x(x_n)$. Дальнейшее основано на некоторых преобразованиях. Сделаем замену переменных $x(t) = x_n + u(t)$. Для u имеем систему

$$\dot{u} - Au(t) = f(x_n + u) - Au(t) \tag{11}$$

с начальными данными $u(t_n) = 0$. Положим, для простоты, $t_n = 0$ и проинтегрируем (11) на интервале $[0, \tau]$, используя оператор $(d/dt - A)^{-1}$. Тогда

$$u(\tau) = e^{A\tau} \int_0^\tau e^{-At} \left[f(x_n + u(t)) - Au(t)\right] dt.$$

Это точная формула.

Приближенную формулу можно получить, взяв в правой части $u(\tau)$ вместо $u(t)$. Основания для этого следующие. Множитель e^{-At} есть величина, быстро возрастающая вправо. Поэтому основной вклад в интеграл дают значения на правом конце интервала интегрирования. Кроме того, по смыслу замены $u(t)$ есть малая величина:

$$f(x_n + u) \approx f(x_n) + f_x(x_n)u + O(\|u\|^2),$$

т.е. выражение в квадратных скобках почти (с точностью до $O(\|u\|^2)$) постоянно. С учетом указанной аппроксимации это выражение выносится из интеграла, после чего он явно вычисляется:

$$u(\tau) \approx e^{A\tau} \int_0^\tau e^{-At}\, dt \left[f(x_n + u(\tau)) - Au(\tau)\right] =$$

$$= A^{-1}(e^{A\tau} - E) \left[f(x_n + u(\tau)) - Au(\tau)\right].$$

Для определения функции $u(\tau)$ получено нелинейное уравнение, которое решается, как показал опыт, простыми итерациями типа $u^{(i+1)} = B\left[f(x_n + u^{(i)}) - Au^{(i)}\right]$. Но сначала нужно вычислить

матрицу $B = A^{-1}(e^{A\tau} - E)$. Дело осложняется тем, что часто A^{-1} не существует. Это естественное свойство тех систем, в которых компоненты x суть относительные концентрации некоторых веществ. Тогда система (1) имеет очевидный первый интеграл — закон сохранения $(x(t), e) = 1$, где $e = \{1, 1, \ldots, 1\}$. Дифференцируя по t, получаем $(\dot{x}, e) = (f(x), e) = 0$. Дифференцируя по x, находим $f_x(x)e = 0$, т.е. нуль есть точка спектра матрицы $f_x(x)$ при всех x, а e — соответствующий собственный вектор.

Определение B корректно (так как $(e^{A\tau} - E)e = 0$), и нужно правильно раскрыть неопределенность $\infty \cdot 0$. Для этого используется специфическая форма алгоритма удвоения аргумента. Она основана на следующих преобразованиях. Обозначим $A^{-1}(e^{A\tau} - E) = B_\tau$. Тогда

$$B_\tau = A^{-1}(e^{A\tau} - E) = A^{-1}(e^{A\tau/2} - E)(e^{A\tau/2} + E) =$$
$$= A^{-1}(e^{A\tau/2} - E)AA^{-1}(e^{A\tau/2} - E + 2E) = B_{\tau/2}(AB_{\tau/2} + 2E).$$

Эта формула позволяет эффективно вычислять B_τ.

Вернемся к алгоритму численного интегрирования. Вычисляется матрица

$$B\tau/2^p = A^{-1}(e^{A\tau/2^p} - E) \approx A^{-1}(E + (\tau/2^p)A - E) = (\tau/2^p)E.$$

Далее за p шагов удвоения аргумента находится B_τ. Итерациями вида, например,

$$u^{(i+1)} = B_\tau[f(x_n + u^{(i)}) - Au^{(i)}]$$

находится с заданной погрешностью $u(\tau)$. Скорость сходимости регулируется выбором шага τ. Если потребовалось слишком много итераций, следующий шаг интегрирования выполняется с меньшим шагом τ. Если итерации сходятся слишком быстро, шаг τ увеличивается.

Перейдем к более сложным задачам.

Сингулярно-возмущенные системы. Рассмотрим класс систем, также относящихся к числу «жестких», для которых уже давно создана асимптотическая теория и качественная картина поведения траекторий достаточно ясна. Это позволяет четко формулировать требования к методам приближенного интегрирования и проверять, в какой мере те или иные методы интегрирования таким требованиям удовлетворяют.

Итак, рассмотрим систему

$$\begin{aligned} &\dot{x} = Lf(x, y) \quad (\text{или } \varepsilon\dot{x} = f,\ \varepsilon = L^{-1} \ll 1), \\ &\dot{y} = \varphi(x, y). \end{aligned} \tag{12}$$

Здесь f, φ и их производные — величины порядка $O(1)$; x, y — векторы размерности I, J соответственно. Спектр вариационной матрицы (12) определяется уравнением

$$\det \begin{pmatrix} Lf_x - \lambda E_I & Lf_y \\ \varphi_x & \varphi_y - \lambda E_J \end{pmatrix} = 0. \tag{13}$$

Легко показать, что жесткая часть спектра определяется спектром матрицы f_x (умноженным на L, если, конечно, f_x не имеет собственных значений $O(L^{-1})$), а соответствующие собственные векторы Φ_i имеют по существу лишь x-компоненту (их y-компонента есть $O(L^{-1})$). Хорошо известна и качественная структура траекторий. Она определяется многообразием Г, уравнением которого является $f(x, y) = 0$. Оно разбивает фазовое пространство на две части: $f(x, y) > 0$ и $f(x, y) < 0$. Система (12) является жесткой в случае «отрицательности» спектра $f_x(x, y)$.

Теория систем вида (12) хорошо развита. На рис. 23 показана типичная картина в случае, когда x и y — скаляры. (В многомерном случае картина, конечно, более сложная, но примерно того же характера.) Вне малой $O(L^{-1})$-окрестности Г поле направлений почти горизонтально, фазовая скорость очень велика (порядка

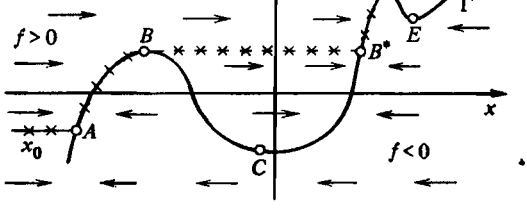


Рис. 23

$O(L)$). Она направлена вправо в области $f > 0$ и влево в области $f < 0$. За короткое время $O(L^{-1})$ система из точки (x_0, y_0) переходит в малую $O(L^{-1})$-окрестность Г. В этой окрестности $\dot{x} = O(1)$, $\dot{y} = O(1)$ (так как $f = O(L^{-1})$) и здесь осуществляется медленное движение фазового вектора $(x(t), y(t))$ вверх или вниз вдоль Г в зависимости от знака φ.

В зависимости от «знака» f_x на многообразии Г выделяются устойчивые и неустойчивые ветви. На рис. 23 устойчивые участки — это (A, B), (C, D), (E, F), неустойчивые — (B, C), (D, E). В окрестности последних система не является жесткой, так как спектр $f_x(x, y)$

не является «отрицательным», появляются собственные значения с положительной действительной частью. Наиболее интересные явления происходят в окрестности точки B (или D), когда теряется устойчивость, т.е. в окрестности B одно из собственных чисел $f_x(x, y)$ переходит в правую полуплоскость. При этом траектория в режиме «внутреннего слоя» за время $O(L^{-1})$ переходит в точку B^*.

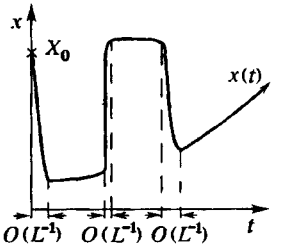


Рис. 24

В зависимости от знака φ на (C, D) траектория либо со скоростью $O(1)$ поднимается вверх до D, после чего быстро (за время порядка $O(L^{-1})$) переходит на устойчивую ветвь Г, либо спускается в точку C и в режиме пограничного слоя переходит в точку A, и т.д. При изображении графика траектории $\{x(t), y(t)\}$ получаем картину, в которой участки медленного движения сменяются быстрыми «скачками» с одного уровня на другой, траектория воспринимается почти как разрывная (рис. 24). Такие же «внутренние слои», разделенные длительными (порядка $O(T)$) промежутками спокойной эволюции, наблюдаются на траекториях уравнений химической кинетики и других аналогичных систем.

Численное интегрирование сингулярно-возмущенной системы с большим шагом. Согласимся, что при численном решении системы (12) с шагом $\tau \ll 1$ и $\tau L \gg 1$ (например, τL порядка 10^3, 10^6 и т.д.) допустимо лишь качественное воспроизведение слоев. Длительность численного слоя может быть порядка $O(\tau)$, что намного больше его действительной длительности порядка L^{-1}. Структуру же слоя численное решение совсем не описывает. Однако важно, чтобы участки медленного движения точки $x(t)$ были воспроизведены достаточно аккуратно. Таким образом, при численном интегрировании жестких систем используется обычное представление о близости приближенного и точного решений при всех t, за исключением малых (порядка $O(\tau)$) окрестностей слоев.

Посмотрим, что дает использование неявной схемы Эйлера:

$$\frac{x_{n+1}-x_n}{\tau} = Lf(x_{n+1}, y_{n+1}), \qquad \frac{y_{n+1}-y_n}{\tau} = \varphi(x_{n+1}, y_{n+1}). \tag{14}$$

Рассмотрим характерные ситуации.

Начало расчета (первый шаг), $n = 0$. Точка (x_0, y_0) находится «далеко» от Г, точка (x_1, y_1) находится из системы уравнений

$$f(x, y) - \frac{1}{L\tau}(x - x_0) = 0, \qquad y - y_0 - \tau\varphi(x, y) = 0. \tag{15}$$

Первое уравнение определяет J-мерное многообразие Γ^*, являющееся (в силу того, что $1/L\tau \ll 1$) слабым возмущением многообразия Γ. Точнее, ограничиваясь невырожденными ситуациями, можно утверждать, что точки Γ^* находятся в $O(1/L\tau)$-окрестности Γ. Второе уравнение определяет I-мерное многообразие, расположенное в $O(\tau)$-окрестности гиперплоскости $y - y_0 = 0$ (рис. 25). Согласно асимптотической теории уравнения (12) первое приближение (имеющее погрешность $O(L^{-1})$) к траектории на интервале пограничного слоя $[0 \leqslant t < O(L^{-1})]$ определяется системой уравнений

$$\dot{x} = Lf(x, y), \quad \dot{y} = 0, \quad x(0) = x_0, \quad y(0) = y_0. \qquad (16)$$

Правая граница пограничного слоя определяется выходом траектории, движущейся в $O(L^{-1})$-окрестности гиперплоскости $y = y_0$, в $O(L^{-1})$-окрестность Γ.

Таким образом, учитывая условность термина «правая граница пограничного слоя», мы все же можем утверждать, что за время $O(L^{-1})$ траектория (12) из точки (x_0, y_0) попадает в $O(L^{-1})$-окрестность корня системы уравнений $f(x, y_0) = 0$. Таких корней может быть много. Траектория же «выберет» из них один, который мы условно назовем первым. Вышеприведенные несложные оценки показывают, что в общем (невырожденном) случае среди корней системы (15) имеется корень, находящийся в $O(\tau)$-окрестности первого («истинного») корня и, тем самым, в $O(\tau)$-окрестности точки $(x(t^*), y(t^*))$, где $t^* = O(L^{-1})$ — правая граница пограничного слоя. И если используемый метод решения системы (15) (а это обычно метод Ньютона) даст именно нужный корень, с принятой здесь точки зрения результат нас вполне удовлетворит.

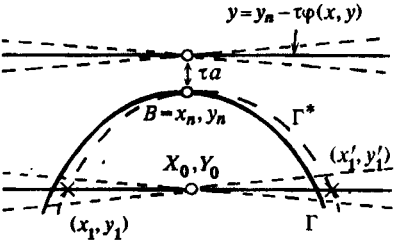


Рис. 25

Но система (15) может иметь и другие корни, в том числе на неустойчивой части Γ. Здесь мы сталкиваемся с потенциальной опасностью, возникающей при прохождении пограничного слоя «за один большой шаг». Наряду с «правильным» решением (x_1, y_1) (см. рис. 25), не исключена возможность получить принципиально неверные значения (x_1', y_1'), (x_1'', y_1'') и т.д., после чего даже точное интегрирование системы даст совершенно неверный результат. Особенно опасными являются корни на неустойчивых ветвях многообразия Γ.

В связи с вышесказанным становится понятным стремление специалистов, занимающихся численным решением жестких систем, интегрировать пограничные слои с малым шагом $\tau^*\|f_x\| \ll 1$. В этом случае поведение численной траектории достаточно аккуратно воспроизводит поведение точной и, затратив определенное машинное время, мы попадем в окрестность именно того решения (x_1, y_1), которое нужно. Однако с такой тактикой связана своя проблема: распознавание «начала» режима типа «слой» и его «конца». Шаг τ^* настолько мал, что преждевременный переход на этот шаг, так же как и запоздалый переход на большой шаг после того, как слой пройден, приводит к расходу машинного времени, часто недопустимому. Конечно, в самом начале расчета следует использовать малый шаг τ^*: практически любая траектория жесткой системы начинается пограничным слоем. Но в дальнейшем заранее неизвестно, когда произойдет очередной скачок. Часто шаг τ так велик, что схема проскакивает очередной слой за один шаг со всеми вытекающими отсюда последствиями.

Простой выход — обнаружив скачок от (x_n, y_n) к (x_{n+1}, y_{n+1}) (а он, очевидно, легко обнаруживается по большому изменению переменных за шаг), вернуться назад и интегрировать от точки (x_n, y_n) с малым шагом τ^* — едва ли может считаться удовлетворительным по причинам, указанным выше. Он может привести к огромному перерасходу машинного времени на чрезмерно точное вычисление траектории задолго до того, как в этом возникает действительная необходимость. Более разумная с практической точки зрения тактика состоит в том, что, если при шаге τ получено большое изменение фазовых переменных, следует вернуться назад, уменьшив шаг всего, скажем, в три раза; в дальнейшем можно поступать таким же образом. Аналогичную проблему представляет переход от малого шага в области слоя к очень большому после выхода траектории из этой области.

Заметим, что проводящиеся иногда «состязания» методов численного интегрирования жестких систем (какой из них быстрее решит ту или иную задачу-тест) являются не столько соревнованием вычислительных формул, сколько соревнованием алгоритмов регулирования шага. Выигрывают те алгоритмы, которые позже других переходят на малый шаг τ^* и раньше возвращаются к большому после прохождения пограничного слоя. К сожалению, задачи-тесты, на которых проводятся подобные состязания, обычно не содержат указанных нами опасностей — других, лишних ветвей многообразия Г. В этом случае излишняя «лихость» оказывается безнаказанной. Видимо, можно сконструировать хорошую задачу-тест, взяв какую-либо хорошо изученную сингулярно-возмущенную систему и «замаскировав» ее какой-то гладкой заменой переменных.

Расчет в окрестности Г. Пусть точка (x_n, y_n) находится в малой окрестности Г, т.е. в области $|\tau L f(x, y)| \ll |x|$. (Это естественное условие аккуратного интегрирования: x мало меняется за один шаг.) Тогда система уравнений неявной схемы (15) имеет решением пересечение многообразий Γ^* (слабое $O(1/\tau L)$-возмущение Г) и γ^*:

$$(x, y) \in \gamma^* = y - y_n - \tau\,\varphi(x, y) = 0$$

(слабое $O(\tau)$- возмущение прямой $y = y_n$). Среди таких точек есть точка, находящаяся на расстоянии $O(\tau)$ от точки (x_n, y_n). Поскольку именно последняя берется в качестве начального приближения в том или ином итерационном методе решения (15), естественно ожидать, что именно близкое к (x_n, y_n) решение будет получено. Остальные корни (15) лежат существенно дальше, и вероятность получить их хотя и существует, но, видимо, в большинстве случаев очень мала.

Что касается точности (в обычном смысле слова) воспроизведения численным решением эволюции системы (12) в окрестности Г, то и здесь ситуация достаточно благополучна. В самом деле, асимптотическая теория решений сингулярно-возмущенной системы (12) приводит к следующему. В первом приближении (с точностью до $O(L^{-1})$) траектория системы (12) совпадает с траекторией вырожденной системы

$$\dot{y} = \varphi(x, y), \quad f(x, y) = 0, \text{ т.е. } (x, y) \in \Gamma. \tag{17}$$

Используя неявную схему, мы в сущности интегрируем почти такую же систему. Единственное отличие состоит в том, что вместо $f(x, y) = 0$ используется условие $f(x, y) - (x - x_n)/\tau L = 0$, т.е., поскольку $x_{n+1} - x_n = O(\tau)$, численная траектория отходит от Г на $O(1/\tau L)$, что укладывается в точность первого приближения системы (17).

Расчет в точке «срыва в режим внутреннего слоя». Это один из наиболее сложных моментов в «жизни» траектории жесткой системы, и здесь мы ограничимся простейшим случаем, когда x и y — скаляры. В этом случае рис. 23 дает представление о поведении траектории. Итак, предположим, что точка (x_n, y_n) совпадает с «последней» точкой B на устойчивой ветви многообразия Г или очень близка к ней, например находится в $O(1/\tau L)$-окрестности точки B. Из дальнейшего станет ясно, в каком смысле можно ослабить это предположение, т.е. расширить окрестность B, в которой может находиться точка (x_n, y_n), с тем чтобы сохранился основной вывод — система (15) не имеет решений в малой окрестности B и точка (x_{n+1}, y_{n+1}) должна совершить большой скачок на другую ветвь Г.

Решение системы типа (15) есть пересечение линии Γ^*, лежащей в $O(1/\tau L)$-окрестности Γ, и линии $y - y_n - \tau\varphi(x, y) = 0$ ($O(\tau)$-возмущение прямой $y = y_n$). До попадания в точку B система двигалась по крайней левой ветви Γ^* вверх, т.е. на этой ветви $\varphi > 0$. В общем случае нет никаких оснований ожидать обращения φ в нуль в окрестности B, т.е. $\varphi(x_n, y_n) = a > 0$. Существенно то, что $a\tau \gg 1/\tau L$. Если φ удовлетворяет условию Липшица по x с постоянной C, то легко показать, что линия $y - y_n - \tau\varphi(x, y) = 0$ находится внутри узкого конуса с вершиной в точке $(x_n, y_n + \tau a)$ и с раствором $C\tau$ (этот конус показан на рис. 25).

В конечной окрестности точки B линия Γ^* лежит в области $y < y_n + O(1/\tau L)$; следовательно, линия Γ^* (вернее, та ее ветвь, которая проходит через точку (x_n, y_n)) не пересекается с конусом на расстоянии a/C. Итак, точка (x_{n+1}, y_{n+1}) может быть найдена лишь на конечном (существенно большем τ) расстоянии от (x_n, y_n). Конечно, можно надеяться, что это будет точка B^*, ближайшая к B, в малую окрестность которой попала бы и траектория системы, но нельзя игнорировать опасности попадания в точки B', B'', ..., что привело бы к конечной погрешности и, быть может, к принципиально недоступной (для траектории, начинающейся из точки (x_0, y_0)) области изменения фазовых переменных.

Расчет на неустойчивой ветви Γ. Нельзя исключить возможности того, что в силу каких-то причин точка (x_n, y_n) окажется в окрестности неустойчивой ветви Γ. Здесь система (12) уже не является жесткой (в принятом смысле), так как среди собственных значений матрицы $f_x(x_n, y_n)$ имеются значения с положительной действительной частью. Такая точка является для системы (12) неустойчивой ($L \gg 1$). Траектория очень быстро (за время $O(L^{-1})$) уходит от этой части Γ, попадая либо на устойчивую ветвь Γ, либо в бесконечность. В какой-то мере здесь мы имеем ситуацию, аналогичную уже рассмотренной на с. 219: ведь проведенный там анализ никак не был связан со свойствами спектра матрицы f_x.

В малой $O(\tau)$-окрестности точки (x_n, y_n) имеется решение системы уравнений неявной схемы (15), и в принципе мы можем получить именно ее, что уже приводит к неверному результату: численная траектория конечное время $O(1)$ будет находиться в окрестности неустойчивой ветви Γ. Реализуется эта возможность или нет, зависит от итерационного процесса, используемого для решения системы нелинейных уравнений (15). Насколько нам известно, ситуация здесь такая. Если используется хороший, быстро сходящийся

процесс, например метод Ньютона, то совершенно неважно, какое из решений системы (15) отыскивается — на устойчивой или на неустойчивой ветви Г. В этом случае мы сталкиваемся с крайне неприятной ситуацией: ветвь Г неустойчивая для дифференциального уравнения, становится устойчивой для разностного уравнения (15) (разумеется, при интегрировании с большим шагом).

Можно предложить итерационный процесс решения (15), различающий устойчивую и неустойчивую ветви Г, т.е. такой, который сходится в первом случае и расходится во втором. Но едва ли он представит какой-нибудь практический интерес, так как сходимость процесса будет столь медленной, что затраты машинного времени на решение (15) приблизят такой метод к интегрированию с малым шагом $\tau^*\|f_x\| \approx 1$. Возможность построения итерационного процесса, быстро сходящегося на устойчивой ветви и автоматически расходящегося на неустойчивой, сомнительна.

А-устойчивые разностные схемы. Переходим к общему случаю, предупредив, что здесь нет полной ясности. Качественный характер решения в этом случае в какой-то мере аналогичен тому, что было в сингулярно-возмущенной системе. В решении выделяются резкие кратковременные скачки, перемежающиеся длительными участками спокойного течения процесса. При этом скачки происходят на коротких отрезках времени, много меньших шага интегрирования на «спокойных» участках. Дело осложняется тем, что характерные объекты, явно выделенные в сингулярно-возмущенных системах (разделение компонент системы на быстрые и медленные, зоны пограничных слоев, уравнение поверхности квазистационарного решения, условие его устойчивости), в общем случае уже не допускают такого выделения; они замаскированы, их аналитическое описание или очень сложно, или даже неизвестно.

При конструировании разностных схем для интегрирования жестких систем с большим шагом τ в настоящее время принято удовлетворять следующим требованиям:

а) схема должна аппроксимировать дифференциальное уравнение в обычном смысле слова (см. § 4—6);

б) схема должна обладать специфической устойчивостью типа A-, $A(\alpha)$-, L-устойчивости (смысл этого требования разъясняется ниже);

в) схема должна пройти практическую проверку решением ряда общепризнанных задач-тестов.

Обратимся к A-устойчивости. Имеется в виду исследование поведения численного решения простейшего уравнения $\dot{x} = \lambda x$, полученного с помощью рассматриваемой схемы с большим шагом τ. Начнем с примера одной из схем типа Адамса:

$$x_{n+2} - \frac{4}{3}x_{n+1} + \frac{1}{3}x_n - \frac{2}{3}\tau f(x_{n+2}) = 0. \qquad (18)$$

Это — неявная схема. Вычисление x_{n+2} требует решения нелинейной системы уравнений. Так как $\tau\|f_x\| \gg 1$, ее нельзя рассматривать как малое возмущение тривиальной системы, получающейся из (18) при $\tau = 0$. Для $f(x) = \lambda x$ решение (18) находится в виде $x_n = C_1 q_1^n + C_2 q_2^n$, где q_1, q_2 — корни характеристического уравнения

$$q^2\left(1 - \frac{2}{3}\tau\lambda\right) - \frac{4}{3}q + \frac{1}{3} = 0.$$

Они легко вычисляются:

$$q_1 = \frac{2+\sqrt{1+2\xi}}{3-2\xi}, \qquad q_2 = \frac{2-\sqrt{1+2\xi}}{3-2\xi}, \qquad \xi = \lambda\tau. \tag{19}$$

Далее нужно исследовать поведение решения (18) для тех значений ξ, которые представляют интерес при интегрировании жестких систем. В плоскости комплексного переменного ξ выделяются две характерные области, которые должны покрыть спектр жесткой системы. Первую область называют «областью точности». В нее входят малые значения ξ, при которых решение разностного уравнения аппроксимирует точное решение $e^{\lambda t}$ в обычном смысле слова. Легко показать, что $q_1(\xi) = e^{\xi}(1 + O(|\xi|^3))$ при $|\xi| \ll 1$. Следовательно, $C_1 q_1^n = C_1 e^{n\tau\lambda}(1 + O(|\xi|^2))$ при $n \leqslant T/\tau$. Второе слагаемое $C_2 q_2^n$ быстро стремится к нулю, так как $q_2 \approx 1/3$. Постоянные C_1 и C_2 определяются заданием дополнительных начальных данных x_1. Эту величину следует задать так, чтобы было $C_1 = X_0 + O(\tau^2)$, только тогда схема имеет второй порядок точности. Обычно зону точности исследуют (привлекая численные методы) подробнее: выделяют зоны, в которых $\ln q_1(\xi)$ совпадает с ξ с погрешностью 1 %, 2 % и т.д. Чем шире подобные зоны, тем точнее схема.

Вторая область — область устойчивости, где $|q_1| < 1$, $|q_2| < 1$. Желательно, чтобы эта область покрывала значительную часть полуплоскости $\operatorname{Re}\xi < 0$. Легко проверить, что $|q_1| \approx |q_2| \approx 1/\sqrt{2|\xi|}$ при $|\xi| \gg 1$. Можно найти такое значение R, что $|q_1| < 1$, $|q_2| < 1$ при $|\xi| > R$. Представляет интерес граница области устойчивости — линия $\max\{|q_1(\xi)|, q_2|\xi|\} = 1$. Считается необходимым, чтобы зона устойчивости содержала какую-то достаточно широкую окрестность линии $\operatorname{Im}\xi = 0$, $\operatorname{Re}\xi < 0$. В частности, если область устойчивости есть полуплоскость $\operatorname{Re}\xi < 0$, схему называют A-устойчивой. Доказана теорема о том, что A-устойчивыми могут быть только неявные схемы не выше второго порядка аппроксимации. Схему называют $A(\alpha)$-устойчивой, если область устой-

чивости содержит конус $|\operatorname{Im}\xi| \leqslant \sin\alpha\,|\operatorname{Re}\xi|$ $(\operatorname{Re}\xi < 0)$. Схему называют L-устойчивой, если в области $\operatorname{Re}\xi < -a^2$ $(a \neq 0)$ решения разностного уравнения убывают, как q^n, где $q < 1$ и не зависит от ξ.

Безытерационные схемы типа схемы Розенброка. В последние годы была предложена некоторая общая конструкция схем интегрирования жестких систем, в которых система нелинейных уравнений не решается. Рассмотрим пример подобной схемы. Стандартный шаг интегрирования состоит из следующих операций. Пусть имеется точка x_n. Вычисляется матрица $A = f_x(x_n)$, и x_{n+1} находится из уравнения

$$(E - \alpha\tau A - \beta\tau^2 A^2)\,\frac{x_{n+1} - x_n}{\tau} = f(x_n + \gamma\tau\, f(x_n)). \qquad (20)$$

Таким образом, шаг стоит двух вычислений f, вычисления A и решения системы линейных уравнений. Параметры α, β, γ подбираются так, чтобы обеспечить возможно более высокий порядок аппроксимации и необходимую устойчивость.

Проиллюстрируем характерную технику подбора параметров. Разложим решение в ряд Тейлора в точке t_n:

$$x(t_n + \tau) = x(t_n) + \tau f_n + \frac{\tau^2}{2}(f_x f)_n + \frac{\tau^3}{6}(f_{xx}ff + f_x f_x f)_n. \qquad (21)$$

Здесь $f_n = \dot{x}(t_n)$, $(f_x f)_n = \ddot{x}(t_n)$, $(f_{xx}ff + f_x f_x f)_n = \dddot{x}(t_n)$. Остальные члены ряда опущены. Для схемы можно написать аналогичное разложение. Из (20) следует, что

$$x_{n+1} = x_n + (E - \alpha\tau A - \beta\tau^2 A^2)^{-1} f(x_n + \gamma\tau\, f(x_n)) = x_n + \tau f_n +$$

$$+ \tau^2(\alpha + \gamma)Af + \tau^3(\beta + \alpha^2 + \alpha\gamma)A^2 f + \tfrac{1}{2}\gamma^2\tau^3 (f_{xx}ff)_n. \qquad (22)$$

Распорядимся параметрами так, чтобы все выписанные члены (21) и (22) совпали. Тем самым будет обеспечен третий порядок аппроксимации. В результате мы получаем систему уравнений

$$\alpha + \gamma = 1/2, \qquad \beta + \alpha(\alpha + \gamma) = 1/6, \qquad \gamma^2 = 1/3,$$

которая легко решается. Приведем числовые значения: $\alpha = 1.077$, $\beta = -0.372$, $\gamma = -0.577$. (Решение с $\gamma = 0.577$ неинтересно.)

Перейдем к анализу устойчивости, используя схему (20) для уравнения $\dot{x} = \lambda x$. После несложных преобразований имеем

$$x_{n+1} = q(\xi)x_n, \qquad \xi = \tau\lambda,$$

где

$$q(\xi) = (1 - 0.077\xi - 0.205\xi^2)/(1 - 1{,}077\xi + 0.37\xi^2).$$

Простой анализ показывает, что $|q(\xi)| < 1$ при $\operatorname{Im}\xi = 0$, $\operatorname{Re}\xi < 0$. Легко проверить, что при $|\xi| > 10$ также $q(\xi) < 1$. Более точные сведения о границе области устойчивости можно получить численно.

Отметим, что схема «устойчива» и в большей части правой полуплоскости. В этом случае качественные поведения траекторий дифференциального и разностного уравнений принципиально различны. Это не относится, разумеется, к области точности.

Регулярные жесткие системы. Изучение жестких систем обнаружило их большое сходство с сингулярно-возмущенными. Сложилось впечатление, что жесткую систему можно получить из сингулярно-возмущенной, сделав гладкую замену переменных. При такой замене теряется четкое разделение переменных на быстрые и медленные (x и y в (12)), маскируется то многообразие Г, около устойчивых ветвей которого происходит медленное движение фазовой точки (из системы (12) мы сразу получаем для Г уравнение $f(x, y) = 0$). Отсутствие асимптотической теории для общих жестких систем, аналогичной теории сингулярно-возмущенных систем, затрудняет разработку и оценку численных методов. Теория объясняет, какой должна быть траектория и чего мы вправе требовать от численного метода.

Опишем возможный вариант такой асимптотической теории. Основным ее объектом является многообразие Г, определяемое уравнениями

$$(f(x), \Phi_i^*(x)) = 0, \quad i = 1, 2, \ldots, I. \tag{23}$$

Здесь $\Phi_i^*(x)$ — собственные векторы матрицы $f_x^*(x)$, соответствующие точкам жесткого спектра. Хотя это определение не конструктивно (так как оно оперирует с трудно вычисляемыми объектами), тем не менее в теоретическом анализе его использовать удалось.

Было показано, что в малой окрестности Г движение происходит так же, как и в окрестности поверхности $f(x, y) = 0$ для системы (12): все траектории очень быстро (со скоростью $O(L)$) входят в $O(L^{-2})$-окрестность Г и движутся в ней со скоростью $O(1)$. Удалось построить «обратную» замену переменных, сводящих общую жесткую систему к сингулярно-возмущенной, разложить x на быструю и медленную компоненты и выписать уравнения их эволюции.

Имея достаточно ясное представление о том, как устроена траектория, оказалось возможным дать достаточно аккуратное обоснование некоторых разностных схем и получить оценку погрешности приближенного решения в терминах двух малых параметров: τ и

$l/L\tau$. Это характерное обстоятельство: теория численного интегрирования жестких систем не может рассматривать предельного перехода при $\tau \to 0$. Поэтому теорема «аппроксимация + устойчивость (A, $A(\alpha)$, L и т.п.) $\Rightarrow$ сходимость» здесь не действует. Точнее, теорема справедлива, но бесполезна, так как нужно обосновать метод не в пределе при $\tau \to 0$, а совсем в другой области значений τ, когда малы оба упомянутых выше параметра.

Теория, оперирующая только с аппроксимацией и A-устойчивостью, принципиально не полна. Это иллюстрирует следующий пример явной A-устойчивой схемы, аппроксимирующей линейную жесткую систему $\dot{x} = Ax$:

$$(x_{n+1} - x_n)/\tau = \beta(x_n, \tau) A x_n, \tag{24}$$

где $\beta(x_n, \tau) = (e^{r(x_n)\tau} - 1)/(r(x_n)\tau)$, а $r(x)$ есть отношение Рэлея в точке x: $r(x) = (Ax, x)/(x, x)$. Легко проверить аппроксимацию: при $\tau \to 0$, очевидно, $\beta = 1 + O(\tau)$. Так же несложно проверить A-устойчивость: если x — скаляр, $Ax \equiv \lambda x$, $\operatorname{Re} \lambda < 0$, то $r(x) = \lambda$, $\beta = (e^{\lambda\tau} - 1)/(\lambda\tau)$ и из (24) получаем $x_{n+1} = e^{\lambda\tau} x_n$.

Можно ли на этом основании утверждать, что хотя бы для линейных жестких систем построен явный A-устойчивый алгоритм численного интегрирования? Видимо, нет. Проанализируем вычисления по схеме (24). Если разложение точки x_n в сумму по собственным векторам A содержит существенную жесткую компоненту (при интегрировании в области пограничного слоя), то $r(x_n) \approx -L$ и при $L\tau \gg 1$ величина $\beta \approx 1/L\tau$, т.е. формула (24) превращается в интегрирование с малым шагом $\tau^* = \beta(x, \tau)\tau \approx 1/L$. Следует только иметь в виду, что и время нужно интегрировать по формуле, аналогичной (24): $t_{n+1} = t_n + \beta\tau$. Пройдя слой, траектория x попадает в область, где в разложении x_n вклад жесткой компоненты пренебрежимо мал. В этом случае $r(x_n) = O(1)$, $\beta = O(1)$ и делается шаг действительно с большим τ. Но это сразу же приводит к росту в x_{n+1} жесткой компоненты, фактический шаг $\beta\tau$ снова падает, и т.д. Трудно оценить эффективность такого адаптирующегося алгоритма. Как показали исследования В. И. Лебедева, явные схемы имеют некоторые возможности и при интегрировании жестких систем. Однако его теория основана на достаточно сложных построениях последовательности чередующихся малых и больших шагов. Она имеет самую непосредственную связь с устойчивыми последовательностями параметров в методе чебышевского ускорения итераций (см. § 14).

B-теория численного интегрирования. Изложим основные понятия и результаты развиваемой в последние годы специальной теории численных методов для жестких систем (2). Класс изучаемых

систем выделяется важной количественной характеристикой правой части, называемой *односторонней константой Липшица*. Предполагается, что f удовлетворяет условию

$$(f(y) - f(x),\ y - x) \leqslant l\|y - x\|^2, \quad l = O(1). \tag{25}$$

Величина l считается имеющей порядок $O(1)$, в то же время классическая константа Липшица или, что почти то же самое, $\|f_x\|$ может быть сколь угодно большой величиной: $L \gg l$.

Целью B-теории является получение таких оценок точности численного решения, которые не зависят от больших констант L, а сформулированы в терминах только односторонней константы Липшица l. Разумеется, эти оценки должны зависеть от гладкости искомого точного решения системы (2). В дальнейшем будем обозначать точное решение $X(t)$. Эта функция предполагается гладкой в том смысле, что

$$X(t) = O(1), \quad \dot{X}(t) = O(1), \quad \ddot{X}(t) = O(1), \quad \ldots \tag{26}$$

Другими словами, столько производных точного решения, сколько нужно при проведении тех или иных оценок, считаются величинами порядка $O(1)$. Следовательно, речь идет об интегрировании системы вне слоев. Это предположение согласуется с предшествующим анализом. Мы имеем дело с гладкой траекторией, со всех сторон окруженной существенно негладкими траекториями. При этом окружающие траектории содержат кратковременные участки, на которых их производные очень большие ($O(L)$), и являются гладкими вне этих тонких слоев.

Одностороннее условие Липшица гарантирует важное свойство множества траекторий системы. Пусть $X(t)$, $Y(t)$ — две такие траектории. Оказывается, они не могут сильно расходиться с течением времени. В самом деле, оценим

$$\frac{d}{dt}\|Y(t) - X(t)\|^2 = \frac{d}{dt}(Y-X,\ Y-X) =$$

$$= 2(\dot{Y}-\dot{X},\ Y-X) = 2(f(Y)-f(X),\ Y-X) \leqslant 2l\|Y - X\|^2.$$

Отсюда

$$\|Y(t) - X(t)\| \leqslant \|Y(0) - X(0)\|\, e^{lt}. \tag{27}$$

Если бы мы использовали классическую константу Липшица, мы имели бы в экспоненте показатель Lt, разрешающий существенно более сильно «разбегание» траекторий. При $l < 0$ системы (2), (25) хорошо известны в теории дифференциальных уравнений под именем «диссипативных». В этом случае все траектории с ростом t сближаются, т.е. обладают свойством «аттрактивности».

Общая теория B-сходимости. Анализ точности численного интегрирования жестких систем оперирует со следующими основными объектами:

1. Разностная схема, записанная, для определенности, в виде

$$x_{n+1} = x_n + \tau\,\Phi(\tau, x_n, x_{n+1}). \tag{28}$$

Здесь x_n — приближенное решение в точке $t_n = n\tau$. Схема неявная; Φ, конечно, тем или иным способом выражается через f.

2. Ограничение на сетку точного решения $X_n = X(t_n)$.

3. Погрешность аппроксимации (невязка α_{n+1}), которая получается при подстановке $\{X_n\}$ в разностное уравнение:

$$\alpha_{n+1} \equiv \frac{X_{n+1} - X_n}{\tau} - \Phi(\tau, X_n, X_{n+1}). \tag{29}$$

4. Погрешность согласования (эта характеристика активно используется в западной литературе по численным методам) определяется следующим образом. Решим разностное уравнение (28), взяв в качестве x_n точное значение X_n. Получим некоторую величину Z. Тогда погрешность согласования есть

$$\gamma_{n+1} \equiv X_{n+1} - Z, \quad \text{т.е.} \quad Z = X_{n+1} - \gamma_{n+1}.$$

Выпишем уравнение для γ_{n+1} в схеме (28):

$$Z = X_n + \tau\,\Phi(\tau, X_n, Z),$$

или

$$X_{n+1} - \gamma_{n+1} = X_n + \tau\,\Phi(\tau, X_n, X_{n+1} - \gamma_{n+1}). \tag{30}$$

Величины α_n и γ_n очень близки друг к другу. Если схема явная, они просто совпадают с точностью до множителя τ (в этом читатель может легко убедиться сам). В обычной, нежесткой, ситуации, когда $|\tau\,\Phi_x| \ll 1$, $\alpha_n\tau$ и γ_n по существу почти совпадают; в жестком случае это достаточно различающиеся объекты. Погрешность аппроксимации α_n сравнительно легко вычисляется и оценивается, погрешность согласования γ_n труднее оценивается, но ее проще использовать в доказательстве сходимости.

5. Глобальная погрешность $\varepsilon_n = X_n - x_n$. Смысл этой величины достаточно ясен, и ее оценка — цель теории.

Теперь дадим определения.

Определение 2. Схема называется B-согласованной порядка p, если для погрешности согласования установлена оценка

$$\|\gamma_n\| \leqslant C_1\tau^{p+1}, \tag{31}$$

причем оценка равномерна на всем изучаемом классе систем, т.е. C_1 зависит от постоянных, ограничивающих производные точного решения (26), и l в оценке (25), т.е. $C_1 = O(1)$.

Определение 3. Разностная схема (вида (28) или еще какого-либо) называется B-устойчивой, если для любых двух решений $\{x_n\}$, $\{y_n\}$ установлено соотношение

$$\|y_{n+1} - x_{n+1}\| \leqslant (1 + C_2\tau)\,\|y_n - x_n\|. \tag{32}$$

Постоянная C_2 зависит от l в (25) и является величиной $O(1)$ независимо от «жесткости» системы $L \approx \|f_x\|$.

Определение 4. Разностная схема является B-сходящейся порядка q, если для приближенного решения установлена оценка

$$\|\varepsilon_n\| \equiv \|x_n - X_n\| \leqslant C_3\tau^q, \quad \forall\, n = 1, 2, \ldots, T/\tau.$$

Постоянная C_3 не должна зависеть от большой константы L, хотя и может содержать множитель типа e^{CT} при $C = O(1) \ll L$.

Определение 5. Схема называется B-аппроксимирующей порядка p, если для погрешности аппроксимации α_n установлена равномерная на классе жестких систем оценка

$$\|\alpha_n\| \leqslant C_4\tau^p.$$

Вышеприведенные определения аналогичны стандартным. Специфической их чертой является только равномерность на классе жестких систем. Теорема Лакса «B-согласованность + B-устойчивость $\Rightarrow$ B-сходимость» является, как и знакомая нам уже теорема Рябенького—Филиппова, почти прямым следствием предположений.

Теорема 1. Пусть схема B-согласована порядка p и B-устойчива. Тогда схема является B-сходящейся с тем же порядком p.

Доказательство. Используя выражение для глобальной погрешности

$$\varepsilon_{n+1} = X_{n+1} - x_{n+1}$$

и связь X_{n+1} с величиной Z, полученной решением уравнения (30): $X_{n+1} = Z - \gamma_{n+1}$, имеем $\varepsilon_{n+1} = Z - x_{n+1} - \gamma_{n+1}$. Следовательно,

$$\|\varepsilon_{n+1}\| \leqslant \|Z - x_{n+1}\| + \|\gamma_{n+1}\|.$$

Но Z и x_{n+1} суть решения разностного уравнения, только в момент времени t_n одно стартует из точки X_n, другое — из точки x_n. В силу B-устойчивости для этих величин справедливо соотношение

$$\|Z - x_{n+1}\| \leqslant (1 + C_2\tau)\|X_n - x_n\| \leqslant (1 + C_2\tau)\|\varepsilon_n\|.$$

Учитывая (31), находим основную оценку для эволюции глобальной погрешности:

$$\|\varepsilon_{n+1}\| \leqslant (1 + C_2\tau)\|\varepsilon_n\| + C_1\tau^{p+1}.$$

Отсюда (как и в § 7) получаем результат:

$$\|\varepsilon_n\| \leqslant \frac{C_1}{C_2} e^{C_2 T} \tau^p, \quad \forall\, n \leqslant T/\tau. \tag{33}$$

Итак, построена абстрактная, достаточно тривиальная теория. Ее смысл — в выделении тех основных свойств (B-согласованности и B-устойчивости), которые надо устанавливать при исследовании конкретных схем.

Обратим внимание на то, что здесь используются несколько иные объекты, чем в теории, изложенной в § 7. Первое отличие уже было отмечено: вместо погрешности аппроксимации используется погрешность согласования. Это более «объективная», хотя и труднее оцениваемая величина. В самом деле, легко понять, что погрешность согласования зависит от метода, но не от формы записи разностных уравнений, которая может быть разной для одного и того же метода. Так, неявная схема $x_{n+1} = x_n + \tau f(x_{n+1})$ может быть записана в явном виде: $x_{n+1} = x_n + \tau F(\tau, x_n)$, с «неявным» определением F.

Другое отличие связано с понятием устойчивости. Если в § 7 используется устойчивость по правой части, то здесь, — так сказать, «устойчивость по начальным данным». При сравнении x_{n+1} с y_{n+1} можно считать, что их различие есть следствие различия «начальных данных» x_n и y_n. Вещи эти очень близкие, в § 12 уже отмечалось, что обычно из устойчивости по начальным данным следует устойчивость по правым частям. В некоторых случаях эти два понятия легко связать. Например, для неявной схемы учет погрешности аппроксимации приводит к сравнению решений двух разностных уравнений

$$x_{n+1} = x_n + \tau f(x_{n+1}), \qquad y_{n+1} = y_n + \tau f(y_{n+1}) + \alpha_{n+1}.$$

Однако можно трактовать α как возмущение «начальных данных для n-го шага», переписав второе уравнение в виде

$$y_{n+1} = (y_n + \alpha_n) + \tau f(y_{n+1}).$$

B-аппроксимация. В стандартной теории численного интегрирования установление аппроксимации, использующее предположение о гладкости искомого решения, является тривиальным упражнением, решаемым простыми разложениями в ряд Тейлора. Но установление B-аппроксимации не тривиально, и некоторые схемы, аппроксимирующие уравнение в обычном смысле, свойством B-аппро-

ксимации просто не обладают. Поясним это парадоксальное на первый взгляд обстоятельство.

Вычисление погрешности аппроксимации состоит в подстановке в разностные уравнения ограничения на сетку точного решения (предполагаемого гладким). Возьмем явную схему Эйлера и вычислим невязку $\alpha = (X_{n+1} - X_n)/\tau - f(X_n)$. Так как

$$X_{n+1} = X(t_n + \tau) = X_n + \tau \dot{X}(t_n) + \frac{\tau^2}{2} \ddot{X}(t_n + \theta\tau), \quad \dot{X}(t_n) = f(X_n),$$

то здесь все в порядке: в оценке участвуют только $\ddot{X}$ в какой-то момент времени, а эта величина по предположению есть $O(1)$.

То же самое относится и к неявной схеме Эйлера. Однако известная схема «трапеция» допускает две в обычном случае равноценные модификации. В одном варианте

$$\alpha = \frac{X_{n+1} - X_n}{\tau} - \frac{f(X_n) + f(X_{n+1})}{2}.$$

В этом случае можно использовать интегральное тождество

$$X_{n+1} - X_n = \int_{t_n}^{t_n+\tau} f[X(t)]\, dt.$$

Так как $f[X(t)] = \dot{X}(t)$ — гладкая функция, то

$$\int_{t_n}^{t_n+\tau} f[X(t)]\, dt = \tau \frac{f[X(t_n)] + f[X(t_{n+1})]}{2} + O(\tau^3);$$

и здесь все в порядке: $\alpha = O(\tau^2)$.

Однако в другом варианте:

$$\alpha = \frac{X_{n+1} - X_n}{\tau} - f\left(\frac{X_n + X_{n+1}}{2}\right),$$

аналогичные оценки не проходят. В самом деле,

$$\frac{X_{n+1} + X_n}{2} = X\left(t_n + \frac{\tau}{2}\right) + O(\tau^2) = X_{n+1/2} + O(\tau^2),$$

где $X_{n+1/2} = X(t_n + \tau/2)$. Обозначение $O(\tau^2)$ мы употребляем в дальнейшем только для величин, допускающих оценку типа $C\tau^2$, где $C = O(1)$ и не зависит от жесткости системы. В данном случае эта величина зависит только от гладкости кривой $X(t)$.

Оценим теперь

$$f\left(\frac{X_n + X_{n+1}}{2}\right) = f(X_{n+1/2} + r),$$

где

$$r = \frac{X(t_n) + X(t_n + \tau)}{2} - X\left(t_n + \frac{\tau}{2}\right) = O(\tau^2)$$

(в силу гладкости $X(t)$). Дальнейшие оценки дают

$$f(X_{n+1/2} + r) = f(X_{n+1/2}) + f_x(X_{n+1/2} + \theta r)\, r.$$

Последнее слагаемое не есть $O(\tau^2)$, хотя $r = O(\tau^2)$; величина f_x вычисляется не на траектории, да и вообще она уже зависит от жесткости системы. Оценки типа $f_x O(\tau^2)$ в B-теории не принимаются. Считается, что такие величины могут быть сколь угодно большими. Конструирование разностных схем, обладающих свойством B-аппроксимации, не так просто, как могло бы показаться.

Схемы интерполяционного типа. Рассмотрим пример схемы, для которой удается сравнительно просто доказать требуемые B-теорией свойства. Опишем стандартный шаг численного интегрирования — переход от x_n в момент t_n к x_{n+1} в момент $t_n + \tau$. Основу схемы составляет аппроксимация очевидного тождества (для точного решения)

$$x(t_n + \tau) = x(t_n) + \int_{t_n}^{t_n+\tau} f[x(t)]\, dt. \tag{34}$$

Обозначим $f[t] \equiv f[x(t)]$ и вычислим интеграл по какой-нибудь хорошей квадратурной формуле, считая пока $f[t]$ известной функцией.

Сделаем замену переменных $t = t_n + \xi\tau$ ($\xi \in [0, 1]$) и введем на $[0, 1]$ сетку узлов $0 < \xi^1 < \xi^2 < \ldots < \xi^s < 1$. Обозначим $t^i = t_n + \xi^i\tau$, $f^i \equiv f[t^i]$, $l_i(\xi)$ — интерполяционный базис на сетке $\{\xi^i\}$, состоящий из полиномов степени $s - 1$, определяемых свойствами $l_i(\xi^j) = \delta_i^j$. Построим интерполяционный полином Лагранжа:

$$\mathscr{L}(\xi) = \sum_{i=1}^{s} f^i\, l_i(\xi), \tag{35}$$

и вычислим приближенно

$$\int_{t_n}^{t_n+\tau} f[t]\, dt \approx \tau \int_0^1 \mathscr{L}(\xi)\, d\xi.$$

Если $f[t]$ — гладкая функция, погрешность квадратуры есть $O(\tau^{s+1})$. (Сделав замену $t = t_n + \xi\tau$, мы строим полином для функ-

ции $\tilde{f}[\xi] \equiv f[t_n + \xi\tau]$. В оценку точности интерполяции полиномом степени $s-1$ входит s-я производная $\tilde{f}[\xi]$ по ξ, которая, очевидно, есть s-я производная по t, умноженная на τ^s. Еще одна степень τ добавляется при интегрировании.) Теперь формула (35) переписывается в виде стандартной механической квадратуры:

$$x_{n+1} = x_n + \tau \sum_{j=1}^{s} b^j f^j, \tag{36}$$

где b^j зависят только от сетки $\{\xi^i\}_{i=1}^s$.

Пока этой формулой мы воспользоваться не можем, так как значения f^j не известны. Но $f^j = f[x(t^j)]$, а промежуточные значения $x(t^j)$ (обозначим их Y^j) могут быть приближенно вычислены точно так же:

$$Y^j = x_n + \int_{t_n}^{t_n+\tau\xi^j} f[t]\,dt. \tag{37}$$

Используя для вычислений интегралов тот же интерполяционный полином, получаем серию соотношений:

$$Y^i = x_n + \tau \sum_{j=1}^{s} a_{ij} f(Y^j), \quad i = 1, 2, \ldots, s. \tag{38}$$

В результате мы имеем следующий алгоритм численного интегрирования:

а) зная x_n, решаем систему $s \dim x$ нелинейных уравнений (38) относительно неизвестных Y^i; здесь же вычисляются и $f(Y^i)$ $(i = 1, 2, \ldots, s)$;

б) вычисляем x_{n+1} по формуле (36).

Подчеркнем, что коэффициенты схемы $\{a_{ij}\}$, $\{b_j\}$ не зависят от вида системы и ее размерности. Они зависят только от выбора узлов ξ^j.

Почти очевидна следующая теорема.

Теорема 2. Схема интерполяционного типа (36), (38) обладает свойством B-аппроксимации порядка s.

Доказательство. Вычислим невязку в (36) при подстановке в это соотношение сеточной функции X_n:

$$\tilde{a} = X_{n+1} - X_n - \tau \sum_{j=1}^{s} b^j f[X(t^j)].$$

Но

$$X_{n+1} - X_n = \int_{t_n}^{t_n+\tau} f[X(t)]\, dt,$$

а $\tau \sum b^j f[X(t^j)]$ есть интеграл от интерполяционного полинома степени $s-1$, построенного для гладкой функции $f[X(t)] = \dot{X}(t)$. Таким образом, эта сумма отличается от $X_{n+1} - X_n$ на величину $O(\tau^{s+1})$ (погрешность интерполяции $O(\tau^s)$ интегрируется по интервалу длиной τ). Итак, невязка $\tilde{\alpha} = O(\tau^{s+1})$. Точно таким же образом оцениваем невязки в соотношениях (38). Обозначая $X^j = X(t^j)$, имеем

$$\alpha^i = X^i - X_n - \tau \sum_j a_{ij} f(X^j), \quad \text{т.е. } \alpha^i = O(\tau^{s+1}).$$

Следующий характерный шаг в B-теории исследования схем — оценка погрешности согласования, причем не всякая схема, обладающая B-аппроксимацией порядка s является B-согласованной порядка s. В принципе порядок согласованности может понизиться и даже стать равным нулю.

Исследование B-согласованности. Схема (36), (38) является характерным примером неявных схем Рунге—Кутты, обладающих свойствами B-согласованности и B-устойчивости. Для таких схем типичны следующие признаки.

а) Вспомогательные величины Y^i вычисляются по формулам типа (38). Если в (38) $a_{ij} = 0$ при $j \geqslant i$, метод оказывается явным. Если $a_{ij} = 0$ при $j > i$, метод называют однократно неявным. В этом случае Y^i вычисляются последовательно. Вычисление каждого вектора Y^i требует решения системы $\dim x$ нелинейных уравнений. Рассматриваемая интерполяционная схема является полностью неявной. Все значения Y^i находятся одновременно решением системы $s \dim x$ нелинейных уравнений.

б) После вычисления величин Y^i, f^i шаг интегрирования осуществляется по формуле (36), в которой сумма аппроксимирует интеграл в (34) с погрешностью $O(\tau^{s+1})$. Таким образом, формальный порядок аппроксимации есть s.

Перейдем теперь к вопросу о B-согласованности. Это исследование требует оценки погрешности согласования γ на гладком решении жесткой системы. Пусть $X(t)$ — точное гладкое решение, а $X_n = X(t_n)$. Сделаем один шаг по схеме (36), (38) и вычислим ве-

личину Z по формуле

$$Z = X_n + \tau \sum_{j=1}^{s} b^j f(Y^j),$$

где Y^j получены обычным образом из (38) (с заменой x_n на X_n). Подлежащая оценке величина есть

$$\gamma_{n+1} = X_{n+1} - Z.$$

Для ее оценки вычислим X_{n+1} тем же самым способом, т.е. выпишем для нее систему соотношений типа (36), (38). Конечно, это будет не в точности та же самая система: она будет отличаться на некоторые погрешности аппроксимации. Возьмем вместо Y^j значения $X(t_n + \xi^j\tau)$, обозначая их через X^j. Подставим X^j в (38), добавляя соответствующие невязки:

$$X^i = X_n + \tau \sum_{j=1}^{s} a_{ij} f(X^j) + \alpha^i, \qquad i = 1, 2, \ldots, s,$$

где $\alpha^i = O(\tau^{s+1})$. В результате мы получили уравнения для величин, подлежащих сравнению ($i = 1, 2, \ldots, s$):

$$\begin{aligned} Y^i &= X_n + \tau \sum_j a_{ij} f(Y^j), & X^i &= X_n + \tau \sum_j a_{ij} f(X^j) + \alpha^i, \\ Z &= X_n + \tau \sum_j b^j f(Y^j), & X_{n+1} &= X_n + \tau \sum_j b^j f(X^j) + \tilde{\alpha}. \end{aligned} \tag{39}$$

Нас интересует оценка $\|Z - X_{n+1}\|$, для $\tilde{\alpha}$ имеем оценку $O(\tau^{s+1})$. Дальше должна работать стандартная схема. Из того, что системы (39) мало отличаются друг от друга, следует заключение, что их решения $\{Y^1, \ldots, Y^s, Z\}$ и $\{X^1, \ldots, X^s, X_{n+1}\}$, которые мы теперь обозначим Y и X соответственно (их размерности, очевидно, равны $(s+1)\dim x$), также мало отличаются друг от друга и можно получить оценку для погрешности согласования:

$$\|\gamma_{n+1}\| = \|Z - X_{n+1}\| \leqslant C\|\alpha\|, \quad \text{т.е.} \quad \|\gamma\| = O(\tau^{s+1}). \tag{40}$$

Эта оценка была бы тривиальной, если бы малый параметр τ был настолько мал, что $\|f_x\|\tau \ll 1$. Но мы имеем дело с принципиально иной ситуацией: $\tau\|f_x\| \gg 1$. Система существенно нелинейная, и нужное соотношение (40) удается получить за счет специальных свойств f: используя диссипативность или одностороннюю константу Липшица $l = O(1)$. Вообще говоря, соотношение (40) справедливо далеко не для всех схем. Это — особое свойство лишь некоторых

схем, которые считаются согласно B-теории пригодными для численного интегрирования жестких систем. Для этого свойства вводится соответствующий термин.

Определение 6. Разностная схема называется BS-устойчивой, если из (39) следует (40).

Итак, BS-устойчивость — это новое для нас свойство разностной схемы, связывающее погрешность согласования с погрешностью аппроксимации. Напомним, что B-теория оперирует именно с погрешностью согласования. Погрешность аппроксимации, вычисление которой тривиально в стандартном случае (когда искомое решение — гладкая функция, а f и все ее необходимые производные суть величины $O(1)$), в жестком случае оценивается на гладком решении далеко не просто. После этого предстоит сложная работа по оценке погрешности согласования.

Исследование BS-устойчивости. Это один из важнейших элементов исследования схемы. Выделим основные факторы, на основе которых удается установить BS-устойчивость. Вычитая уравнения (39), получаем ($i = 1, 2, \ldots, s$)

$$\begin{aligned} X^i - Y^i &= \tau \sum_j a_{ij} \{f(X^j) - f(Y^j)\} + \alpha^i, \\ X_{n+1} - Z &= \tau \sum_j b^j \{f(X^j) - f(Y^j)\} + \tilde{\alpha}. \end{aligned} \tag{41}$$

Введем векторы

$$V = \{X^1 - Y^1, \ldots, X^s - Y^s\},$$

$$W = \{f(X^1) - f(Y^1), \ldots, f(X^s) - f(Y^s)\}.$$

Каждая компонента этих векторов есть вектор размерности $N = \dim x$; таким образом, $V, W \in R^{Ns}$. Введем еще единичную матрицу e в N-мерном пространстве. Определим матрицы

$$A = \{a_{ij}e\}_{i,j=1}^{s}, \qquad B = \{b^1 e, b^2 e, \ldots, b^s e\}. \tag{42}$$

Очевидно, A — матрица $Ns \to Ns$, B — матрица $Ns \to N$. Теперь уравнения (41) можно записать в виде

$$V = \tau AW + \alpha, \qquad \gamma = \tau BW + \tilde{\alpha}, \qquad \alpha = \{\alpha^1, \alpha^2, \ldots, \alpha^s\}. \tag{43}$$

Ключевым фактом, используемым для установления BS-устойчивости, является соотношение

$$(V, W) \leqslant l \|V\|^2 \tag{44}$$

(l — односторонняя константа Липшица для f). Оно устанавливается прямой проверкой:

$$(V, W) = \sum_{j=1}^{s} (X^j - Y^j, f(X^j) - f(Y^j)) \leqslant \sum_{j=1}^{s} l \|X^j - Y^j\|^2 = l \|V\|^2.$$

Перепишем (43) в виде

$$A^{-1}V = \tau W + A^{-1}\alpha. \tag{45}$$

Умножая (45) скалярно на V, получаем

$$(A^{-1}V, V) = \tau(V, W) + (A^{-1}\alpha, V).$$

С учетом (44) имеем оценку

$$(A^{-1}V, V) \leqslant \tau l \|V\|^2 + \|A^{-1}\| \, \|\alpha\| \, \|V\|.$$

Выше неявно использовалось предположение о невырожденности A. Примем теперь более жесткое предположение о том, что матрица A^{-1} определяет метрику в том смысле, что существуют положительные C_1, C_2, для которых справедливо

$$C_1 \|V\|^2 \leqslant (A^{-1}V, V) \leqslant C_2 \|V\|^2, \quad \forall V. \tag{46}$$

Перестановкой строк и столбцов матрицу A можно привести к блочно-диагональному виду, в котором матрица представляет собой N стоящих на диагонали блоков, каждый из которых есть матрица $\{a_{ij}\}$ (размером $s \times s$). Очевидно, свойство (46) достаточно проверять только для такого блока, т.е. справедливость (46) зависит только от узлов ξ^j и не зависит от размерности N решаемой системы.

Используя (46), получаем оценку

$$C_1 \|V\|^2 \leqslant \tau l \, \|V\|^2 + C_2 \|\alpha\| \, \|V\|.$$

Отсюда следует $\|V\|(C_1 - \tau l) \leqslant C_2 \|\alpha\|$ и при $\tau l < C_1$ имеем

$$\|V\| \leqslant \frac{C_2}{C_1 - \tau l} \|\alpha\|. \tag{47}$$

Теперь можно оценить $\tau\|W\|$ из (45):

$$\tau \|W\| \leqslant \|A^{-1}\| \, \|V\| + \|A^{-1}\| \, \|\alpha\| \leqslant C_3 \|\alpha\|.$$

И наконец, получаем окончательный результат:

$$\|\gamma_{n+1}\| \leqslant \|B\| \, \tau \, \|W\| + \|\tilde{\alpha}\| = O(\|\alpha\| + \|\tilde{\alpha}\|).$$

Итак, при условии (46) схема является BS-устойчивой в том смысле, что погрешность согласования есть величина $O(\tau^{p+1})$, где

p — наиболее низкий порядок B-аппроксимации (его называют стадийным порядком аппроксимации). Выше была использована и ограниченность $\|B\|$, но это есть факт общего характера в отличие от требования (46), которое для одних схем выполняется, для других — нет.

Перейдем к следующему моменту B-теории — к исследованию B-устойчивости. Здесь тоже есть некоторая общая схема анализа.

Исследование B-устойчивости. Оно состоит в оценке поведения величины $\|x_n - \tilde{x}_n\|$ для двух разных решений разностной схемы. Выпишем уравнения для x_n и $\tilde{x}_n$ $(i = 1, 2, \ldots, s)$:

$$Y^i = x_n + \tau \sum_j a_{ij} f(Y^j), \qquad \tilde{Y}^i = \tilde{x}_n + \tau \sum_j a_{ij} f(\tilde{Y}^j),$$
$$x_{n+1} = x_n + \tau \sum_j b^j f(Y^j), \qquad \tilde{x}_{n+1} = \tilde{x}_n + \tau \sum_j a_{ij} f(\tilde{Y}^j). \tag{48}$$

Вычитая их, получаем $(i = 1, 2, \ldots, s)$

$$\tilde{Y}^i - Y^i = (\tilde{x}_n - x_n) + \tau \sum_j a_{ij} \{f(\tilde{Y}^j) - f(Y^j)\},$$
$$\tilde{x}_{n+1} - x_{n+1} = (\tilde{x}_n - x_n) + \tau \sum_j b^j \{f(\tilde{Y}^j) - f(Y^j)\}. \tag{49}$$

Используем определенные выше векторы V, W и матрицы A, B. Введем $\delta_n = \tilde{x}_n - x_n$, а также матрицы E и $\tilde{B}$ (типа $N \to Ns$ и $Ns \to Ns$ соответственно):

$$E = \{e, e, \ldots, e\}^T, \qquad \tilde{B} = \begin{pmatrix} b^1 e & & & 0 \\ & b^2 e & & \\ & & \ddots & \\ 0 & & & b^s e \end{pmatrix}.$$

В этих терминах можно записать (49) в виде

$$\text{а)}\ V = E\delta_n + \tau AW, \qquad \text{б)}\ \delta_{n+1} = \delta_n + \tau BW. \tag{50}$$

Приступим к оценке. Из (50б) имеем

$$\|\delta_{n+1}\|^2 = \|\delta_n + \tau BW\|^2 = (\delta_n + \tau BW, \delta_n + \tau BW) =$$
$$= \|\delta_n\|^2 + 2\tau (\delta_n, BW) + \tau^2 (BW, BW).$$

Проделаем простые и очевидные преобразования:

$$\|\delta_{n+1}\|^2 = \|\delta_n\|^2 + 2\tau\,(B^*\delta_n, W) + \tau^2\,(BW, BW) =$$

$$= \|\delta_n\|^2 + 2\tau\,(B^*\delta_n - \tilde{B}V + \tilde{B}V, W) + \tau^2\,(BW, BW) =$$

$$= \|\delta_n\|^2 + 2\tau\,(\tilde{B}V, W) - 2\tau\,(\tilde{B}V - B^*\delta_n, W) + \tau^2\,(BW, BW).$$

Для слагаемого $(\tilde{B}V, W) = \sum_j B^j\,(\tilde{Y}^j - Y^j, f(\tilde{Y}^j) - f(Y^j))$, ограничиваясь, для простоты, диссипативными системами (с неположительной односторонней константой Липшица $l \leqslant 0$) и принимая существенное, но весьма естественное предположение о положительности всех коэффициентов квадратурной формулы ($b^j \geqslant 0$, $\forall\, j$), получаем очевидную оценку

$$(\tilde{B}V, W) \leqslant 0. \tag{51}$$

Если для схемы имеет место оценка

$$\|\tilde{Y}^i - Y^i\| \leqslant C_4\,\|\tilde{x}_n - x_n\|, \quad i = 1, 2, \ldots, s, \tag{52}$$

с постоянной $C_4 = O(1)$ на всем рассматриваемом классе жестких систем, вместо (51) имеем оценку типа

$$(\tilde{B}V, W) \leqslant C_5\,\|\delta_n\|^2, \quad C_5 = O(1), \tag{53}$$

которая тоже может быть использована для установления B-устойчивости.

Из соотношения (50а), предполагая обратимость A, находим

$$W = \frac{1}{\tau}\,A^{-1}(V - E\delta_n). \tag{54}$$

Подставим это выражение в полученную для $\|\delta_{n+1}\|^2$ формулу:

$$\|\delta_{n+1}\|^2 = \|\delta_n\|^2 + 2\tau(\tilde{B}V, W) - 2(\tilde{B}V - B^*\delta_n, A^{-1}(V - E\delta_n)) +$$

$$+ (BA^{-1}(V - E\delta_n), BA^{-1}(V - E\delta_n)). \tag{55}$$

Используя почти очевидное соотношение $B^* = \tilde{B}E$ и преобразуя третий член правой части (обозначим $z = V - E\delta_n$) имеем

$$2(\tilde{B}V - B^*\delta_n, A^{-1}(V - E\delta_n)) = 2(\tilde{B}V - \tilde{B}E\delta_n, A^{-1}(V - E\delta_n)) =$$

$$= 2(\tilde{B}z, A^{-1}z) = (\tilde{B^*}A^{-1}z, z) + ((A^{-1})^*\tilde{B}z, z).$$

Наконец, представим последний член правой части (55) в форме

$$(BA^{-1}z, BA^{-1}z) = ((A^{-1})^*B^*BA^{-1}z, z).$$

Используя эти преобразования, запишем окончательное выражение:

$$\|\delta_{n+1}\|^2 = \|\delta_n\|^2 + 2\tau\,(\tilde{B}V, W) - (Qz, z), \tag{56}$$

где

$$Q = \tilde{B}^* A^{-1} + (A^{-1})^* \tilde{B} - (A^{-1})^* B^* B A^{-1}. \tag{57}$$

Итак, если система диссипативна, коэффициенты квадратуры $b^j \geqslant 0$, матрица A имеет обратную матрицу A^{-1} и самосопряженная матрица Q неотрицательна, то разностная схема обладает свойством аттрактивности. Для любых двух решений x_n и $\tilde{x}_n$, полученных по этой схеме, имеет место соотношение

$$\|\tilde{x}_{n+1} - x_{n+1}\| \leqslant \|\tilde{x}_n - x_n\|,$$

что сильнее требуемого свойства B-устойчивости. Если, кроме того, имеет место соотношение (52), то можно получать B-устойчивость на более широком классе систем — с положительной односторонней константой Липшица $l = O(1)$. В этом случае, очевидно, мы придем к соотношению

$$\|\tilde{x}_{n+1} - x_{n+1}\| \leqslant (1 + C\tau)\,\|\tilde{x}_n - x_n\|. \tag{58}$$

Интерполяционная схема с гауссовыми узлами. Выше был указан общий способ построения полностью неявных разностных схем типа Рунге—Кутты. Отмечалось, что выбор узлов интерполяционной формулы ξ^j является резервом, разумно распоряжаясь которым можно получить полезные свойства схемы. Подтвердим это положение, выбирая в качестве узлов ξ^j так называемые гауссовы узлы.

Как известно, при любом выборе узлов ξ^j ($j = 1, 2, \ldots, s$) квадратурная формула точна в классе полиномов степени $s - 1$. Имея в своем распоряжении s свободных параметров ξ^j, можно выбрать их так, что квадратурная формула станет точной в классе полиномов степени $(s-1) + s = 2s - 1$. Такие узлы называются гауссовыми (для них существуют таблицы). Если использовать схему интерполяционного типа именно с этими узлами, B-устойчивость такой схемы доказывается просто и не требуется проверять свойства сложно определяемой матрицы Q. Другими словами, B-устойчивость оказывается закономерным свойством схемы интерполяционного типа с гауссовыми узлами.

Предварительно докажем одно полезное свойство схем интерполяционного типа, справедливое при любых, в сущности, узлах ξ^j. Напомним, что в такой схеме фигурируют значения Y^j и

$f^j = f(Y^j)$ $(j = 1, 2, \ldots, s)$, определяемые решением системы уравнений (38), и интерполяционный полином $\mathscr{L}(t)$, вычисляемый по узлам t^j и значениям f^j. Расширим эту информацию, добавив узел $t^0 = t_n$, и зададим значения $Y^0 = x_n$, Y^j $(j = 1, 2, \ldots, s)$. По сеточной функции $\{t^j, Y^j\}$ $(j = 0, 1, \ldots, s)$ построим интерполяционный полином $p(t)$ степени s: $p(t^j) = Y^j$ $(j = 0, 1, \ldots, s)$. Имеет место следующее утверждение.

Утверждение. Полином $p(t)$ в точках t^j $(j = 1, 2, \ldots, s)$ удовлетворяет исходному дифференциальному уравнению:

$$\dot{p}(t^j) = f[p(t^j)], \quad j = 1, 2, \ldots, s. \tag{59}$$

Доказательство. Имеем очевидное тождество

$$p(t^i) - p(t^0) = \int_{t^0}^{t^i} \dot{p}(t)\, dt.$$

Но $p(t^i) = Y^i$, $P(t^0) = x_n$, $\dot{p}(t)$ — полином степени $s - 1$; интеграл точно вычисляется по квадратурной формуле с коэффициентами τa_{ij}. Итак,

$$Y^i = x_n + \tau \sum_{j=1}^{s} a_{ij}\, \dot{p}(t^j), \quad i = 1, 2, \ldots, s.$$

Видно, что величины $\dot{p}(t^j)$ и $f^j = f(Y^j) = f[p(t^j)]$ удовлетворяют одной и той же системе уравнений (38) (линейной относительно этих величин). Предполагая невырожденность матрицы A, получаем совпадение:

$$\dot{p}(t^j) = f^j = f[p(t^j)], \quad j = 1, 2, \ldots, s.$$

Кроме того, очевидно, что $\dot{p}(t) \equiv \mathscr{L}(t)$, так как $\dot{p}$ и $\mathscr{L}$ — полиномы степени $s - 1$, совпадающие в s точках t^j $(j = 1, 2, \ldots, s)$.

Таким образом, схему интерполяционного типа можно получить так называемым методом коллокации. Решение на интервале $[t_n, t_n + \tau]$ ищется (приближенно) в виде полинома степени s, для определения которого используются условия коллокации (т.е. выполнения уравнения в нескольких точках t^j, число которых, естественно, связывается с числом свободных параметров полинома):

$$\dot{p}(t^j) = f[p(t^j)], \quad j = 1, 2, \ldots, s.$$

К этому добавляется, конечно, условие $p(t^0) = x_n$.

Пусть теперь x_n и $\tilde{x}_n$ — два разных решения, полученных по схеме (36), (38) с гауссовыми узлами t^j. Используем полиномы $p(t)$ и $\mathscr{L}(t)$, определенные выше для решения x_n (аналогичные полиномы для $\tilde{x}_n$ обозначим $\tilde{p}(t)$ и $\tilde{\mathscr{L}}(t)$). Как оказалось, для этих полиномов точки t^j являются точками коллокации, т.е. выполняются соотношения ($j = 1, 2, \ldots, s$)

$$p(t_n) = x_n, \qquad p(t_n + \xi^j\tau) = Y^j, \qquad \dot{p}(t_n + \xi^j\tau) = f(Y^j).$$

Такие же соотношения, естественно, выполняются и для второго решения.

Рассмотрим величину

$$r(t) = \|\tilde{p}(t) - p(t)\|^2 = (\tilde{p}(t) - p(t),\ \tilde{p}(t) - p(t)).$$

Нас интересует сравнение величин $r(t_n)$ и $r(t_n + \tau)$. B-устойчивость будет установлена, если будет показано (ограничимся классом диссипативных систем), что $r(t_n + \tau) \leqslant r(t_n)$. Очевидно,

$$r(t_n + \tau) - r(t_n) = \int_{t_n}^{t_n+\tau} \frac{dr}{dt}\, dt$$

и надо оценить интеграл от $\dot{r}$:

$$\dot{r}(t) = 2(\dot{\tilde{p}}(t) - \dot{p}(t),\ \tilde{p}(t) - p(t)).$$

Заметим, что p и $\tilde{p}$ — полиномы степени s; следовательно, $r(t)$ — полином степени $2s$, а $\dot{r}(t)$ — полином степени $2s - 1$. Для таких полиномов гауссова квадратура точна и интеграл $\int \dot{r}\, dt$ можно вычислить по квадратурной формуле:

$$\int_{t_n}^{t_n+\tau} \dot{r}\, dt = \sum_{j=1}^{s} b^j\, \dot{r}(t_n + \xi^j\tau) = \sum_{j=1}^{s} b^j\, (\dot{\tilde{p}} - \dot{p},\ \tilde{p} - p)\,\Big|_{\,t_n+\xi^j\tau}.$$

В силу (59) имеем ($j = 1, 2, \ldots, s$)

$$\dot{r}(t_n + \xi^j\tau) = 2(\dot{\tilde{p}} - \dot{p},\ \tilde{p} - p)\,\Big|_{\,t_n+\xi^j\tau} = (f(\tilde{Y}^j) - f(Y^j),\ \tilde{Y}^j - Y^j) \leqslant 0.$$

Здесь, конечно, использовалась диссипативность системы уравнений $\dot{x} = f(x)$. Так как коэффициенты гауссовых квадратур положительны, получаем требуемый результат — аттрактивность разностной схемы.

§ 18. Жесткие линейные краевые задачи

Рассмотрим численное решение специального класса линейных краевых задач, неявно содержащих большой параметр. Формально задача ставится стандартно: на интервале $[0, T]$ решается линейная задача

$$\frac{dx}{dt} = Ax + a(t), \quad x \in R^n, \tag{1}$$

с краевыми условиями (при $t = 0$ и $t = T$ соответственно)

$$\begin{aligned} (l_i, x(0)) &= b_i, \quad i = 1, 2, \ldots, k < n, \\ (l_i, x(T)) &= b_i, \quad i = k+1, k+2, \ldots, n, \end{aligned} \tag{2}$$

$$l_i \in R^n, \qquad \|l_i\| = 1, \qquad i = 1, 2, \ldots, n.$$

Матрицу A в теоретических рассуждениях будем считать постоянной, хотя все последующее относится и к случаю, когда A зависит от t, но изменяется со временем в некотором смысле (ниже будет уточнено, в каком именно) «медленно».

Жесткость системы (1) означает, что собственные значения матрицы A можно разделить на три части:

1) левый жесткий спектр (собственные значения Λ_i^-) характеризуется соотношениями

$$\operatorname{Re} \Lambda_i^- \leqslant -L, \qquad |\operatorname{Im} \Lambda_i^-| < L, \quad i = 1, 2, \ldots, I^-;$$

2) правый жесткий спектр (собственные значения Λ_i^+) характеризуется неравенствами

$$\operatorname{Re} \Lambda_i^+ \geqslant +L, \qquad |\operatorname{Im} \Lambda_i^+| < L, \quad i = 1, 2, \ldots, I^+;$$

3) мягкий спектр образуют собственные значения λ_j, удовлетворяющие условиям

$$|\lambda_j| \leqslant l, \quad j = 1, 2, \ldots, J, \qquad I^- + I^+ + J = n.$$

Число $L/l \gg 1$ является характеристикой жесткости системы. Кроме того, мы считаем, что $Tl = O(1)$, $TL \gg 1$. Качественная структура решения в этом случае хорошо известна, ее легко угадать из общих соображений: решение содержит два пограничных слоя — левый и правый. Это следует хотя бы из вида общего решения однородной задачи

$$x(t) = \sum_i C_i^- e^{\Lambda_i^- t} \Phi_i^- + \sum_i C_i^+ e^{\Lambda_i^+ t} \Phi_i^+ + \sum_j c_j e^{\lambda_j t} \varphi_j,$$

где Φ_i^-, Φ_i^+, φ_j — собственные векторы A, соответствующие выделенным выше частям ее спектра. При обобщении анализа на случай

переменной матрицы $A(t)$ мы предполагаем, что спектр не претерпевает существенных изменений при разных t, т.е. число точек в каждой части спектра остается постоянным.

Область приложения численных алгоритмов, о которых пойдет речь, конечно, намного шире. В частности, точки спектра могут, так сказать, «непрерывно» заполнять интервал $[-L, L]$, и разделение спектра на жесткую и мягкую части становится достаточно условным. Проблемы, которые рассматриваются ниже, отличаются от изложенных в § 17. Предположим, что величина $\|A\|T$ большая, но не очень. Использование в расчетах сетки с шагом $\|A\|\tau \ll 1$ (на практике «много меньше единицы» означает величину 0.1 или 0.01, например) приемлемо с точки зрения объема необходимой памяти и количества арифметических действий. Будем считать, что именно такой шаг используется в простой разностной схеме (второго порядка)

$$\frac{x_{m+1}-x_m}{\tau} = A_{m+1/2}\frac{x_m+x_{m+1}}{2} + a_{m+1/2}, \tag{3}$$

$$m = 0, 1, \ldots, M-1, \qquad M\tau = T.$$

Суть проблемы в том, как решать систему (2), (3).

Жесткие краевые задачи возникают, например, в расчетах процессов прохождения излучения через слой большой оптической толщины. При этом разные компоненты вектора x относятся к частицам разной энергии. Матрица A содержит члены, описывающие как поглощение излучения, так и переход частиц из одной энергетической группы в другую («замедление»). В некоторых случаях может присутствовать и «рождение» частиц под воздействием их поглощения («цепные реакции», характерные для процессов в ядерных реакторах). Итак, величину $\|A\|T$ (примерно совпадающую с LT) мы считаем не слишком большой, а вот величину e^{LT} считаем очень большой.

Характерной особенностью рассматриваемых задач является тот факт, что их решения суть ограниченные функции. Более аккуратно это можно сформулировать в виде требования

$$\|x(t)\| \leqslant C\,(\|a\| + \|b\|), \tag{4}$$

где $\|a\|$ — обычная норма правой части $a(t)$, $\|b\|$ — норма правой части в краевых условиях. Постоянная $C = O(1)$. Здесь необходимы пояснения. Оценка (4), если не оговаривать требований к C, тривиальна: она выполняется всегда (за редким исключением задач, как говорят, «на спектре», т.е. когда нарушены условия единственности и существования решения при любых правых частях). Однако «универсальная» оценка (4) в общем случае имеет постоянную $C \approx \exp(\|A\|T)$. Мы же рассматриваем класс задач, в котором $C = O(1) \ll \exp(\|A\|T)$. Такие жесткие краевые задачи играют осо-

бенно большую роль в приложениях (им присвоено специальное наименование «вычислительно корректные задачи»).

Приведем самые общие сведения из теории, позволяющей выделять эти задачи. Оказывается, что далеко не все формально возможные постановки задач для жесткой системы (1) приводят к вычислительно корректной задаче. Определяющую роль играет такой сравнительно простой и легко контролируемый фактор, как число краевых условий на левом и правом концах интервала $[0, T]$. Это число должно находиться в определенном соотношении с числом точек в разных частях спектра. Точнее, необходимым условием вычислительной корректности краевой задачи являются следующие неравенства:

а) $k \geqslant I^-$, т.е. число краевых условий на левом конце интервала должно быть не меньше числа сильно убывающих вправо решений;

б) $n - k \geqslant I^+$, т.е. число краевых условий на правом конце интервала должно быть не меньше числа сильно убывающих влево решений.

При нарушении этих условий краевая задача (1) оказывается вычислительно некорректной, т.е. в оценке (3) постоянная C имеет недопустимую с принятой здесь точки зрения величину $O(e^{LT})$. Докажем этот факт. (Доказательство технически простое; оно поучительно, так как вскрывает механизм вычислительной некорректности.) Построим конечное ($O(1)$) возмущение решения краевой задачи, приводящее к очень малому возмущению краевых условий. Будем истолковывать эту конструкцию «в обратном направлении» — как демонстрацию того, что очень малое возмущение краевых условий приводит к конечному возмущению решения.

Итак, пусть $x(t)$ — решение краевой задачи (1), (2) и пусть $k < I^-$. Рассмотрим множество всех сильно убывающих вправо решений однородной системы (1). Оно имеет структуру

$$\sum_{i=1}^{I^-} \alpha_i \Phi_i^- e^{t\Lambda_i^-},$$

где α_i — произвольные постоянные. Таким образом, мы имеем I^--мерное пространство. Поскольку число левых краевых условий k меньше размерности этого пространства, в нем можно найти элемент, удовлетворяющий однородным левым краевым условиям. Другими словами, существует нетривиальное решение системы k уравнений с I^- неизвестными α_i:

$$\left(\sum \alpha_i \Phi_i^-, l_j\right) = 0, \quad j = 1, 2, \ldots, k.$$

Пусть α_i — решение этой системы и $\|\alpha\| = 1$. Рассмотрим «возмущенное» решение

$$y(t) = x(t) + \sum \alpha_i \Phi_i^- e^{t\Lambda_i^-}.$$

Вектор-функция $y(t)$, очевидно, удовлетворяет уравнению (1) и левым краевым условиям. Правые краевые условия, конечно, нарушаются. Но так как возмущение при $t = T$ очень мало ($O(e^{-LT})$), то такого же порядка будут и невязки в правых краевых условиях.

Приведенное несложное построение доказывает необходимость сформулированных требований к числу краевых условий для вычислительной корректности задачи. Более сложный анализ, который мы не воспроизводим, показывает, что эти условия являются и почти достаточными. Точнее, если по числу краевых условий задача «правильная», это еще не гарантия вычислительной устойчивости задачи. В принципе среди «правильных» краевых задач могут встретиться и вычислительно неустойчивые, но такие задачи — редкое исключение, требующее некоторых случайных совпадений. В общем случае задача является вычислительно устойчивой.

Выше описана ситуация, в которой требуется построить вычислительно устойчивый способ решения разностной краевой задачи (3) (краевые условия для (3) имеют в точности тот же вид, что и для (1)). Характер вычислительных проблем уже разъяснен в более простой ситуации (см. § 10). Напомним, что при попытке решить краевую задачу в виде комбинации решений подходящим образом подобранных задач Коши мы приходим к необходимости вычислять конечную величину, суммируя функции типа e^{Lt}, что, как известно, приводит к значительной потере точности из-за сокращения главных знаков. Кроме того, само численное интегрирование задач Коши в этих условиях сопровождается сильным накоплением вычислительных погрешностей: множитель e^{Lt} в оценке погрешности численного решения неустраним.

Итак, проблема поставлена. Перейдем к ее решению. Заметим только, что в некоторых случаях мы будем описывать алгоритм прогонки в дифференциальной (а не в разностной) форме, т.е. будем сводить решение линейной краевой задачи (1) к последовательности нелинейных, но зато устойчивых задач Коши. Проблема же численного интегрирования этих задач решается самыми простыми средствами, так как выбор шага из соотношения $\|A\|\tau \ll 1$ допустим. Конечно, задача предполагается вычислительно корректной. В противном случае едва ли имеет смысл искать вычислительно устойчивые методы ее решения.

Ортогональная прогонка. Методы численного решения жестких линейных краевых задач основаны на следующем соображении. Рассмотрим множество решений дифференциального уравнения

$\dot{x} = Ax + a$, удовлетворяющих только левым краевым условиям. Обозначим это многообразие через $R^-(t)$. При каждом фиксированном значении t многообразие $R^-(t)$ является просто линейным (точнее, афинным) подпространством n-мерного фазового пространства. Его размерность, очевидно, есть $n - k$. Рассмотрим аналогичное подпространство, образованное всеми траекториями, удовлетворяющими правым краевым условиям. Обозначим его через $R^+(t)$.

Получим «явное выражение», например, для $R^-(t)$. Теория линейных дифференциальных уравнений дает простой рецепт:

1. Находим частное решение неоднородной задачи (1), удовлетворяющее неоднородным левым краевым условиям. Для этого находим любую точку $X^0(0) \in R^n$, удовлетворяющую системе k линейных уравнений (2), и с такими данными Коши проинтегрируем систему (1). Полученное решение обозначим через $X^0(t)$.

2. Построим $n - k$ линейно-независимых решений однородной системы $\dot{x} = Ax$, удовлетворяющих однородным левым краевым условиям. Обозначим эти решения через $X^i(t)$. Их можно получить решением соответствующих задач Коши. Данные Коши для этих решений строятся просто. Пусть в матрице, составленной из k n-мерных строк l_i $(i = 1, 2, \ldots, K)$, не вырождена матрица из первых k столбцов. Тогда $X^i(0) = \{\ldots, 0, \ldots, 1_{k+i}, 0, \ldots\}$, т.е. все компоненты с номерами $k + 1, k + 2, \ldots, n$ равны нулю, кроме $(k + i)$-й, равной единице. Первые k компонент получаются решением системы k линейных уравнений $(l_i, X^i(0)) = \emptyset$ $(i = 1, 2, \ldots, k)$.

Явное представление $R^-(t)$ имеет вид

$$R^-(t) = X^0(t) + \sum_{i=1}^{n-k} \alpha_i X^i(t). \qquad (5)$$

Таким образом, гиперплоскость $R^-(t)$ задается точкой $X^0(t)$ и «репером» из векторов $X^i(t)$, α_i — произвольные множители. В такой же форме можно представить и $R^+(t)$.

Очевидно, искомое решение краевой задачи есть пересечение этих многообразий:

$$x(t) = R^-(t) \cap R^+(t).$$

(Именно так рекомендует решать краевые задачи общая теория, не имеющая в виду задачи с большим параметром.) При каждом фиксированном t это есть пересечение $(n - k)$-мерной и k-мерной гиперплоскостей n-мерного пространства. Каждое многообразие $R^-(t)$, $R^+(t)$ состоит из траекторий уравнения $\dot{x} = Ax + a$. Каждая отдель-

ная траектория включает в себя сильно растущие (как вправо, так и влево) компоненты типа e^{Lt}, e^{-Lt}, т.е. R^-, R^+ в представлении (5) образованы из «неустойчивых» элементов. Однако сами многообразия, если задача вычислительно корректна, оказываются устойчивыми. Если рассматривать, например, $R^-(t)$ как некоторую поверхность в $(n+1)$-мерном пространстве $[0, T] \times R^n$, то при малой вариации правой части уравнения или значений b_i, входящих в краевые условия, это многообразие подвергнется, соответственно, малой деформации. (Хотя составляющие его отдельные траектории при этом изменяются очень сильно, малое изменение начальной точки $X^i(0)$ такой траектории порождает отклонение, которое с ростом t растет, как e^{Lt}.)

Вышеприведенные качественные соображения подсказывают путь, на котором следует искать устойчивый метод решения такой краевой задачи: надо работать не с индивидуальными решениями, а с многообразиями. Что это означает практически? Стандартный метод решения краевых линейных задач, описанный в § 8, — пример метода, основанного именно на использовании подходящего набора индивидуальных решений дифференциального уравнения. Они неустойчивы, вычислительно неустойчивым оказывается и такой «школьный» алгоритм.

Перейдем к методам, в которых используются многообразия $R^-(t)$, $R^+(t)$. Есть два стандартных, двойственных друг другу способа описания линейных подпространств $R(t)$ размерности r (где t — параметр, $R(t)$ — r-мерная гиперплоскость в R^n при каждом t).

1. Достаточно знать некоторую точку $X(t) \in R(t)$ и r векторов (обозначим их $e_1, e_2, \ldots, e_r(t)$), образующих базис в $R(t)$, и тогда $R(t)$ есть множество точек $x \in R^n$ вида

$$x = X(t) + \sum_{i=1}^{r} \alpha_i e_i(t), \tag{6}$$

где α — произвольные числа.

2. Можно выделить $R(t)$ системой $n - r$ линейных уравнений, имеющих ту же форму, в которой заданы краевые условия

$$(x, l_i(t)) = b_i(t), \quad i = 1, 2, \ldots, n - r, \tag{7}$$

где $l_i(t)$ — система $n - r$ линейно-независимых векторов в R^n, $b_n(t)$ — некоторые числа.

Способ вычисления X, e_i был указан выше. Почему же он в такой форме не пригоден? Причина в том, что не всякие формально правильные базисы приводят к устойчивому представлению многообразия. Хорошими являются ортогональные и близкие к ним базисы, плохими — сильно «сплюснутые» базисы, векторы которых хотя

и линейно-независимы, но очень близки друг к другу. Рисунок 26 поясняет сказанное. Каждый из двух базисов, изображенных на рисунке, описывает в трехмерном пространстве плоскость рисунка. Если вектор $\mathbf{e}_1$ подвергнуть малому возмущению, добавив малый вектор, ортогональный плоскости рисунка, то в первом случае определяемая базисом новая плоскость лишь на малый угол отклонится от первоначального положения, во втором случае (если возмущение по модулю сопоставимо с разностью $\mathbf{e}_2 - \mathbf{e}_1$) новая плоскость может стать почти ортогональной по отношению к плоскости рисунка.

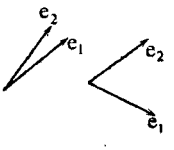


Рис. 26

При интегрировании задачи Коши (слева-направо) с начальными значениями $x(0) = e_i$, образующими линейно-независимый, близкий к ортогональному базис, по мере роста времени t базис $e_1(t), e_2(t), \ldots, e_r(t)$ начинает вырождаться. В каждом решении выделяются быстро растущие вправо компоненты, на фоне которых остальные «теряются» и даже пропадают из конечной разрядной сетки машинного представления чисел: происходит «сплющивание» базиса, и он уже определяет нужные нам многообразия неустойчивым образом.

Однако нам нужны сами подпространства $R^-(t)$, $R^+(t)$, а не их частные представления. Поэтому можно бороться с вырождением, исправляя базис через какие-то отрезки времени. Следует иметь в виду, что хороший базис при интегрировании не сразу становится плохим. Для того чтобы растущие экспоненты «задавили» все остальные, нужно некоторое время. Эти соображения и лежат в основе метода прогонки с ортогонализацией, предложенного С. К. Годуновым.

Перейдем к более аккуратному описанию метода. Прежде всего предположим, что векторы l_i $(i = 1, 2, \ldots, k)$, входящие в левые краевые условия, ортогональны друг к другу (если это не так, можно перейти к эквивалентной системе краевых условий, ортогонализовав векторы l_i). Более того, дополним совокупность векторов $l_i^-(0) = l_i$ $(i = 1, 2, \ldots, k)$ — нам удобно именно такое обозначение этих векторов — до полной ортонормированной системы векторов $l_i^-(0)$ $(i = 1, 2, \ldots, n)$. Левые краевые условия теперь можно записать в виде $(l_i^-(0), x(0)) = b_i$ $(i = 1, 2, \ldots, k)$.

Начнем интегрировать слева-направо $n - k + 1$ задач Коши. Их решения обозначим через $X^0, X^1, \ldots, X^{n-k}(t)$. Вектор $X^0(t)$ — это «какое-то» решение неоднородного уравнения. Определим его следующими данными Коши: $X^0(0) = \sum_{i=1}^{k} b_i\, l_i(0)$. Очевидно, выполняются условия

$$\dot{X}^0 = AX^0 + a, \qquad (X^0(0), l_i^-(0)) = b_i, \quad i = 1, 2, \ldots, k.$$

Остальные $n-k$ векторов $X^j(t)$ определяются данными Коши для однородной системы:

$$X^j(0) = l^-_{k+j}(0), \quad j = 1, 2, \ldots, n-k.$$

Тогда конструкция (5) при произвольных α_j дает все решения системы (1), удовлетворяющие левым краевым условиям, и описывает многообразие $R^-(t)$.

Однако это плохое описание, так как точка $X^0(t)$ состоит в основном из растущих со скоростью e^{Lt} решений, т.е. она находится далеко от искомого решения, которое в силу (3) ограничено ($O(1)$). При этом погрешности в вырождающемся при росте t базисе $X^j(t)$ ($j = 1, 2, \ldots, n-k$) приводят к очень большому отклонению представления (5) от точного многообразия $R^-(t)$. Мы заинтересованы и в качестве базиса, и в том, чтобы «начало координат» $X^0(t)$ находилось на расстоянии $O(1)$ от искомого решения $x(t)$. Как уже отмечалось, эффект жесткости системы не сразу приводит к столь неприятным последствиям. При малых временах t разница между e^{Lt} и e^{-Lt} еще не очень велика.

Итак, назначим некоторое число Δ, такое, что $\|A\|\Delta \approx 0.1 \div 1$, и проинтегрируем систему для $X^0, X^1, \ldots, X^{n-k}$ так, как было указано выше. В момент времени $t_1 = \Delta$ мы имеем представление $R^-(t)$ вида (5). Перейдем к другому, более удобному для наших целей представлению. Для этого систему векторов $X^1(t_1), X^2(t_1), \ldots, X^{n-k}(t_1)$ подвергнем стандартной процедуре ортонормировки и превратим ее в ортонормированный базис $e^1_1, e^1_2, \ldots, e^1_{n-k}$.

Таким образом, конструкции

$$X^0(t_1) + \sum_{j=1}^{n-k} \alpha_j X^j(t_1), \qquad X^0(t_1) + \sum_{j=1}^{n-k} \alpha_j e^1_j$$

описывают одну и ту же $(n-k)$-мерную гиперплоскость. Теперь в этой гиперплоскости следует найти более «удобную» точку $\tilde{X}^0(t_1)$, например наиболее близкую к началу координат, с тем чтобы расстояние от искомого решения (о котором мы знаем только, что $\|x(t)\| = O(1)$) до $\tilde{X}^0(t_1)$ было $O(1)$. Такой пересчет $X(t_1)$ в $\tilde{X}(t_1)$ легко осуществляется:

$$\tilde{X}(t_1) = X^0(t_1) - \sum_{j=1}^{n-k} (X^0(t_1), e^1_j)\, e^1_j.$$

На интервале $(t_1, t_1 + \Delta)$ интегрируется неоднородная система (1) с начальными данными при $t = t_1$: $\tilde{X}^0(t_1)$ дает «продолжение»

решения X^0 на новый интервал. В качестве базиса на этом интервале используются векторы $X^j(t)$ $(j=1, 2, \ldots, n-k)$, являющиеся решениями однородной системы $\dot{x}=Ax$ с начальными данными $X^j(t_1)=e_j^1$. Этот процесс продолжается с периодической ортонормировкой базиса и смещением точки X^0. Интегрированием справа-налево с периодической ортонормировкой базиса получаем представление многообразия $R^+(t)$.

Имея $R^-(t)$ и $R^+(t)$, решение $x(t)$ в каждый момент времени t находим как единственную точку пересечения $(n-k)$-мерной гиперплоскости $R^-(t)$ с k-мерной гиперплоскостью $R^+(t)$. Фактически это сводится к системе n линейных уравнений с n неизвестными $\alpha_1, \alpha_2, \ldots, \alpha_n = \{\alpha_1^-, \alpha_2^-, \ldots, \alpha_{n-k}^-, \alpha_1^+, \alpha_2^+, \ldots, \alpha_k^+\}$:

$$X_-^0(t) + \sum_{j=1}^{n-k} \alpha_j^- X_-^j(t) = X_+^0(t) + \sum_{j=1}^{k} \alpha_j^+ X_+^j(t).$$

Непрерывная ортонормировка. А. А. Абрамовым был предложен иной подход к построению устойчивого алгоритма решения краевой задачи (1) с использованием методов численного интегрирования задач Коши. Он основан на втором способе описания многообразий R^-, R^+ (см. (7)). Имея левые краевые условия $(l_i, x(0)) = b_i$ $(i=1, 2, \ldots, k)$, попытаемся «распространить» эти соотношения на весь интервал, т.е. получить для решения $x(t)$ соотношения (7). Если функции $l_i(t)$, $b_i(t)$ будут вычислены (и для правых краевых условий тоже), краевая задача в сущности будет решена.

На первый взгляд вычисление требуемых функций несложно. В самом деле, продифференцируем по t соотношение (7), опуская для простоты индекс i: $(\dot{l}, x) + (l, \dot{x}) = \dot{b}$. Используя при этом уравнение $\dot{x} = Ax + a$, получаем $(\dot{l}, x) + (l, Ax) + (l, a) = \dot{b}$ или, в другой форме, $(\dot{l} + A^* l, x) + (l, a) = \dot{b}$. Очевидно, цель будет достигнута, если в качестве $l(t)$ взять решение задачи Коши:

$$\dot{l} = -A^* l, \qquad \dot{b} = (l, a) \tag{8}$$

($l(0)$ и $b(0)$ берутся из краевого условия).

Однако это решение не может нас устроить: спектр A^* аналогичен спектру A, и интегрирование задачи Коши для $l(t)$ содержит те же проблемы, что и интегрирование задач Коши для уравнения $\dot{x} = Ax$; в $l(t)$ также происходит выделение быстро растущих экспонент. Можно нормировать $l(t)$, т.е. записать (7) в виде

$$(l(t)/\|l(t)\|, x(t)) = b(t)/\|l(t)\|,$$

или, вводя новые функции $\lambda(t) = l(t)/\|l(t)\|$ и $\beta(t) = b(t)/\|l(t)\|$, в форме $(\lambda(t), x(t)) = \beta(t)$, где растущие функции уже не фигурируют.

Нетрудно проверить, что $\lambda(t)$ является решением задачи Коши

$$\frac{d\lambda}{dt} = -A^*\lambda + \frac{(A\lambda, \lambda)}{(\lambda, \lambda)}.$$

Уравнение для β не выписываем, так как этот прием не снимает основной неприятности: при независимых, например, взаимно-ортогональных векторах $\lambda_i(0)$ векторы $\lambda_i(t)$ по мере роста t вырождаются, выходят на общую асимптотику, определяемую наиболее быстро растущей экспонентой.

Другими словами, «базис» из векторов $\lambda_i(t)$ сплющивается, становится плохим, его использование приводит к резкому возрастанию влияния малых вычислительных погрешностей. Переход от $l_i(t)$ к $\lambda_i(t)$ только маскирует ситуацию, делая ее внешне более благополучной: из расчета устраняются быстро растущие функции, иногда приводящие к выходу чисел из класса машинных (выход в машинную бесконечность, т.е. «авост»). Выход состоит в том, чтобы в процессе интегрирования всех уравнений для $l_i(t)$ подправлять векторы, беря их специальные линейные комбинации (краевые условия $(l_i, x) = b_i$ можно заменять некоторыми их линейными комбинациями, не меняя самой краевой задачи).

Перейдем к систематическому изложению варианта прогонки, в котором ортогонализация производится непрерывно и, так сказать, автоматически. Представим левые краевые условия в компактной форме:

$$L^- x(0) = b^-.$$

Здесь L^- — матрица $n \to k$, имеющая k строк, n столбцов; ее строками являются векторы l_i $(i = 1, 2, \ldots, k)$, входящие в левые краевые условия. Вектор b^- составлен из чисел b_i $(i = 1, 2, \ldots, k)$, входящих в те же левые краевые условия. Аналогичное обозначение примем и для правых краевых условий:

$$L^+ x(T) = b^+,$$

где L^+ — матрица $n \to n - k$.

Как было сказано выше, решение краевой задачи является пересечением двух интегральных многообразий системы (1): $R^-(t)$ и $R^+(t)$. Используем второе представление этих многообразий в виде пересечения k и $n - k$ гиперплоскостей (зависящих от параметра t) соответственно:

$$R^-(t)\colon\ L^-(t)x = b^-(t), \qquad R^+(t)\colon\ L^+(t)x = b^+(t). \tag{9}$$

Повторим (в матричной форме) вывод уравнений для $L(t)$ и $b(t)$. Дифференцируя по t соотношение $Lx = b$, получаем $\dot{L}x + L\dot{x} = \dot{b}$. Замена $\dot{x}$ на $Ax + a$ дает $(\dot{L} + LA)x + La = \dot{b}$. Примем для L и b уравнения

$$\dot{L} + LA = 0, \qquad \dot{b} = La. \tag{10}$$

Начальные данные Коши $L^-(0)$, $b^-(0)$ дают левые краевые условия, $L^+(T)$, $b^+(T)$ — правые.

Пусть $L^-(t)$, $L^+(t)$, $b^-(t)$, $b^+(t)$ найдены интегрированием соответствующих задач Коши. Тогда решение $x(t)$ в каждый момент времени находится решением системы n линейных уравнений:

$$\begin{aligned} &L^-(t)\, x(t) = b^-(t) \quad (k \text{ уравнений}), \\ &L^+(t)\, x(t) = b^+(t) \quad (n - k \text{ уравнений}). \end{aligned} \tag{11}$$

Эта процедура, называемая иногда алгоритмом встречной прогонки, как было сказано, в нашем случае (для жесткой краевой задачи) вычислительно неустойчива. Однако она полезна как теоретический ориентир. В частности, полезна следующая теорема.

Теорема. Пусть $L^-(t)$, $L^+(t)$, $b^-(t)$, $b^+(t)$ получены решением соответствующих задач Коши для уравнения (10) и $x(t)$ находится из системы (11), которая не вырождена ни при каком t. Тогда $x(t)$ является решением краевой задачи.

Заметим, что предположение о невырожденности системы (11) при всех t можно, конечно, заменить предположением о невырожденности при одном каком-то t или, иначе, предположением о невырожденности самой краевой задачи.

Доказательство. Пусть (11) выполняется при всех t. Дифференцируя по t, получаем $\dot{L}x + L\dot{x} = \dot{b}$. Используя (10) для $\dot{L}$ и $\dot{b}$, имеем

$$-LAx + L\dot{x} = La, \quad \text{или} \quad L(\dot{x} - Ax - a) = 0.$$

Таким образом, мы приходим к системе

$$L^-(\dot{x} - Ax - a) = 0, \qquad L^+(\dot{x} - Ax - a) = 0,$$

из которой в силу ее невырожденности получаем уравнение $\dot{x} - Ax - a = 0$. Выполнение краевых условий следует из данных Коши для L^-, L^+, b^-, b^+.

Перейдем к построению «устойчивых» описаний многообразий $R^-(t)$, $R^+(t)$. Рассмотрим, для определенности, описание R^-. Поскольку для R^+ все делается точно так же, индексы «—» и «+» опустим. Итак, опишем многообразие $R(t)$ в форме

$$G(t)\, x(t) = \beta(t), \tag{12}$$

где G есть матрица $n \to k$, β — k-вектор. При этом мы должны позаботиться о том, чтобы описание (12) было эквивалентно описа-

нию (9), а строки $G(t)$ образовывали хорошо обусловленный базис. Лучше всего, чтобы они были взаимно-ортогональными. (Требования к β сформулируем несколько позже.)

Будем искать $G(t)$ в форме $G(t) = M(t)L(t)$, где $M(t)$ — матрица $k \to k$, пока не определенная. При любой матрице $M(t)$ строки G суть некоторые линейные комбинации строк L. Если M — невырожденная матрица (т.е. из $Mr = 0$ следует $r = 0$), то линейные подпространства, натянутые на строки L и G, совпадают, или, другими словами, L и G отображают n-мерное подпространство в одно и то же k-мерное, определенное лишь разными базисами. Требование равномерной по t хорошей обусловленности базиса из строк G можно оформить в виде требования постоянства матрицы $G(t)\,G^*(t)$. Это — матрица $k \to k$, элементами которой являются всевозможные скалярные произведения строк G.

Предполагая, что строки l_i, входящие в левое краевое условие, предварительно ортонормированы, будем считать, что $M(0) = E_k$ (так будем обозначать единичную матрицу $k \to k$). Постараемся определить $M(t)$ таким образом, чтобы GG^* при всех t оставалась единичной, т.е. чтобы строки $G(t)$ образовывали ортонормированный базис. Для этого нужно обеспечить равенство

$$\frac{d}{dt}(GG^*) = \dot{G}G^* + G\dot{G}^* = \dot{G}G^* + (\dot{G}G^*)^* = 0. \tag{13}$$

Из (10) имеем

$$\dot{G} = (ML)_t = \dot{M}L + M\dot{L} = \dot{M}L - MLA.$$

Используя выражение для $\dot{G}$, а также соотношение $L = M^{-1}G$, преобразуем $\dot{G}G^*$:

$$\dot{G}G^* = (\dot{M}L - MLA)G^* = (\dot{M}M^{-1}G - MM^{-1}GA)G^* =$$
$$= \dot{M}M^{-1}GG^* - GAG^*.$$

Определим M таким образом, чтобы это выражение было равно нулю. (Второй член (13) при этом тоже автоматически обратится в нуль, и будет обеспечено $G(t)G^*(t) = \text{const}$.)

Итак, после несложных формальных преобразований для M получаем задачу Коши:

$$\dot{M} = GAG^*(GG^*)^{-1}M, \quad M(0) = E. \tag{14}$$

Это уравнение решать не придется. Мы используем его при выводе уравнения для G:

$$\dot{G} = (ML)_t = \dot{M}L + M\dot{L} = \dot{M}L - MLA = GAG^*(GG^*)^{-1}ML - MLA.$$

И наконец, учитывая, что $ML = G$, $GG^* = E_k$, получаем

$$\dot{G} = GAG^*G - GA, \quad G(0) = L(0). \tag{15}$$

Теперь приступим к выводу уравнения для $\beta(t)$. Из (9) имеем $L(t)\,x(t) = b(t)$, где $b(t)$ — решение задачи Коши (10). Умножение на M дает

$$ML\,x(t) = M\,b(t), \quad \text{т.е.} \quad G(t)\,x(t) = M(t)\,b(t).$$

Однако брать $\beta(t) = M(t)\,b(t)$ нельзя, так как мы не собираемся вычислять M, L, b.

Выведем уравнение для $\beta(t)$, используя то соотношение, которое мы хотим получить: $G(t)\,x(t) = \beta(t)$, где $G(t)$ — уже известная матрица $n \to k$, $x(t)$ — решение краевой задачи. Дифференцируя по t, имеем

$$\dot{\beta} = \dot{G}x + G\dot{x} = (\dot{G} + GA)x + Ga.$$

Учитывая (15) и связь $\beta = Gx$, преобразуем первый член:

$$(\dot{G} + GA)x = GAG^*Gx = GAG^*\beta.$$

Окончательно для $\beta(t)$ получаем задачу Коши:

$$\dot{\beta} = GAG^*\beta + Ga, \quad \beta(0) = b^-(0). \tag{16}$$

Отметим важное обстоятельство. Так как $\beta(t) = G(t)\,x(t)$, а строки $G(t)$ образуют ортонормированную (хотя и не полную в n-мерном пространстве) систему, то $\beta(t)$ — величина того же порядка, что и $x(t)$. Если задача вычислительно корректна и искомое решение $x(t)$ ограничено в смысле (3), то $\beta(t)$ — такая же величина. Но пока мы получили только часть уравнений для $x(t)$, порожденную переносом («прогонкой») левого краевого условия на весь интервал $[0, T]$. Точно так же можно перенести правые краевые условия, интегрируя задачи Коши справа-налево. Сами же уравнения для G^+ и β^+ по форме в точности совпадают с уравнениями для G^- и β^- (выше мы выводили их в общей форме (15), (16)).

Итак, в каждой точке мы получаем систему n уравнений

$$\begin{aligned} G^-(t)\,x(t) &= \beta^-(t), \\ G^+(t)\,x(t) &= \beta^+(t), \end{aligned} \qquad \text{или} \quad G(t)\,x(t) = \beta(t),$$

где теперь $G(t)$ — матрица $n \to n$, $\beta(t)$ — n-вектор. Заметим, что $\|\beta(t)\| = O(\|x(t)\|)$. Строки матрицы G, вообще говоря, ортонормированной системы не образуют, так как ее части G^- и G^+ получены независимо. Можно исправить и этот недостаток, если сначала решить задачу для $G^-(t)$ и при ортогонализации векторов l_i, входящих в пра-

вые краевые условия, потребовать еще и ортогональности к строкам матрицы $G^-(T)$. Если краевая задача не вырождена, это требование выполняется.

В принципе при ортогонализации может возникнуть большая потеря точности, если пространства, натянутые на «правые» векторы l_i $(i = k+1, k+2, \ldots, n)$, и строки $G^-(T)$ образуют очень малый угол, хотя и остаются еще формально независимыми, т.е. дают в сумме все n-мерное пространство. Такая ситуация возникает тогда, когда исходная краевая задача почти вырождена, т.е. среди точек спектра однородной задачи $\dot{x} - Ax = \lambda x$ с однородными краевыми условиями (2) имеется точка, близкая к нулю.

Если нуль принадлежит спектру, краевая задача вырождена, теряются стандартные свойства существования и единственности решения. Такую задачу называют «задачей на спектре» (мы здесь ее не рассматриваем). Однако чем меньше по модулю ближайшее к нулю собственное значение спектральной задачи, тем хуже обусловлена исходная задача, тем больше постоянная C в оценке (4). Это почти очевидно.

Если $y(t)$ — собственная функция ($\|y\| = 1$), соответствующая малому собственному числу λ, то функция $\tilde{x}(t) = x(t) + y(t)$ удовлетворяет уравнению

$$\dot{\tilde{x}} = A\tilde{x} + \tilde{a}, \quad \tilde{a} = a + \lambda y, \quad \|\tilde{x} - x\| = O(1),$$

с малым возмущением $O(|\lambda|)$ правой части ($\tilde{x}$, очевидно, удовлетворяет невозмущенным краевым условиям). Таким образом, постоянная C в оценке (4) не может быть меньше $1/|\lambda|$.

Что касается обоснования алгоритма, то оно пока носит качественный характер: при его реализации мы не имеем дела с решениями типа e^{Lt}, как это было бы, если бы мы попытались решать задачу обычным методом «фундаментальных решений» (см. § 8). Однако полной ясности все-таки нет. Не очень ясен механизм преодоления влияния большого параметра. Он как был в исходной задаче, так и остался в уравнениях (15), (16), в которых присутствует матрица A. В § 9 для простейшего случая было проведено достаточно полное исследование и, в частности, было обнаружено, что метод прогонки требует интегрирования задач Коши с большим параметром, но устойчивых. Здесь, видимо, механизм носит несколько иной характер.

Ниже в более простой и прозрачной ситуации мы попробуем прояснить этот вопрос. Он состоит в следующем: как происходит накопление вычислительных погрешностей при численном решении задач Коши (15), (16) по стандартным схемам типа Рунге—Кутты, например? Ведь если ориентироваться на общие оценки, то для (15) имеем

$$\|(GAG^*G)_G\| \approx \|GAG^*\| \approx \|A\|$$

(это тривиальная оценка, не учитывающая возможных тонких «компенсаций» больших величин при вычислении правой части (15)). Как было сказано выше, выбор шага численного интегрирования τ из условия $\|A\|\tau \ll 1$ считается здесь вполне приемлемым. Однако почему при оценке погрешности численного интегрирования не возникает стандартной и в общем случае неулучшаемой величины $\tau^p e^{\|A\|T}$ (p — порядок аппроксимации), не очень понятно.

Тригонометрическая прогонка. Рассмотрим часто встречающуюся в приложениях задачу Штурма—Лиувилля

$$-\frac{d}{dt}\left[p(t)\frac{dy}{dt}\right] + q(t)\,y(t) + f(t) = 0, \quad 0 \leqslant t \leqslant T, \tag{17}$$

с краевыми условиями, которые запишем в удобной для дальнейшего форме:

$$\begin{aligned} y(0)\cos\alpha_1 + \dot{y}(0)\sin\alpha_1 &= b_1, \quad t = 0, \\ y(T)\cos\alpha_2 + \dot{y}(T)\sin\alpha_2 &= b_2, \quad t = T, \end{aligned} \tag{18}$$

где $p(t)$, $q(t)$, $f(t)$, α_1, α_2, b_1, b_2, T — заданные функции и числа.

Запишем (17), (18) в стандартной форме — в виде системы двух уравнений первого порядка; суть дела от этого не меняется. Обозначая $x_1 = y$, $x_2 = p\dot{y}$ (разумеется, предполагается $p(t) \geqslant p_0 > 0$), получаем систему

$$\begin{aligned} \dot{x}_1 &= x_2/p, \\ \dot{x}_2 &= qx_1 + f, \end{aligned} \quad \text{или} \quad \dot{x} = \begin{pmatrix} 0 & 1/p \\ q & 0 \end{pmatrix} x + \begin{pmatrix} 0 \\ f \end{pmatrix}. \tag{19}$$

Краевые условия имеют стандартную форму с векторами $l_i = (\cos\alpha_i, \sin\alpha_i)$. Алгоритм прогонки состоит в том, что соотношение, имеющее форму краевого условия, продолжается (в силу уравнения) на весь интервал $[0, T]$.

Рассмотрим «прогоночное соотношение слева»:

$$x_1(t)\cos\varphi(t) + x_2(t)\sin\varphi(t) = \beta(t). \tag{20}$$

Для $\varphi(t)$ и $\beta(t)$ имеем данные Коши: $\varphi(0) = \alpha_1$, $\beta(0) = b_1$. Выведем уравнения для них, дифференцируя (20) по t и заменяя производные от x из (19):

$$\dot{x}_1\cos\varphi - x_1\sin\varphi\,\dot{\varphi} + \dot{x}_2\sin\varphi + x_2\cos\varphi\,\dot{\varphi} = \dot{\beta}.$$

Преобразуем это выражение, используя (20) и исключая $\dot{x}_1$, $\dot{x}_2$:

$$(1/p)x_2\cos\varphi - x_1\sin\varphi\,\dot{\varphi} + qx_1\sin\varphi + x_2\cos\varphi\,\dot{\varphi} = \dot{\beta} - f\sin\varphi.$$

Приводя подобные члены, имеем

$$x_1 \sin\varphi\,(q-\dot\varphi)+x_2\cos\varphi\,(1/p+\dot\varphi)=\dot\beta-f\sin\varphi. \tag{21}$$

Умножая выражение (21) на $\sin\varphi$ и вычитая из него (20), умноженное на $(1/p+\dot\varphi)\cos\varphi$, исключим члены с x_2:

$$x_1[(q-\dot\varphi)\sin^2\varphi-(1/p+\dot\varphi)\cos^2\varphi]=$$
$$=\dot\beta\sin\varphi-f\sin^2\varphi-\beta(1/p+\dot\varphi)\cos\varphi.$$

Оно выполняется при любом x_1, если потребовать одновременного обращения в нуль коэффициента при x_1 и свободного члена.

В итоге мы получаем уравнения для φ и β:

$$\dot\varphi=q\sin^2\varphi-(1/p)\cos^2\varphi,\quad \varphi(0)=\alpha_1, \tag{22}$$

$$\dot\beta=\beta(1/p+\dot\varphi)\cos^2\varphi/\sin^2\varphi+f\sin\varphi,\quad \beta(0)=b_1. \tag{23}$$

Деление на $\sin\varphi$ не приводит к неприятностям, так как из (22) следует

$$(1/p+\dot\varphi)=(q+1/p)\sin^2\varphi.$$

Уравнение для β лучше использовать в виде

$$\dot\beta=f\sin\varphi+\beta\sin\varphi\cos\varphi\,(q+1/p). \tag{24}$$

Заметим, что нас интересуют задачи с большим параметром. Большой величиной считаем q, причем знак $q(t)$ не определяем однозначно. При $q>0$ однородное уравнение (17) имеет решения экспоненциального типа (одно быстро убывающее, другое быстро растущее вправо). При $q<0$ уравнение (17) имеет решения колебательного характера, причем, если $|q|T\gg 1$ (1000, например), на $[0, T]$ укладываются десятки колебаний. В некоторых задачах (см. § 15) $q(t)$ может иметь разные знаки на разных частях интервала $[0, T]$.

Что можно сказать в этих условиях об уравнении (22) для φ и о возможности достаточно аккуратного численного интегрирования его на большом интервале времени? А интервал действительно большой, так как величина

$$T\,[q\sin^2\varphi+(1/p)\cos^2\varphi]_\varphi\approx 2Tq\sin\varphi\cos\varphi$$

в принципе может быть достаточно большой, если только функция $\varphi(t)$ не «застревает» надолго в окрестности таких значений, где $\sin\varphi\cos\varphi\approx 0$.

Пытаясь разобраться в характере уравнения (22), заметим, что $|\dot{\varphi}| \leqslant q$. (Мы пользуемся не очень строгими оценками; речь идет о грубом анализе, в котором величиной $(1/p)\cos^2\varphi \approx O(1)$ можно пренебречь.) Следовательно, все траектории (22) проходят в конусе с раствором, определяемым величиной $|q|$, т.е. в целом траектории (22) не имеют экспоненциального роста, траекторий типа $e^{|q|t}$ среди решений (23) нет, хотя на малых частях интервала $[0, T]$ такие решения могут и появиться.

В § 7 специально подчеркивалось, что процесс накопления погрешностей при численном интегрировании задачи Коши существенным образом зависит от устойчивости вычисляемой траектории. Устойчивость же определяется уравнением в вариациях, которое для (22) имеет вид ($\delta\varphi$ — малое возмущение траектории φ)

$$\delta\dot{\varphi} = \sin 2\varphi\,(q - 1/p)\,\delta\varphi. \tag{25}$$

Устойчивость траектории зависит от знака и модуля $q \sin 2\varphi$. В принципе возможна ситуация, когда $\sin 2\varphi \approx 1$ и траектория сильно неустойчива: решение (25) ведет себя, как $e^{|q|t}$. Однако из уравнения (22) видно, что φ быстро уходит от такого значения и «неустойчивый» участок на траектории не может быть длительным. Ограничимся этими простыми соображениями, которые, видимо, можно превратить в достаточно строгий анализ.

Периодическая прогонка. Опишем полезный в приложениях алгоритм решения специальной системы уравнений высокого порядка, возникающей при решении краевой задачи для уравнения Штурма—Лиувилля (10.1) с периодическими краевыми условиями

$$x(0) = x(T), \qquad \dot{x}(0) = \dot{x}(T).$$

Разностное уравнение (10.2) можно считать определенным при всех значениях $n = 0, 1, \ldots, N$, если реализовать условие периодичности, отождествив выходящие за пределы сетки значения с сеточными: $x_{N+1} = x_0$, $x_{-1} = x_N$.

Итак, мы приходим к системе уравнений, аналогичной (10.3):

$$a_0 x_N - b_0 x_0 + c_0 x_1 = f_0,$$

$$a_n x_{n-1} - b_n x_n + c_n x_{n+1} = f_n,$$

$$a_N x_{N-1} - b_N x_N + c_N x_0 = f_N,$$

где $n = 1, 2, \ldots, N-1$. Матрица системы отличается от знакомой нам трехдиагональной наличием двух ненулевых элементов: на последнем месте первой строки и на первом месте последней. Для экономного (требующего $O(N)$ операций) решения такой системы А. А. Абрамо-

вым построен алгоритм, обобщающий классическую прогонку. Он часто используется в практической вычислительной работе.

Решение ищем в форме «прогоночного соотношения»

$$x_{n-1} = P_n x_n + Q_n + R_n x_N.$$

Очевидно, первое уравнение системы можно записать в этой форме, и мы получаем явные выражения для стартовых значений прогоночных коэффициентов:

$$P_1 = c_0/b_0, \qquad Q_n = -f_0/b_0, \qquad R_1 = a_0/b_0.$$

Теперь стандартная процедура позволяет получить рекуррентную формулу. Пусть P_n, Q_n, R_n известны. Исключая из уравнения с индексом n значение x_{n-1}, имеем уравнение

$$a_n(P_n x_x + Q_n + R_n x_N) - b_n x_n + c_n x_{n+1} = f_n,$$

связывающее x_n, x_{n+1}, x_{N+1}. Этому уравнению можно придать стандартную прогоночную форму, разрешив его относительно x_n. В результате мы получаем искомые соотношения:

$$A = b_n - a_n P_n,$$

$$P_{n+1} = \frac{c_n}{A}, \qquad R_{n+1} = \frac{a_n R_n}{A}, \qquad Q_{n+1} = \frac{a_n Q_n - f_n}{A}.$$

Эту операцию можно продолжить вплоть до значения $n = N - 1$. Прогоночное соотношение

$$x_{N-1} = P_N x_N + Q_N + R_N x_N$$

после подстановки в N-е уравнение системы даст соотношение, связывающее x_N с x_0. Придадим ему форму $x_N = \alpha_N x_0 + \beta_N$ и будем искать решение в виде $x_n = \alpha_n x_0 + \beta_n$.

Новые прогоночные коэффициенты α_n, β_n $(n = N, N-1, \ldots, 1)$ находим по рекуррентным формулам справа-налево, имея их стартовые значения α_N, β_N. Для этого из прогоночного соотношения $x_{n-1} = P_n x_n + Q_n + R_n x_N$, считая, что значения α_n и β_n известны, исключим x_n, x_N:

$$x_{n-1} = P_n(\alpha_n x_0 + \beta_n) + Q_n + R_n(\alpha_N x_0 + \beta_N).$$

Приводя подобные члены, получаем рекуррентные соотношения

$$\alpha_{n-1} = P_n \alpha_n + R_n \alpha_N, \qquad \beta_{n-1} = Q_n + P_n \beta_n + R_n \beta_N.$$

Последнее такое соотношение имеет вид

$$x_0 = \alpha_0 x_0 + \beta_0, \quad \text{т.е.} \quad x_0 = \beta_0/(1 - \alpha_0).$$

Остальные значения x_n найдем по формуле $x_n = \alpha_n x_0 + \beta_n$.

Пятиточечная прогонка. Опишем алгоритм решения системы уравнений с пятидиагональной матрицей. Такие системы возникают при численном решении разностных уравнений, аппроксимирующих краевую задачу для уравнения четвертого порядка:

$$\frac{d^4x}{dt^4} + p\frac{d^2x}{dt^2} + q(x) = f, \quad 0 \leqslant t \leqslant T,$$

$$x(0) = A_0, \quad \dot{x}(0) = B_0, \quad x(T) = A_1, \quad \dot{x}(T) = B_1.$$

Ограничимся этой простейшей задачей. Вводя сетку, аппроксимируем уравнение разностным:

$$\frac{1}{h^4}(x_{n-2} - 4x_{n-1} + 6x_n - 4x_{n+1} + x_{n+2}) + \\ + \frac{1}{h^2}p_n(x_{n-1} - 2x_n + x_{n+1}) + q_n x_n = f_n,$$

$$n = 2, 3, \ldots, N-2, \qquad h = T/N.$$

Краевые условия аппроксимируем самым простым способом:

$$x_0 = A_0, \quad x_1 - x_0 = hB_0, \quad x_N = A_1, \quad x_N - x_{N-1} = hB_1.$$

Придадим системе уравнений стандартную пятидиагональную форму:

$$c_0x_0 - d_0x_1 + e_0x_2 = f_0,$$

$$-b_1x_0 + c_1x_1 - d_1x_2 + e_1x_3 = f_1,$$

$$a_nx_{n-2} - b_nx_{n-1} + c_nx_n - d_nx_{n+1} + e_nx_{n+2} = f_n,$$

$$a_{N-1}x_{N-3} - b_{N-1}x_{N-2} + c_{N-1}x_{N-1} - d_{N-1}x_N = f_{N-1},$$

$$a_Nx_{N-2} - b_Nx_{N-1} + c_Nx_N = f_N$$

($n = 2, 3, \ldots, N-2$). Формулы для коэффициентов системы очевидны. Прогоночное соотношение имеет вид

$$x_{n-1} = P_nx_n + R_nx_{n+1} + Q_n.$$

После несложных преобразований первые два уравнения (левые краевые условия) дают стартовые значения прогоночных коэффициентов $(P, R, Q)_1$ и $(P, R, Q)_2$. Стандартный вывод рекуррентных соотношений для прогоночных коэффициентов проводится в предположении, что в процессе прямой прогонки (слева-направо) коэффициенты $(P, R, Q)_{n-1}$ и $(P, R, Q)_{n-2}$ (и все предшествующие) уже найдены. С их помощью из стандартного n-го уравнения можно исключить x_{n-2} и x_{n-1} и получить связь между x_n, x_{n+1}, x_{n+2}, которая разрешается относительно x_n.

Несложные преобразования дают рекуррентные формулы:

$$b_n' = b_n - a_n P_{n-1}, \qquad c_n' = c_n + a_n R_{n-1}, \qquad f_n' = f_n - a_n Q_{n-1},$$

$$A = c_n' - b_n' P_n,$$

$$P_{n+1} = -\frac{d_n + b_n' R_n}{A}, \qquad R_{n+1} = -\frac{e_n}{A}, \qquad Q_{n+1} = \frac{f_n' + b_n' Q_n}{A}.$$

Эта операция продолжается стандартно до значения $n = N - 2$, т.е. последнее прогоночное соотношение имеет вид

$$x_{N-2} = P_{N-1} x_{N-1} + R_{N-1} x_N + Q_{N-1}.$$

Вместе с двумя последними уравнениями (правыми краевыми условиями) оно дает нам три линейных уравнения с тремя неизвестными x_{N-2}, x_{N-1}, x_N. Решив эту систему, процессом «обратной прогонки» мы вычислим все x_n последовательно справа-налево.

Предоставим читателю в качестве полезного упражнения внести необходимые изменения в том случае, когда краевые условия заданы с использованием вторых и третьих производных. Несколько бóльших изменений требует алгоритм в том случае, когда на одном конце задано одно краевое условие, на другом — три.

§ 19. Осреднение быстрых вращений

Рассмотрим важный в приложениях метод интегрирования специального класса обыкновенных дифференциальных уравнений. Приложения его столь разнообразны, что имеет смысл начать с абстрактной постановки задачи. Пусть имеется система уравнений

$$\dot{z} = f(z), \qquad z(0) = q_0, \qquad t > 0, \tag{1}$$

описывающая некоторое физическое явление «в главном» (факторами, мало влияющими на эволюцию системы, пренебрегаем). Известно общее решение — функция $z(t, q_0)$ (точнее, вектор-функция, но размерность z в дальнейшее явно входить не будет).

И наконец (это существенное предположение, выделяющее узкий, но важный класс задач), пусть все траектории (1) периодичны с периодом $T(q_0)$, своим на каждой траектории. Итак, нам известна функция $z(t, q_0)$, удовлетворяющая соотношениям

$$\begin{aligned} &\text{а)}\ z_t(t, q_0) = f[z(t, q_0)], \quad \forall\, t, q_0, \\ &\text{б)}\ z(0, q_0) = q_0, \quad \forall\, q_0, \\ &\text{в)}\ z(T(q_0), q_0) = q_0, \quad \forall\, q_0. \end{aligned} \tag{2}$$

Если бы начальные данные были заданы в момент t_0, общим решением (в силу независимости f от t) была бы функция $z(t-t_0, q_0)$. Систему (1) будем называть «невозмущенной», ее решение — «невозмущенной траекторией».

Пусть более полное описание явления, учитывающее влияние малых сил, приводит к системе, именуемой в дальнейшем «возмущенной»:

$$\dot{x} = f(x) + \varepsilon F(x), \quad x(0) = q_0, \quad \varepsilon \ll 1. \tag{3}$$

Нас интересует это более адекватное действительности описание явления. Здесь ε — малый параметр, функции f, F, T будем считать «гладкими», т.е. они сами и их используемые в выкладках производные суть величины $O(1)$ (без этой оговорки предположение $\varepsilon \ll 1$ не имело бы смысла). Система (3) не имеет явного решения и возникает вопрос: нельзя ли узнать что-либо о траектории (3), используя ее близость к «интегрируемой» системе (1)?

Если нас интересует ограниченный отрезок времени (например, три-пять периодов), ответ очевиден и ничего интересного в задаче нет. Из общего курса дифференциальных уравнений известна теорема о непрерывности решения задачи Коши по правой части, т.е. $x(t, q_0) = z(t, q_0) + O(\varepsilon)$. (Для этого периодичность z не нужна.) Но что произойдет за «большой» интервал времени? Как «накопятся» последствия малого возмущения $\varepsilon F(x)$ за время порядка $1/\varepsilon$? Здесь очевидного ответа нет. Изложенная ниже достаточно сложная теория позволяет производить соответствующие расчеты.

Речь идет о теории малых возмущений на больших временах. В задаче имеется малый параметр ε и большой параметр t — время процесса $O(1/\varepsilon)$. Именно это последнее обстоятельство определяет нетривиальный характер проблемы, решение которой удается продвинуть за счет использования важного свойства невозмущенной системы (1) — периодичности всех ее траекторий. Что касается «согласованности» параметров ε и $t \approx O(1/\varepsilon)$, то она связана не с существом задачи, а просто с тем, что удается построить аппарат решения задачи (3), работающий эффективно именно на временах $O(1/\varepsilon)$. В частных случаях удается распространить его действие на времена $O(1/\varepsilon^2)$, а иногда и на весь интервал $[0, \infty]$. В некоторых ситуациях удается построить метод, работающий на временах $O(1/\sqrt{\varepsilon})$, и это тоже представляет интерес.

Содержательные примеры. Рассмотрим пример, исторически положивший начало развитию и применению метода осреднения. Движение планет Солнечной системы достаточно точно описывается системой уравнений вида

$$\frac{dx_i}{dt} = f_i(x_i) + \sum_j \varepsilon_{ij} F_{ij}(x_i, x_j), \quad i = 1, 2, \ldots, I. \tag{4}$$

Здесь I — число планет, i — номер планеты, x_i — шестимерный вектор, описывающий состояние планеты-точки в фазовом пространстве, f_i — сила притяжения Солнца, действующая на i-ю планету, $\varepsilon_{ij} F_{ij}(x_i, x_j)$ — сила взаимного тяготения i-й и j-й планет, ε_{ij} — соответствующий малый параметр.

Невозмущенная система

$$\dot{z}_i = f_i(z_i), \quad i = 1, 2, \ldots, I, \tag{5}$$

имеет известное решение — движение по эллипсам. Каждая планета имеет свой период T_i и, строго говоря, то, что будет излагаться ниже, неприменимо к данному примеру. Хорошую и эффективную теорию удается построить для одночастотной задачи, когда все компоненты невозмущенной траектории имеют общий период. Обобщение этой теории на многочастотный случай (а именно таким является Солнечная система) связано с принципиальными и до настоящего времени еще не преодоленными трудностями. Тем не менее именно для расчета движения планет впервые без строгого обоснования («эвристически») стали использоваться методы осреднения, которые берут начало в трудах классиков небесной механики, в частности Гаусса.

Второй пример задачи (3) — расчет движения искусственного спутника Земли. В этом случае f — сила притяжения Земли, εF — малые силы, связанные с нестрогой сферичностью Земли, с сопротивлением крайне разреженной на высоте орбиты спутника атмосферы, с притяжением Луны и т.п. Наконец, третий пример — дрейф электрона в «скрещенных» магнитном и слабом электрическом полях.

Может показаться, что для современного специалиста, вооруженного мощными ЭВМ, нижеследующее особого значения не имеет. В конце концов это обычная задача Коши, с которой «все ясно», существуют хорошие стандартные программы и можно «пробить» задачу мощью современных компьютеров. Однако речь идет об интегрировании задачи Коши на очень большом интервале времени, и здесь остро стоит вопрос об оценке накопления вычислительных погрешностей. Надежно выделить на этом фоне влияние малого возмущения не так-то просто. Нужно учитывать и то, что в расчетных формулах численного интегрирования типа

$$x_{n+1} = x_n + \tau f(x_n) + \varepsilon\tau F(x_n) \tag{6}$$

при достаточно малом ε величина $\varepsilon\tau F$ может выйти за пределы точности машинного представления x. В таком расчете возмущение просто игнорируется.

Не следует забывать и «экономическую» сторону: при интегрировании шаг τ должен быть таким, чтобы погрешность аппроксимации была существенно меньше возмущения εF. Пусть, для опреде-

ленности, это достигается при $\tau \approx 10^{-2}T$ (T — период невозмущенной системы). Если ε очень мало и время $O(1/\varepsilon)$ содержит, допустим, 10^4 периодов (а это не такой уж нереальный случай), прямое интегрирование может оказаться на пределе возможностей ЭВМ.

Стробоскопический метод. Качественные соображения. Приступим к предварительному качественному анализу движения возмущенной системы, который приведет нас к разработке аппарата асимптотического интегрирования на большом интервале времени. Существуют разные подходы к построению такого аппарата. Мы предпочтем так называемый стробоскопический метод из-за его простоты и наглядности.

Рассмотрим траектории $z(t, q_0)$ и $x(t, q_0)$ систем (1) и (3), стартующие из одной и той же точки q_0. Через время $T(q_0)$ траектория $z(t, q_0)$ возвратится в ту же самую точку q_0, траектория $x(t, q_0)$ попадает в точку q_1, лежащую в $O(\varepsilon)$-окрестности q_0. Однако нам нужно более точное соотношение:

$$q_1 \equiv x(T(q_0), q_0) = q_0 + \varepsilon P(q_0) + O(\varepsilon^2). \tag{7}$$

В дальнейшем мы докажем его и опишем методы вычисления функции $P(q)$. Пока же примем это соотношение и достаточно естественное предположение о гладкости функции $P(q)$. Обоснование (7) будет результатом не очень сложной теории малых возмущений на конечном интервале времени, равном только одному периоду $T(q_0)$.

Применим тот же аппарат, рассматривая траектории, стартующие в момент времени $t_1 = T(q_0)$ из точки q_1 (которая известна с точностью до $O(\varepsilon^2)$). Траектория $z(t - t_1, q_1)$ через время $T(q_1)$ вернется в точку q_1, а траектория $x(t - t_1, q_1)$, являющаяся продолжением траектории $x(t, q_0)$, попадает в точку $q_2 \equiv x(T(q_1), q_1) = x(t_2, q_0)$, где $t_2 = T(q_0) + T(q_1)$, и для нее имеем

$$q_2 = q_1 + \varepsilon P(q_1) + O(\varepsilon^2).$$

Продолжая далее, получаем последовательность моментов времени t_k и положений возмущенной траектории $q_k \equiv x(t_k, q_0)$, связанных соотношениями

$$\begin{aligned} q_1 &= q_0 + \varepsilon P(q_0) + O(\varepsilon^2), & t_1 &= t_0 + T(q_0), \\ q_2 &= q_1 + \varepsilon P(q_1) + O(\varepsilon^2), & t_2 &= t_1 + T(q_1), \\ &\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ q_{k+1} &= q_k + \varepsilon P(q_k) + O(\varepsilon^2), & t_{k+1} &= t_k + T(q_k). \end{aligned} \tag{8}$$

Рисунок 27 дает качественную иллюстрацию этой конструкции. Попробуем соединить точки $q_0, q_1, \ldots, q_k, \ldots$ некоторой плавной линией $q(\tau)$, где τ — пока просто параметр. Она «устроена» гораздо проще траектории $x(t, q_0)$ — их можно сравнить с прямой и спиралью соответственно.

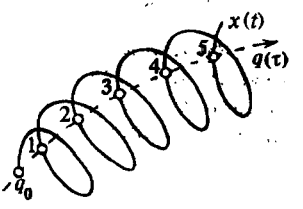


Рис. 27

Итак, следя не за всеми положениями точки $x(t, q_0)$, а «высвечивая» ее только в специальные дискретные моменты времени (это и есть стробоскопия), мы получаем существенно более простую кривую $q(\tau)$, которая несет достаточную информацию о траектории $x(t, q_0)$. Вопрос в том: как найти кривую $q(\tau)$? Внимательного взгляда на разностные соотношения (8) достаточно, чтобы возникла следующая догадка. Точки q_k можно получить в процессе численного интегрирования (методом Эйлера с шагом ε) дифференциального уравнения

$$\frac{dQ}{d\tau} = P(Q), \quad Q(0) = q_0. \tag{9}$$

Наличие «лишнего» слагаемого $O(\varepsilon^2)$ не принципиально. Точнее, если известно решение (9), введем на оси τ сетку с шагом ε и обозначим $Q_k = Q(\tau_k)$, $\tau_k = k\varepsilon$. Тогда величины Q_k будут удовлетворять разностным уравнениям

$$Q_{k+1} = Q_k + \varepsilon P(Q_k) + O(\varepsilon^2) \tag{10}$$

— тем же самым, что и уравнения (8) для q_k (конечно, $O(\varepsilon^2)$ у них разные, но это не существенно).

Таким образом, кривую $q(\tau)$ можно приближенно вычислить, интегрируя уравнение (9), именуемое *уравнением в медленном времени*. Однако нужно еще установить связь между «медленным временем» τ и физическим временем t. Она следует из соотношений

$$\tau_{k+1} = \tau_k + \varepsilon, \quad t_{k+1} = t_k + T(q_k), \quad k = 0, 1, 2, \ldots,$$

которые можно рассматривать как приближенное интегрирование дифференциального уравнения

$$\frac{dt}{d\tau} = \frac{1}{\varepsilon} T(Q(\tau)), \tag{11}$$

т.е. грубо говоря, время τ меняется в ε^{-1} раз медленнее времени t. Изменению медленного времени τ на «шаг» ε соответствует изменение физического времени на $T(q) = O(1)$. Следовательно, если будет установлена близость величин Q_k и q_k для $k \approx \varepsilon^{-1}$, то тем самым, зная $Q(\tau)$, мы получим информацию о траектории $x(t, q_0)$

за ε^{-1} периодов, т.е. за физическое время $O(\varepsilon^{-1})$. По существу выше были изложены все основные идеи метода осреднения. Перейдем к их оформлению с технической стороны.

Элементарная теория малых возмущений. Начнем с теории, позволяющей рассчитывать движение возмущенной системы на ограниченном интервале времени. В этой теории используется малость возмущения и известное решение невозмущенной системы. Речь идет о стандартной в таких вопросах технике. Решение ищется в виде ряда Пуассона по степеням ε:

$$x(t, q_0) = X_0(t, q_0) + \varepsilon X_1(t, q_0) + \varepsilon^2 X_2(t, q_0) + \ldots, \qquad (12)$$

где коэффициенты X_0, X_1, X_2, ... подлежат определению.

Подставляя конструкцию (12) в уравнение $\dot{x} = f + \varepsilon F$, имеем

$$\dot{X}_0 + \varepsilon \dot{X}_1 + \varepsilon^2 \dot{X}_2 + \ldots = \\ = f(X_0 + \varepsilon X_1 + \varepsilon^2 X_2 + \ldots) + \varepsilon F(X_0 + \varepsilon X_1 + \varepsilon^2 X_2 + \ldots). \qquad (13)$$

Разлагая правую часть (13) в ряд Тейлора, получаем

$$\dot{X}_0 + \varepsilon \dot{X}_1 + \varepsilon^2 \dot{X}_2 + \ldots = f(X_0) + \varepsilon f_x(X_0) X_1 + \frac{\varepsilon^2}{2} f_{xx}(X_0) X_1 X_1 + \\ + \varepsilon^2 f_x(X_0) X_2 + \ldots + \varepsilon F(X_0) + \varepsilon^2 F_x(X_0) X_1 + \ldots$$

Обычная техника приравнивания членов при одинаковых степенях ε дает последовательные уравнения (для ε^0, ε^1, ε^2 соответственно)

$$\begin{aligned} &\dot{X}_0 = f(X_0), && X_0(0) = q_0, \\ &\dot{X}_1 = f_x(X_0) X_1 + F(X_0), && X_1(0) = 0, \\ &\dot{X}_2 = f_x(X_0) X_2 + \tfrac{1}{2} f_{xx}(X_0) X_1 X_1 + F_x(X_0) X_1, && X_2(0) = 0. \end{aligned} \qquad (14)$$

Начальные данные Коши к уравнениям (14) получены точно таким же образом из начальных данных

$$x(0) = X_0(0) + \varepsilon X_1(0) + \varepsilon^2 X_2(0) + \ldots = q_0.$$

Мы не будем здесь расшифровывать формального выражения $f_{xx} X_1 X_1$ (f_{xx} — вектор, дифференцируемый дважды по вектору x). Не будем выписывать и членов более высокого по ε порядка. Это достаточно сложные выражения, в которых появляются f_{xxx} и т.п. Обратим внимание на специфику уравнений и опишем процедуру их решения.

Уравнение для функции X_0 — это просто уравнение (1). Оно нелинейно, но мы предполагаем, что его решение нам известно.

Итак, $X_0(t, q_0) = z(t, q_0)$. Следующее уравнение — это уравнение для определения X_1. Оно является линейным неоднородным уравнением с переменными коэффициентами. Матрица $f_x(X_0)$ может рассматриваться как известная функция от t. Ее точное значение есть $A(t) = f_x[z(t, q_0)]$. Правая часть этого уравнения — тоже известная функция времени $F[z(t, q_0)]$. Таким образом, задача Коши однозначно определяет функцию $X_1(t, q_0)$.

Уравнение для $X_2(t, q_0)$ имеет ту же структуру, что и уравнение для X_1, отличаясь лишь правой частью. Но после определения X_0, $X_1(t)$ правая часть этого уравнения вычисляется и может считаться известной. Все последующие уравнения для коэффициентов формального ряда имеют одну и ту же структуру:

$$\dot{X}_k = A(t)X_k + R_k(t), \quad X_k(0) = 0, \tag{15}$$

где R_k — сложное выражение из производных f, F по x и функций $X_0(t), X_1(t), \ldots, X_{k-1}(t)$.

Таким образом, функции $X_0, X_1, X_2, \ldots$ могут быть вычислены последовательно. Конечно, фактическое вычисление этих функций является сложным делом. Прежде всего мы сталкиваемся с резким возрастанием сложности аналитических выражений при последовательных дифференцированиях по x, причем сложность зависит от выбора переменных. Удачный их выбор может существенно упростить процедуру, и этому уделяли большое внимание классики небесной механики. В настоящее время проведение таких громоздких, но алгоритмически четко определенных выкладок все чаще поручается ЭВМ. Что касается решения уравнения (15), то эта проблема имеет достаточно эффективное решение.

Решение уравнения в вариациях. Теорема Пуанкаре. Последовательное решение цепочки уравнений (14), определяющей коэффициенты ряда Пуассона, в принципе является чисто технической задачей. Это следует из теоремы Пуанкаре.

Теорема 1. Пусть известно полное решение $z(t, q_0)$ невозмущенного уравнения (1). Тогда решение уравнения в вариациях

$$\dot{y} = f_z[z(t, q_0)]\, y + R(t), \quad y(0) = 0, \tag{16}$$

сводится к дифференцированию и взятию квадратур.

Доказательство. То, что $z(t, q_0)$ является полным решением задачи (1), означает выполнение тождеств (2а) и (2б). Что касается решения линейной системы (16), то, как известно, нужно прежде всего иметь фундаментальную систему решений однородного уравнения — матрицу $\Psi(t)$, каждый столбец которой удовлетворя-

нородному уравнению в вариациях, т.е. Ψ должна быть решением уравнения

$$\frac{d\Psi}{dt} = f_z[z(t, q_0)]\,\Psi, \quad \Psi(0) = E. \tag{17}$$

Утверждение 1. $\Psi(t) = z_q(t, q_0)$.

В самом деле, дифференцируя по q_0 тождество (2а), получаем (меняем порядки дифференцирования по t и q_0)

$$\frac{\partial}{\partial t} z_q(t, q_0) = f_z[z(t, q_0)]\, z_q(t, q_0).$$

Дифференцируя по q тождество (2б), имеем

$$z_q(0, q) = \frac{\partial q}{\partial q} = E.$$

Утверждение доказано: $z_q(t, q_0)$ удовлетворяет уравнениям (17). Используя обозначение $A(t) = f_z[z(t, q_0)]$, сформулируем следующее утверждение.

Утверждение 2. Вектор-функция $\psi(t) = \Psi(t)a$ (где a — произвольный вектор) является решением задачи Коши $\dot{\psi} = A(t)\psi$, $\psi(0) = a$.

В самом деле, $\dot{\psi} = \dot{\Psi}a$. Но $\dot{\Psi} = A\Psi$; следовательно, мы имеем $\dot{\psi} = A\Psi a = A\psi$. Кроме того, $\psi(0) = \Psi(0)a = a$. Утверждение доказано.

Решение неоднородного уравнения (16) можно найти методом вариации произвольных постоянных. Ищем решение в виде $y(t) = \Psi(t)\, a(t)$, где $a(t)$ — подлежащая определению вектор-функция. Подставим эту конструкцию в (16): $\dot{\Psi}a + \Psi\dot{a} = A\Psi a + R$. В силу $\dot{\Psi} = A\Psi$ имеем $\Psi\dot{a} = R$, т.е.

$$\dot{a} = \Psi^{-1}(t)\, R(t).$$

Это уравнение интегрируется в квадратурах:

$$a(t) = \int_0^t \Psi^{-1}(\tau)\, R(\tau)\, d\tau, \qquad y(t) = \Psi(t) \int_0^t \Psi^{-1}(\tau)\, R(\tau)\, d\tau.$$

Подводя итог, получаем «явное» выражение для решения задачи (16):

$$y(t) = z_q(t, q_0) \int_0^t z_q^{-1}(\tau, q_0)\, R(\tau)\, d\tau. \tag{18}$$

Итак, мы рассмотрели проблемы, связанные с вычислением формального ряда Пуассона. Обсудим содержательный вопрос: что дает ряд Пуассона для описания траектории возмущенной системы? В частности, нельзя ли его использовать для оценки $x(t, q_0)$ на большом интервале времени?

Оценка ряда Пуассона. Рассмотрим частичные суммы ряда Пуассона, опуская аргумент q_0:

$$x_0(t) = X_0(t) = z(t),$$

$$x_1(t) = X_0(t) + \varepsilon X_1(t),$$

$$x_2(t) = X_0(t) + \varepsilon X_1(t) + \varepsilon^2 X_2(t),$$

. . . . . .

Оценим их уклонение от траектории $x(t)$. Начнем с оценки $x_0(t) - x(t)$.

Теорема 2. Решения систем уравнений

$$\dot{x}_0 = f(x_0), \quad x_0(0) = q_0, \qquad \dot{x} = f(x) + \varepsilon F(x), \quad x(0) = q_0,$$

удовлетворяют неравенству $r(t) \equiv \|x(t) - x_0(t)\| \leqslant (B/C)e^{Ct}$, где C — константа Липшица функции f, B — оценка функции F: $\|F(x)\| \leqslant B$.

Будем предполагать, что условие Липшица и ограниченность F выполнены «всюду», хотя на самом деле достаточно потребовать этого лишь в некоторой окрестности множества точек x, пробегаемых траекторией $x(t)$.

Доказательство. Вычитая уравнения для x и x_0, получаем уравнение для их разности:

$$\frac{d}{dt}(x - x_0) = f(x) - f(x_0) + \varepsilon F(x), \quad (x - x_0)|_{t=0} = 0.$$

Выпишем уравнение для $r^2(t) \equiv (x - x_0, x - x_0)$:

$$2r\dot{r} = 2(\dot{x} - x_0, x - x_0) \leqslant 2\|\dot{x} - \dot{x}_0\| \, \|x - x_0\| \leqslant$$
$$\leqslant 2r\{\|f(x) - f(x_0)\| + \varepsilon\|F(x)\|\} \leqslant 2r\{C\|x - x_0\| + \varepsilon B\}.$$

Итак, имеем оценку (дифференциальное неравенство) $\dot{r} \leqslant Cr + \varepsilon B$. Отсюда утверждение теоремы получается с помощью леммы, полезной и в других вопросах.

Лемма 1 (Гронуола). Если гладкая функция $r(t) \geqslant 0$ удовлетворяет неравенству $\dot{r} \leqslant Cr + A$ $(r(0) = 0)$, то $r(t) \leqslant (A/C)e^{Ct}$.

Доказательство. Введем функцию $R(t)$ как решение дифференциального уравнения $\dot{R} = CR + A$ $(R(0) = 0)$. Очевидно, $R(t) = (A/C)(e^{Ct} - 1) \leqslant (A/C)e^{Ct}$. Покажем, что $r(t) \leqslant R(t)$. Это есть простое следствие (при $r(t) = R(t)$) соотношений

$$\frac{d}{dt}[r(t) - R(t)] \leqslant 0, \qquad [r(0) - R(0)] = 0.$$

Поэтому функция $r(t)$ не может обогнать в росте $R(t)$.

Из доказанной теоремы следует, что траектории возмущенной и невозмущенной систем ε-близки друг к другу. Но крайне неприятный множитель e^{Ct} ограничивает действие этого утверждения такими временами, для которых $e^{Ct} \ll 1/\varepsilon$. Это соотношение выполняется для любого t при достаточно малом ε, однако t и ε в оценку входят «неравноправно» (ε линейно, t экспоненциально). Поэтому полученная оценка теряет смысл при $t = O(\ln \varepsilon)$. Правда, она получена при очень грубой информации о функции f: используется только условие Липшица или, что более или менее то же самое, ограниченность $\|f_x\|$. В § 7 мы видели, что привлечение некоторых дополнительных свойств матрицы f_x может существенно улучшить оценки подобного рода (см. ниже утверждение 8).

Перейдем к оценке первого приближения по ряду Пуассона, т.е. к оценке $\|x_1 - x\|$. Выпишем уравнение для $x_1(t)$:

$$\dot{x}_1 = \dot{X}_0 + \varepsilon \dot{X}_1 = f(X_0) + \varepsilon f_x[z(t)]X_1 + \varepsilon F[z(t)].$$

Поскольку $\varepsilon X_1 = x_1 - x_0 = x_1 - z$, запишем уравнение в форме

$$\dot{x}_1 = f(z) + f_z(z)(x_1 - z) + \varepsilon F(z).$$

Уравнение $\dot{x} = f(x) + \varepsilon F(x)$ для x после простых тождественных преобразований примет вид

$$\dot{x} = f(x) + \varepsilon F(x) = f(z + x - z) + \varepsilon F(z + x - z) = f(z) + \\ + f_x(z)(x - z) + O(\|x - z\|^2) + \varepsilon F(z) + \varepsilon F_z(z)(x - z) + O(\varepsilon^3).$$

Здесь мы использовали уже доказанное (для конечного интервала времени) соотношение $\|x - z\| = O(\varepsilon)$.

Итак, мы имеем

$$\dot{x}_1 = f(z) + f_z(z)(x_1 - z) + \varepsilon F(z), \quad x_1(0) = q_0,$$

$$\dot{x} = f(z) + f_z(z)(x - z) + \varepsilon F(z) + O(\varepsilon^2).$$

Эти два уравнения отличаются друг от друга наличием члена $O(\varepsilon^2)$ во втором. Применяя те же оценки, что и при доказательстве предыдущей теоремы, получаем аналогичную оценку:

$$\|x_1(t) - x(t)\| \leqslant e^{Ct}\, O(\varepsilon^2).$$

Здесь постоянная Липшица (по переменным типа x, x_1) правой части относится к линейной правой части, т.е. по существу совпадает с $\|f_z(z)\|$. Разумеется, мы неявно использовали предположение о гладкости функций большей, чем этого требовала теорема 2.

Тем же способом можно доказать теоремы об $O(\varepsilon^3)$-точности второго приближения $x_2(t)$ и т.д. Так как мы предполагаем исполь-

зовать несколько членов ряда Пуассона только на интервале времени, равном одному периоду, множитель e^{Ct} нас не стеснит, и мы ограничимся этими сравнительно грубыми оценками.

Подведем итог. Для приближенного решения возмущенной системы

$$\dot{x} = f(x) + \varepsilon F(x), \quad x(t_0) = q,$$

может быть построен формальный ряд Пуассона

$$z(t - t_0, q) + \varepsilon X_1(t - t_0, q) + \varepsilon^2 X_2(t - t_0, q) + \ldots,$$

частичные суммы которого приближают $x(t)$ (при соответствующих требованиях к гладкости f и F) с точностью $O(\varepsilon)$, $O(\varepsilon^2)$, ... Нас интересует смещение за период $T(q)$.

Запишем формальный ряд:

$$x(T(q), q) = z(T(q), q) + \varepsilon X_1(T(q), q) + \varepsilon^2 X_2(T(q), q) + \ldots$$

Вводя обозначение $P_i(q) = X_i(T(q), q)$, получаем используемую в дальнейшем формулу

$$x(T(q), q) = q + \varepsilon P_1(q) + \varepsilon^2 P_2(q) + \varepsilon^3 P_3(q) + \ldots$$

Обрывая ее на каком-то члене, имеем формулу соответствующей точности.

«Остаток» ряда Пуассона заменим выражением $O(\varepsilon^k)$, считая, что эта величина равномерно (во всем интересующем нас диапазоне изменения q) оценивается следующим образом:

$$\|O(\varepsilon^k)\| \leqslant C_k \varepsilon^k.$$

Равномерность приведенной оценки является, конечно, следствием равномерности оценок тех или иных производных функций f, F. Эту сторону проблемы мы не будем оформлять с должной строгостью, но о ней все-таки стоит помнить.

Полученные для ряда Пуассона оценки не позволяют применить его для расчета влияния возмущения на больших временах. Подобные оценки сверху, как неоднократно подчеркивалось, грубы, они получены без использования конкретных свойств $f(z)$. Но, может быть, более аккуратные оценки привели бы к другим выводам, тем более что в рассматриваемой ситуации есть принципиальные доводы в пользу обязательного наличия у f_z определенных положительных свойств? Имеется в виду следующее. Предположение о периодичности траекторий невозмущенной системы есть свойство, близкое по существу к нейтральности системы, т.е. к неотрицательности матрицы

$$\operatorname{Re} f_z(z) = 0.5(f_z + f_z^*).$$

(Из этого в § 7 удалось получить в аналогичном вопросе ослабление влияния экспоненциального множителя в оценке.)

Тем не менее ряд Пуассона для расчета на далекие времена не годится. Это следует, в частности, из простых примеров. Так, для системы

$$\dot{x} = y, \quad \dot{y} = -x + \varepsilon x^3, \quad x(0) = a, \quad y(0) = 0,$$

ряд Пуассона в первом приближении дает

$$x_1(t) = a\cos t - \frac{3}{8}\varepsilon a^3 t \sin t + \frac{1}{32}\varepsilon a^3 \cos 3t.$$

Слагаемое $(3/8)\varepsilon a^3 t \sin t$ при $t = O(\varepsilon^{-1})$ есть $O(1)$, что противоречит известному интегралу энергии

$$0.5y^2 + 0.5x^2 + 0.25\varepsilon x^4 = \text{const} = 0.5a^2 + 0.25\varepsilon a^4.$$

Содержащие степени t члены ряда Пуассона типичны. Они сильно снижали ценность этого аппарата в небесной механике, где получили специальное название «секулярные» члены (т.е. «вековые», влияние которых растет с ростом времени). Борьба с такими членами в конце концов привела к разработке методов осреднения.

Теперь мы имеем в своем распоряжении технический аппарат, с помощью которого можно обосновать стробоскопический метод.

Обоснование стробоскопического метода. Для исследования возмущенного движения рассмотрим моменты $t_0, t_1, t_2, \ldots$ стробоскопии и положения системы в эти моменты времени

$$q_0, \quad q_1 = x(t_1, q_0), \quad q_2 = x(t_2, q_0), \quad \ldots$$

Эти величины связаны соотношениями (ограничимся пока самым грубым приближением)

$$q_{k+1} = q_k + \varepsilon P_1(q_k) + O(\varepsilon^2).$$

Вводя уравнение в медленном времени:

$$dQ/d\tau = P_1(Q), \quad Q(0) = q_0,$$

приходим к разностным соотношениям для $Q_k = Q(k\varepsilon)$:

$$Q_{k+1} = Q_k + \varepsilon P_1(Q_k) + O(\varepsilon^2)$$

(разумеется, при соответствующей гладкости P_1).

Итак, для величин q_k и Q_k мы имеем разностные уравнения, отличающиеся друг от друга только членами $O(\varepsilon^2)$. Из теорем об устойчивости разностных уравнений (см. § 7) получаем следующие утверждения.

Утверждение 3. Пусть $\|\partial P_1/\partial Q\| \leqslant C$ и $\varepsilon C < 1$. Тогда имеем оценку $\|Q_k - q_k\| \leqslant O(\varepsilon) e^{Ck\varepsilon}$.

Эта оценка имеет ценность при $k = O(\varepsilon^{-1})$.

Утверждение 4. Пусть матрица $\operatorname{Re}(\partial P_1/\partial Q) \leqslant -a^2 < 0$. Тогда $\|Q_k - q_k\| \leqslant O(\varepsilon)$ при всех $k > 0$.

Иными словами, если траектория $Q(\tau)$ уравнения в медленном времени асимптотически устойчива, аппарат осреднения работает при всех $t > 0$.

Перейдем к более точным (по ε) вариантам теории. Используем два члена ряда Пуассона. При этом разностные соотношения для q_k примут форму

$$q_{k+1} = q_k + \varepsilon P_1(q_k) + \varepsilon^2 P_2(q) + O(\varepsilon^3).$$

Уравнение в медленном времени следует уточнить таким образом, чтобы аналогичное соотношение для его решения (а это просто отрезок ряда Тейлора) давало тоже самое разностное соотношение с точностью до $O(\varepsilon^3)$. Естественно взять это уравнение в виде

$$\frac{dQ}{d\tau} = P_1(Q) + \varepsilon R(Q),$$

где $R(Q)$ — подлежащая определению поправка.

Разложение $Q(k\varepsilon + \varepsilon)$ в ряд Тейлора по ε дает

$$Q_{k+1} = Q_k + \varepsilon \frac{dQ}{d\tau} + \frac{\varepsilon^2}{2} \frac{d^2Q}{d\tau^2} + O(\varepsilon^3) =$$

$$= Q_k + \varepsilon [P_1(Q_k) + \varepsilon R(Q_k)] + \frac{\varepsilon^2}{2} \frac{\partial P_1}{\partial Q} P_1(Q_k) + O(\varepsilon^3).$$

Здесь мы использовали соотношения

$$Q_\tau = P_1 + \varepsilon R, \quad Q_{\tau\tau} = \frac{\partial}{\partial Q} [P_1(Q) + \varepsilon R(Q)] \frac{dQ}{d\tau} = \frac{\partial P_1}{\partial Q} P_1(Q) + O(\varepsilon).$$

Совпадение разностных уравнений для q_k и Q_k с точностью до $O(\varepsilon^3)$ будет обеспечено при

$$R(Q) + \frac{1}{2} \frac{\partial P_1}{\partial Q} P_1(Q) = P_2(Q).$$

Итак, получено уравнение в медленном времени, имеющее второй порядок точности:

$$\frac{dQ}{d\tau} = P_1(Q) + \varepsilon \left[P_2(Q) - \frac{1}{2} \frac{\partial P_1}{\partial Q} P_1(Q) \right].$$

Относительно связи q_k с Q_k при использовании решения этого уравнения можно высказать утверждения, аналогичные утверждениям 1

и 2 с заменой в них $O(\varepsilon)$ на $O(\varepsilon^2)$. Кроме того, справедливо следующее специальное утверждение.

Утверждение 5. Пусть матрица $\mathrm{Re}(\partial P_1/\partial Q) \leqslant 0$. Тогда имеет место оценка $\|q_k - Q_k\| \leqslant O(\varepsilon)\, e^{Ck\varepsilon^2}$. С потерей одного порядка в ε оценка сохраняет смысл для $k = O(\varepsilon^{-2})$, т.е. на отрезках физического времени длиной $O(\varepsilon^{-2})$. Это утверждение следует из теоремы об устойчивости разностных схем (см. § 7). Строго говоря, прямое их применение потребовало бы предположения неположительности матрицы $\mathrm{Re}\,\frac{\partial}{\partial Q}(P_1 + \varepsilon R)$, но почти очевидная модификация доказательства позволяет формулировать предположение в терминах только $P_1(Q)$.

Аналогичным образом можно строить уравнения в медленном времени и последующих порядков. Они повышают точность аппарата по ε, но не снимают ограничений времени, на котором он работает ($O(\varepsilon^{-1})$ в общем случае). Возможность распространения оценок на большие времена существенно связана с характером получившегося уравнения в медленном времени. Играет роль и его конкретная траектория. В общем случае это уравнение нелинейно и свойства матрицы $\partial P_1/\partial Q$ могут быть разными в окрестности разных траекторий.

Что дает решение уравнения в медленном времени? Предположим, что уравнения (9) и (11) проинтегрированы и известны функции $Q(\tau)$, $t(\tau)$, $\tau(t)$. Пусть задан момент физического времени t. Что можно сказать о точке $x(t, q_0)$? Не претендуя на строгость, можно сформулировать такой ответ. Вычислим $\tau(t)$ и период $T = T[Q(\tau(t))]$. Тогда в момент физического времени t', лежащий где-то на интервале $|t' - t(\tau)| < T/2$, точка $x(t', q_0)$ попадает в $O(\varepsilon)$-окрестность точки $Q(\tau(t))$. Этой информации часто бывает достаточно для физических приложений. Иначе можно сказать и так. Если не обращать внимания на величины $O(\varepsilon)$, то точка $Q(\tau(t))$ определяет положение $x(t, q_0)$ «с точностью до положения на орбите невозмущенного движения, проходящей через точку $Q(\tau(t))$».

С вычислительной точки зрения основное преимущество перехода от описания траектории исходным уравнением $\dot{x} = f + \varepsilon F$ к описанию уравнением в медленном времени $Q_\tau = P_1(Q)$ состоит в том, что траектория $Q(\tau)$ является гладкой относительно интервалов физического времени тем больших периода невозмущенного движения, чем меньше ε. Численное интегрирование уравнения для Q может осуществляться шагом, не зависящим от ε и включающим в себя сразу много периодов физического времени.

Что касается термина «осреднение быстрых вращений», то он связан с характером вычисления функции $P_1(Q)$ по формуле Пуанкаре:

$$P_1(q) = \int_0^{T(q)} z_q(T(q), q)\, z_q^{-1}(t, q)\, F[z(t, q)]\, dt,$$

т.е. P_1 получается специфическим осреднением возмущения F вдоль траектории невозмущенного движения за его период. В сложных задачах $P_1(q)$ может определяться приближенным интегрированием на интервале времени $T(q)$.

Выбор медленных переменных. Роль периодичности невозмущенного движения. Изложенная выше теория, приводящая к уравнениям в медленном времени, основана на следующих важных свойствах рассматриваемой задачи.

1. Рассматривается влияние малых возмущений на систему, невозмущенное движение которой считается известным.

2. В случае, когда $z(t - t_0, q)$ — периодическое при всех q движение, удается найти «медленные» переменные, т.е. величины, которые на траектории $x(t)$ за период $T(q)$ изменяются на величины $O(\varepsilon)$.

Однако роль периодичности невозмущенного движения этим не исчерпывается. Покажем, что медленные переменные всегда есть и их выбор более или менее очевиден (для этого периодичность z не нужна). Существенно то, что периодичность z позволяет построить уравнения для медленных переменных, сформулированные в терминах тех же самых медленных переменных, т.е. уравнения в медленном времени оказываются «замкнутыми».

Посмотрим, как далеко можно продвинуться по этому пути, не используя периодичности. Ради простоты мы фиксируем $t_0 = 0$. Общее решение системы (2) будем писать в виде $z(t, q)$. Итак, первый вопрос: существуют ли медленные переменные в системе (3) и каковы они? Точный смысл этого вопроса: можно ли найти замену переменных $y = Y(x)$, в результате которой систему (1) удастся записать в виде $\dot{y} = \varepsilon R(y)$? Ответ почти очевиден. В качестве «медленных переменных» нужно взять величины, которые на траектории невозмущенной системы остаются постоянными, т.е. любую полную систему первых интегралов системы (1). Таковыми, в частности, являются величины q_0.

Итак, нужно перейти от переменных (t, x) к переменным (t, q). Что это значит? Очевидно, осью t в новой системе координат (т.е. линий $q = \text{const}$) будет траектория $z(t, q)$. Для того чтобы точке (t, x) сопоставить точку (t, q), нужно проинтегрировать систему $\dot{z} = f$ с начальными данными $z(t) = x$ в обратном направлении (от

t до нуля), тогда требующееся значение $q = z(0)$. Этим алгоритмом и определяется отображение $(t, x) \to (t, q)$. В исходной системе координат траектория $x(t)$ возмущенной системы меняется сильно (так же как и $z(t)$), а в системе координат (t, q) — медленно. Хотя такая замена переменных кажется не очень эффективной, на самом деле все не так уж сложно, если (это очень существенно!) нам известно общее решение $z(t, q)$.

Нетрудно понять, что если мы станем искать решение возмущенной системы в виде $x(t) = z(t, q(t))$, то $q(t)$ будут меняться медленно. Получим уравнение для эволюции $q(t)$. Подставляя x в уравнение (3), имеем

$$z_t(t, q(t)) + z_q(t, q(t))\, \dot{q} = f[z(t, q(t))] + \varepsilon\, F[z(t, q(t))].$$

Но $z_t = f$; следовательно,

$$\dot{q} = \varepsilon\, z_q^{-1}(t, q(t))\, F(z(t, q(t)). \tag{19}$$

Итак, мы получили (не используя предположения о периодичности z) уравнение для медленно меняющихся переменных. Эта процедура носит название «метод вариации произвольных постоянных». Осталось ввести медленное время $\tau = \varepsilon t$ и записать систему (19) в виде $q_\tau = \mathscr{P}(q, \tau/\varepsilon)$. Наличие в $\mathscr{P}$ зависимости от «быстрого» переменного τ/ε существенно осложняет дело.

Рассмотрим процедуру численного интегрирования уравнения (19) с малым шагом Δ (при этом $dt = \Delta/\varepsilon$ может быть очень большим!):

$$q(\tau + \Delta) = q(\tau) + \int_\tau^{\tau+\Delta} \mathscr{P}[q(\tau'), \tau'/\varepsilon]\, d\tau'.$$

Упростим эту формулу, заменив ее приближенной. Воспользуемся тем, что на интервале $[\tau, \tau + \Delta]$ величина q изменяется на $O(\Delta)$; поэтому

$$q(\tau + \Delta) = q(\tau) + \int_\tau^{\tau+\Delta} \mathscr{P}[q(\tau), \tau'/\varepsilon]\, d\tau' + O(\Delta^2),$$

или

$$q(\tau + \Delta) = q(\tau) + \Delta P(q(\tau)) + O(\Delta^2),$$

где

$$P(q) = \frac{1}{\Delta} \int_\tau^{\tau+\Delta} \mathscr{P}(q, \tau'/\varepsilon)\, d\tau',$$

т.е. функция $P(q)$ получена операцией усреднения по явно входящему времени функции $\mathscr{P}(q, t)$.

Вышеприведенные выкладки приводят к уравнению в медленном времени в том случае, если существует не зависящий от τ и Δ предел

$$P(q) = \lim_{\varepsilon \to 0} \left\{ \frac{1}{\Delta} \int_{\tau}^{\tau+\Delta} \mathscr{P}(q, \tau/\varepsilon)\, d\tau' \right\}. \tag{20}$$

В этом случае мы получаем уравнение в медленном времени, содержащее только медленно меняющиеся переменные: $q_\tau = P(q)$. Существование предела (20) типично при периодической или почти периодической зависимости $\mathscr{P}(q, t)$ от t. Но согласно (19) функция $\mathscr{P}(q, t) = z_q^{-1}(t, q)\, F(z(t, q))$. Отсюда ясно, какую роль играет периодичность невозмущенного движения в получении замкнутых уравнений в медленном времени.

Простейшая двухчастотная задача. До сих пор мы изучали наиболее простой случай, когда в невозмущенной системе был только один, общий для всех компонент, период $T(q)$. Как отмечалось, наличие в задаче разных периодов (хотя бы двух) сильно осложняет ситуацию. Некоторое представление об этом даст следующий анализ.

Рассмотрим задачу о малом возмущении периодического движения периодической силой. Имеется невозмущенная система (1) и ее общее решение $z(t - t_0, q_0)$, периодическое с периодом $T(q_0)$. Предположим, что возмущенная система описывается уравнением $\dot{x} = f(x) + \varepsilon F(x, t)$. Возмущающую силу F будем считать π-периодической по t (период силы постоянен).

Окрестность точки резонанса. Пусть в некоторой окрестности точки q_0 имеет место «почти резонанс». Существуют некоторые, не очень большие целые числа m и n, такие, что $n\,T(q) - m\pi = \eta(q)$. При этом «рассогласование» $\eta(q)$ мало в окрестности q_0, т.е. $|\eta(q)| \ll \pi$. Это может быть величина, сравнимая с ε или $\sqrt{\varepsilon}$, — в зависимости от этого теория движения «в медленном времени» будет иметь ту или иную точность.

Можно построить ряд Пуассона, позволяющий рассчитывать возмущенное движение на конечном отрезке времени (период или несколько периодов). Если мы непосредственно используем стробоскопический метод (первого порядка точности, для простоты), то ничего хорошего не получится. Прежде всего заметим, что члены ряда Пуассона теперь надо обозначить $X_i(t, t_0, q_0)$, так как в возмущенную систему явно входит время.

Рассмотрим последовательность моментов t_k стробоскопии и положений $x(t)$ в эти моменты:

$$q_{k+1} = q_k + \varepsilon X_1(t_k + T_k, t_k, q_k) + O(\varepsilon^2), \quad t_{k+1} = t_k + T_k,$$

где $T_k = T(q_k)$, $q_k = x(t_k)$. Эти разностные соотношения, однако, нельзя рассматривать как процедуру приближенного интегрирования некоторой системы дифференциальных уравнений «в медленном времени» (считая ε шагом интегрирования). Дело в том, что аргументы $t_k + T_k$, t_k за один шаг меняются на $O(1)$; такого же порядка, следовательно, и изменение X_1 за один шаг. Если бы система была точно резонансной ($\eta \equiv 0$), то мы имели бы одночастотный случай: нужно только рассматривать «большой» период $nT = m\pi$. Это замечание подсказывает и путь анализа «почти резонансной» ситуации: надо использовать стробоскопический метод с таким большим периодом.

Итак, рассмотрим последовательность моментов t_k и положений системы $q_k \equiv x(t_k, t_0, q_0)$:

$$q_{k+1} = q_k + \varepsilon X_1(t_k + nT_k, t_k, q_k) + O(\varepsilon^2), \quad t_{k+1} = t_k + nT_k.$$

В дальнейшем нам будет полезно следующее почти очевидное утверждение.

Утверждение 6. Система $\dot{x} = f(x) + \varepsilon F(x, t)$ инвариантна относительно сдвига времени на π.

Уточним аналитический смысл этого утверждения. Пусть $x(t, t_0, q_0)$ — общее решение системы, т.е.

$$x_t(t, t_0, q_0) = f[x(t, t_0, q_0)] + \varepsilon F[x(t, t_0, q_0), t], \tag{21}$$

$$x(t_0, t_0, q_0) = q_0, \quad \forall\, t, t_0, q_0. \tag{22}$$

Введем функцию $y(t, t_0, q_0) \equiv x(t+\pi, t_0 + \pi, q_0)$. Тогда $y(t, t_0, q_0) \equiv \equiv x(t, t_0, q_0)$. Доказательство состоит в том, что для функции y проверяются условия (21), (22), определяющие функцию x однозначно. Опустим эти простые выкладки, в которых используется π-периодичность F по t.

Так как мы используем ряды Пуассона в моменты $t_{k+1} = = t_k + n\,T(q_k)$, то в рекуррентном соотношении остаются только два существенных аргумента: t_k и q_k. В этом случае рекуррентное соотношение можно записать в виде

$$q_{k+1} = q_k + \varepsilon \mathscr{P}(t_k + nT_k, q_k) + O(\varepsilon^2),$$

положив $\mathscr{P}(t, q) = X_1(t, t - n\,T(q), q)$.

В силу доказанной выше инвариантности возмущенной системы относительно сдвига времени на π, $\mathscr{P}$ является π-периодической по t функцией. Это, впрочем, почти очевидно и без доказательства. Строя ряды Пуассона в точках (t_0, q_0) и $(t_0 + \pi, q_0)$, мы будем иметь дело с одними и теми же объектами: траектория невозмущен-

ной системы вообще не реагирует на такой сдвиг по времени, а сила F π-периодична. В результате рекуррентное соотношение можно переписать в виде

$$q_{k+1} = q_k + \varepsilon \mathscr{P}(t_k + nT_k - m\pi, q_k) + O(\varepsilon^2).$$

Вводя рассогласование фаз $\eta(q)$, имеем

$$q_{k+1} = q_k + \varepsilon \mathscr{P}(t_k + \eta(q_k), q_k) + O(\varepsilon^2).$$

Теперь осталось учесть еще одну «медленную переменную» — фазу $\alpha_{k+1} = \alpha_k + \eta(q_k)$, в терминах которой цепочка разностных уравнений запишется в форме

$$t_{k+1} = t_k + n\,T(q_k), \qquad \alpha_{k+1} = \alpha_k + \eta(q_k),$$

$$q_{k+1} = q_k + \varepsilon \mathscr{P}(\alpha_k, q_k) + O(\varepsilon^2).$$

Так как мы предполагаем рассогласование $\eta(q)$ малым (сравнимым с ε или $\sqrt{\varepsilon}$), то поставленная цель достигнута: в цепочке разностных уравнений все аргументы за один шаг изменяются медленно (q на $O(\varepsilon)$, α на η). Можно перейти к уравнениям в медленном времени τ:

$$\frac{dt}{d\tau} = \frac{T(q)}{\varepsilon}, \qquad \frac{d\alpha}{d\tau} = \frac{\eta(q)}{\varepsilon}, \qquad \frac{dq}{d\tau} = \mathscr{P}(\alpha, q).$$

Случай несоизмеримости периодов. Пусть периоды $T(q)$ и π «несоизмеримы», т.е. $nT \neq m\pi$ при любых целых числах m и n. Практически нет особой разницы между несоизмеримостью и «резонансом»: $nT \approx m\pi$ при очень больших значениях m и n. Разумеется, понятие «большие m и n» должно быть согласовано с величиной ε. Но в чистой теории все объекты фиксируются, а величина ε считается настолько малой, насколько этого требует доказательство оценки. В этом смысле, конечно, термин «несоизмеримость» нужно понимать буквально, как он трактуется в теории чисел.

Для задачи о возмущении T-периодического решения малой π-периодической силой из (19) имеем следующее «разностное» соотношение с шагом Δ (малым для медленного времени, но большим с точки зрения «физического» времени):

$$q(\tau + \Delta) = q + \int_{\tau}^{\tau+\Delta} z_q^{-1}(q, \tau'/\varepsilon)\, F[z(q, \tau'/\varepsilon), \tau'/\varepsilon]\, d\tau' + O(\varepsilon^2). \qquad (23)$$

Напомним, что здесь $q = q(\tau)$ (эта величина остается постоянной при интегрировании).

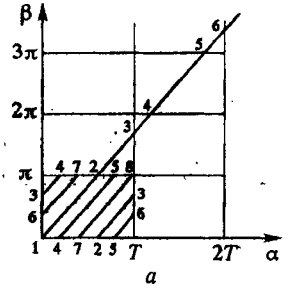
Нас будет интересовать оценка интеграла, который можно записать и в терминах физического времени $t = \tau/\varepsilon$:

$$\varepsilon \int_{t}^{t+\Delta/\varepsilon} z_q^{-1}(q, t')\, F[z(q, t'), t']\, dt'. \qquad (24)$$

Обратим внимание на то, что в подынтегральную функцию t' входит тремя разными способами. Первые два вхождения t' связаны с решением невозмущенной системы, по этим t' подынтегральная функция T-периодична. Третье вхождение t' связано с зависимостью F от t, по этому аргументу подынтегральная функция π-периодична.

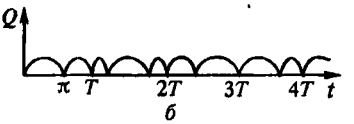
Подынтегральную функцию можно записать в виде функции $Q(\alpha, \beta)$, игнорируя аргумент q, который при интегрировании остается постоянным. Итак, речь идет о приближенном вычислении интеграла от двояко-периодической функции Q на линии $\alpha = \beta = t'$, $t' \in [t, t + \Delta/\varepsilon]$. Интервал интегрирования большой в том смысле, что на нем укладывается большое число периодов (как π, так и T, которые мы считаем величинами одного порядка).

На рис. 28*а* изображена плоскость (α, β), разделенная на прямоугольники $T \times \pi$, и линия интегрирования. В силу свойств $Q(\alpha, \beta)$ можно ограничится только одним прямоугольником, отождествив точки его противоположных сторон. Такой прямоугольник называют тором. Изображая линию интегрирования на торе, при ее выходе на границу прямоугольника скачком следует перейти на противоположную сторону.



а

Образ линии на торе образует так называемую обмотку тора. Если периоды соизмеримы, то через время $nT = m\pi$ линия на торе замкнется. Если периоды несоиз-



б

Рис. 28

меримы, линия равномерно заполняет тор. При достаточной длине линии (т.е. если ε достаточно малая величина) среднее значение Q вдоль линии почти совпадает со средним значением по тору:

$$\frac{\varepsilon}{\Delta} \int_{t}^{t+\Delta/\varepsilon} Q(t', t')\, dt' \approx \frac{1}{T\pi} \int_{0}^{\pi} \int_{0}^{T} Q(\alpha, \beta)\, d\alpha\, d\beta \qquad (25)$$

(в пределе при $\varepsilon \to 0$ эти величины совпадают).

Интеграл в правой части (25) вычислить легче, чем интеграл на линии. Поясним это. Пусть $Q(\alpha, \beta)$ — простая функция, для наглядности обращающаяся в нуль на границах прямоугольника. Функция на линии $Q(t', t')$ показана на рис. 28*б*. Это сложная функция: вычисление интеграла по линии с помощью какой-либо квадратурной формулы требует большого числа узлов (порядка $O(\varepsilon^{-1})$). Обозначим среднее значение $Q(\alpha, \beta)$ по тору $P(q)$. (Напомним, что Q зависит от q). Тогда уравнение в медленном времени принимает вид $dq/d\tau = P(q)$. В случае резонанса среднее значение Q вдоль линии $\alpha = \alpha_0 + t$, $\beta = t$ превращается в среднее по периоду $mT = n\pi$, но зависит еще и от начальной фазы α_0. В случае несоизмеримых периодов эта линия «равномерно» (при достаточно большом Δ/ε) покрывает тор независимо от того, в какой точке (α_0, β_0) она начинается.

Замечание. Смысл соотношения (25) можно пояснить так. Построим около линии $\alpha = \beta$ параллелограмм малой ширины h. Подберем h таким образом, чтобы площадь параллелограмма равнялась πT, т.е. $h(\Delta/\varepsilon) = \pi T$. Интеграл по этой узкой «ленте» при очень малой ширине $h = \varepsilon\pi T/\Delta$ почти совпадает с интегралом по линии, умноженным на h. При «обмотке» «лента» заполнит ячейку πT (площади перекрытий и пустот стремятся к нулю при $\varepsilon \to 0$). Отсюда и следует (25).

Ряд Пуассона в специальном случае. При построении метода осреднения для многочастотных задач мы заинтересованы в использовании ряда Пуассона на возможно большем интервале времени. Это требует улучшения стандартных оценок за счет привлечения дополнительных предположений о функции f. Рассмотрим одно из таких уточнений, в котором используется предположение $A = 0.5(f_z(z) + f_z^*(z)) \leqslant 0$. В известном смысле можно говорить о том, что траектории невозмущенной системы «нейтральны», или «не неустойчивы», и что матрица f_z не имеет собственных чисел в правой части плоскости комплексного переменного, но может иметь их на мнимой оси.

Оценим расхождение траекторий возмущенной и невозмущенной систем за время $t = O(\varepsilon^{-1})$.

Утверждение 7. В рассматриваемом случае для траекторий систем

$$\dot{x} = f(x) + \varepsilon F(x), \quad \dot{z} = f(z), \quad x(0) = z(0) = q_0,$$

справедлива оценка $\|x(t) - z(t)\| \leqslant O(\varepsilon t)$.

Доказательство. Имеем уравнение

$$\frac{d}{dt}(x - z) = f(x) - f(z) + \varepsilon F(x), \quad (x - z)|_{t=0} = 0.$$

Введя $r^2(t) \equiv (x-z, x-z)$, вычислим и оценим производную:

$$2r\dot{r} = 2(\dot{x}-\dot{z}, x-z) = 2(f(x)-f(z), x-z) + 2\varepsilon(F(x), x-z).$$

В § 7 в аналогичном случае (при $A \leqslant 0$) было показано, что $(f(x)-f(z), x-z) \leqslant 0$. Предположим, что $\|F(x)\| \leqslant B$ при всех x. Оценивая обычным образом $(F(x), x-z) \leqslant Br$, получаем $\dot{r} \leqslant \varepsilon B$, откуда и следует утверждение.

Теперь перейдем к оценке первого приближения по ряду Пуассона $x_1(t) = z(t) + \varepsilon X_1(t)$. Для x_1 имеем уравнение

$$\dot{x}_1 = f(z) + f_z(z)(x_1 - z) + \varepsilon F(z), \quad x_1(0) = q_0.$$

Преобразуем уравнение для x так, как это делалось раньше, но с более подробным представлением членов $O(\varepsilon^2)$:

$$\dot{x} = f(x) + \varepsilon F(x) = f(z + (x-z)) + \varepsilon F(z + (x-z)) =$$
$$= f(z) + f_z(z)(x_1 - z) + O(\|f_{zz}\| \, \|x-z\|^2) +$$
$$+ \varepsilon F(z) + \varepsilon O(\|F_z\| \, \|x-z\|).$$

Таким образом, уравнение для разности имеет вид

$$\frac{d}{dt}(x_1 - x) = f_z(z)(x_1 - z) + O(\|f_{xx}\| \, \|x-z\|^2) + \varepsilon O(\|F_z\| \, \|x-z\|).$$

Используя $A \leqslant 0$ и оценку $\|x(t) - z(t)\| \leqslant B\varepsilon t$, получаем

$$\|x_1(t) - x(t)\| \leqslant \varepsilon^2 t^3 O(\|f_{xx}\|) + \varepsilon^2 t^2 O(\|F_z\|). \tag{26}$$

Из оценки (26) непосредственно вытекают два следующих утверждения.

Утверждение 8. На временах $t = O(1/\sqrt{\varepsilon})$ первое приближение по ряду Пуассона сохраняет точность $O(\sqrt{\varepsilon})$.

Это утверждение следует из оценки первого члена правой части (26). Отметим любопытное обстоятельство. При $t = O(1/\varepsilon)$ оценка погрешности первого приближения $O(1/\varepsilon)$ хуже оценки «нулевого» приближения $O(1)$. Это — действие пресловутых «секулярных» членов: уточняя решение при малых t, они ухудшают его при больших.

Утверждение 9. Пусть невозмущенная система линейна, т.е. $f_{xx} = 0$. Тогда первое приближение имеет погрешность $O(\varepsilon)$ на временах $t \approx O(1/\sqrt{\varepsilon})$ и погрешность $O(\sqrt{\varepsilon})$ на временах $t \approx O(1/\varepsilon^{3/4})$.

Это утверждение следует из оценки величины $\varepsilon^2 t^2$. Оно представляет интерес, например, для задачи, в которой невозмущенная система распадается на несколько гармонических осцилляторов (задача об эволюции слабо связанных осцилляторов).

§ 20. Одномерные уравнения газовой динамики и их численное интегрирование

Уравнения газовой динамики сами по себе представляют большой интерес, так как ими описывается очень важные явления. Вместе с уравнениями теплопроводности, распространения электромагнитных волн и т.п. эти уравнения входят в описания большого числа сложных явлений, интересующих современную физику. Развитие вычислительной физики в значительной мере определялось задачами газовой динамики. Необходимость их эффективного решения стимулировала разработку многих новых вычислительных конструкций, которые затем успешно использовались и в других областях. Поэтому специалисту в современной вычислительной математике нужно знать основные математические факты газовой динамики, чтобы понимать возникающие вычислительные трудности и способы их преодоления.

Ниже мы опишем необходимый минимум знаний в этой области. Мы начнем с одномерной газовой динамики. Уже этот простой случай содержит характерные трудности, связанные с необходимостью расчета разрывных решений — ударных волн, контактных разрывов. Для расчета одномерных течений газа были разработаны эффективные методы, специальные приемы, которые в дальнейшем обобщались на случай более сложных (двумерных и в настоящее время даже трехмерных) течений газа.

Перейдем к формулировке задачи. Будем рассматривать модель, в которой состояние среды (газа) описывается следующими функциями, зависящими от двух независимых переменных t (время) и x (пространственная координата): $u(t, x)$ — скорость газа, $\rho(t, x)$ — плотность, $p(t, x)$ — давление, $e(t, x)$ — внутренняя энергия (удельная).

Величины e, ρ, p не являются независимыми: они связаны соотношением, называемым уравнением состояния. Это уравнение мы будем употреблять либо в форме $p = P(e, \rho)$, либо в форме $e = E(p, \rho)$. Иногда используются и другие величины, однозначно вычисляемые через любую пару термодинамических переменных (e, ρ), (p, ρ), ... (энтропия, энтальпия и т.д.). Используя эти термодинамические соотношения, газовая динамика, таким образом, ограничивается описанием явлений, протекающих в условиях локального термодинамического равновесия. Время свободного пробега молекул и его длина считаются «бесконечно малыми» по сравнению с временами и длинами, на которых происходят заметные (с точки зрения газовой динамики) изменения основных величин, описывающих состояние газа.

Уравнения газовой динамики имеют вид законов сохранения импульса, массы и полной энергии соответственно:

$$\begin{aligned} &\text{а)}\ \frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + \frac{1}{\rho}\frac{\partial p}{\partial x} = 0,\\ &\text{б)}\ \frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} + \rho\frac{\partial u}{\partial x} = 0,\\ &\text{в)}\ \frac{\partial}{\partial t}\left(e + \frac{u^2}{2}\right) + u\frac{\partial}{\partial x}\left(e + \frac{u^2}{2}\right) + \frac{1}{\rho}\frac{\partial (pu)}{\partial x} = 0. \end{aligned} \tag{1}$$

Эти дифференциальные уравнения для четырех функций u, ρ, e, p замыкаются уравнением состояния $p = P(e, \rho)$.

Уравнения газовой динамики допускают разные формы записи; они эквивалентны, если предположить непрерывную дифференцируемость функций. Из них мы отметим важную для дальнейшего *дивергентную* форму уравнений:

$$\begin{aligned} &\text{а)}\ \frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(\rho u^2 + p) = 0,\\ &\text{б)}\ \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) = 0,\\ &\text{в)}\ \frac{\partial}{\partial t}\left[\rho\left(e + \frac{u^2}{2}\right)\right] + \frac{\partial}{\partial x}\left[\rho u\left(e + \frac{u^2}{2} + \frac{p}{\rho}\right)\right] = 0. \end{aligned} \tag{2}$$

Уравнение (2а) есть сумма (1а) и (1б), умноженных на ρ и u соответственно. Уравнение (2б) прямо получено из (1б). Уравнение (2в) есть сумма (1а) и (1в), умноженных на $(e + u^2/2)$ и ρ. Каждое из этих уравнений имеет форму

$$R_t + Q_x = 0,$$

где R, Q — функции от u, ρ, e, p. Именно это обстоятельство служит основанием для термина «дивергентная форма уравнения». Она очень важна, так как из нее непосредственно следует запись уравнений в так называемой *интегральной* форме. Последняя приводит к определению *обобщенных решений уравнений газовой динамики.*

В газовой динамике нельзя обойтись классическими решениями. Напомним, что это функции, имеющие непрерывные производные и удовлетворяющие уравнениям в прямом смысле этого слова. При этом несущественно, в какой форме записаны уравнения. Многие задачи газовой динамики классических решений не имеют. Необходимо рассматривать функции, имеющие на некоторых линиях в пространстве (t, x) разрывы не только производных, но и самих функций. В этом случае понятие «решение» должно быть соответствующим образом обобщено.

Обобщенные решения уравнений газовой динамики. Пусть функции $u(t, x)$, $\rho(t, x)$, $e(t, x)$, $p(t, x)$ являются классическими решениями уравнений, записанных в дивергентной форме. Тогда они удовлетворяют и уравнениям в интегральной форме. Выведем их. Рассмотрим в плоскости (t, x) произвольный замкнутый контур Γ, ограничивающий односвязную, для простоты, область Ω. Вычислим

$$0 = \iint\limits_{\Omega} (R_t + Q_x)\, dt\, dx = \oint\limits_{\Gamma} (R\, dx - Q\, dt), \quad \forall\, \Gamma. \tag{3}$$

Равенство нулю интеграла по любому замкнутому контуру Γ есть интегральная форма уравнений. Таким образом, классические решения являются и решениями уравнений в интегральной форме. Однако эта интегральная форма может быть принята за основную, определяющую.

Итак, обобщенными решениями уравнений газовой динамики назовем функции $u(t, x)$, $\rho(t, x)$, $e(t, x)$, $p(t, x)$ удовлетворяющие интегральным соотношениям (3). При этом

$$R(\rho, u, e) = \begin{cases} \rho u, \\ \rho, \\ \rho\left(e + \frac{u^2}{2}\right), \end{cases} \qquad Q(\rho, u, e) = \begin{cases} \rho u^2 + p, \\ \rho u, \\ \rho u\left(e + \frac{u^2}{2} + \frac{p}{\rho}\right). \end{cases} \tag{4}$$

И наоборот, если для любого Γ имеет место $\oint\limits_{\Gamma} (R\, dx - Q\, dt) = 0$ и R, Q имеют производные, то почти всюду $R_t + Q_x = 0$. Проверка того, что функции u, ρ, e являются решениями уравнений газовой динамики в обобщенном смысле, носит не очень обозримый характер (нужно проверить соотношения (3) для всех Γ), но зато не требует дифференцируемости этих функций.

Другие формы уравнений газовой динамики. Разные формы записи уравнений подчеркивают тот или иной аспект описываемого ими явления. Эти формы используются для построения разностных аппроксимаций и приводят к отличающимся разностным схемам, каждая из которых может оказаться предпочтительной при расчете какого-то специального класса течений. В дальнейшем мы специально коснемся этого вопроса еще раз. Нам потребуется другая форма уравнения энергии. Из (1в) вычтем (1а), умноженное на u:

$$\frac{\partial e}{\partial t} + u\frac{\partial e}{\partial x} + \frac{p}{\rho}\frac{\partial u}{\partial x} = 0. \tag{5}$$

Эта недивергентная запись уравнения для внутренней энергии часто оказывается полезной по соображениям, которые мы подробно обсудим в § 22.

Другие формы уравнений мы получим, ограничившись простым, но очень важным в приложениях случаем *идеального газа*. Этот термин связан с конкретной формой уравнения состояния

$$e = \frac{p/\rho}{\gamma - 1}, \quad \text{или} \quad p = (\gamma - 1)e\rho,$$

где γ — постоянная. С учетом этого соотношения преобразуем уравнение (5) для e в уравнение для p:

$$\frac{\partial p}{\partial t} + u\frac{\partial p}{\partial x} + \gamma p\frac{\partial u}{\partial x} = 0. \tag{6}$$

Используя еще одну термодинамическую величину — *адиабатическую скорость звука* $c = \sqrt{\gamma p/\rho}$ (она может быть выражена через любую пару термодинамических величин, принятых за «основные»), получаем уравнение в форме

$$p_t + up_x + c^2\rho u_x = 0. \tag{7}$$

Выведем уравнение «для энтропии». Вычтем уравнения (5) и (1б), умножив их на ρ и $p/\rho = (\gamma - 1)e$ соответственно. Умножая результат на $1/(\rho e)$, группируя отдельные члены (члены с u_x, очевидно, взаимно уничтожаются) и вводя в качестве термодинамической величины энтропию идеального газа

$$S = \ln(e\rho^{1-\gamma}) = \ln\frac{p/\rho^{\gamma}}{\gamma - 1},$$

получаем уравнение «для энтропии»:

$$S_t + uS_x = 0, \quad \text{или} \quad \left(\frac{\partial}{\partial t} + u\frac{\partial}{\partial x}\right)S = 0. \tag{8}$$

Из него следует вывод: энтропия сохраняется вдоль «траектории частицы», т.е. на траектории уравнения $\dot{X} = u(t, X)$, $X(0) = X_0$.

Сформулируем это важное обстоятельство более аккуратно. Прежде всего подчеркнем, что вышеизложенные выкладки были проведены формально, в предположении, что используемые производные существуют. Другими словами, все эти уравнения равносильны (из одних следуют другие) только в случае классических решений. Пусть мы имеем классическое решение уравнений газовой динамики. Обозначим траектории частиц более аккуратно $X(t, X_0)$. Имея решение $u(t, x)$, $\rho(t, x)$, $e(t, x)$, $p(t, x)$, мы имеем и энтропию $S(t, x)$. Тогда $S(t, X(t, X_0)) = S(0, X_0)$. В частности, если в начальных данных энтропия была постоянной, она остается постоянной всюду (изоэнтропическое течение) и одно уравнение оказывается уже проинтегрированным. Еще раз подчеркнем, что все это справедливо лишь для гладких решений. Ударные волны (разрывы), которые могут возникнуть при сколь угодно гладких начальных данных, приводят к изменению энтропии.

Римановы инварианты. Характеристики. Следующая форма уравнений также оказывается очень полезной как при аналитических исследованиях, так и при конструировании приближенных методов. Сложим уравнение (1а) с умноженным на $1/(\rho c)$ уравнением (7). После очевидной группировки членов имеем

$$[u_t + (u+c)u_x] + \frac{1}{\rho c}[p_t + (u+c)p_x] = 0. \tag{9}$$

Обозначая через $(d/dt)^+$ оператор дифференцирования по направлению $dx : dt = (u+c) : 1$, можно записать это уравнение в виде

$$\left(\frac{d}{dt}\right)^+ u + \frac{1}{\rho c}\left(\frac{d}{dt}\right)^+ p = 0. \tag{10}$$

Такие же выкладки с заменой c на $-c$ дают аналогичные уравнения. В результате система уравнений газовой динамики принимает так называемую характеристическую форму:

$$\left(\frac{d}{dt}\right)^0 S = 0, \qquad \left(\frac{d}{dt}\right)^+ u + \frac{1}{\rho c}\left(\frac{d}{dt}\right)^+ p = 0, \qquad \left(\frac{d}{dt}\right)^- u + \frac{1}{\rho c}\left(\frac{d}{dt}\right)^- p = 0, \tag{11}$$

где $(d/dt)^0 = d/dt + u\, d/dx$.

Систему (11) можно сделать более прозрачной, если предположить течение изоэнтропическим. В этом случае вся термодинамика определяется одним переменным параметром, в качестве которого удобно взять скорость звука c. Выражение $dp/(\rho c)$ становится, очевидно, дифференциалом некоторой «новой» термодинамической переменной, которую мы сейчас вычислим, а уравнения газовой динамики становятся (внешне) совсем простыми. Изоэнтропичность означает, что $p(t, x) = A\,\rho^\gamma(t, x)$, $A = \text{const}$. Тогда $c^2 = A\gamma\,\rho^{\gamma-1}$. После несложных преобразований получаем $\frac{1}{\rho c}\,dp = \frac{2}{\gamma-1}\,dc$.

После внесения множителя $1/\rho c$ под знак дифференцирования уравнения (11) принимают следующую форму:

$$\begin{aligned}
&\text{а)}\ \frac{\partial S}{\partial t} + u\frac{\partial S}{\partial x} = 0, \quad \text{или} \quad \left(\frac{d}{dt}\right)^0 S = 0,\\
&\text{б)}\ \frac{\partial R^+}{\partial t} + (u+c)\frac{\partial R^+}{\partial x} = 0, \quad \text{или} \quad \left(\frac{d}{dt}\right)^+ R^+ = 0,\\
&\text{в)}\ \frac{\partial R^-}{\partial t} + (u-c)\frac{\partial R^-}{\partial x} = 0, \quad \text{или} \quad \left(\frac{d}{dt}\right)^- R^- = 0.
\end{aligned} \tag{12}$$

Здесь использованы новые переменные: $R^+ = u + 2c/(\gamma-1)$ и $R^- = u - 2c/(\gamma-1)$. Они называются *римановыми инвариантами*, так как в изоэнтропическом течении их значения сохраняются на

траекториях уравнений $\dot{X}^{\pm} = u \pm c$ в том же смысле, в каком энтропия сохраняется на «траектории частицы».

Нетрудно видеть, что через значения новых переменных S, R^-, R^+ можно вычислить все остальные величины (u, ρ, e, ...), описывающие течение газа. Теперь уравнения «интегрируются» почти очевидным образом ($S(t, x)$, $R^-(t, x)$, $R^+(t, x)$ постоянны вдоль траекторий систем):

$$\dot{X}^0 = u, \qquad \dot{X}^- = u - c, \qquad \dot{X}^+ = u + c. \tag{13}$$

К сожалению, u и c сами суть функции S, R^-, R^+, поэтому «явного» решения мы здесь не получили. Однако интересен частный случай — газ с показателем адиабаты $\gamma = 3$. В этом случае уравнения интегрируются «до конца» (так как $R^{\pm} = u \pm c$) и семейства траекторий $X^-(t, X_0^-)$, $X^+(t, X_0^+)$ суть просто семейства прямых (вдоль линий этих семейств сохраняются значения их наклонов!).

Случай $\gamma = 3$, как ни странно, реализуется физически. Ему соответствует газ, известный под названием «продукты взрыва». Но нам интереснее другое: этот случай позволяет пояснить механизм образования разрывных решений из гладких начальных данных. Здесь он особенно прозрачен. В самом деле, проинтегрируем уравнения газовой динамики. Имея начальные данные

$$S(0, x) = S_0 = \text{const}, \qquad R^-(0, x) = R_0^-(x), \qquad R^+(0, x) = R_0^+(x),$$

находим траектории X^-, X^+. Дополняя уравнения (13) начальными данными Коши $X^{\pm}(0) = X_0^{\pm}$, получаем

$$X^{\pm}(t, X_0^{\pm}) = X_0^{\pm} + R_0^{\pm}(X_0^{\pm})t.$$

Для того чтобы иметь «явное» решение уравнений газовой динамики, нужно уметь вычислять в каждой данной точке (t, x) значения $R^-(t, x)$, $R^+(t, x)$. Решим (относительно X_0^-, X_0^+) систему нелинейных уравнений:

$$x = X_0^- + R_0^-(X_0^-)t, \qquad x = X_0^+ + R_0^+(X_0^+)t.$$

Тогда

$$R^-(t, x) = R_0^-(X_0^-), \qquad R^+(t, x) = R_0^+(X_0^+)$$

($S(t, x) = S_0$, так как течение изоэнтропическое).

Все было бы хорошо, если бы отображение $(t, x) \rightleftharpoons X_0^-, X_0^+$ было взаимно однозначным, т.е. через каждую точку (t, x) проходило бы только по одной прямой из семейств линий $X_0^{\pm} + R_0^{\pm}(X_0^{\pm})t$. К сожалению, этого нельзя гарантировать никакой гладкостью начальных данных. Если, например, $R_0^+(x') > R_0^+(x'')$ при $x' < x''$, линия

$x' + R_0^+(x')t$ догоняет линию $x'' + R_0^+(x'')t$, и в момент времени $t = (x'' - x')/(R_0^+(x') - R_0^-(x''))$ они пересекутся. В этом случае описанный выше аппарат интегрирования уравнений газовой динамики отказывает. А что можно сказать о решении уравнений газовой динамики, о течении газа, которое эти уравнения описывают? Что случается с течением в этот момент? Ответ прост: в течении образуется разрыв — так называемая ударная волна. Мы еще вернемся к этому в дальнейшем.

Семейства линий $X^0(t, X_0^0)$, $X^-(t, X_0^-)$, $X^+(t, X_0^+)$ играют большую роль в газовой динамике, хотя в общем случае они не определяются явно начальными данными (как в случае продуктов взрыва), а могут быть найдены либо после того, как проинтегрированы уравнения газовой динамики, либо процессом совместного интегрирования уравнений газовой динамики и уравнений этих линий. Они называются *характеристиками*: X^0 — это *энтропийная характеристика*, X^- — *левая звуковая характеристика*, X^+ — *правая звуковая характеристика*. Эти термины связаны с тем, что по этим характеристикам распространяется «звуковой» сигнал: $\pm c$ — скорость звука относительно газа, $u \pm c$ — скорость звука относительно геометрического пространства, в котором газ движется со скоростью u.

Краевые задачи для уравнений газовой динамики. Характеристики позволяют разобраться в правильной постановке краевых условий в конкретных задачах. Для того чтобы решение было полностью определено, уравнения следует дополнить заданием начальных данных и краевых условий. И дополнить так, чтобы не возникло противоречие (т.е. чтобы решение существовало) и чтобы постановка задачи была полной (т.е. решение должно быть единственным).

Несколько слов о физическом смысле характеристик. Пусть имеется некоторое решение $u(t, x)$, $\rho(t, x)$, $e(t, x)$ уравнений (1). Рассмотрим решение, которое при $t = 0$ отличается от невозмущенного малым искажением функций $u(0, x)$, $\rho(0, x)$, $e(0, x)$ на очень малом локальном участке. Тогда и в последующие моменты времени возмущенное течение будет мало отличаться от невозмущенного, но начальное финитное возмущение распадется на три финитных возмущения римановых инвариантов, распространяющиеся со скоростями $u - c$, u, $u + c$ соответственно.

Теперь можно описать постановки тех краевых задач, для которых, как все убеждены (но это не доказано!), справедливы теоремы существования и единственности решений. Мы ограничимся рассмотрением простой области: $0 \leqslant t \leqslant T$, $0 \leqslant x \leqslant L$. При $t = 0$ следует задать начальные данные, т.е. значения всех функций $u(0, x)$, $\rho(0, x)$, $e(0, x)$.

Рассмотрим границу области (для определенности, левую). Из каждой ее точки $(t, 0)$ исходят три характеристики с наклонами $\dot{X}$, равными $u-c$, u, $u+c$ соответственно. Те из них, наклоны которых положительны, назовем входящими в область. На левой границе $(x=0)$ следует задать столько краевых условий (например, независимых соотношений между u, ρ, e), сколько характеристик входит в область. На правой границе $(x=L)$ следует задать столько краевых условий, сколько характеристик входит в область.

Вышеизложенное поясняет рис. 29, на котором схематически показаны характеристики и обозначено число краевых условий в каждой из возможных ситуаций, причем каждая ситуация на левой границе может (в данный момент времени t) сочетаться с любой ситуацией на правой границе. С течением времени ситуации могут меняться. Наклоны характеристик зависят от искомого решения и не всегда могут быть определены заранее (даже качественно: сколько характеристик идут вправо, сколько — влево).

Рис. 29

Таким образом, постановки краевых условий в газовой динамике — дело довольно тонкое. Обоснованием приведенного выше рецепта по постановке краевых условий является анализ уравнений в характеристической форме. Из нее следует, что каждое из трех уравнений является, так сказать, обыкновенным дифференциальным уравнением вдоль соответствующей характеристики, и все дело только в том, что каждое такое уравнение должно быть замкнуто соответствующими данными Коши. Неявно здесь используется принцип причинности: состояние в момент t' определяет состояние при $t > t'$.

Каждая характеристика начинается либо при $t=0$, либо на одной из боковых границ и должна быть «замкнута» соответствующими данными Коши. При $t=0$ из каждой точки в область входят три характеристики, заданы три величины, все три характеристики имеют свои данные Коши. Если характеристика «рождается» на боковой границе, то приведенный выше рецепт также приводит к замкнутой и не переопределенной системе уравнений. Поясним это несколько иначе: если в данную точку грачицы, например в точку $(t^*, 0)$, приходит $k < 3$ характеристик из области, то $3-k$ характеристик выходит из этой точки внутрь области.

Интегрирование (в направлении роста $t \leqslant t^*$) по каждой из приходящих характеристик (оно ведется изнутри области) определяет в точке $(t^*, 0)$ соответствующее число k соотношений между описыва-

ющими состояние газа значениями u, ρ, e. Если в этой точке будет задано больше чем $3 - k$ условий, возникнет (в общем случае) противоречие и такого решения не существует. Если будет задано меньшее число краевых условий, можно добавить еще одно произвольное и нарушится единственность поставленной задачи.

Это рассуждение выглядит «почти доказательством», но не следует упускать из вида, что сами характеристики — это объект, однозначно определенный лишь на решении уравнений газовой динамики. И только для гиперболических линейных систем, когда характеристики определяются коэффициентами уравнений и не зависят от решений, приведенные выше соображения можно оформить в виде точных теорем. Тем не менее в нелинейной газовой динамике эти соображения используются с успехом. Более того, используется и более тонкий факт: выясняется запрет на некоторые формы краевых условий.

Грубо говоря, в качестве задаваемого краевого условия (соотношения между значениями u, ρ, e) нельзя использовать то соотношение, которое «приносится» по приходящей из области (снизу) характеристике. Например, если в точку $(t^*, 0)$ приходит левая звуковая характеристика (а правая и энтропийная входят в этой точке в область), то в качестве одного из двух требуемых в этом случае условий нельзя задавать значение риманова инварианта R^-. Его значение определяется однозначно состоянием при $t < 0$, возникает противоречие, и решения уже не существует. Вообще, приносимые на границу по приходящим характеристикам соотношения, дополненные заданным краевым условием, должны составлять систему уравнений (относительно u, ρ, e), допускающую однозначную разрешимость.

Уравнения газовой динамики в форме Лагранжа. Выше были описаны уравнения газовой динамики в так называемой форме Эйлера. Она характеризуется тем, что в качестве независимых переменных выбираются время t и декартова координата x, связанная с геометрическим пространством. Очень удобна во многих задачах другая система независимых переменных — так называемая лагранжева система, в которой одной из независимых переменных остается время t, вторая же (назовем ее ξ) определяется так, что она остается постоянной вдоль траектории частиц. Траектория — это кривая в пространстве (t, x), описываемая выделенной частицей газа. Каждой частице соответствует своя траектория — решение уравнения $\dot{X} = u(t, X(t))$.

Отметим все частицы газа параметром ξ. Это и будет «лагранжева» координата частицы. Теперь все траектории будут описываться функцией $X(t, \xi)$, удовлетворяющей уравнению

$$X_t(t, \xi) = u(t, X(t, \xi)).$$

Здесь $u(t, x)$ берется из решения уравнений газовой динамики. В качестве параметра ξ можно взять, например, координату x частицы в начальный момент времени $t = 0$. Тогда уравнение дополняется данными Коши $X(0, \xi) = \xi$.

Для того чтобы для данной точки (t^*, x) узнать ее координаты (t^*, ξ), нужно проинтегрировать «назад» (от t^* к нулю) уравнение $\dot{Y} = u(t, Y)$, $Y(t^*) = x$. Тогда $\xi = Y(0)$. Взаимная однозначность отображения (t, x) в (t, ξ) следует из того, что траектории не пересекаются.

Следующей нашей задачей будет вывод уравнений газовой динамики в лагранжевых координатах. Пусть имеется некоторая функция эйлеровых переменных $f(t, x)$. Превратим ее в функцию лагранжевых переменных $\tilde{f}(t, \xi) \equiv f(t, X(t, \xi))$. Вычислим производные $\tilde{f}$:

$$\tilde{f}_\xi = f_x X_\xi, \qquad \tilde{f}_t = f_t + f_x X_t = f_t + u f_x.$$

Пусть $u(t, x)$, $\rho(t, x)$, $e(t, x)$ — решение уравнений газовой динамики (1), каждое из которых содержит так называемый оператор субстанциальной производной — производной по t вдоль траектории частицы $(\partial/\partial t + u\partial/\partial x)$. Определим функции $\tilde{u}(\xi, t)$, $\tilde{\rho}(\xi, t)$, $\tilde{e}(\xi, t)$ заменой переменных. Они, очевидно, и будут решением уравнений газовой динамики в лагранжевых координатах. Перепишем уравнения:

эйлерова форма	*лагранжева форма*
$u_t + uu_x + \frac{1}{\rho} p_x = 0,$	$\tilde{u}_t + \frac{1}{\tilde{\rho}} X_\xi^{-1} \tilde{p}_\xi = 0,$
$\rho_t + u\rho_x + \rho u_x = 0,$	$\tilde{\rho}_t + \rho X_\xi \tilde{u}_\xi = 0,$
$\left(e + \frac{u^2}{2}\right)_t + u\left(e + \frac{u^2}{2}\right)_x + \frac{1}{\rho}(pu)_x = 0,$	$\left(\tilde{e} + \frac{\tilde{u}^2}{2}\right)_t + \frac{1}{\tilde{\rho}} X_\xi^{-1} (\tilde{p}\tilde{u})_\xi = 0,$
	$X_t(\xi, t) = \tilde{u}(\xi, t).$

Массовые лагранжевы координаты. Особенно простую и удобную для аналитических исследований и организации расчетов форму имеют уравнения газовой динамики при специальном выборе лагранжевой координаты. Чтобы пояснить его смысл, рассмотрим выражение $\tilde{\rho} X_\xi$ (пока мы имеем дело с определенной выше лагранжевой координатой $X(\xi, 0) = \xi$). Величина $\tilde{\rho} X_\xi \, d\xi = \rho \, dx$ равна массе вещества, заключенной между траекториями, соответствующими частицам ξ и $\xi + d\xi$. Она, естественно, остается постоянной; следовательно,

$$[\tilde{\rho}(t, \xi) \, X_\xi(t, \xi)]_t = 0,$$

$$\tilde{\rho}(t, \xi) \, X_\xi(t, \xi) = \tilde{\rho}(0, \xi) \, X_\xi(0, \xi) = \rho(0, x).$$

Теперь уравнения в форме Лагранжа можно записать так:

$$\frac{\partial \tilde{u}}{\partial t} + \frac{1}{\tilde{\rho}(\xi, 0)} \frac{\partial \tilde{p}}{\partial \xi} = 0, \quad \text{и т.д.}$$

Введем *массовую лагранжеву* координату $m(\xi)$, связанную с ξ дифференциальным уравнением $m_\xi = \tilde{\rho}(\xi, 0)$. В массовых лагранжевых координатах уравнения газовой динамики принимают совсем простую форму (вместо плотности ρ используем удельный объем $v = \rho^{-1}$). Опуская тильду в обозначениях, имеем

$$\frac{du}{dt} + \frac{\partial p}{\partial m} = 0, \quad \frac{dv}{dt} - \frac{\partial u}{\partial m} = 0, \quad \frac{d}{dt}\left(e + \frac{u^2}{2}\right) + \frac{\partial (pu)}{\partial m} = 0. \quad (14)$$

Обозначение d/dt вместо $\partial/\partial t$ принято в лагранжевой системе координат. Система лагранжевых уравнений обязательно дополняется уравнением связи лагранжевой (m) и эйлеровой (x) координат:

$$x_t(t, m) = u(t, m), \quad \partial x(0, m)/\partial m = v(0, m).$$

Разрывные решения уравнений газовой динамики. Рассмотрим возможные, в принципе, разрывы в решениях уравнений газовой динамики. Предположим, что функции $u(t, x)$, $\rho(t, x)$, $e(t, x)$ рвутся на некоторой линии в пространстве (t, x). Пусть эта линия гладкая и по обе стороны от линии разрыва функции u, ρ, e — классическое решение уравнений газовой динамики. Хотя в этой ситуации уравнения выполнены «почти всюду» (всюду, за исключением линии разрыва, являющейся множеством меры нуль), такая произвольная «склейка» двух решений не может считаться решением. Оказывается, между величинами u, ρ, e в точках на левом и правом краях разрыва должны выполняться некоторые соотношения. Выведем их, используя определение обобщенного решения (уравнения в интегральной форме).

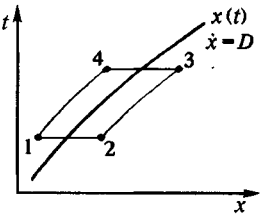


Рис. 30

Возьмем на линии разрыва некоторую точку и построим около нее малую область Ω — параллелограмм со сторонами, параллельными линии разрыва (рис. 30). Длины горизонтальных сторон считаем существенно меньшими длин боковых сторон. Чтобы разрывные функции $u(t, x)$, $\rho(t, x)$, $e(t, x)$ могли считаться обобщенными решениями, нужно, чтобы для каждой области типа Ω выполнялось соотношение (мы исходим из эйлеровой дивергентной формы)

$$\oint_{\partial\Omega} [\rho u\, dx - (\rho u^2 + p)\, dt] = 0.$$

(Аналогично для двух остальных уравнений.)

Введем обозначения: D — скорость распространения разрыва, т.е. $X_t = D$ ($X(t)$ есть уравнение линии разрыва); u_1, ρ_1, e_1 — значения функций справа от разрыва (область Ω столь мала, что изменениями функций вдоль контура справа от разрыва можно пренебречь). В дальнейшем мы перейдем к пределу при стягивании Ω в точку. Переменность функций можно было бы без труда учесть, но она дала бы вклад в малые более высокого порядка по сравнению с основными членами. То же самое предполагается и слева от разрыва. Соответствующие предельные значения обозначим u_2, ρ_2, e_2. Интегралами по отрезкам *12* и *34*, как уже указывалось, можно пренебречь. Итак, в главных членах соотношение дает

$$\oint_{\partial\Omega} [\rho u\, dx - (\rho u^2 + p)\, dt] = \int_2^3 [\rho u\, dx - (\rho u^2 + p)\, dt] + \int_4^1 \ldots$$

Учтем, что отрезок *23* есть вектор $(dx, dt) = (D, 1)\, dt$, а отрезок *41* есть вектор $-(D, 1)\, dt$. Сокращая на dt, имеем

$$(D\rho_1 u_1 - \rho_1 u_1^2 - p_1) - (D\rho_2 u_2 - \rho_2 u_2^2 - p_2) = 0.$$

Такие соотношения на разрыве принято обозначать в виде

$$[D\rho u - \rho u^2 - p]_1^2 = 0.$$

Точно таким же образом получаем

$$[(D-u)\rho]_1^2 = 0, \qquad \left[(D-u)\rho\left(e + \frac{u^2}{2}\right) - up\right]_1^2 = 0.$$

Удобно привести эти соотношения к другой форме (сохранение массы, импульса и энергии соответственно):

$$\begin{aligned} &\text{а)}\ [\rho(u-D)]_1^2 = 0, \\ &\text{б)}\ [\rho(u-D)^2 + p]_1^2 = 0, \\ &\text{в)}\ \left[\rho(u-D)\left(e + \frac{p}{\rho} + \frac{(u-D)^2}{2}\right)\right]_1^2 = 0. \end{aligned} \tag{15}$$

Они носят название соотношений Гюгонио.

Два типа разрывов в газовой динамике. Рассмотрим одно из простейших решений соотношений Гюгонио. Пусть $u_1 = D$. Тогда из (15а) имеем $\rho_2(u_2 - D) = 0$. Так как $\rho_2 \neq 0$, то $u_2 = D$. Из (15б) следует $p_2 = p_1$. Соотношение (15в) автоматически выполняется. На этом разрыве, называемом *контактным*, имеют место следующие факты:

контактный разрыв совпадает с какой-то траекторией, так как его скорость D совпадает со скоростью газа u;

скорость u и давление p на контактном разрыве не рвутся;

плотность на контактном разрыве рвется произвольным образом.

Если два сорта газа граничат друг с другом, находятся при одном и том же давлении и движутся с одной и той же скоростью u, никаких событий в среде не происходит. Просто граница раздела движутся с той же скоростью, что и весь газ.

Пусть $u_1 \neq D$. В этом случае рвутся все функции: $u_2 \neq u_1$, $\rho_2 \neq \rho_1$, $p_2 \neq p_1$. Такой разрыв называют *ударной волной.*

Итак, мы имеем три примера точных решений уравнений газовой динамики.

Константное решение:

$$u(t, x),\ \rho(t, x),\ e(t, x) = \text{const}.$$

Чистый контактный разрыв:

$$u(t, x),\ \rho(t, x),\ e(t, x) = \begin{cases} u_1, \rho_1, e_1, & x > u_1 t, \\ u_2, \rho_2, e_2, & x < u_1 t. \end{cases}$$

При этом $P_1(e_1, \rho_1) = P_2(e_2, \rho_2)$, так как контактный разрыв часто разделяет вещества с разными уравнениями состояния.

Чистая ударная волна:

$$u(t, x),\ \rho(t, x),\ e(t, x) = \begin{cases} u_1, \rho_1, e_1, & x > Dt, \\ u_2, \rho_2, e_2, & x < Dt. \end{cases}$$

Значения u_1, ρ_1, e_1 произвольны (разумеется, $\rho_1 > 0$, $e_1 > 0$). Произвольно и значение D. Значения u_2, ρ_2, e_2 находятся из трех соотношений Гюгонио.

Очевидно, в ударной волне можно поменять местами значения с индексами 1 и 2, условия Гюгонио будут выполнены. Если $\rho_1 < \rho_2$, ударную волну называют волной сжатия, в противоположном случае — волной разрежения. Это относится к волне, идущей вправо ($D > 0$), когда ρ_1 есть плотность до прохождения волны. Волна разрежения в природе не реализуется. Ее существование отрицается на основании как физических, так и чисто математических соображений.

Физическая аргументация состоит в том, что, как показывает анализ, при прохождении газа через ударную волну сжатия энтропия скачком растет, а при прохождении через ударную волну разрежения — падает. Поэтому физика признает лишь ударные волны сжатия. С математической точки зрения различие в этих формальных решениях уравнений газовой динамики вносится анализом устойчивости. Волна сжатия устойчива относительно малых возмуще-

ний. Волна разрежения неустойчива, она не может долго существовать и быстро «разваливается».

Что же произойдет, если мы зададим в начальных данных кусочно-постоянные значения, соответствующие ударной волне разрежения? Оказывается, существует еще одно обобщенное (и уже устойчивое) решение уравнений. Чтобы дать о нем представление, рассмотрим еще более общую ситуацию. Пусть в начальных данных заданы кусочно-постоянные значения u_1, ρ_1, e_1 при $x > 0$ и u_2, ρ_2, e_2 при $x < 0$. Можно ли найти точное решение уравнений газовой динамик в этом простом случае? Оказывается, да. Решение этой задачи (так называемой задачи о распаде произвольного разрыва в начальных данных) было найдено в сороковых годах Н. Е. Кочиным. Чтобы качественно описать его, нам понадобится еще одно «чистое» решение уравнений газовой динамики.

Центрированная волна разрежения. В уравнения газовой динамики входят только производные по t и x первого порядка. Поэтому они инвариантны относительно преобразования подобия независимых переменных $t = \alpha t'$, $x = \alpha x'$. Точнее, если u, ρ, $e(t, x)$ — решения уравнений, то и функции u', ρ', $e'(t', x') \equiv u$, ρ, $e(\alpha t', \alpha x')$ удовлетворяют уравнениям. В самом деле,

$$\partial u'/\partial x' = \alpha\, \partial u/\partial x, \quad \partial u'/\partial t' = \alpha\, \partial u/\partial t', \quad \ldots$$

Видно, что функции u', ρ', e' уравнениям удовлетворяют. Мы рассматриваем безграничную задачу Коши с данными

$$u, \rho, e(0, x) = \begin{cases} u_1, \rho_1, e_1, & x > 0, \\ u_2, \rho_2, e_2, & x < 0. \end{cases}$$

Поэтому функции u', ρ', $e'(0, x')$ имеют точно такие же значения.

Таким образом мы имеем бесконечное множество (при любом $\alpha > 0$) решений одной и той же задачи Коши. Неявно опираясь на единственность ее решения, мы получаем тождество

$$u, \rho, e(\alpha t, \alpha x) = u, \rho, e(t, x).$$

Это возможно лишь в случае, когда решение зависит не от двух переменных t, x, а лишь от одной *автомодельной переменной* $\xi = x/t$. Итак,

$$u, \rho, e(t, x) = U, R, E(x/t).$$

Уравнение газовой динамики становятся обыкновенными дифференциальными уравнениями, которые допускают достаточно обозримый анализ. Эти уравнения выводятся после замены операторов

$$\frac{\partial}{\partial t} = \frac{d}{d\xi}\frac{\partial \xi}{\partial t} = -\frac{\xi}{t}\frac{d}{d\xi}, \qquad \frac{\partial}{\partial x} = \frac{d}{d\xi}\frac{\partial \xi}{\partial x} = \frac{1}{t}\frac{d}{d\xi}.$$

Используя уравнения в форме (14), после простых преобразований получаем систему уравнений для автомодельного решения:

$$P' = \xi U', \quad U' = -\xi V', \quad -\xi E' + PU' = 0$$

(штрих — символ производной по ξ). Для уравнения состояния $E = PV/(\gamma - 1)$ они легко интегрируются:

$$V(\xi) = V_0\, \xi^{-2/(\gamma+1)}, \quad P(\xi) = \frac{V_0}{\gamma}\, \xi^{2\gamma/(\gamma+1)},$$
$$u(\xi) = u_0 - \frac{2V_0}{\gamma - 1}\, \xi^{(\gamma-1)/(\gamma+1)}. \tag{16}$$

В результате мы имеем общее решение с двумя произвольными постоянными.

Отметим важное соотношение $C^2(\xi) = \gamma P(\xi)/V(\xi) = \xi^2$. Здесь C есть скорость звука (отличие от формулы $c^2 = \gamma p/\rho$ связано с тем, что C — это «массовая» скорость звука, так как мы используем уравнение с массовой лагранжевой переменной). Таким образом, линия $\xi = C$, т.е. $x = Ct$, является характеристикой. Заметим, что при $\xi = 0$ решение имеет особенность, которой можно избежать, используя это решение только в интервале $[\xi', \xi'']$ при $\xi' < \xi'' < 0$ или $0 < \xi' < \xi''$.

Построим еще одно точное решение типа «центрированной волны разрежения» (для определенности, идущей вправо). Пусть u_1, ρ_1, e_1 произвольны. Вычислим $\xi'' = \sqrt{\gamma p_1 \rho_1}$ (это будет правая граница волны). Речь идет о непрерывном решении, поэтому нам известны значения $V(\xi'') = v_1 = 1/\rho_1$ и $U(\xi'') = u_1$, что позволяет без труда вычислить постоянные V_0, u_0 в (16). При $\xi < \xi''$ решение описывается формулами (16). Левую границу волны $0 < \xi' < \xi''$ можно назначить произвольно. Вычислим $u_2 = U(\xi')$, $v_2 = V(\xi')$, $p_2 = P(\xi')$. Эти значения определяют константное решение при $\xi < \xi'$. Заметим, что и ударная волна, и контактный разрыв входят в семейство автомодельных обобщенных решений — это константные решения, рвущиеся при некотором значении ξ.

Теперь можно вернуться к вопросу о распаде произвольного разрыва в начальных данных. Решение является автомодельным и состоит (в общем случае) из контактного разрыва, справа и слева от которого расположена ударная волна или центрированная волна разрежения, причем возможны четыре сочетания: все определяется расположением точек u_1, ρ_1, e_1 и u_2, ρ_2, e_2.

Основные трудности, возникающие при численном решении уравнений газовой динамики, связаны с наличием разрывов в искомых решениях. При конструировании численных методов обычно выделяют характерные особенности решений, т.е. строят задачи, в

которых наиболее трудные особенности существуют в чистом виде, без взаимодействия с не очень трудными для расчетов гладкими течениями. В таких задачах известно точное решение и качество расчетной схемы оценивается по тому, как оно справляется с решением «модельной задачи».

Метод Годунова. Для расчета разрывных решений широко используется метод, основанный на решении задачи о распаде разрыва. Пусть начальные данные являются кусочно-постоянными на некоторой сетке $\{x_m\}_{m=0}^{M}$, т.е. $u, \rho, e(0, x) = \{u^0_{m+1/2}, \rho^0_{m+1/2}, e^0_{m+1/2}\}$ при $x \in (x_m, x_{m+1})$. Оказывается, эта задача имеет точное явное решение. Оно строится так: в каждой точке x_m нужно решить задачу о распаде разрыва независимо от всех остальных разрывов. Такое решение можно использовать до того момента времени, когда при каком-то значении m правая волна, образовавшаяся от разрыва в точке x_m, встретится с левой волной, идущей от распада в точке x_{m+1}.

Перейдем к описанию схемы. Основными счетными величинами являются $u^n_{m+1/2}$, $\rho^n_{m+1/2}$, $e^n_{m+1/2}$, где n — номер временного слоя. Величина $u^n_{m+1/2}$, например, представляет собой приближенное значение функции $u(t_n, (x_m + x_{m+1})/2)$. Пусть начальные данные $u^0_{m+1/2}$, $\rho^0_{m+1/2}$, $e^0_{m+1/2}$ превращены в кусочно-постоянные на сетке $\{x_m\}$ функции. Построим точное решение уравнений газовой динамики и определим малое (порядка $\min(x_{m+1} - x_m)$) время τ, в течение которого это явное решение существует.

Один шаг численного интегрирования соответствует продвижению по t на τ. Для того чтобы сделать второй шаг, нужно и при $t = \tau$ иметь кусочно-постоянное решение. Точное решение, конечно, таковым не является. Поэтому оно аппроксимируется кусочно-постоянным с сохранением в пределах каждого интервала основных физических величин: массы, импульса и энергии. Например,

$$(x_{m+1} - x_m)\rho^1_{m+1/2} = \int_{x_m}^{x_{m+1}} \rho(\tau, x)\, dx.$$

Аналогично вычисляются

$$\rho^1_{m+1/2}\, u^1_{m+1/2}, \qquad \rho^1_{m+1/2}[e^1_{m+1/2} + (u^1_{m+1/2})^2/2].$$

Однако u, ρ, e при $t = \tau$ являются сложными функциями x, вычислять интегралы от них трудно. Это препятствие обходится следующим образом.

Рассмотрим ячейку $(x_m, x_{m+1}) \times (0, \tau)$ и запишем уравнения в интегральной форме (3), взяв ячейку в качестве Ω. В результате

получим соотношение

$$\int_{x_m}^{x_{m+1}} R[0, x]\,dx - \int_0^{\tau} Q[t, x_{m+1}]\,dt - \int_{x_m}^{x_{m+1}} R[\tau, x]\,dx + \int_0^{\tau} Q[t, x_m]\,dt = 0, \tag{17}$$

где $R[t, x] \equiv R(u(t, x), \rho(t, x), e(t, x))$. В (17) третий интеграл есть то, что нам требуется. Первый интеграл вычисляется элементарно, так как $R[0, x] = R(u^0_{m+1/2}, \rho^0_{m+1/2}, e^0_{m+1/2})$ в силу постоянства начальных данных на (x_m, x_{m+1}).

Так же легко вычисляются второй и четвертый интегралы. В силу автомодельности решения задачи о распаде произвольного разрыва на линии $x = x_m$ (т.е. $\xi = \text{const}$) все функции постоянны. Обозначим их u_m, ρ_m, e_m. Тогда второй интеграл есть $Q(u_m, \rho_m, e_m)\tau$.

Формально схему Годунова можно записать в виде

$$\frac{R^{n+1}_{m+1/2} - R^n_{m+1/2}}{\tau} + \frac{Q_{m+1} - Q_m}{x_{m+1} - x_m} = 0.$$

Следует иметь в виду, что «поток» Q_m вычисляется решением задачи о распаде разрыва. Она сводится к решению системы нелинейных уравнений. Это относительно «дорогая» операция (ведь она проводится при всех m, n). Значительные усилия прилагаются к тому, чтобы снизить ее трудоемкость. В частности, используется то, что в большинстве узлов (n, m) величины слева и справа от x_m мало отличаются друг от друга. Разработанная в середине пятидесятых годов схема до сих пор применяется в расчетах; при этом она, разумеется, обобщается и совершенствуется.

Расчет контактного разрыва. Проблемы расчета течения, содержащего контактный разрыв, рассмотрим, используя известное нам точное решение уравнений газовой динамики типа «чистый контактный разрыв». В этом случае из всех уравнений газовой динамики нетривиально только одно: $\rho_t + u\rho_x = 0$ (здесь $u = \text{const}$). Будем решать его методом сеток.

Построим равномерную сетку с шагом h по x и τ по t. Узлы сетки $x_m = mh$, $t_n = n\tau$. Приближенное решение ищем в виде сеточной функции ρ^n_m. Используем простейшую явную схему (предполагая $u > 0$):

$$(\rho^{n+1}_m - \rho^n_m)/\tau + u(\rho^n_m - \rho^n_{m-1})/h = 0. \tag{18}$$

Как известно, эта схема устойчива при условии Куранта $u\tau/h \leqslant 1$ (см. § 12). (Заметим, что такая схема весьма популярна при расчете задач в эйлеровых координатах: во все уравнения в этом случае

входит характерный оператор «субстанциальной производной» $\partial/\partial t + u\,\partial/\partial x$.) При $u < 0$ используется аппроксимация с шаблоном, ориентированным в противоположную сторону (против потока):

$$(\rho_m^{n+1} - \rho_m^n)/\tau + u(\rho_{m+1}^n - \rho_m^n)/h = 0.$$

Когда решается полная система уравнений газовой динамики, функция $u(t, x)$ может менять знак. Соответственно, и разностные формулы строятся в точках (n, m) в зависимости от знака u_m^n. Что же получается при расчете контактного разрыва по такой схеме? Происходит неприятное явление. В расчете контактный разрыв размывается, его «ширина» растет с ростом времени. На рис. 31 показана характерная картина расчета: точное решение («ступенька») и приближенное.

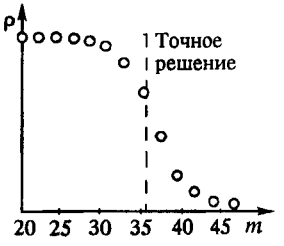


Рис. 31

Это явление особенно неприятно в тех задачах, в которых контактный разрыв разделяет среды с разными уравнениями состояния. Нужно знать достаточно точно границу между разными газами, чтобы пользоваться в данной точке нужным уравнением состояния. На рис. 31 представлены результаты, полученные в вычислительном эксперименте. Но для того чтобы бороться с «размазыванием контактного разрыва», нужно иметь хоть какую-то теорию. Этим мы сейчас и займемся. Заметим, что излагаемый ниже аппарат имеет достаточно общее значение, не ограничивающееся только задачей о контактном разрыве. Он может применяться при построении разностных схем, лучших в том или ином отношении (смотря по тому, что нам нужно в задаче).

Исследование дисперсионного соотношения для разностной схемы. Подчеркнем, что этот аппарат, строго говоря, работает только в случае линейных уравнений с постоянными коэффициентами. Он применяется для линеаризованных моделей реальных уравнений, однако полученные в линейной модели рекомендации затем используются и в реальных задачах. Хорошо известно, что́ такое дисперсионное соотношение для линейного дифференциального уравнения с постоянными коэффициентами. В нашем случае для уравнения $\rho_t + u\rho_x = 0$ дисперсионное соотношение появляется, когда мы ищем решение вида $e^{\lambda(k)t + ikx}$, где $\lambda(k) = -uki$.

Оказывается, разностные уравнения (линейные, однородные, с постоянными коэффициентами) имеют решения того же вида, но, конечно, с другой функцией $\lambda(k)$, зависящей от шагов τ, h и вида разностной схемы. Найдем дисперсионное соотношение для напи-

санной выше схемы, полагая $\rho_m^n = e^{\lambda(k)n\tau + ikmh}$. Подставляя эту функцию в разностное уравнение, после очевидных преобразований получаем соотношение

$$(e^{\lambda\tau} - 1)/\tau + u(1 - e^{-ikh})/h = 0.$$

Из него легко вычислить дисперсионное соотношение для схемы «против потока»:

$$\lambda(k, h, \tau) = \frac{1}{\tau} \ln \left[\left(1 - u\frac{\tau}{h}\right) + u\frac{\tau}{h} e^{-ikh} \right].$$

Естественно оценивать качество разностной схемы по степени совпадения дисперсионных функций для дифференциального и разностного уравнений. Идеальным было бы их совпадение. Оно обеспечивается условием $u\tau/h = 1$. Этот идеальный случай, к сожалению, практически не интересен. Соотношение $u\tau/h = 1$ в реальной задаче, когда значение u не является постоянным, а меняется во времени и в пространстве, во всех точках сетки выдержать нельзя. Поэтому связанными с ним преимуществами воспользоваться в практической работе не удается. Разумеется, функция $\lambda(k, h, \tau)$ должна аппроксимировать $\lambda(k)$. Параметр h есть малый параметр, и аппроксимация, естественно, тем лучше, чем меньше волновое число k (чем глаже по x рассматриваемое частное решение); на сетке с шагом h волновые числа $k > 2\pi/h$ уже не реализуются.

Суждения о качестве разностной схемы можно делать, сравнивая графики $\lambda(k)$ и $\lambda(k, h, \tau)$. Некоторые выводы можно получить, считая $kh \ll 1$ (в смысле $kh \leqslant 0.5$, например) и разлагая в ряд Тейлора хотя бы с точностью до второго члена:

$$\lambda(k, h, \tau) \approx -iuk - \frac{uhk^2}{2}\left(1 - u\frac{\tau}{h}\right).$$

Сравним частные решения дифференциального и разностного уравнений:

$$e^{ik(mh - un\tau)}, \qquad e^{ik(mh-un\tau)}\, e^{-uk^2(h-u\tau)n\tau/2}.$$

Сделаем некоторые качественные выводы из полученной формулы.

1. Как отмечалось, малоинтересен специальный случай $u\tau = h$.

2. При $u < 0$ схема непригодна: решения разностного уравнения отличаются от решений дифференциального множителем порядка $e^{|u|k^2hn\tau}$, который при $k \approx 1/h$ катастрофически (при $h, \tau \to 0$) растет, как $e^{|u|t/h}$.

3. При $h - u\tau < 0$, т.е. $u\tau/h > 1$, схема непригодна по тем же причинам.

4. При $u > 0$ и $u\tau \leqslant h$ решения отличаются множителем, затухающим при росте t, причем темп затухания тем выше, чем больше волновое число k (т.е. чем меньше длина волны частного решения

типа e^{ikx}). Таким образом в отличие от решения дифференциального уравнения, в котором все гармоники с течением времени сохраняют свою амплитуду, в решении разностного уравнения происходит затухание коротковолновых гармоник. Это приводит к тому, что разрыв в начальных данных с течением времени сглаживается.

Приведем еще две популярные схемы аппроксимации «уравнения переноса» $\rho_t + u\rho_x = 0$, имеющие второй порядок аппроксимации по τ и h, и их дисперсионные соотношения. Схема «квадрат» имеет вид

$$\frac{1}{\tau}\left(\frac{\rho_m^{n+1}+\rho_{m+1}^{n+1}}{2} - \frac{\rho_m^n+\rho_{m+1}^n}{2}\right) + \frac{u}{h}\left(\frac{\rho_{m+1}^{n+1}+\rho_{m+1}^n}{2} - \frac{\rho_m^{n+1}+\rho_m^n}{2}\right) = 0. \quad (19)$$

Ее дисперсионное соотношение таково:

$$\lambda(k, h, \tau) = \frac{1}{\tau}\ln\left[\left(1 - iu\frac{\tau}{h}\operatorname{tg}\frac{kh}{2}\right)\Big/\left(1 + iu\frac{\tau}{h}\operatorname{tg}\frac{kh}{2}\right)\right].$$

Асимптотика при $kh \ll 1$:

$$\lambda(k, h, \tau) \approx -iku + \frac{1}{12}iuk^3(u^2\tau^2 - h^2).$$

Другая схема (называемая *характеристической схемой второго порядка*):

$$\frac{\rho_m^{n+1}-\rho_m^n}{\tau} + u\frac{\rho_m^n-\rho_{m-1}^n}{h} + u\frac{h}{2}\left(1 - u\frac{\tau}{h}\right)\frac{\rho_{m+1}^n - 2\rho_m^n + \rho_{m-1}^n}{h^2} = 0, \quad (20)$$

имеет дисперсионную функцию

$$\lambda(k, h, \tau) = \frac{1}{\tau}\ln\left[1 + u\frac{\tau}{h}(e^{-ikh} - 1) + 4u\frac{h}{2}\left(1 - u\frac{\tau}{h}\right)\sin^2\frac{kh}{2}\right].$$

Ее асимптотика при $|kh| \ll 1$ такова:

$$\lambda(k, h, \tau) \approx -iku + \frac{1}{6}ik^3u(h^2 - u^2\tau^2).$$

В обоих случаях $\lambda(k, h, \tau)$ совпадает с $\lambda(k)$ с точностью до $O(k^3)$, а не $O(k^2)$, как в первом случае (эти схемы имеют второй порядок аппроксимации, а аппроксимация «против потока» — только первый).

Какие же выводы можно сделать из полученных формул? Решения дифференциального уравнения имеют вид волн с «частотой» k, движущихся равномерно вправо со скоростью $u > 0$. Волны (для всех k) движутся с одной и той же скоростью. Поэтому график $\rho(t, x)$, заданный в начальный момент времени, просто движется со скоростью u, не меняя своей формы. Решения разностного уравнения имеют вид

$$\exp\left[ik\left(x - \frac{\lambda(k, h, \tau)}{k}t\right)\right], \quad x = x_m, \quad t = t_n,$$

причем

$$-\frac{1}{k}\lambda(k, h, \tau) = u\left[1 + \frac{k^2}{6}(h^2 - u^2\tau^2)\right]$$

(для характеристической схемы).

Таким образом, каждая элементарная волна движется со своей собственной скоростью $u_k = -\lambda/k$, которая мало отличается от u при малых частотах k. Высокочастотные же волны движутся с существенно отличной от u скоростью. Заметим, что в схемах второго порядка точности гармоники не затухают с течением времени. Различие в скоростях u_k приводит к тому, что первоначальный «волновой пакет», определяющий форму начального профиля ρ_m^0, деформируется за счет «рассогласования фаз». В расчетах это сказывается в том, что график ρ_m^n теряет монотонность.

Наличие таких немонотонностей очень не нравится вычислителям, так же как и сильное размазывание контактной границы. Существует специальный термин *монотонная схема.* Если в начальных данных задана произвольная монотонная сеточная функция ρ_m^0 и разностное решение ρ_m^n, полученное по какой-то схеме, остается монотонным, то схему называют монотонной. Схема «против потока» монотонна, но сильно «мажет» контактную границу. Схемы второго порядка размазывают границу существенно меньше, но они не монотонны. С. К. Годунов доказал, что среди явных схем второго (и выше) порядка аппроксимации не существует монотонных.

Разработчики разностных схем прикладывают определенные усилия для создания схем, в которых оба дефекта — размазывание и немонотонность — были бы возможно меньшими. В частности, автором в 1962 г. была предложена схема, в которой использовалась схема «против потока» (18) (первого порядка) или схема (19) — в зависимости от «локальных дифференциальных свойств решения», т.е. в зависимости от величины

$$\eta_m^n = \left|\frac{\rho_{m+1}^n - 2\rho_m^n + \rho_{m-1}^n}{\rho_m^n - \rho_{m-1}^n}\right|.$$

Если эта величина не очень велика ($\eta < 3$), в данном узле (m, n) используется схема второго порядка, в противном случае — первого. Этот прием позволил устранить осцилляции в профиле ρ_m^n и сохранить размазывание разрыва, характерное для схемы второго порядка. Такие схемы теперь называют «гибридными».

На рис. 31, 32 показаны результаты расчета задачи о движении контактного разрыва. Представленные на момент времени $t = 35$ (т.е. разрыв прошел 35 счетных точек) результаты получены по следующим схемам:

а) по схеме первого порядка (разрыв сглаживается, с ростом времени ширина размазывания растет, как $\sqrt{t}$; см. рис. 31);

б) по схеме второго порядка (появляются паразитические осцилляции, но ширина зоны размазывания разрыва уменьшается);

в) по «гибридной» схеме первого и второго порядков (см. рис. 32*а*);

г) по схеме третьего порядка (разрыв выражен резче, но видны, хотя и не очень значительные, осцилляции);

д) по «гибридной» схеме первого, второго и третьего порядков (см. рис 32*б*; кружки — «гибридная» схема третьего порядка).

Видно, что «гибридность» позволяет устранить осцилляции, сохраняя ширину размазывания, характерную для схемы наибольшего используемого порядка.

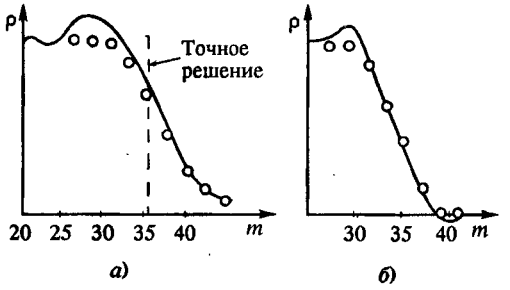


Рис. 32

Однако в наиболее сложных ситуациях при расчетах двумерных течений (в t, x, y) на сетках с относительно умеренным числом точек и при сложной деформации первоначальной формы контактных границ такие методы проблемы не решают. В этих случаях используются так называемые методы типа PIC (см. § 23). Отметим только, что создание и развитие таких методов существенно связано именно с проблемой расчета контактных разрывов.

Характеристические схемы. Схема «против потока» и ее уточнение (20) являются примерами так называемых характеристических схем. Поясним простой принцип их конструирования, широко применяющийся и в более сложных задачах. Оператор $\partial/\partial t + u\,\partial/\partial x$ есть производная по t вдоль направления $dt : dx = 1 : u$, а уравнение $\rho_t + u\rho_x = 0$ означает, что значение ρ переносится без изменений вдоль этого направления. Для того чтобы, зная величины ρ_m^n ($m = 0, 1, 2, \ldots$), вычислить значение ρ_m^{n+1}, нужно найти значение ρ в момент t_n в точке $x_m - u\tau$. Так как эта точка не совпадает с узлом сетки, следует проинтерполировать в эту точку значения u из ближайших узлов n-го слоя.

Используя линейную интерполяцию значений ρ_{m-1}^n и ρ_m^n, получаем схему (14):

$$\rho_m^{n+1} = \alpha\rho_m^n + (1 - \alpha)\rho_{m-1}^n, \quad \alpha = u\tau/h.$$

Только при условии устойчивости (когда $u > 0$, а $u\tau < h$) точка $x_m - u\tau \in (x_{m-1}, x_m)$ и ρ_m^{n+1} вычисляется интерполяцией. Если эта формула реализует экстраполяцию, она становится неустойчивой. Схема второго порядка (20) получается точно так же, но используется квадратичная интерполяция значений ρ_{m-1}^n, ρ_m^n, ρ_{m+1}^n. Если интерполируемые значения не имеют нужного запаса гладкости, формальное повышение точности интерполяции может привести к худшему результату.

Схему (20) можно получить и другим способом, который часто используется при конструировании схем повышенной точности. Строится простейшая схема (14) и аккуратно вычисляется главный член погрешности аппроксимации. Для схемы (14) получаем

$$\frac{\rho_m^{n+1} - \rho_m^n}{\tau} + u\frac{\rho_m^n - \rho_{m-1}^n}{h} = \frac{\partial p}{\partial t} + u\frac{\partial p}{\partial x} + \frac{u}{2}(u\tau - h)\frac{\partial^2 p}{\partial x^2} + O(\tau^2 + h^2).$$

Заметим, что при прямом вычислении появляется член $0.5\tau\rho_{tt}$, но в силу уравнения $\rho_{tt} = u^2\rho_{xx}$, и такая замена произведена. Теперь главный член погрешности переносится в левую часть, производная ρ_{xx} заменяется конечной разностью и получается схема (20), имеющая второй порядок аппроксимации на решениях уравнения переноса. Таким же способом можно получить характеристическую схему третьего порядка. Расчеты по этой схеме представлены на рис. 32*б*. Характеристические схемы для уравнений газовой динамики основаны на их записи в форме (11). Для вычисления величин в узле $(n+1, m)$ из него проводятся три характеристики, в точках их пересечения с линией $t = t_n$ интерполируются величины из узлов (n, m').

В лагранжевых переменных скорость контактного разрыва равна нулю, и нет проблемы его размывания. Именно это обстоятельство делает лагранжевы переменные весьма удобными и популярными при построении расчетных методов решения задач гидродинамики. К сожалению, это свойственно только одномерным задачам газовой динамики (задачи в t, x). При переходе к двумерным задачам использование лагранжевых переменных оказывается весьма трудным и во многих случаях просто невозможным. Подробнее об этом см. в § 23.

Расчет ударных волн. Искусственная вязкость. Перейдем к проблеме численного решения задач с ударными волнами. Будем использовать уравнения в массовых лагранжевых координатах в дивергентной форме (14). Разностная аппроксимация дифференциальных уравнений основана на предположении об определенной гладкости искомых решений. Когда этой гладкости нет, нужно вводить соответствующие усложнения вычислительной схемы.

Расчет ударных волн основан на предложенном фон-Нейманом и Рихтмайером приеме — на введении в уравнения искусственной вязкости. Она вводится таким образом, чтобы уравнения мало искажались вне зоны ударной волны. Сама же ударная волна при этом «размазывается» на пренебрежимо малую ширину. Таким образом мы заменяем уравнения газовой динамики на слабо возмущенные, но уже не имеющие разрывных решений уравнения. Конкретно, новые уравнения имеют вид

$$u_t + (p+q)_x = 0, \qquad v_t - u_x = 0,$$
$$\left(e + \frac{u^2}{2}\right)_t + [(p+q)u]_x = 0, \tag{21}$$

где q — малая искусственная вязкость.

Нейман и Рихтмайер предложили очень удобную конструкцию (она наиболее популярна и чаще других используется в расчетах):

$$q = \frac{\varepsilon}{2v}(u_x - |u_x|)u_x. \tag{22}$$

Здесь ε — малый параметр, выбор которого мы в дальнейшем уточним. Очевидно, $q \neq 0$, если $u_x < 0$. Это условие выделяет те участки течения, на которых происходит сжатие вещества (плотность растет: из $u_x < 0$ следует $v_t < 0$, т.е. $\rho_t > 0$). Наоборот, там, где течение сопровождается «разрежением» ($u_x > 0$, $q = 0$), вязкость «выключается». Ударная волна — это как раз участок течения, на котором происходит ударное, скачкообразное, сжатие вещества.

Достоинства неймановской искусственной вязкости проще всего продемонстрировать, сравнив точные решения простых модельных задач для исходной системы уравнений (14) и для уравнений с вязкостью (21). Эти простые решения относятся к важному классу автомодельных решений, в которых характерные особенности реальных решений (в данном случае, ударные волны) проявляются, так сказать, «в чистом виде», без взаимодействия с какими-то гладкими течениями.

Забегая вперед, опишем результат. Оказывается, решением исходной системы будет ударная волна — «ступенька», движущаяся с постоянной скоростью, а решением системы уравнений с искусственной вязкостью (21) будет движущаяся с той же скоростью «размазанная ступенька», причем ширина зоны размазывания фиксирована, она зависит, разумеется, от величины ε. Вне зоны «размазанной ударной волны» все функции u, v, p, e для обеих систем уравнений совпадают.

Итак, рассмотрим следующую ситуацию. Пусть имеется решение уравнений (14) типа чистой ударной волны, т.е.

$$u, v, e(t, x) = \begin{cases} u_1, v_1, e_1, & x > Dt, \\ u_2, v_2, e_2, & x < Dt, \end{cases}$$

где D — скорость ударной волны, и выполнены соотношения Гюгонио (в массовых лагранжевых переменных):

$$\begin{aligned} &\text{а)}\ -Du_1 + p_1 = C_1 = -Du_2 + p_2, \\ &\text{б)}\ Dv_1 + u_1 = C_2 = Dv_2 + u_2, \\ &\text{в)}\ -D\left(e_1 + \frac{u_1^2}{2}\right) + p_1u_1 = C_3 = -D\left(e_2 + \frac{u_2^2}{2}\right) + p_2u_2. \end{aligned} \tag{23}$$

Построим аналогичное решение для уравнений с искусственной вязкостью. Это решение будем искать в классе автомодельных решений типа «бегущей волны», т.е. когда все функции u, v, e зависят от одного аргумента $\xi = x - Dt$. Уравнения в частных производных (14) превращаются в обыкновенные после замены операторов $\partial/\partial t = -D\, d/d\xi$, $\partial/\partial x = d/d\xi$. Выписывая эту систему и интегрируя ее один раз, мы получаем систему «первых интегралов»:

$$\begin{aligned} -Du_\xi + (p+q)_\xi = 0 &\Rightarrow -Du + p + q = C_1, \\ -Dv_\xi - u_\xi = 0 &\Rightarrow Dv + u = C_2, \\ -D\left(e + \frac{u^2}{2}\right)_\xi + [(p+q)u]_\xi = 0 &\Rightarrow -D\left(e + \frac{u^2}{2}\right) + u(p+q) = C_3. \end{aligned} \tag{23*}$$

Так как мы не решаем какую-то определенную задачу, а просто конструируем нужное нам решение, постоянными интегрирования можем распоряжаться так, как нам будет удобно. В частности, здесь C_1, C_2, C_3 — те же самые, что и в соотношениях Гюгонио (23).

Дальнейшие выкладки будем проводить для идеального газа: $e = pv/(\gamma - 1)$. Руководящей идеей последующих преобразований является стремление получить уравнение только для $v(\xi)$, остальные переменные будем исключать через v. Следующие ниже выкладки оправданы лишь при $u_x < 0$, т.е. при $v_\xi > 0$. В этом случае $q = \varepsilon D^2 (v_\xi)^2/v$. Из полученных «первых интегралов» имеем

$$u = C_2 - Dv, \qquad (p+q) = C_4 - D^2 v, \qquad p = C_4 - D^2 v - q.$$

Точные значения постоянных C_4 и других (A, B; см. ниже) нас пока не интересуют.

Из «интеграла энергии» для идеального газа

$$-D\left(\frac{pv}{\gamma - 1} + \frac{u^2}{2}\right) + (p+q)u = C_3$$

исключим p, u, $(p+q)$ по уже найденным формулам. После простых преобразований получаем уравнение для v:

$$qv = -\frac{\gamma+1}{2} D^2 v^2 + Av + B. \tag{24}$$

Утверждение. Многочлен в правой части (24) обращается в нуль при $v = v_1$ и $v = v_2$; поэтому он может быть записан в виде

$$qv = -\frac{\gamma+1}{2} D^2 (v - v_1)(v - v_2). \tag{25}$$

Доказательство. Соотношение (24) является следствием соотношений (23*). Покажем, что полагая в этих соотношениях $v = v_1$, можно получить в качестве следствий равенства $p = p_1$, $u = u_1$, $q = 0$.

В самом деле, из $Dv + u = C_2 = Dv_1 + u_1$ при $v = v_1$ следует, что $u = u_1$. Из $-Du + p + q = C_1 = -Du_1 + p_1$ при $u = u_1$ имеем $p + q = p_1$. Из

$$-D\left(\frac{pv}{\gamma-1} + \frac{u^2}{2}\right) + (p+q)u = C_3 = -D\left(\frac{p_1 v_1}{\gamma-1} + \frac{u_1^2}{2}\right) + p_1 v_1$$

при $v = v_1$, $u = u_1$, $p + q = p_1$ получаем $p = p_1$ и, следовательно, $q = 0$. Таким образом, при $v = v_1$ левая, а следовательно, и правая части (24) обращаются в нуль. Точно так же рассматривается и случай $v = v_2$.

Используя выражение для q, получаем дифференциальное уравнение для v:

$$\varepsilon (v_\xi)^2 = 0.5(\gamma+1)(v - v_1)(v_2 - v) \tag{26}$$

(и условие $v_\xi > 0$). После замены переменных $\xi = \sqrt{2\varepsilon/(\gamma+1)}\,\eta$ и $v = (v_1 + v_2)/2 + z(v_1 - v_2)/2$ уравнение (26) принимает вид

$$(z_\eta)^2 = 1 - z^2.$$

Решение угадывается: $z(\eta) = \pm 1$ или $z(\eta) = \sin \eta$. Последним решением можно пользоваться (в силу условия $v_\xi > 0$, т.е. $z_\eta > 0$) лишь при $\eta \in [-\pi/2, \pi/2]$.

Возвращаясь к прежним переменным, получаем решение (продолжая его постоянным за пределами выделенного интервала η):

$$v(\xi) = \begin{cases} v_2, & \xi = x - Dt < -\sqrt{\frac{2\varepsilon}{\gamma+1}}\,\frac{\pi}{2}, \\ \frac{v_1+v_2}{2} + \frac{v_1-v_2}{2} \sin \sqrt{\frac{\gamma+1}{2\varepsilon}}(x - Dt), & |\xi| < \sqrt{\frac{2\varepsilon}{\gamma+1}}\,\frac{\pi}{2}, \\ v_1, & \xi = x - Dt > \sqrt{\frac{2\varepsilon}{\gamma+1}}\,\frac{\pi}{2}. \end{cases}$$

Такие же формулы легко получить и для функций $u(\xi)$ и $p(\xi) + q(\xi)$ с заменой v_1, v_2 на u_1, u_2 или p_1, p_2 соответственно.

Итак, получено решение уравнений с вязкостью, совпадающее с решением типа «чистая ударная волна» всюду, за исключением узкой полосы вдоль фронта ударной волны $|x - Dt| \leqslant (\pi/2)\sqrt{2\varepsilon/(\gamma+1)}$. Ширина «размазанной ударной волны» есть $\pi\sqrt{2\varepsilon/(\gamma+1)}$. Опыт показал, что хорошие результаты дает выбор ε, при котором волна разре-

шается четырьмя-пятью счетными точками. Например,

$$5h = \pi\sqrt{2\varepsilon/(\gamma+1)}, \quad \text{т.е.} \quad \varepsilon = 25h^2(\gamma+1)/(2\pi^2) \approx 2h^2.$$

Отметим, что можно выписать формулы для $p(\xi)$ в зоне волны. Детали нам не очень нужны, отметим лишь, что фактически область плавного перехода от p_1 к p_2 примерно в два раза уже, чем для остальных функций v, u, $p+q$. При расчетах в лагранжевых координатах по x дифференцируются только функции u и $p+q$, каждая из которых имеет стандартную ширину зоны размазывания. Функции p и q отдельно по x не дифференцируются. Поэтому сокращение фактической ширины волны для p не играет роли. Правда, мы должны еще обеспечить должное размазывание фронта волны на четыре-пять точек сетки по времени. Практика расчетов, проводившихся в пятидесятых годах привела к такому рецепту. Ударная волна за один шаг времени должна проходить примерно половину интервала по массовой координате, т.е. по времени зона размазанной волны захватывает примерно в два раза больше счетных интервалов.

Обратим внимание на то, что по t дифференцируются все функции. Таким образом, и более крутой график p также «разрешается» четырьмя-пятью точками сетки по t. Этот рецепт (половина шага по пространству за шаг по времени) не был связан с условием устойчивости, так как мы применяли неявные абсолютно устойчивые схемы. Он был связан с тем, что для необходимой точности разностной аппроксимации нужно разметить на крутых профилях функции четыре-пять счетных точек сетки (как по времени, так и по пространству). Попытки расчетов с бо́льшими шагами по t (расчет, например, со скоростью один шаг по пространству за шаг по времени) приводили к ухудшению результатов: на графиках появлялись осцилляции явно счетного происхождения.

Заметим, что все проведенные выкладки можно повторить и в эйлеровых координатах. Однако ситуация осложняется тем, что в эйлеровой форме уравнений по x дифференцируются все функции. Поэтому приходится брать в два раза более широкую зону размазывания (по x), чтобы фактическая ширина зоны размазывания p была покрыта четырьмя-пятью интервалами сетки. Это уже не очень приятно, так остальные величины при этом размазываются на десять точек. В 1962 г. автором проводился расчет ударной волны в эйлеровых координатах. Чтобы избежать слишком широкого счетного фронта волны, была выбрана довольно экстравагантная форма записи уравнений: в качестве основных функций были взяты u, ρ, c где c — скорость звука. Уравнения (в недивергентной форме) оказались такими, что по x дифференцировались только функции, профили которых имели в зоне волны стандартный синусоидальный вид. Это позволило вести расчеты с шириной размазывания порядка $4h$. (Подробнее об этом см. § 22 в связи с применением гибридной схемы для решения уравнений газовой динамики в эйлеровых координатах.)

§ 21. Нелинейное уравнение теплопроводности

Рассмотрим некоторые вопросы, возникающие при численном решении нелинейного уравнения теплопроводности.

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left[\varkappa(x, t, u)\frac{\partial u}{\partial x}\right] + Q(x, t, u), \qquad (1)$$

которое решается в простой области $0 \leqslant t \leqslant T$, $0 \leqslant x \leqslant X$ с начальными данными $u(0, x) = u^0(x)$ и краевыми условиями, для простоты, первого рода: $u(t, 0) = u(t, X) = 0$.

Нужно иметь в виду, что нелинейность уравнений приводит к сложностям не сама по себе, а лишь в тех случаях, когда она порождает сложные, описываемые негладкими функциями, явления. Чтобы сделать это более конкретным, рассмотрим важный в приложениях случай, когда коэффициент теплопроводности $\varkappa$ зависит от u степенным образом:

$$u_t = [u^k u_x]_x. \qquad (2)$$

Уравнения такого типа встречаются при описании процессов в высокотемпературном веществе (лучистая теплопроводность), например в звездах. Аналогичные уравнения описывают и процессы фильтрации.

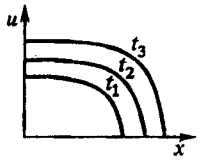


Рис. 33

Рассмотрим характерное и очень важное в приложениях явление, описываемое этим уравнением, — так называемую тепловую волну, или, иначе, тепловой фронт. На рис. 33 изображены графики функций $u(t_i, x)$ для трех моментов времени $t_1 < t_2 < t_3$. В этом случае мы имеем процесс распространения высокой температуры по «нулевому фону» (перед фронтом тепловой волны $u = 0$; в действительности, конечно, перед фронтом температура не нулевая, но очень маленькая по сравнению с температурой за фронтом).

Тепловой фронт. Будем искать «автомодельное» решение уравнения, т.е. решение, зависящее не от t и x, а от их комбинации, в данном случае от $\xi = x - Dt$, где D — некоторая постоянная, смысл которой потом станет ясным. Тогда

$$\frac{\partial u}{\partial t} = \frac{du}{d\xi}\frac{\partial \xi}{\partial t} = -D\frac{du}{d\xi}, \qquad \frac{\partial u}{\partial x} = \frac{du}{d\xi}.$$

Нам следует найти решение обыкновенного дифференциального уравнения $-Du' = [u^k u']'$. Интегрируя, получаем $-Du = u^k u'$. (Так как мы ищем какое-нибудь решение, постоянную интегрирования положим равной нулю.) Уравнение $u^{k-1}u' = -D$, или $(u^k)' = -kD$, интегрируется и дает $u^k(\xi) = -kD\xi$.

Итак, мы получили решение $u(\xi) = \sqrt[k]{kD}(-\xi)^{1/k}$. Но нас интересует вещественное и положительное u. Поэтому это решение имеет смысл только при $\xi \leqslant 0$, т.е. при $x \leqslant Dt$. Положим $u(\xi) = 0$ при $\xi > 0$. Легко видеть, что функция $u(\xi) \equiv 0$ является решением. Вопрос только в том, можно ли эти два решения склеить, точнее говоря, можно ли и в каком смысле говорить, что функция

$$u(\xi) = \begin{cases} \sqrt[k]{kD}(-\xi)^{1/k}, & \xi \leqslant 0, \\ 0, & \xi > 0 \end{cases} \tag{3}$$

есть решение уравнения теплопроводности.

Естественно обратиться к понятию обобщенного решения, так как в любой точке (t, x), кроме линии $x = Dt$ (фронт тепловой волны), эта функция удовлетворяет уравнению теплопроводности в классическом смысле. Обобщенное же решение вводится как функция, удовлетворяющая интегральному тождеству (закону сохранения). Для любой области Ω

$$\iint\limits_{\Omega} \left(\frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\left(\varkappa \frac{\partial u}{\partial x}\right)\right) dx\, dt = 0, \tag{4}$$

или для любого контура $\partial\Omega$

$$\oint\limits_{\partial\Omega} \left(u\, dx + \varkappa \frac{\partial u}{\partial x}\, dt\right) = 0. \tag{5}$$

Именно в этой форме и проверяется, является ли u обобщенным решением. Если контур не пересекает фронта, проблемы нет; там, где функция $u(t, x)$ гладкая, соотношения (2), (4), (5) эквивалентны.

Рассмотрим элементарный контур $\partial\Omega$, пересекающий фронт $x = Dt$ (рис. 34). Проведем линии *12* и *34*, параллельные фронту и находящиеся от него на расстоянии ε. Тогда

$$\int\limits_{\partial\Omega} = \lim_{\varepsilon \to 0} \left\{ \int\limits_{1A2} + \int\limits_{3B4} \right\}.$$

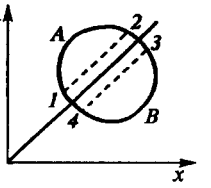


Рис. 34

Заменим эти части контура штриховыми линиями, показанными на рис. 34. Обозначим замкнутые контуры через *1A21* и *3B43*. Каждый из этих контуров лежит в области, в которой составное решение (5) является классическим. Поэтому, например, справедливо соотношение

$$\int\limits_{1A21} = \int\limits_{1A2} + \int\limits_{21} = 0, \quad \text{т.е.} \quad \int\limits_{1A2} = -\int\limits_{21}.$$

Итак,

$$\oint = -\lim_{\varepsilon \to 0} \left\{ \int\limits_{21} + \int\limits_{43} \right\}.$$

Очевидно, $\int_{43} = 0$. Осталось оценить

$$\int_{1}^{2} (u\,dx + u^k \frac{\partial u}{\partial x}\,dt).$$

На линии *12* имеем $|x - Dt| = \varepsilon$, $u(x - Dt) = O(\varepsilon^{1/k})$; следовательно, интеграл от u стремится к нулю. На этой же линии $u^k = O(\varepsilon)$, $u_x = du/d\xi = O(\varepsilon^{1/k-1})$. Таким образом, $u^k u_x = O(\varepsilon^{1/k})$ на линии *12*, и интеграл от этой величины также стремится к нулю. Итак, составное решение (3) является обобщенным решением нелинейного уравнения теплопроводности.

Обсудим несколько вопросов, возникающих при построении разностных схем для уравнения теплопроводности.

Аппроксимация на контактном разрыве. Рассмотрим уравнение (1) в случае, если $Q = 0$ и коэффициент теплопроводности $\varkappa(x)$ разрывен. Пусть среда имеет контактную границу, т.е. при $x < 0$ одно вещество с коэффициентом теплопроводности $\varkappa_1$, при $x > 0$ другое вещество с коэффициентом теплопроводности $\varkappa_2$. Рассмотрим разностную схему, в которой по каким-то причинам удобно поместить контактный разрыв в «целую» счетную точку, а температуру u определить в полуцелых. Это бывает в задачах, в которых, кроме теплопроводности, учитываются и другие процессы, т.е. уравнение для u входит в более сложную систему, например в систему гидродинамики с теплопроводностью (см. § 22). Если счетная точка, в которой определена температура u (и другие термодинамические параметры), совпадает с контактным разрывом, возникают сложности с уравнением состояния.

Следуя § 11, построим в пространстве (x, t) сетку с узлами (m, n), приписывая им координаты x_m, t_n $(x_{m+1} = x_m + h_{m+1/2})$. Введем сеточную функцию $u^n_{m+1/2}$, считая ее определенной в точке $x_{m+1/2} = 0.5(x_m + x_{m+1})$. Разностная аппроксимация строится так:

$$\frac{u^{n+1}_{m+1/2} - u^n_{m+1/2}}{\tau} = \frac{\Pi_{m+1} - \Pi_m}{h_{m+1/2}},$$

где Π_m — аппроксимация теплового потока $\varkappa u_x$ через границу x_m. Если в точке x_m свойства среды непрерывны, для Π_m имеем очевидную аппроксимацию:

$$\Pi_m = (\varkappa_m/h_m)(u^n_{m+1/2} - u^n_{m-1/2}) \quad \text{(явная схема)},$$

$$\Pi_m = (\varkappa_m/h_m)(u^{n+1}_{m+1/2} - u^{n+1}_{m-1/2}) \quad \text{(неявная схема)}.$$

Разберем вопрос о том, как следует поступать в точке $m = 0$, в которой рвутся $\varkappa$ и u_x, т.е. $u_{xx} \approx \delta(x)$ ($\delta(x)$ — дельта-функция), и, стало быть, не выполняются предположения, на которых базируется стандартная техника построения разностных аппроксимаций. В подобных ситуациях необходимо привлечь более точную информацию о дифференциальных свойствах искомого решения. В данном случае следует использовать физические предположения о процессе распространения теплоты. В точке $x = 0$:

а) функция $u(t, x)$ непрерывна, т.е. $u(t, -0) = u(t, +0)$;

б) непрерывен тепловой поток, т.е. $\varkappa_1 u_x(t, -0) = \varkappa_2 u_x(t, +0)$.

Введем (временно) в точке $x = 0$ температуру u_0 и запишем разностную аппроксимацию условия непрерывности теплового потока (справа и слева от разрыва функция u гладкая, только точка $x = 0$ является точкой нарушения гладкости) в виде

$$\varkappa_1 \frac{u_0 - u_{-1/2}}{h_1/2} = \varkappa_2 \frac{u_{1/2} - u_0}{h_2/2}.$$

(Временной индекс не пишем, он может быть и n, и $n+1$.) Отсюда

$$u_0 = \left(\frac{\varkappa_1}{h_1} u_{-1/2} + \frac{\varkappa_2}{h_2} u_{1/2}\right) \Big/ \left(\frac{\varkappa_1}{h_1} + \frac{\varkappa_2}{h_2}\right).$$

Опуская простые выкладки, вычисляем Π_0:

$$\Pi_0 = 2(u_{1/2} - u_{-1/2}) \Big/ \left(\frac{h_1}{\varkappa_1} + \frac{h_2}{\varkappa_2}\right). \tag{6}$$

Такую аппроксимацию теплового потока в точке контактного разрыва иногда называют «наилучшей». (Ниже будет показано, как опасно придавать этому термину слишком универсальное значение.) Эта же формула применяется не только в случае разрыва коэффициента $\varkappa$, но и при переменном коэффициенте теплопроводности. В частности, в рассмотренной выше задаче «газодинамика + теплопроводность», коэффициент $\varkappa$ зависит от термодинамических величин u, ρ (ρ — плотность вещества), а эти величины определены в полуцелых точках. Таким образом, в каждой «целой» точке m тепловой поток аппроксимируется формулой

$$\Pi_m = 2(u_{m+1/2} - u_{m-1/2}) \Big/ \left(\frac{h_{m+1/2}}{\varkappa_{m+1/2}} + \frac{h_{m-1/2}}{\varkappa_{m-1/2}}\right). \tag{7}$$

Аппроксимация при расчете тепловой волны. Выведенная выше «наилучшая» формула (7), однако, не пригодна для расчета тепловой волны. И вот почему. Пусть в начальных данных фронт тепловой волны находится в точке x_l, т.е. $u_{m+1/2} = 0$ при $m \geqslant l$. В этом случае $\Pi_m = 0$ при $m \geqslant l$ (так как $h_{m+1/2}/u^k_{m+1/2} = 1/0$). Таким обра-

зом, тепловой поток через точку x_l равен нулю и температура при $x_m \geqslant x_l$ останется нулевой и в дальнейшем, т.е. фронт тепловой волны не продвигается, а застревает в начальном положении.

Для правильного расчета тепловой волны используется аппроксимация, учитывающая структуру функции $u(t, x)$ в окрестности фронта тепловой волны (3). Обратим внимание на то, что $u^k(t, x)$ — линейная функция x в окрестности фронта. Учитывая этот факт, в расчетах применяют аппроксимацию типа

$$\Pi_m = \frac{u^k_{m-1/2} + u^k_{m+1/2}}{2} \frac{u_{m+1/2} - u_{m-1/2}}{0.5(h_{m-1/2} + h_{m+1/2})}. \qquad (8)$$

Для линейной функции u^k линейная интерполяция является точной. В свое время автору приходилось решать задачи гидродинамики с теплопроводностью в ситуации, когда коэффициент теплопроводности $\varkappa$ имел вид $\varkappa(u, \rho, x) = f(x, \rho)u^k$, причем функция f была разрывной по x; искомая функция $\rho(t, x)$ тоже была разрывной (контактные разрывы). При этом были плохи обе формулы (7) и (8): первая препятствовала правильному расчету тепловой волны (а это явление играло важную роль в проводившихся расчетах), вторая приводила к погрешностям на контактных разрывах.

Решение было найдено в виде компромиссной формулы

$$\Pi_m = \frac{u^k_{m-1/2} + u^k_{m+1/2}}{2} \frac{u_{m+1/2} - u_{m-1/2}}{0.5[(h/f)_{m-1/2} + (h/f)_{m+1/2}]}, \qquad (9)$$

в которой разрывная часть коэффициента теплопроводности учитывалась так, как это рекомендуется теорией для разрывного коэффициента, а множитель u^k, ответственный за фронт тепловой волны, усреднялся с учетом типичного графика $u(t, x)$ в окрестности фронта.

Уравнение теплопроводности с нелинейным источником. Рассмотрим уравнение (1) в случае, если $\varkappa = \varkappa(u)$, $Q = Q(u)$. Допустим, мы используем неявную схему. Возникает вопрос: что делать с нелинейностями в $\varkappa$ и Q? Есть два варианта. Можно оставить их «на нижнем слое» и получить простую схему

$$\frac{u^{n+1}_m - u^n_m}{\tau} = \frac{1}{h}\left[\varkappa^n_{m+1/2} \frac{u^{n+1}_{m+1} - u^{n+1}_m}{h} - \varkappa^n_{m-1/2} \frac{u^{n+1}_m - u^{n+1}_{m-1}}{h}\right] + Q(u^n_m), \qquad (10)$$

где $\varkappa^n_{m+1/2} = \varkappa(u^n_m + u^n_{m+1})/2$. Уравнение «на верхнем слое» (для u^{n+1}) линейное; оно решается прогонкой.

Второй вариант (нелинейность «на верхнем слое») отличается от схемы (10) только тем, что в нем используются значения $Q(u^{n+1}_m)$. В этом случае уравнения «на верхнем слое» нелинейные. Их прихо-

дится решать итерациями с линеаризацией по методу Ньютона и прогонкой для линеаризованных уравнений. Это, конечно, намного сложнее, чем при аппроксимации (10). Из общих соображений трудно понять, зачем нужна такая трудно реализуемая схема. Однако в литературе часто встречаются указания на предпочтительность именно более сложной схемы. В чем дело? Попробуем немного прояснить этот вопрос. Все дело в характере нелинейности и в шаге τ по времени. Грубо говоря, дело обстоит так. Если в рассчитываемом процессе шаг τ таков, что $|\tau Q_u(u)| \ll 1$, то обе схемы более или менее равносильны, и следует отдать предпочтение более простой схеме (10).

Поясним это положение следующими оценками. За один шаг τ температура изменится на τQ (предполагаем, что $\varkappa u_{xx}$ — величина того же порядка), т.е. $u^{n+1} \approx u^n + \tau Q$. Вычислим

$$Q(u^{n+1}) \approx Q(u^n + \tau Q) \approx Q(u^n) + \tau Q_u Q = (1 + \tau Q_u) Q(u^n).$$

Если $|\tau Q_u| \ll 1$, то $Q(u^n)$ и $Q(u^{n+1})$ почти совпадают и схема, в которой Q вычисляется на верхнем слое, мало чем отличается от схемы с вычислением Q на нижнем слое. Но бывают задачи, например связанные с расчетом тепловых явлений в звездах и близких к ним объектах, когда вычисления с шагом τ, таким, что $\tau|Q_u| \ll 1$, немыслимы. Это слишком малый шаг. С таким шагом τ за приемлемое время работы ЭВМ не удается провести расчет на заданном интервале времени $[0, T]$ (число шагов T/τ слишком велико) и нужно считать с шагом $\tau \gg 1/|Q_u|$.

При определенных условиях выходом из положения является аппроксимация источника на верхнем слое. Нужно сказать, что эта ситуация внешне и по существу очень близка к тому кругу вопросов, которые мы обсуждали при описании жестких систем и методов их интегрирования (там тоже решающую роль играли неявные схемы). Естественным условием применимости схемы с большим шагом τ является «устойчивость»: $Q_u < 0$.

Описанная выше ситуация часто встречается в задачах астрофизики, когда $Q = Q_1 - Q_2$, где Q_1 и Q_2 определяют выделение (за счет ядерных реакций, например) и поглощение энергии соответственно. Оба этих процесса очень интенсивны и «почти сбалансированы», т.е. $|Q_1 - Q_2| \ll |Q_1| + |Q_2|$.

Другими словами, выделившаяся в какой-то точке энергия поглощается почти в этом же месте. Разумеется, термин «почти в этом же месте» означает, что энергия поглощается на расстоянии от места выделения, меньшем шага счетной сетки. В задачах, связанных с расчетом процессов в звездах, когда по радиусу звезды вводится $10^2 \div 10^3$ точек, шаг h достигает тысяч километров.

Если в описываемой ситуации $Q_u > 0$, нужно считать с шагом $\tau \ll Q_u^{-1}$ или придумывать что-то другое, более сложное, чем переход к неявной схеме.

Зачем нужны неявные схемы? Ответ на этот вопрос кажется совершенно очевидным. Неявные схемы нужны для того, чтобы считать задачи с шагом по времени τ, существенно бо́льшим, чем это позволяет известное условие Куранта. Для уравнения $u_t = u_{xx}$, $0 \leqslant x \leqslant 1$, $t \geqslant 0$, $u(t, 0) = u(t, 1) = 0$, на примере которого мы ниже рассмотрим некоторые вопросы, условие Куранта, как известно, имеет вид $\tau \leqslant 0.5h^2$.

Итак, неявные схемы позволяют проводить устойчивый счет при $\tau \gg h^2/2$, например при $\tau \approx h$. Но всегда ли следует пользоваться этим преимуществом? Обсудим это. Результат будет примерно такой. Для уравнения теплопроводности соотношение $\tau \approx h^2$, в известной мере, естественное. Его не следует нарушать очень уж сильно, счет с $\tau \gg h^2$ требует большой осторожности. Это связано с другим важным понятием — с фактической погрешностью аппроксимации.

Перейдем к конкретному анализу. Допустим, мы проводим расчет с шагом $h = 1/100$. Имеет ли смысл расчет с шагом $\tau = h = 1/100$? Вообще говоря, нет. Используем неявную схему:

$$\frac{u_m^{n+1} - u_m^n}{\tau} = \frac{u_{m-1}^{n+1} - 2u_m^{n+1} + u_{m+1}^{n+1}}{h^2}, \quad m = 1, 2, \ldots, M-1.$$

Разложим решение в ряд Фурье:

$$u_m^n = \sum_k c_k^n \sin(k\pi m h).$$

Тогда для коэффициентов Фурье c_k^n без труда можно получить разностное уравнение:

$$\frac{c_k^{n+1} - c_k^n}{\tau} = -\frac{4}{h^2} \sin^2 \frac{k\pi h}{2}\, c_k^{n+1}, \quad \text{т.е.} \quad c_k^n = c_k^0 \left(1 + 4\frac{\tau}{h^2} \sin^2 \frac{k\pi h}{2}\right)^{-n}.$$

Точное решение дифференциальной задачи имеет вид

$$u(t, x) = \sum_k c_k^0 e^{-\lambda_k t} \sin(k\pi x), \quad \lambda_k = k^2\pi^2.$$

О точности разностного решения можно судить, сравнивая величины $e^{-\lambda_k n\tau}$ и $\left(1 + 4\frac{\tau}{h^2} \sin^2 \frac{k\pi h}{2}\right)^{-n}$. При $|k\pi h| \ll 1$ можно положить

$$1 + 4\frac{\tau}{h^2} \sin^2 \frac{k\pi h}{2} \approx 1 + 4\frac{\tau}{h^2} k^2\pi^2 h^2 = 1 + \tau k^2 \pi^2.$$

Если еще принять, что $\tau k^2\pi^2 \ll 1$, то $1 + \tau k^2\pi^2 \approx e^{\tau k^2\pi^2}$ и

$$\left(1 + 4\frac{\tau}{h^2}\sin^2\frac{k\pi h}{2}\right)^{-n} \approx e^{-n\tau k^2\pi^2}.$$

Этой простой оценкой мы выяснили следующие обстоятельства. Достаточно правильно вычисляются те Фурье-компоненты решения (т.е. те члены ряда $\sum c_k^n \sin(k\pi mh)$), у которых $k\pi h \ll 1$. Так как $h = 0.01$, то это, грубо говоря, первые 20 гармоник ($20\pi \cdot h \approx 0.6$). В общем случае, при $h = 1/N$, это примерно 1/5 всех присутствующих в разностном решении гармоник, так как $(N/5)\pi h \approx 0.6$. Считаем, что $0.6 \ll 1$, так как $\sin^2\alpha/\alpha^2 \approx 0.89$ при $\alpha = 0.6$.

Итак, на данном этапе рассмотрения мы отказались от претензий правильно рассчитывать эволюцию 4/5 всех гармоник, т.е. в рассчитываемом явлении они не должны играть существенной роли. Интересующее нас решение не должно претерпеть существенных изменений, если мы ограничимся отрезком ряда Фурье

$$\tilde{u}(t, x) = \sum_{k=1}^{20} c_k^0 e^{-\lambda_k t} \sin k\pi x.$$

Если это не так, то ста точек для расчета недостаточно. Но и это еще не все: точность численного решения задачи зависит от шага по времени τ.

Выше было введено условие, при котором разностный коэффициент Фурье c_k^n воспроизводит правильное значение $c_k^0 e^{-\lambda_k n\tau}$ с точностью, зависящей как от k, так и от τ. Проверим условие $\tau k^2\pi^2 \ll 1$, полагая $k = N/5$:

$$\tau\pi^2(N/5)^2 \leqslant 0.5, \quad \text{т.е.} \quad \tau \leqslant 1/N^2 = h^2.$$

(Здесь мы считаем, что $0.5 \ll 1$, так как $e^{-0.5} = 0.61 \approx 1 - 0.5$.) Таким образом, даже при не очень высоких требованиях к точности, мы пришли почти к тому же соотношению между τ и h, которое следует из требования устойчивости для явной схемы!

Можно взять схему второго порядка точности по τ:

$$\frac{u_m^{n+1} - u_m^n}{\tau} = \frac{1}{2}\frac{u_{m-1}^{n+1} - 2u_m^{n+1} + u_{m+1}^{n+1}}{h^2} + \frac{1}{2}\frac{u_{m-1}^n - 2u_m^n + u_{m+1}^n}{h^2}.$$

Для c_k^n получаем выражение

$$c_k^n = c_k^0 \left[\left(1 - 2\frac{\tau}{h^2}\sin^2\frac{k\pi h}{2}\right) \Big/ \left(1 + 2\frac{\tau}{h^2}\sin^2\frac{k\pi h}{2}\right)\right]^n.$$

При том же условии $k\pi h \ll 1$ имеем

$$c_k^n \approx c_k^0 \left[\left(1 - \frac{\tau k^2\pi^2}{2}\right) \Big/ \left(1 + \frac{\tau k^2\pi^2}{2}\right)\right]^n.$$

Сравним функции $(1 - 0.5x)/(1 + 0.5x)$ и e^{-x}. При $x = 1$ погрешность достигает 10 %, при $x = 0.5$ она порядка 1 %. В этом случае мы можем претендовать на приличную точность при $\tau k^2\pi^2 \leqslant 0.5$, т.е. при $\tau \leqslant h^2$. (Это может породить неверное впечатление, будто бы схема второго порядка точности не имеет преимуществ. Конечно, имеет, но они сказываются в точности расчета более гладких компонент решения, соответствующих меньшим k. Не надо также упускать из вида, что как в точном решении, так и в разностном негладкая компонента решения быстро стремится к нулю с ростом t.)

В каких же задачах применение неявных схем дает существенный выигрыш по сравнению с явными? Это — задачи с нелинейной теплопроводностью и с решением типа «тепловой фронт». В решениях таких задач можно выделить три характерных области (см. рис. 33).

1. Зона *перед* фронтом тепловой волны (или «фон», по которому распространяется тепловая волна) характеризуется очень малыми температурами u_Φ. Коэффициент теплопроводности u_Φ^k в уравнении (2) так мал, что практически не происходит никаких перетоков теплоты. Это утверждение имеет смысл лишь относительно тех характерных времен T, которые нас интересуют в данной задаче. Оно означает, что $Tu_\Phi^k/L^2 \ll 1$, где L — характерное для задачи расстояние по x.

2. За фронтом тепловой волны температура u_+ очень велика и реализуется почти изотермический режим: температура почти не меняется по x, но может меняться по t. Опять-таки дело в том, что очень велик коэффициент теплопроводности u_+^k, точнее (в безразмерных терминах) $\tau u_+^k/L^2 \gg 1$. Из этого соотношения следует, что за малое с точки зрения характерных времен задачи время $\tau \ll T$ в изотермической зоне успевает выравняться температура.

Пусть в начальных данных в изотермической зоне имеется какой-то профиль температуры. Разложим его в ряд Фурье, учитывая, что характерный масштаб по x есть L:

$$u(0, x) = c_0 + \sum_{m=\pm 1}^{\pm\infty} c_m\, e^{im\pi x/L}.$$

(Для иллюстрации примем модель линейной задачи: $u_t = u_+^k u_{xx}$.) Тогда через время τ решение будет

$$u(\tau, x) = c_0 + \sum_m c_m\, e^{im\pi x/L}\, e^{-m^2\pi^2 u_+^k \tau/L^2}.$$

Даже для самой гладкой и наиболее медленной первой гармоники ($m = \pm 1$) временной множитель $\exp(-\pi^2 u_+^k \tau/L^2) \ll 1$, т.е. фактически $u(\tau, x) = c_0$. Возмущения и неровности, наложенные на изотермический профиль, мгновенно (с точки зрения времени T) выровнялись. В этой зоне u_+^k играет роль большого параметра, и решение определяется «квазистационарным» уравнением $[u^k u_x]_x = 0$, или $u^k u_x = c(t)$, т.е. тепловой поток почти постоянен по x, но, вообще говоря, может меняться по времени.

3. И наконец, есть еще переходная зона — зона тепловой волны, в которой u переходит от u_+ к $u_ф$ и профиль $u(t, x)$ носит характер, близкий к автомодельному. Передний фронт тепловой волны рассчитывается при условии Куранта $\tau u^k \approx h^2$.

Нельзя заранее разделить всю область переменных (t, x) на эти три зоны. Мы хотим их рассчитывать по единой схеме, не вводя разных формул в разных зонах. Здесь проявляется решающее преимущество неявной схемы, которая выдерживает сильное изменение критерия Куранта (безразмерной величины $\tau u^k/h^2$) и не теряет устойчивости.

Теперь уместно вспомнить, что мы только что скомпрометировали расчет с очень большим «курантом», показав, что он не обеспечивает точности в передаче временной эволюции коэффициентов Фурье. Это так, но в изотермической зоне точный темп временной эволюции разных гармоник нам и не нужен. Важно только, чтобы разностная схема правильно передавала качественный характер — почти полное их исчезновение за время τ, и это она обеспечивает. Вспомним еще раз жесткие системы уравнений: уравнение теплопроводности является жесткой системой в бесконечномерном пространстве. Именно решение задач с описанным выше качественным поведением решения и было основным стимулом, приведшим к активному использованию неявных схем для уравнений теплопроводности и к «изобретению» метода прогонки.

Метод потоковой прогонки. При расчете решений, содержащих изотермический участок с очень большим коэффициентом теплопроводности, вычислители встретились с характерной трудностью, преодоление которой, в частности, привело к созданию специального варианта прогонки, названного потоковым. В чем же было дело? На этом участке, как уже отмечалось, поток $u^k u_x \approx c(t)$ был почти постоянным по x, а величина $c = O(1)$. Стало быть, $u_x = c/u^k$ — величина очень малая и в некоторых ситуациях настолько малая, что ее невозможно правильно вычислить по разностной формуле типа $(u_m - u_{m-1})/h$.

В самом деле, величины u_m в машинном представлении заменяются на $\tilde{u}_m = u_m(1+\varepsilon_m)$, где ε — погрешность машинного представления чисел. Фактически ЭВМ вычисляет

$$\frac{\tilde{u}_m - \tilde{u}_{m-1}}{h} = \frac{u_m - u_{m-1}}{h} + O\left(\frac{u\varepsilon}{h}\right).$$

В некоторых случаях мы сталкиваемся с ситуацией, когда погрешность $O(u\varepsilon/h)$ много больше основной величины $(u_m - u_{m-1})/h$ и ничего хорошего ожидать не приходится.

Поясним сказанное несколько иначе. Конечное значение теплового потока c получается, так сказать, раскрытием неопределенности «$c = 0 \cdot \infty$», причем $u_x \approx 0$, $u^k \approx \infty$. Но на ЭВМ с конечным числом разрядов в представлении числа значение разностной производной $(\tilde{u}_m - \tilde{u}_{m-1})/h$ не может стремиться непрерывно к нулю: оно либо нуль (при $\tilde{u}_m \equiv \tilde{u}_{m-1}$), либо не меньше $|u|\varepsilon/h$.

Первый случай реализуется при совпадении u_m и u_{m-1} со всеми машинными знаками, второй — при различии их хотя бы в последнем знаке мантиссы машинного числа. Если модуль $|c| \ll u^k u_x \varepsilon/h$, поток $u^k u_x$ либо нуль, либо много больше c. В этом источник трудностей. Его можно преодолеть переходом к расчету с двойной точностью, но можно поступить иначе, применяя метод потоковой прогонки. Его основу составляет представление уравнения теплопроводности (1) в виде системы

$$\frac{\partial u}{\partial t} = \frac{\partial \Pi}{\partial x} + Q, \qquad \varkappa \frac{\partial u}{\partial x} = \Pi.$$

Запишем разностные уравнения на сетке $\{x_m\}_{m=0}^{M} = 0$ $(x_m = mh)$:

$$\begin{aligned} &\frac{u_m^{n+1} - u_m^n}{\tau} = \frac{\Pi_{m+1/2} - \Pi_{m-1/2}}{h} + Q_m, \quad m = 1, 2, \ldots, M-1, \\ &\varkappa_{m+1/2} \frac{u_{m+1}^{n+1} - u_m^{n+1}}{h} = \Pi_{m+1/2}, \quad m = 0, 1, \ldots, M-1. \end{aligned} \tag{11}$$

Прогоночное соотношение имеет вид

$$u_m^{n+1} = P_m \Pi_{m+1/2} + R_m, \quad m = 0, 1, \ldots, M-1,$$

где P_m, R_m — прогоночные коэффициенты, которые должны быть определены. Первые прогоночные коэффициенты P_0, R_0 определяются из левого краевого условия. Пусть оно имеет вид $u(t, 0) = \varphi(t)$. Тогда $P_0 = 0$, $R_0 = \varphi(t_{n+1})$. Предоставим читателю вывести формулы для P_0, R_0 в случае, когда поставлено общее краевое условие $u_x + \alpha u = \varphi$.

Рекуррентное соотношение получается после исключения u_m^{n+1} из второго уравнения (11):

$$u_{m+1}^{n+1} - P_m \Pi_{m+1/2} - R_m - (h/\varkappa_{m+1/2})\Pi_{m+1/2} = 0,$$

т.е.

$$u_{m+1}^{n+1} = (P_m - h/\varkappa_{m+1/2})\Pi_{m+1/2} + R_m.$$

В этом выражении исключим $\Pi_{m+1/2}$ через $\Pi_{m+3/2}$ и u_{m+1}^{n+1}, используя первое уравнение (11):

$$\Pi_{m+1/2} = \Pi_{m+3/2} - \frac{h}{\tau} u_{m+1}^{n+1} + \frac{h}{\tau} u_{m+1}^{n} - hQ_{m+1}.$$

Разрешая полученное соотношение относительно u_{m+1}^{n+1}, имеем

$$u_{m+1}^{n+1} = P_{m+1}\Pi_{m+3/2} + R_{m+1},$$

где

$$P_{m+1} = \frac{1}{A}\left(P_m + \frac{h}{\varkappa_{m+1/2}}\right),$$

$$R_{m+1} = \frac{1}{A}\left[\left(P_m + \frac{h}{\varkappa_{m+1/2}}\right)\left(\frac{h}{\tau} u_{m+1}^{n} - hQ_{m+1}\right) + R_m\right],$$

$$A = 1 + \frac{h}{\tau}\left(P_m + \frac{h}{\varkappa_{m+1/2}}\right).$$

Это и есть формулы потоковой прогонки.

Получив значения P_{M-1}, R_{M-1} и разрешив правые краевые условия, т.е. определив u_M^{n+1}, обратную прогонку реализуем, вычисляя поочередно $\Pi_{m-1/2}$, u_{M-1}^{n+1}, $\Pi_{M-3/2}$ и т.д. Стандартный шаг имеет вид

$$\Pi_{m-1/2} = \Pi_{m+1/2} - \frac{h}{\tau} u_m^{n+1} + \frac{h}{\tau} u_m^{n} - hQ_m,$$

$$u_{m-1}^{n+1} = P_{m-1}\Pi_{m-1/2} + R_m.$$

Механизм преодоления трудностей, связанных с конечной разрядностью машинных чисел, тот же, что был указан в § 5: переход от уравнения высокого порядка к системе уравнений первого порядка.

Возможен и другой способ расчета области с очень большим коэффициентом теплопроводности, позволяющий обойтись стандартной прогонкой. Нужно лишь скорректировать формулу для $\varkappa$:

$$\varkappa(u) = \{u^k \text{ при } u < u^*, \ (u^*)^k \text{ при } u \geqslant u^*\}.$$

Значение u^* зависит от разрядности машинных чисел. Вычислительный эксперимент показывает, что точное значение $\varkappa(u)$ не сущест-

венно, важно только то, что это очень большая величина. Конечно, значение u_x от u^* зависит сильно, но физически существенная величина — тепловой поток $\varkappa u_x$ — при росте u^* быстро выходит на предельное, асимптотическое, значение.

§ 22. Реализация разностной схемы для уравнений газовой динамики с теплопроводностью

При создании алгоритма численного интегрирования уравнений газовой динамики возникает необходимость решения большого числа относительно «мелких» непринципиальных вопросов, относящихся, так сказать, к вычислительной технологии. Однако квалифицированное их решение существенным образом влияет на успех дела. В этом параграфе на примере одной конкретной схемы мы постараемся выделить эти вопросы и покажем, на каком уровне они решаются. Это, в основном, — уровень качественных соображений, теоретических исследований упрощенных моделей и, конечно же, проверка принятых решений математическим экспериментом.

Удобным примером представляется схема, разработанная в 1953—1954 гг. авторским коллективом под руководством И. М. Гельфанда (это, видимо, была одна из первых схем подобного рода). Выбор этой схемы оправдан еще и тем, что ее реализация затрагивает достаточно полный набор наиболее важных моментов.

Математическая постановка задачи. *Область расчета.* Решение ищется в прямоугольной области $0 \leqslant x \leqslant X$, $0 \leqslant t \leqslant t^*$, где x — массовая лагранжева переменная, t — время, т.е. рассчитываются события, происходящие в выделенном объеме вещества. Интервал $[0, X]$ разбит на части точками $0 = X_0 < X_1 < \ldots < X_I = X$, причем каждый из интервалов $[X_i, X_{i+1}]$ заполнен газом того или иного сорта. Другими словами, уже геометрия задачи определяет наличие некоторых контактных разрывов (в процессе решения могут появиться и другие).

Искомые функции. Расчет состоит в определении функций, описывающих состояние газа: u, r, p, v, e, T (они имеют физический смысл скорости, эйлеровой координаты, давления, удельного объема, удельной внутренней энергии и температуры). Из этих функций основными являются u, r, v, T. Функции p, e связаны с v, T уравнением состояния, которое имеет свою форму в каждом веществе (т.е. на каждом из интервалов $[X_i, X_{i+1}]$. Для простоты и определенности можно иметь в виду идеальный газ, параметры которого различны для разных газов, хотя программы в таких ситуациях

обычно пишутся в терминах заданных функций $P_i(T, v)$, $E_i(T, v)$, где i — номер вещества. Начальные данные при $t = 0$ задаются значениями $u(0, x)$, $r(0, x)$, $v(0, x)$, $T(0, x)$.

Уравнения. Схема строится на основе уравнений газовой динамики в массовых лагранжевых координатах с добавлением теплопроводности и искусственной вязкости:

$$u_t + (p+q)_x = 0, \quad r_t = u, \quad v_t - u_x = 0,$$
$$(e + u^2/2)_t + [(p+q)u]_x = [\varkappa_i(T, v)\, T_x]_x, \tag{1}$$

где $q = (\varepsilon/v)u_x(u_x - |u_x|)$ — вязкость Неймана, $\varkappa(T, v)$ — заданный коэффициент теплопроводности. Для дальнейшего существенна его следующая форма, явно выделяющая степенную зависимость $\varkappa$ от T: $\varkappa(T, v) = T^\alpha a(T, v)$, где $\alpha > 1$, $a(T, v)$ — гладкая функция, ограниченная неравенствами $0 < a^- \leqslant a(T, v) \leqslant a^+$. Разумеется, параметры и вид функции a зависят от вещества. Ради простоты мы ограничимся «плоским» вариантом задачи, когда r не входит явно в уравнения.

Краевые условия. Они могут иметь различную форму. Ради определенности ограничимся такими: при $x = 0$ заданы скорость $u(t, 0) = \tilde{u}(t)$ и поток энергии $\varkappa T_x = Q(t)$; при $x = X$ заданы температура $T(t, X)$ и давление $p(t, X)$.

Основные особенности решений. Сложность приближенного решения дифференциальных уравнений определяется прежде всего свойствами гладкости искомых функций. Ниже имеются в виду задачи, решения которых были кусочно-гладкими функциями. Точнее, область счета некоторыми линиями разбивалась на большое число подобластей, в каждой из которых решение было достаточно гладким. Число этих линий и их форма не задавались заранее, они определялись в процессе решения. Линии, на которых нарушалась гладкость решения, являются хорошо известными особенностями решений уравнений газовой динамики и нелинейной теплопроводности. Это ударные волны, границы волн разрежения (линии разрыва производных), фронты тепловых волн и фиксированные в лагранжевых координатах линии разрыва плотности и формул уравнений состояния.

Особенно сложный характер имеет течение в окрестностях точек пересечения линий нарушения гладкости (т.е., например, прохождение ударных и тепловых волн через контактные границы X_i, сопровождающиеся «рождением» новых линий нарушения гладкости). Определенные трудности возникают тогда, когда разные подобласти состоят из существенно разных веществ, например если подобласти из очень тяжелых веществ разделяются значительной по эйлеровым размерам областью из очень легкого вещества, имеющей ничтожный

в массовых координатах размер (такую область условно назовем «вакуумом»).

Сложным является, например, прохождение ударной волны через вакуум. При выходе ударной волны на внутреннюю границу вакуума, она исчезает, сменяясь волной разрежения. Одновременно начинается быстрое движение этой границы вакуума в сторону другой его границы. В какой-то момент эти границы встречаются, происходит «удар», снова рождающий ударную волну.

Интересный класс течений создается в следующей ситуации. На границу холодного покоящегося газа подается мощный поток энергии (задается либо большой поток на одной из границ, либо высокая температура). Возникает характерная картина — прогрев газа в режиме тепловой волны, фронт которой движется с конечной скоростью. Расчет такого режима затруднен тем, что температура существенно негладкая около точки фронта (см. § 21).

Градиент температуры порождает градиент давления, и в дальнейшем возможно образование ударной волны, причем могут осуществиться два разных предельных режима: либо ударная волна обгоняет тепловую, либо ударная волна отстает от тепловой и распространяется как изотермическая по сильно нагретому веществу. Такого рода процессы протекают при облучении сферических мишеней мощным потоком лазерного излучения.

Разностная аппроксимация задачи. Введем основные объекты, появляющиеся при конструировании метода приближенного решения.

Сетка и счетные величины. Интервал $[0, X]$ покрывается сеткой $\{x_m\}_{m=0}^{M}$, сетка $\{t_n\}$ формируется в процессе решения, так как шаг по времени $\tau_{n+1/2}$ выбирается в зависимости от полученного на n-м временном шаге решения. Узлы сетки с координатами $\{n, m\}$ образуют множество «целых» счетных точек. В них определены «механические» величины u_m^n и r_m^n. Кроме того, вводятся «полуцелые» счетные точки с координатами $x_{m+1/2} = (x_m + x_{m+1})/2$, t_n. В этих «полуцелых» точках определены «термодинамические» величины $\{T_{m+1/2}^n, v_{m+1/2}^n\}$ $(m = 0, 1, \ldots, M)$. Границы подобластей X_i совпадают с какими-то из «целых» точек x_m.

Особое положение занимают граничные точки. В точках $(0, n)$ (левая граница), кроме механических величин, могут быть определены некоторые термодинамические, используемые для реализации краевого условия и для нестандартной аппроксимации некоторых уравнений. В точках $(M + 1/2, n)$ (правая граница) могут быть определены механические величины $u_{M+1/2}^n$, $r_{M+1/2}^n$, используемые в тех же целях. Мы таких нестандартных счетных точек использовать не будем (нужда в них появляется, например, при иных краевых условиях).

Разностная аппроксимация в стандартных точках. Сначала опишем стандартные формулы разностной аппроксимации, т.е. те, в которых не используются «термодинамические» величины в граничных узлах. Введем следующие обозначения: $h_{m+1/2} = x_{m+1} - x_m$, $h_m = x_{m+1/2} - x_{m-1/2}$ — шаги сетки. Численное интегрирование проводим по стандартной для эволюционных задач схеме счета «по слоям». Шаг интегрирования состоит в том, что значения на n-м слое $(u, r, T, v)^n$ уже известны и надо вычислить $(n+1)$-й слой $(u, r, T, v)^{n+1}$, решая систему уравнений на верхнем слое.

Ради простоты рассмотрим полностью неявную схему, хотя можно использовать и схему, в которой пространственные производные аппроксимируются взвешенной (обычно с весами 0.55 и 0.45) суммой аппроксимаций на верхнем и нижнем слоях. В любом случае приходим к системе нелинейных уравнений относительно неизвестных величин на верхнем слое, которая решается специальным итерационным алгоритмом. Итерации строятся на основе неполного метода Ньютона.

В уравнения на верхнем слое входят близкие величины трех видов. Поясним это на примере $T_{m+1/2}$ (то же самое относится и к $v_{m+1/2}$, u_m). Мы имеем дело с $T^n_{m+1/2}$, $T^{(i)}_{m+1/2}$, $T^{(i+1)}_{m+1/2}$. Величины $T^n_{m+1/2}$ (с нижнего слоя) считаются уже известными. При вычислении $T^{n+1}_{m+1/2}$ методом итераций фигурируют уже найденные значения $T^{(i)}_{m+1/2}$ (i-е приближение к $T^{n+1}_{m+1/2}$) и неизвестные значения $T^{(i+1)}_{m+1/2}$. В пределе величины $T^{(i)}_{m+1/2} \to T^{n+1}_{m+1/2}$. Мы используем обозначения $\widetilde{T}_{m+1/2} = T^{(i)}_{m+1/2}$, $T_{m+1/2} = T^{(i+1)}_{m+1/2}$. Именно по отношению к неизвестным $T_{m+1/2}$ производится линеаризация при выполнении очередной итерации.

Аппроксимация уравнений движения:

$$\frac{u^{n+1}_m - u^n_m}{\tau} + \frac{(p+q)^{n+1}_{m+1/2} - (p+q)^{n+1}_{m-1/2}}{h_m} = 0 \qquad (2)$$

(стандартная аппроксимация применима при $m = 1, 2, \ldots, M-1$),

$$\frac{r^{n+1}_m - r^n_m}{\tau} = \frac{u^{n+1}_m + u^n_m}{2} \qquad (3)$$

(формула применима при всех $m = 0, 1, \ldots, M$).

Уравнения для удельного объема:

$$\frac{v^{n+1}_{m+1/2} - v^n_{m+1/2}}{\tau} - \frac{u^{n+1}_{m+1} - u^{n+1}_m}{h_{m+1/2}} = 0 \qquad (4)$$

(формула имеет смысл при $m = 0, 1, \ldots, M-1$).

Уравнение для энергии:

$$\frac{1}{\tau}\left\{e^{n+1}_{m+1/2}+\frac{(u^{n+1}_m)^2+(u^{n+1}_{m+1})^2}{4}-e^n_{m+1/2}-\frac{(u^n_m)^2+(u^n_{m+1})^2}{4}\right\}+$$

$$+\frac{1}{h_{m+1/2}}\left\{(p+q)^{n+1}_{m+1}\,u^{n+1}_{m+1}-(p+q)^{n+1}_m\,u^{n+1}_m\right\}=$$

$$=\frac{1}{h_{m+1/2}}\left\{\widetilde{\varkappa}_{m+1}\frac{T^{n+1}_{m+3/2}-T^{n+1}_{m+1/2}}{h_{m+1}}-\widetilde{\varkappa}_m\frac{T^{n+1}_{m+1/2}-T^{n+1}_{m-1/2}}{h_m}\right\} \qquad (5)$$

(уравнение имеет смысл при $m=1, 2, \ldots, M-1$).

К уравнениям (2)—(5) следует добавить выражения для величин $q_{m+1/2}$, p_m, $\varkappa_m$, не входящих в число основных счетных величин.

а) Неймановская вязкость $q_{m+1/2}$ имеет вид ($m=0, 1, \ldots, M-1$)

$$q_{m+1/2}=(\varepsilon/v_{m+1/2})\left\{(\tilde{u}_{m+1}-\tilde{u}_m)-|\tilde{u}_{m+1}-\tilde{u}_m|\right\}(u_{m+1}-u_m).$$

При этом $q_{M+1/2}=0$ («фиктивное» граничное условие).

б) Интерполяция p_m осуществляется по естественной формуле

$$p_m=\frac{h_{m-1/2}p_{m+1/2}+h_{m+1/2}p_{m-1/2}}{h_{m-1/2}+h_{m+1/2}}.$$

Это — линейная интерполяция, имеющая формально погрешность аппроксимации $O(h^2)$, т.е. минимальную, обеспечивающую при численном дифференцировании погрешность $O(h)$. Она была введена на заключительном этапе эксплуатации программы. Первоначально использовалась рассчитанная на равномерную сетку формула $p_m=(p_{m-1/2}+p_{m+1/2})/2$. В дальнейшем сетка стала неравномерной, а формула осталась, что привело к определенным трудностям, о которых будет сказано подробнее ниже. Заметим, что формально ошибочная формула с полусуммой рекомендуется и сейчас. Это приводит к существенным ограничениям сетки x_m: она должна быть «почти равномерной», т.е. $h_{m+1/2}=h_{m-1/2}(1+O(h))$.

в) Коэффициент теплопроводности $\varkappa_m$ вычисляется по формуле, в которой совмещаются идеи линейной и «гармонической» интерполяции (см. § 21).

Обсуждение разностной схемы. Приведем соображения, на основании которых были выбраны вышеприведенные формулы аппроксимации.

Два типа узлов. Разделение счетных точек на целые (механические) и полуцелые (термодинамические) является очевидным следствием различного вхождения соответствующих величин в уравнения. В каждое уравнение входят производные по времени от вели-

чин одного сорта и производные по x от величин другого сорта. Это кажется нарушенным для уравнения энергии, но выполняется и для него, если использовать эквивалентную недивергентную форму

$$e_t + (p+q)u_x = (\varkappa T_x)_x. \tag{6}$$

В дальнейшем мы заменим приведенную выше аппроксимацию уравнения для энергии (5) на эквивалентную, но столь же компактную, как и очевидная аппроксимация уравнения в форме (6).

Используемая сетка позволяет при минимальном шаблоне получить (на равномерной сетке) второй порядок аппроксимации по x, а при аппроксимации пространственных производных комбинацией аппроксимаций по верхнему и нижнему слоям с весами 0.55 и 0.45, например, можно получить «почти второй» порядок аппроксимации по τ. При весах 0.5 и 0.5 порядок был бы вторым, однако схема стала бы (по спектральному признаку) нейтральной, т.е. точки спектра, соответствующие параметру $\varphi = \pi$, оказываются на единичной окружности и высокочастотные паразитические возмущения хотя и не нарастают катастрофически, но и не затухают.

Кроме того, формальный разностный порядок производных совпадает с формальным их дифференциальным порядком. В связи с этим схема не требует дополнительных краевых условий (или «односторонних» разностных аппроксимаций в ближайших к краям счетных точках), необходимых при превышении разностным порядком схемы истинного порядка дифференциальных выражений.

Дивергентность схемы. Все разностные уравнения имеют так называемый дивергентный вид, т.е. они могут быть записаны в следующей форме (для термодинамических величин и скорости уравнение имеет ту же форму со сдвигом индекса m на 1/2):

$$\frac{P^{n+1}_{m+1/2} - P^{n}_{m+1/2}}{\tau} + \frac{Q^{n+1}_{m+1} - Q^{n+1}_{m}}{h_{m+1/2}} = 0,$$

где P и Q — функции от основных счетных величин. (Такой вид имеют разностные уравнения в предположении, что уравнения на верхнем слое решены точно. Фактически же они выполняются с точностью до погрешности итерационного процесса, обычно пренебрежимо малой.)

Следствием дивергентности схемы является выполнение разностных аналогов известной интегральной формы уравнений газовой динамики (см. § 20). Именно она является основой определения обобщенных решений, и это обстоятельство весьма существенно для расчетов, так как программа предназначалась в первую очередь для решения задач с разрывами.

Разностный аналог интегральных уравнений можно получить, суммируя разностные уравнения по прямоугольной (для простоты) области:

$$\sum_{m=M_1}^{M_2}\sum_{n=N_1}^{N_2-1}\left(\frac{P^{n+1}_{m+1/2}-P^{n}_{m+1/2}}{\tau}+\frac{Q^{n+1}_{m+1}-Q^{n+1}_{m}}{h_{m+1/2}}\right)\tau h_{m+1/2}=$$

$$=\sum_m P^{N_2}_{m+1/2}h_{m+1/2}-\sum_m P^{N_1}_{m+1/2}h_{m+1/2}+\sum_n Q^{n+1}_{M_2+1}\tau-\sum_n Q^{n+1}_{M_1}\tau=0. \tag{7}$$

Значение подобных соотношений обычно обосновывают ссылкой на законы сохранения, разностным аналогом которых они являются. Это, конечно, справедливо, но мы постараемся привести более четкие соображения.

Как уже отмечалось, наиболее важна интегральная форма уравнений для расчета разрывных решений. Рассмотрим (в весьма упрощенной и идеализированной форме) расчет изолированной ударной

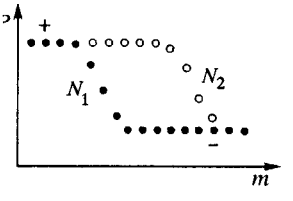


Рис. 35

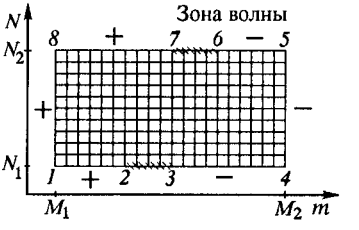


Рис. 36

волны. Существенным для таких расчетов является определение правильной скорости волны и правильных скачков величин при прохождении фронта волны. Пусть в расчетах получена следующая характерная картина: графики всех величин выглядят примерно так, как это изображено на рис. 35, где они показаны для моментов времени t_{N_1}, t_{N_2}.

Отметим основные свойства численного решения: есть некоторые значения u_1, v_1, e_1 перед фронтом волны, они постоянны (по m и n) и есть аналогичные величины u_2, v_2, e_2 за фронтом волны, они тоже постоянны. Наконец, есть «размазанная» волна, точечный график которой за время от t_{N_1} до t_{N_2} просто сместился на несколько узлов. Картина, конечно, идеализирована, но достаточно хорошо отражает свойства численного решения, которое получается по описываемой программе при расчете такого «чистого» течения, как движение первоначально покоящегося газа под действием равномерно движущегося поршня. Преобразуем соотношение (7), используя рис. 36.

Запишем разностное интегральное соотношение, учитывая, что на участках *12*, *87*, *18* величины $P^n_{m+1/2}$, Q^n_m имеют постоянные значения P_2, Q_2, а на участках *34*, *45*, *56* — значения P_1, Q_1. Обозначая $T = t_{N_2} - t_{N_1}$ и $l_{12}, l_{23}, \ldots$ — длины соответствующих участков контура, из (7) получаем

$$l_{87}P_2 + \sum_{76} P^{N_2}_{m+1/2} + l_{65}P_1 - l_{12}P_2 - \sum_{23} P^{N_1}_{m+1/2} - l_{34}P_1 + T(Q_1 - Q_2) = 0.$$

Предполагая, что графики в зонах волн при t_{N_1} и t_{N_2} одинаковы, т.е. $\sum_{23} = \sum_{76}$, и учитывая, что $l_{12} + l_{23} + l_{34} = l_{87} + l_{76} + l_{65}$, $l_{23} = l_{76}$, имеем

$$(L/T)(P_2 - P_1) - (Q_2 - Q_1) = 0,$$

где $L = l_{87} - l_{12} = l_{34} - l_{65}$ — расстояние, пройденное счетной ударной волной, а $D = L/T$ — ее счетная скорость.

Полученные соотношения суть соотношения Гюгонио, связывающие состояния до и после фронта волны и скорость волны. Подчеркнем, что этот результат является следствием не только дивергентной формы разностных уравнений, но и правильного характера численного решения. Последний факт устанавливается экспериментально. Поэтому возникающие при расчетах по некоторым схемам явно нефизические высокочастотные осцилляции с заметной, но не очень большой амплитудой подрывают доверие к расчету, хотя, конечно, они и не являются безусловным признаком его ошибочности. Стремление получить решения, свободные от такого «эстетического» недостатка входит в цели современной практики конструирования разностных схем.

Формула для кинетической энергии. При аппроксимации полной энергии $(e + u^2/2)_{m+1/2}$ используется линейная интерполяция u^2, хотя в принципе можно использовать величину $(u_m + u_{m+1})^2/4$. Это решение оправдывается следующими соображениями. Во-первых, квадрат полусуммы нечувствителен к возмущениям вида $(-1)^m$, т.е. сеточные функции u_m и $u_m + (-1)^m$ в этом случае имеют одну и ту же кинетическую энергию. Это дает основание ожидать появления в численном решении подобных возмущений, и эксперимент часто подтверждает эти ожидания. Вторая причина предпочесть именно интерполяцию квадрата скорости выяснится ниже.

Полностью консервативные схемы. Кроме классических и теоретически обоснованных обязательных качеств разностной схемы (аппроксимация и устойчивость), в современной практике вырабо-

тана система дополнительных требований, улучшающих численное решение в тех или иных ситуациях. В первую очередь это относится к ситуациям критического характера, когда искомое решение содержит нарушения гладкости, сосредоточенные в сравнительно узких зонах в плоскости (t, x), причем шаги сетки τ, h не могут быть взяты столь малыми, чтобы упомянутые узкие зоны разрешались достаточно большим числом счетных узлов сетки. Дивергентность схемы, роль которой мы выше разъяснили, — это одно из таких дополнительных свойств.

Перейдем к обсуждению другого свойства, получившего название «полная консервативность». Оно связано с некоторыми деталями аппроксимации уравнения для энергии. Начнем с критики построенной выше аппроксимации (5), основанной на прямом использовании уравнения в дивергентной форме. Уравнение (6) для e аппроксимируется проще и компактнее (выбросим пока теплопроводность, не в ней дело):

$$\frac{e^{n+1}_{m+1/2} - e^{n}_{m+1/2}}{\tau} + (p+q)^{n+1}_{m+1/2} \frac{u^{n+1}_{m+1} - u^{n+1}_{m}}{h_{m+1/2}} = 0. \qquad (8)$$

Это уравнение имеет определенное достоинство, хотя мы потеряли дивергентность. В чем же оно?

В различных руководствах по этому поводу можно встретить рассуждение такого типа. Дивергентное уравнение (5) «правильно» описывает эволюцию полной энергии $e + u^2/2$, однако распределение приращений отдельных ее видов (внутренней e и кинетической $u^2/2$) может оказаться нарушенным. Уравнение (8) «правильно» описывает приращение собственно внутренней энергии, но оно, к сожалению, недивергентно.

Из этого естественно возникает задача: построить такую аппроксимацию уравнения энергии, которая была бы и дивергентной, и «правильно» описывала бы изменение e. Такие схемы были построены, их называют полностью консервативными. Попробуем разобраться в этом вопросе. Не случайно слово «правильно» взято в кавычки. Оно не имеет сколько-нибудь определенного смысла. Требование полной консервативности уже вошло в практику конструирования разностных схем, и все более или менее однозначно понимают ее, хотя это свойство, видимо, не имеет четкого определения.

Можно дать такое объяснение. Уравнение для e получается (в дифференциальной форме) из уравнений для $e + u^2/2$ и u простой формальной выкладкой. Аналогичную выкладку можно проделать и для разностных аппроксимаций этих уравнений. Результат легко предвидеть: получается разностная аппроксимация уравнения для e типа (8), но с точностью до каких-то членов, пропорциональных τ, h (с точностью до «аппроксимационных источников»). Это есть очевидное следствие простого факта: разностное уравнение совпада-

ет с дифференциальным с точностью до погрешностей аппроксимации.

Схема называется полностью консервативной, если упомянутое формальное преобразование приводит к разностной аппроксимации уравнения для e, не содержащей упоминавшихся выше аппроксимационных источников и, следовательно, «правильно» описывающей эволюцию e во времени. Это почти определение, но мешает маленькая деталь. А что, собственно говоря, означают слова «уравнение содержит аппроксимационные источники или не содержит»? И почему, если содержит, это нехорошо?

Вычислим эти «источники» для описываемой схемы. Умножим (2) на $(u_m^{n+1} + u_m^n)/4$ и вычтем из (5). То же самое сделаем, используя (2) при значении $m+1$. После несложных преобразований получаем в качестве следствия формул (2), (5) разностную аппроксимацию уравнения (6) для e:

$$\frac{e_{m+1/2}^{n+1} - e_{m+1/2}^n}{\tau} + (p+q)_{m+1/2}^{n+1} \frac{u_{m+1}^{n+1} - u_m^{n+1}}{h_{m+1/2}} = \tau r_{m+1/2}^n, \qquad (9)$$

где

$$r_{m+1/2}^n = (1/4\tau^2)\,[(u_{m+1}^{n+1} - u_m^{n+1})^2 + (u_m^{n+1} - u_m^n)^2].$$

Это и есть пресловутый «аппроксимационный источник», превращающий «правильную» аппроксимацию (8) уравнения для e в якобы неправильную (9). Конечно, если бы кому-либо предложили на данной сетке аппроксимировать уравнение для e, едва ли кто-нибудь так сразу написал бы формулу (9), а формулу (8) написал бы всякий. С этой точки зрения аппроксимация (9) неестественна.

Но все это имеет смысл лишь при простейшей технике составления разностных схем (производные заменяются самыми простыми, наглядными разностными аппроксимациями). Однако в настоящее время по мере усложнения задач, вида уравнений, сеток и т.п. все чаще в практику входят гораздо более сложные методы составления уравнений, в том числе и чисто формальные, когда выбирается шаблон, пишется общая комбинация величин в узлах шаблона с неопределенными коэффициентами. Значения таких коэффициентов затем определяются требованиями аппроксимации, устойчивости и какими-то дополнительными требованиями, совокупность которых делает выбор схемы однозначным (это так называемый метод неопределенных коэффициентов в построении разностных схем).

При такой технике (а к ней приходится прибегать все чаще) в полученных выражениях не так-то просто выделить члены, относящиеся к тому или иному члену дифференциального уравнения. Поэтому понятие «аппроксимационный источник», казалось бы, очевидное в данном случае, в действительности особого смысла не имеет. Тем не менее некоторый смысл есть, и мы попробуем его сейчас выявить. Начнем с того, что предъявим более определенные

претензии к дивергентной форме аппроксимации (4). Все нижеследующее основано на опыте автора и его коллег, входивших в группу И. М. Гельфанда. Серийные расчеты задач такого типа, однако, продолжались в сущности только до 1960 г., поэтому наша точка зрения отражает опыт тех лет.

Явные дефекты аппроксимации уравнения для полной энергии проявляются обычно в зонах сильного разрежения, когда происходит интенсивная «перекачка» внутренней энергии в кинетическую и $u^2/2$ существенно превосходит e. В этой ситуации неизбежные погрешности приближенного метода решения могут привести к отрицательным значениям e. Это может произойти в одной-двух счетных точках, и этого можно было бы даже и не заметить, если бы не следующее крайне неприятное явление. В области $e < 0$ уравнения газовой динамики теряют физический смысл. С математической точки зрения они меняют тип, превращаясь из гиперболических в эллиптические: $e < 0$ означает «отрицательный квадрат скорости звука».

Уравнения распространения звука превращаются в систему уравнений Коши—Римана, для которой, как известно, задача Коши некорректна. Но программа и в этом случае продолжает решать задачу Коши (счет по слоям)! И вот эта некорректность, существующая сначала на очень небольшом участке (в двух-трех точках), начинает «разрушать» течение в соседних точках. Процесс приобретает катастрофический характер, и численное решение быстро теряет физический смысл. С этим иногда удается справляться, искусственно полагая $e^{n+1}_{m+1/2} = 0$, если расчет привел к отрицательному значению. Но это — скверный выход: он маскирует явные признаки неблагополучия, и расчет может продолжаться внешне благопристойно, потеряв, в сущности, точность. К таким мерам следует прибегать очень осторожно.

Чем же лучше в этом отношении недивергентная форма аппроксимации (8)? Дело в том, что p можно считать пропорциональным e. Следовательно, уравнение (8) можно записать в виде $e_t = Ae$ (где $A = -(p/e)u_x$). Решение этого уравнения не может перейти через ось $e = 0$. Это в дифференциальной форме очевидно. В разностной форме аналогичное свойство не гарантируется, но его можно обеспечить достаточно малым шагом τ. В самом деле, для явной и неявной схем имеем

$$(p^{n+1} - p^n)/\tau = Ap^n, \qquad p^{n+1} = (1 + \tau A)p^n,$$

$$(p^{n+1} - p^n)/\tau = Ap^{n+1}, \qquad p^{n+1} = p^n/(1 - \tau A).$$

Внимательный читатель заметит, что и v может изменить знак, что тоже приведет в нефизическую область. Здесь ситуация контролируется выбором шага:

$$v^{n+1}_{m+1/2} = v^n_{m+1/2} + (\tau/h)(u^{n+1}_{m+1} - u^{n+1}_m).$$

Конечно, шаг начинается лишь при известных величинах на n-м слое. Анализ этих данных позволяет выбрать шаг τ, учитывая, например, условие типа $\tau < 0.5h v^n_{m+1/2}/|u^n_{m+1} - u^n_m|$, $\forall m$. В большинстве случаев такой шаг τ обеспечивает положительность $v^{n+1}_{m+1/2}$. В противном случае переход с n-го слоя на $(n+1)$-й повторяется после уменьшения шага τ вдвое, и т.д. Заметим, что это не единственные критерии, по которым шаг τ ограничивается сверху. Итак, в расчетах по формуле (8) у нас есть средства, обеспечивающие положительность e.

Обратим внимание на то, что в описываемой схеме (а она, таким образом, не является полностью консервативной) положение с этой точки зрения еще более благоприятное, так как источники неотрицательны. Можно было бы предположить, что постоянный знак источников приведет к систематическому завышению значения внутренней энергии. Может быть это и так, но тут все-таки нужна более основательная аргументация. В самом деле, по сравнению с чем будет это систематическое завышение? Ведь даже утверждать, что завышение будет по сравнению с расчетом по схеме (8), не содержащей источников, нельзя.

Если же сравнить с точным решением, то и тут ситуация далеко неоднозначная. Подстановка в разностные уравнения точного решения дает хорошо известный нам результат. Точное решение уравнений газовой динамики (точнее, его ограничение на сетку) удовлетворяет разностным уравнениям с «источниками» в правой части (эти источники — погрешность аппроксимации)! Если бы мы знали эти источники, то включив их явно в правую часть схемы, мы получили бы точное совпадение разностного и точного решений. Так что сам по себе факт наличия «источников аппроксимационного типа» не является безоговорочным дефектом разностной схемы.

Аккуратное определение «полностью консервативной» схемы, должно учитывать следующее. Аппроксимацию (8), не содержащую источников, можно записать в виде (обозначая $\pi = p + q$)

$$\frac{e^{n+1}_{m+1/2} - e^n_{m+1/2}}{\tau} + 0.5\left(\pi^{n+1}_{m+1/2} + \pi^n_{m+1/2}\right)\frac{u^{n+1}_{m+1} - u^{n+1}_m}{h} + \\ + 0.5\left(\pi^{n+1}_{m+1/2} - \pi^n_{m+1/2}\right)\frac{u^{n+1}_{m+1} - u^{n+1}_m}{h} = 0.$$

Последний член можно трактовать как типичный аппроксимационный источник, имеющий (формально) величину $O(\tau)$.

Другой пример. Имеется аппроксимация уравнения для e вида $(e^{n+1}_{m+1/2} - e^n_{m+1/2})/\tau + A = 0$, где A — некоторая аппроксимация члена pu_x, содержащая источники. Пусть схема $(e^{n+1}_{m+1/2} - e^n_{m+1/2})/\tau + B = 0$ таких источников не содержит. Запишем «плохую» схему в виде

$(e^{n+1}_{m+1/2} - e^n_{m+1/2})/\tau + \gamma B = 0$, где $\gamma = A/B = 1 + O(\tau, h)$. Почему нельзя считать это уравнение хорошей аппроксимацией, не содержащей источников?

Уравнения на верхнем слое. Перейдем к аккуратному выписыванию уравнений, решая которые можно определить величины u, v, T, r. Напомним, что в разностные уравнения входят уже известные величины (с n-го слоя и с $(i-1)$-й итерации). Искомые величины на i-й итерации мы условились обозначать без верхнего индекса. Для разработки метода решения уравнений на верхнем слое нам прежде всего важна структура уравнений. Поэтому займемся именно структурой, т.е. выяснением того, какие именно неизвестные входят в то или иное уравнение (напомним, что и тех, и других очень много).

Начнем с уравнения (4) для v. Оно явно разрешается относительно $v_{m+1/2}$, и мы в дальнейшем будем использовать формулу

$$v_{m+1/2} = V_{m+1/2}(u_m, u_{m+1}), \quad m = 0, 1, \ldots, M-1. \qquad (10)$$

При этом мы будем описывать уравнения именно в такой форме, указывая явно только неизвестные величины; наличие известных величин мы будем отмечать индексом $m + 1/2$. Конечно, конкретные формы зависимостей должны быть однозначно и безошибочно запрограммированы, но в данный момент мы этого технического вопроса не рассматриваем. В (10) мы получили не все «уравнения для v». Величина $v_{M+1/2}$ пока не имеет «своего» уравнения. Таковым является краевое условие. Мы приняли заданными значения $T_{M+1/2}$, $p_{M+1/2}$. Уравнение состояния позволяет вычислить и $v_{M+1/2}$.

Рассмотрим уравнения (2) для u_m. Они записываются для $m = 1, 2, \ldots, M-1$, т.е. в системе пока не хватает двух уравнений. Представим эти уравнения в общей форме:

$$\tilde{U}_m(u_{m-1}, u_m, u_{m+1}, p_{m+1/2}, p_{m-1/2}) = 0$$

(величины u_{m-1}, u_{m+1} вошли в $\tilde{U}$ через формулы для $q_{m-1/2}$, $q_{m+1/2}$). Это пока предварительная формула.

Уравнение состояния позволяет исключить $p_{m-1/2}$ через $T_{m-1/2}$, и $v_{m-1/2}$, которое, в свою очередь, исключается через u_{m-1}, u_m (по формуле (10)). Аналогичным образом $p_{m+1/2}$ исключается через $T_{m+1/2}$, u_m, u_{m+1}. Легко проверить (и это нужно сделать обязательно), что при $m = 1$, $M-1$ мы не выходим за пределы действия формулы (10).

Заметим еще, что упомянутое «исключение» не следует трактовать буквально как подстановку в конкретные формулы вместо аргументов соответствующих, часто громоздких, формул. Современ-

ная техника программирования позволяет оперировать с описаниями таких сложных функциональных зависимостей в виде суперпозиции относительно простых.

Итак, запишем стандартные уравнения для u_m в виде

$$U_m(u_{m-1}, u_m, u_{m+1}, T_{m-1/2}, T_{m+1/2}) = 0, \quad m = 1, 2, \ldots, M-1.$$

Уравнение для заданного u_0 представим в общей форме, позволяющей использовать схему вычислений и для иных краевых условий, например $U_0(u_0, u_1, T_{1/2}) = 0$.

Наконец, уравнение для u_M можно получить, полагая $q_{M+1/2} = 0$. Это по существу есть дополнительное краевое условие. Необходимость в нем возникает из-за введения искусственной вязкости. Она не является физически обоснованным фактором, но проводит к повышению порядка дифференцирования по x (член q_x, грубо говоря, аналогичен члену u_{xx}). Поэтому первичная (физическая) постановка задачи не содержит требуемого краевого условия и оно вводится искусственно. Конечно, эта «произвольная» операция требует осторожности: она не должна оказывать заметного влияния на численное решение. В данном случае, поскольку q есть величина $O(h^2)$, условие $q_{M+1/2} = 0$ достаточно естественно.

Таким образом, уравнение (2) для u_M можно представить в виде

$$\tilde{U}_M(u_{M-1}, u_M, p_{M-1/2}, p_{M+1/2}) = 0.$$

Учитывая, что $p_{M+1/2}$ задано краевым условием, запишем его в окончательной форме:

$$U_M(u_{M-1}, u_M, T_{M-1/2}, T_{M+1/2}) = 0$$

($T_{M+1/2}$ добавлено «для общности»). Теперь все значения u_m обеспечены «своими» уравнениями.

Перейдем к уравнениям для T:

$$\frac{e_{m+1/2} - e^n_{m+1/2}}{\tau} + (\tilde{p} + \tilde{q})_{m+1/2} \frac{u_{m+1} - u_m}{h_{m+1/2}} - \tilde{r}_{m+1/2} =$$

$$= \frac{1}{h_{m+1/2}} \left[\tilde{\varkappa}_{m+1} \frac{T_{m+3/2} - T_{m+1/2}}{h_{m+1}} - \tilde{\varkappa}_m \frac{T_{m+1/2} - T_{m-1/2}}{h_m} \right].$$

Эти уравнения можно использовать при $m = 1, 2, \ldots, M-1$, т.е. в системе пока не хватает двух уравнений. Исключая $e_{m+1/2}$ через $T_{m+1/2}$, $v_{m+1/2}$ (уравнение состояния) и $v_{m+1/2}$ через u_m, u_{m+1}, придадим уравнениям форму

$$E_{m+1/2}(T_{m-1/2}, T_{m+1/2}, T_{m+3/2}, u_m, u_{m+1}) = 0.$$

Уравнение для $T_{1/2}$ получаем, используя краевое условие (задан поток $\varkappa T_x = Q$ при $x = 0$):

$$\frac{e_{m+1/2} - e^n_{m+1/2}}{\tau} + (\tilde{p} + \tilde{q})_{1/2} \frac{u_1 - u_0}{h_{1/2}} = \frac{1}{h_{1/2}} \left[\tilde{\varkappa}_1 \frac{T_{3/2} - T_{1/2}}{h_1} - Q \right] + \tilde{r}_{1/2}.$$

Это уравнение можно представить в виде

$$E_{1/2}(T_{1/2}, T_{3/2}, u_0, u_1) = 0$$

(u_0 включено тоже «для общности»). Включив в уравнение для $e_{m+1/2}$ аппроксимационный источник, $r_{m+1/2}$ (вычисляемый по значениям u_m, а не $\tilde{u}_m$), получим уравнение $E_{m+1/2} = 0$ точно такой же структуры.

Уравнение для $T_{M+1/2}$ в рассматриваемом случае тривиально — эта величина просто задана. Запишем это уравнение в общей форме, имея в виду и более сложные краевые условия:

$$E_{M+1/2}(T_{M-1/2}, T_{M+1/2}, u_M) = 0.$$

Подведем итог, выписав все уравнения, которые предстоит решать:

$$\begin{gathered} U_0(u_0, u_1, T_{1/2}) = 0, \\ U_m(u_{m-1}, u_m, u_{m+1}, T_{m-1/2}, T_{m+1/2}) = 0, \\ U_M(u_{M-1}, u_M, T_{M-1/2}, T_{M+1/2}) = 0; \end{gathered} \tag{11}$$

$$\begin{gathered} E_{1/2}(T_{1/2}, T_{3/2}, u_0, u_1) = 0, \\ E_{m+1/2}(T_{m-1/2}, T_{m+1/2}, T_{m+3/2}, u_m, u_{m+1}) = 0, \\ E_{M+1/2}(T_{M-1/2}, T_{M+1/2}, u_M) = 0, \end{gathered} \tag{12}$$

где $m = 1, 2, \ldots, M - 1$. Перейдем к алгоритму их решения.

Метод раздельной прогонки. В этом методе сначала величины T фиксируются как $\tilde{T}$ (т.е. как уже найденные приближения к T^{n+1}), затем решаются уравнения (11) относительно u (линеаризацией по Ньютону). В результате получается линейная система уравнений относительно u, имеющая ту же структуру, т.е. система с трехдиагональной матрицей. Она легко решается методом прогонки (см. § 10). Фиксируя $u^{(i+1)}$, можно линеаризовать вторую группу уравнений относительно T. Линейная система с такой же трехдиагональной матрицей решается прогонкой. Далее эти процедуры повторяются до достижения требуемой точности.

Метод векторной прогонки. В методике, которая описывается в этом параграфе, система уравнений на верхнем слое (11), (12) решалась методом векторной прогонки. (Раздельная прогонка была предложена позднее.) В методе векторной прогонки одновременно линеаризуются обе системы уравнений. Эта операция приводит к следующим линейным уравнениям:

$$A_0^0 u_0 + A_0^1 u_1 + A_0^{1/2} T_{1/2} = A_0,$$

$$B_{1/2}^0 T_{1/2} + B_{1/2}^1 T_{3/2} + B_{1/2}^{-1/2} u_0 + B_{1/2}^{1/2} u_1 = B_0;$$

$$A_m^{-1} u_{m-1} + A_m^0 u_m + A_m^1 u_{m+1} + A_m^{-1/2} T_{m-1/2} + A_m^{1/2} T_{m+1/2} = A_m,$$

$$B_{m+1/2}^{-1} T_{m-1/2} + B_{m+1/2}^0 T_{m+1/2} + B_{m+1/2}^1 T_{m+3/2} + \\ + B_{m+1/2}^{-1/2} u_m + B_{m+1/2}^{1/2} u_{m+1} = B_{m+1/2};$$

$$A_M^{-1} u_{M-1} + A_M^0 u_M + A_M^{-1/2} T_{M-1/2} + A_M^{1/2} T_{M+1/2} = A_M,$$

$$B_{M+1/2}^{-1} T_{M-1/2} + B_{M+1/2}^0 T_{M+1/2} + B_{M+1/2}^{-1/2} u_M = B_{M+1/2},$$

где $m = 1, 2, \ldots, M-1$.

Вводя вектор $z_m = \{u_m, T_{m+1/2}\}$, запишем эти уравнения в матричной форме ($m = 1, 2, \ldots, M-1$):

$$\mathscr{B}_0 z_0 + \mathscr{C}_0 z_1 = \mathscr{D}_0$$

$$\mathscr{A}_m z_{m-1} + \mathscr{B}_m z_m + \mathscr{C}_m z_{m+1} = \mathscr{D}_m, \qquad (13)$$

$$\mathscr{A}_M z_{M-1} + \mathscr{B}_M z_M = \mathscr{D}_M.$$

Здесь использованы обозначения

$$\mathscr{A}_m = \begin{pmatrix} A_m^{-1} & A_m^{-1/2} \\ 0 & B_{m+1/2}^{-1} \end{pmatrix}, \qquad \mathscr{B}_m = \begin{pmatrix} A_m^0 & A_m^{1/2} \\ B_{m+1/2}^{-1/2} & B_{m+1/2}^0 \end{pmatrix},$$

$$\mathscr{C}_m = \begin{pmatrix} A_m^1 & 0 \\ B_{m+1/2}^{1/2} & B_{m+1/2}^1 \end{pmatrix}, \qquad \mathscr{D}_m = \begin{pmatrix} A_m \\ B_{m+1/2} \end{pmatrix}.$$

Формулы вычисления элементов этих матриц через значения функций U, E и их производных очевидны, но громоздки. Нет необходимости их воспроизводить.

Система уравнений (13) имеет «трехдиагональную» форму и решается несложным обобщением алгоритма прогонки. Вывод формул алгоритма отличается от вывода, изложенного в § 10, только тем, что теперь мы работаем с матрицами (некоммутативная алгебра) и надо аккуратно следить за порядком множителей. Решение ищется в форме

$z_{m-1} = X_m z_m + Y_m$, где X_m — матрица 2×2, Y_m — вектор. Опуская простые выкладки, приведем результат ($m = 1, 2, \ldots, M-1$):

$$X_1 = -\mathscr{B}_0^{-1}\mathscr{C}_0, \qquad X_{m+1} = -(\mathscr{B}_m + \mathscr{A}_m X_m)^{-1}\mathscr{C}_m,$$

$$Y_1 = \mathscr{B}_0^{-1}\mathscr{D}_0, \qquad Y_{m+1} = (\mathscr{B}_m + \mathscr{A}_m X_m)^{-1}(\mathscr{D}_m - \mathscr{A}_m Y_m).$$

Теперь последнее уравнение (13) и прогоночное соотношение $z_{M-1} = X_M z_M + Y_M$ можно разрешить относительно z_M. Эта величина вычисляется и позволяет начать «обратную прогонку» — вычисление справа-налево искомых величин z_m.

Мы не будем здесь обсуждать проблем разрешимости всех встречающихся в алгоритме задач (существования обратных матриц) и сходимости итерационного процесса. Укажем лишь, что легко угадать тривиальный результат: при достаточно малом τ все обстоит благополучно. Это естественное следствие вырождения в пределе $\tau \to 0$ всех уравнений в тривиальные. Теоретические оценки того малого τ, начиная с которого гарантируется успех вычислений, в практике расчетов не используются — это привело бы к неоправданно заниженному шагу по времени. Однако сам факт зависимости, например, скорости сходимости итераций от τ (она тем выше, чем меньше τ) используется в режиме обратной связи. Считается, что требуемая точность должна достигаться за три-пять итераций.

Если итераций потребовалось больше, следующий шаг интегрирования уравнений выполняется с уменьшенным шагом τ. Если точность достигается за меньшее число итераций, τ увеличивается в пределах, определяемых другими критериями выбора шага. Что касается сопоставления скорости сходимости методов раздельной и векторной прогонок, то преимущество имеет последняя. Это естественно: оба метода являются комбинацией метода линеаризации (Ньютона) и метода простой итерации (Пикара).

Общая картина в таких алгоритмах такова, что метод оказывается тем быстрее сходящимся, чем больше в нем доля метода Ньютона. Однако итерация метода векторной прогонки требует больших вычислений. Вообще, следует отметить, что основное время работы ЭВМ связано не с прогонкой, а с вычислением коэффициентов системы (13). Одним из ресурсов экономии вычислительной работы является алгоритм с однократным вычислением этих коэффициентов; при этом в процессе итераций пересчитываются только D_m в (13).

Поясним это на промере решения нелинейного уравнения $f(x) = 0$. Упрощенный вариант метода Ньютона с однократным вычислением f_x имеет форму

$$x^{i+1} = x^i - f_x(x^0)\, f(x^i).$$

При достаточно хорошем начальном приближении x^0 он сходится «линейно», т.е. $\|f(x^i)\| \approx q^i$, где $q \approx \|E - f_x(x)\, f_x^{-1}(x^0)\|$ (x — решение системы). Эту оценку предоставим вывести читателю.

И наконец, подчеркнем, что приведенные выше формы организации решения уравнений на верхнем слое образуют некоторую общую схему, в рамках которой возможны различные варианты. Они появляются при различных способах отнесения тех или иных неизвестных к i-й итерации (по таким неизвестным проводится линеаризация) или к $(i-1)$-й, а в иных случаях и к n-му слою. Выбор того или иного варианта диктуется особенностями решаемой задачи и здесь, естественно, не конкретизируется.

Расчет ударной волны по недивергентной гибридной схеме. Сказанное выше может создать у читателя впечатление, что для правильного расчета ударных волн дивергентная форма разностных уравнений является существенным фактором. В общем это верно. Не следует только возводить это положение в ранг абсолютного, безусловного требования к используемым в расчетах схемам. Обсудим этот вопрос подробнее, опираясь на результаты вычислительного эксперимента, проведенного автором в 1962 г.

Рассмотрим численное решение задачи о распаде произвольного разрыва в начальных данных. Кроме трудностей расчета ударной волны, мы имеем проблему расчета контактного разрыва, так как задача решается в переменных Эйлера. Итак, уравнения имеют вид

$$u_t + uu_x + v\left(\frac{c^2+q}{v}\right)_x = 0,$$

$$v_t + uv_x - vu_x = 0,$$

$$c_t + uc_x + \frac{\gamma-1}{2}\frac{c^2+q}{v}\,u_x = 0.$$

В качестве основных термодинамических величин берутся удельный объем v и величина $c = \sqrt{pv}$ (которая с точностью до множителя совпадает с адиабатической скоростью звука; рассматривается идеальный газ с $\gamma = 5/3$). Вязкость берется в форме фон-Неймана: $q = (Lh)^2 u_x(u_x - |u_x|)$, где $L \approx 4 \div 5$.

Выбор «экзотической» переменной c объясняется просто. В зоне размазанной волны (см. § 20) переменная c ведет себя так же, как функции u и v, тогда как p и e фактически размазываются на вдвое меньшую длину. Поскольку профиль волны должен быть разрешен четырьмя-пятью счетными точками, при счете в терминах e или p пришлось бы увеличить L раза в два, что приводит к слишком большому размазыванию u и v. При решении задач в лагранжевых координатах этой проблемы нет, так как переменные e и p по x не дифференцируются, а шаг по времени по разным причинам таков,

что временное размазывание волны заметно больше пространственного (содержит больше шагов τ).

Начальные данные имеют вид

$$u_m^0,\ v_{m+1/2}^0,\ c_{m+1/2}^0 = \begin{cases} 2.0,\ 0.25,\ 1.688, & m \leqslant 0, \\ 0,\ 1.0,\ 0, & m > 0. \end{cases}$$

Можно найти точное решение задачи. Оно состоит из:

а) волны разрежения, левой и правой границей которой являются линии $x_1(t) = -0.89t$, $x_2(t) = 0.39t$;

б) контактного разрыва на линии $x_3(t) = 2.92t$;

в) ударной волны на линии $x_4 = 3.9t$.

При $x_2(t) \leqslant x < x_4(t)$ значения $u(t, x) = 2.92$, $p(t, x) = 11.40$. В этой области v рвется на контактном разрыве:

$$v(t, x) = 0.352 \quad \text{при } x_2(t) < x < x_3(t),$$
$$v(t, x) = 0.250 \quad \text{при } x_3(t) < x < x_4(t).$$

Мы имеем дело с так называемой «сильной ударной волной», идущей по «холодному газу». В этом случае скачок плотности при переходе через волну максимален (сжатие в $(\gamma + 1)/(\gamma - 1) \approx 4$ раза).

Используем явную схему:

$$\left[\frac{c^{n+1} - c^n}{\tau}\right]_{m+1/2} + \left[u\frac{\Delta c}{h}\right]_{m+1/2}^n + \frac{\gamma - 1}{2}\left[\frac{c^2 + q}{v}\right]_{m+1/2}^n \frac{u_{m+1}^n - u_m^n}{h} = 0,$$

$$\left[\frac{v^{n+1} - v^n}{\tau}\right]_{m+1/2} + \left[u\frac{\Delta v}{h}\right]_{m+1/2}^n - v_{m+1/2}^n \frac{u_{m+1}^n - u_m^n}{h} = 0,$$

$$\left[\frac{u^{n+1} - u^n}{\tau}\right]_m + \left[u\frac{\Delta u}{h}\right]_m^n + v_m^{n+1} \frac{\pi_{m+1/2} - \pi_{m-1/2}}{h} = 0.$$

Поясним некоторые обозначения: общие индексы вынесены за квадратные скобки; $u_{m+1/2} = 0.5(u_m + u_{m+1})$; $v_m = 0.5(v_{m-1/2} + v_{m+1/2})$; $\pi_{m+1/2}$ есть значение $(c^2 + q)/v$, вычисленное по очевидной разностной аппроксимации q, причем значения v и c берутся с $(n+1)$-го слоя (сначала эти величины находятся из двух первых уравнений, затем считается u^{n+1}); $[\Delta c]_{m+1/2} = c_{m+1/2} - c_{m-1/2}$ при $u_{m+1/2} > 0$ и $[\Delta c]_{m+1/2} = c_{m+3/2} - c_{m+1/2}$ при $u_{m+1/2} < 0$. Таким же образом («против потока») берутся и разности Δv, Δu.

Назовем вышеприведенную схему схемой I. Ее основной дефект — первый порядок аппроксимации конвективной производной $f_t + uf_x$. Эта величина (при $u > 0$) аппроксимируется разностью типа

$$\tau^{-1}[f(t_n + \tau, x_m) - f(t_n, x_m - u\tau)].$$

Значение $x_n - u\tau$ не попадает в узел сетки, поэтому в эту точку значение f интерполируется линейно по ближайшим узлам $(n, m-1)$ и (n, m). Можно заранее предвидеть (см. § 20), что схема I приводит к размазыванию контактного разрыва.

Уточним схему в этом месте, вычисляя $f(t_n, x_m - u\tau)$ квадратичной интерполяцией значений $f^n_{m-1}, f^n_m, f^n_{m+1}$. Это будет схема II. Можно и здесь предвидеть неприятности, связанные с нефизическими осцилляциями. Наконец, рассмотрим гибридную схему (схему

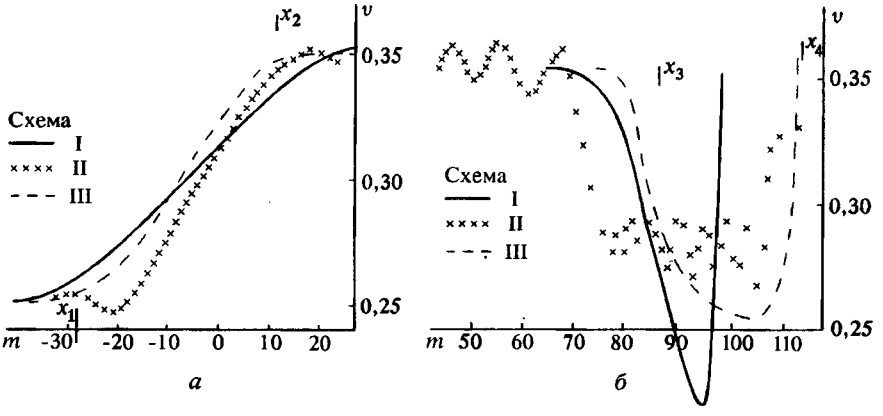


Рис. 37

III), в которой используется линейная или квадратичная интерполяция в зависимости от дифференциальных свойств решения в данной точке (см. § 20). Они характеризуются отношением второй и первой разностей, например $|f_{m-1} - 2f_m + f_{m+1}|/|f_m - f_{m-1}|$.

На рис. 37 показаны фрагменты численных решений, полученных по всем трем схемам. Они соответствуют моменту $t = 30$ (при $h = 1$), т.е. ударная волна прошла 117 счетных точек, контактный разрыв — 88 точек. Положения точных границ x_1, x_2, x_3, x_4 изображены на рис. 37. Обсудим результаты.

Схема I. Дефекты численного решения очевидны: сильно размазанный контактный разрыв, скорость ударной волны занижена примерно на 15 % (3.3 вместо 3.9), заметно размыты и слабые разрывы (границы волны разрежения).

Схема II. Повышение формального порядка аппроксимации привело к существенному ухудшению результатов: графики функций искажены сильными осцилляциями явно нефизического характера.

Схема III (гибридная). Существенное улучшение качества решения очевидно, хотя и не все дефекты численного решения ликвидированы. В частности, контактный разрыв размыт больше, чем хо-

телось бы. При $x_2 < x < x_4$ давление $p_{m+1/2} = 11.35 \pm 0.05$ (точное значение 11.40), $u_m = 2.93 \pm 0.01$ (точное значение 2.92).

В последние годы в вычислительной газовой динамике ведется активная работа по конструированию схем с улучшенными свойствами решений. Целью этой работы является получение таких схем, в которых контактные разрывы размываются как можно меньше и не проявляются нефизические осцилляции. Отличительной чертой таких конструкций является использование различных анализаторов локальной гладкости решения в каждой точке (n, m). В зависимости от показателя гладкости решения используется либо схема первого порядка аппроксимации, либо второго, либо некоторая промежуточная («гибрид» схем разного порядка аппроксимации). О качестве схемы судят по качеству решения задачи о распаде разрыва в начальных данных и других задач-тестов.

§ 23. Приближенное решение двумерных задач газовой динамики

Прикладные задачи газовой динамики, как правило, не допускают явных решений, поэтому важное значение имеют методы приближенного решения. В настоящее время ведется интенсивная разработка таких методов. Их создано уже достаточно много, тем не менее работа продолжается. Это объясняется тем, что одни и те же уравнения газовой динамики описывают (в зависимости от тех или иных краевых условий, значений входящих в уравнения физических постоянных) качественно разные явления. Они часто очень сложны, и эффективный метод решения должен учитывать характерные особенности подлежащего расчету явления. Именно стремлением учесть специфику явления при конструировании расчетной схемы определяется содержание научной работы в области численных методов газовой динамики.

Если полагаться на простейшие разностные схемы, мощность существующих ЭВМ окажется явно недостаточной для решения задач, которые выдвигаются современной техникой и достаточно успешно решаются специализированными методами. При изучении различных схем решения уравнений газовой динамики нужно прежде всего четко представлять себе, каков класс задач, в которых эффективен именно тот, а не другой из многих известных методов. Эту сторону вопроса мы постараемся разъяснить в процессе изложения.

Формулировка задачи газовой динамики. В дальнейшем мы будем иметь дело с так называемыми двумерными задачами, т.е. с задачами, в которых искомые функции зависят от времени t и двух пространственных координат x, y. Конечно, реальные задачи газовой

динамики трехмерны; мы ограничимся двумерными ради простоты изложения. Основные идеи построения методов можно объяснить уже в двумерном случае. Переход к трехмерному случаю вносит осложнения, в основном, технического характера (в то же время переход от одномерного случая к двумерному вносит ряд принципиальных осложнений). Другая причина состоит в том, что большая часть современных расчетов в газовой динамике — пока что двумерные; освоение массовых трехмерных расчетов по существу только начинается.

Итак, мы имеем дело с некоторой областью пространства D, разные части которой заполнены разными газами. Заметим, что термин «газ» не следует понимать слишком узко. В определенных условиях (при высоких температурах) металлы ведут себя, как газы, и описываются теми же уравнениями газовой динамики. Короче, мы имеем дело со сплошной средой, состояние которой описывается следующими функциями: компоненты скорости $u(t, x, y)$ и $v(t, x, y)$, плотность $\rho(t, x, y)$, давление $p(t, x, y)$, удельная внутренняя энергия $e(t, x, y)$. Они удовлетворяют уравнениям газовой динамики. Существуют разные формы этих уравнений, удобные в тех или иных ситуациях. Начнем с уравнений в форме Эйлера:

$$\begin{aligned}
&\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} + \frac{1}{\rho}\frac{\partial p}{\partial x} = 0,\\
&\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} + \frac{1}{\rho}\frac{\partial p}{\partial y} = 0,\\
&\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} + v\frac{\partial \rho}{\partial y} + \rho\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right) = 0,\\
&\frac{\partial e}{\partial t} + u\frac{\partial e}{\partial x} + v\frac{\partial e}{\partial y} + \frac{p}{\rho}\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right) = 0.
\end{aligned} \tag{1}$$

Система (1) замыкается конечным соотношением — уравнением состояния, связывающим термодинамические характеристики среды p, ρ, e в каждой точке (t, x, y). Уравнение состояния используется в виде $e = E(p, \rho)$ или $p = P(e, \rho)$, где E, P — известные функции. Например, идеальный газ определяется соотношением $E(p, \rho) = (1/(\gamma - 1))p/\rho$, где γ — постоянная, характеризующая данную среду (разные газы имеют разные значения γ). Уравнение состояния может иметь и более сложную форму. Разумеется, уравнения (1) дополняются начальными данными, заданными функциями u, v, p, $\rho(0, x, y)$, и краевыми условиями на границе D. Эти вопросы мы пока не рассматриваем.

Обратим внимание на то, что во всех уравнениях присутствует характерный оператор $\partial/\partial t + u\partial/\partial x + v\partial/\partial y$. Он называется субстанциальной производной и обозначается d/dt в связи со следующей важной физической интерпретацией. Пусть фиксированная («окрашенная») частица газа в момент времени $t = 0$ находится в точке (X_0, Y_0). В по-

следующие моменты времени она будет находится в точках $(X(t), Y(t))$. Уравнения движения выделенной частицы суть

$$\dot{X} = u(t, X(t), Y(t)), \quad \dot{Y} = v(t, X(t), Y(t)). \tag{2}$$

Рассмотрим функцию $f(t, x, y)$. На данной траектории $(X(t), Y(t))$ она является функцией только от t: $f(t, X(t), Y(t))$. Вычислим ее производную по времени:

$$\frac{df}{dt} = f_t + f_x\dot{X} + f_y\dot{Y} = f_t + uf_x + vf_y.$$

Таким образом, субстанциальная производная — это производная по t вдоль траектории частицы.

Уравнения газовой динамики в дивергентной форме. Простыми преобразованиями уравнения (1) можно привести к важной в приложениях дивергентной форме:

$$\frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(\rho u^2 + p) + \frac{\partial}{\partial y}(\rho uv) = 0,$$

$$\frac{\partial}{\partial t}(\rho v) + \frac{\partial}{\partial x}(\rho uv) + \frac{\partial}{\partial y}(\rho v^2 + p) = 0, \tag{3}$$

$$\frac{\partial}{\partial t}\rho + \frac{\partial}{\partial x}(\rho u) + \frac{\partial}{\partial y}(\rho v) = 0,$$

$$\frac{\partial}{\partial t}\left[\rho\left(e + \frac{u^2+v^2}{2}\right)\right] + \frac{\partial}{\partial x}\left\{u\left[p + \rho\left(e + u\frac{u^2+v^2}{2}\right)\right]\right\} + \frac{\partial}{\partial y}\left\{v\left[p + \rho\left(e + \frac{u^2+v^2}{2}\right)\right]\right\} = 0.$$

Эти уравнения могут быть записаны в компактной форме:

$$\frac{\partial W}{\partial t} + \frac{\partial R}{\partial x} + \frac{\partial Q}{\partial y} = 0, \tag{4}$$

которая часто служит исходной при построении разностных методов, так как из нее непосредственно вытекают важные, имеющие фундаментальное физическое значение, соотношения (законы сохранения).

Интегрируя (4) по параллелепипеду $[t_0, t_1] \times [a, A] \times [b, B]$, получаем

$$\int_{t_0}^{t_1}\int_{a}^{A}\int_{b}^{B}(W_t + R_x + Q_y)\,dt\,dx\,dy =$$

$$= \int_{a}^{A}\int_{b}^{B} W\,dx\,dy\,\Big|_{t_0}^{t_1} + \int_{t_0}^{t_1}\int_{b}^{B} R\,dy\,dt\,\Big|_{a}^{A} + \int_{t_0}^{t_1}\int_{a}^{A} Q\,dt\,dx\,\Big|_{b}^{B} = 0. \tag{5}$$

Соотношение (5) имеет следующий смысл: $\int_a^A \int_b^B W\,dx\,dy$ — общее количество величины W (компоненты импульса, массы или полной энергии) в объеме $[a, A] \times [b, B]$, $\iint R\,dt\,dy$, $\iint Q\,dt\,dy$, — потоки за время $t_1 - t_0$ через границу этого объема. Таким образом, изменение в данном объеме количества W связано с перетеканием его через границу этого объема.

Решения уравнений газовой динамики нужно искать среди «обобщенных решений», т.е. среди функций, удовлетворяющих тождеству (5) для всех параллелепипедов. Обратим внимание на то, что проверка тождества (5) не требует дифференцирования функций u, v, ρ, p и может быть осуществлена даже при наличии разрывов в этих функциях. Действительно, решения газодинамических задач могут содержать поверхности, на которых рвутся функции u, v, ρ, p.

В двумерных задачах имеются те же два основных типа разрывов: ударные волны и контактные разрывы. Соотношения на разрывах имеют ту же форму, что и в одномерных задачах, если использовать систему координат, в которой поверхность разрыва ортогональна (в рассматриваемой точке) оси x, а u и v — проекции скорости на оси локальной системы координат. На контактном разрыве непрерывны p и u (нормальная к разрыву компонента скорости); ρ и v (касательная к разрыву компонента скорости) могут иметь произвольный разрыв. Если v рвется, разрыв называют тангенциальным (кстати, такое течение неустойчиво). На ударной волне u, ρ, p по разные стороны от разрыва связаны одномерными соотношениями Гюгонио. Касательная к разрыву компонента скорости v на ударной волне непрерывна. Однако в двумерных задачах линии разрывов в плоскости $t = \text{const}$ могут иметь угловые точки.

Уравнения газовой динамики в форме Лагранжа. Другая форма уравнений газовой динамики связана с точкой зрения Лагранжа. Она отличается от рассмотренной выше тем, что искомые функции u, v, ρ, e считаются не функциями декартовых координат t, x, y, а функциями лагранжевых переменных t, ξ, η, где t — то же время, что и в эйлеровой форме, а координаты ξ, η выбираются так, что они остаются постоянными вдоль каждой траектории системы (2).

Введем функции $X(t, \xi, \eta)$, $Y(t, \xi, \eta)$, являющиеся эйлеровыми координатами частицы (ξ, η). Они удовлетворяют уравнениям

$$\begin{aligned} X_t(t, \xi, \eta) &= u(t, X(t, \xi, \eta), Y(t, \xi, \eta)), \\ Y_t(t, \xi, \eta) &= v(t, X(t, \xi, \eta), Y(t, \xi, \eta)). \end{aligned} \tag{6}$$

К ним следует присоединить начальные данные. Обычно берут в качестве лагранжевых координат частицы ее декартовы координаты в момент времени $t = 0$, т.е.

$$X(0, \xi, \eta) = \xi, \quad Y(0, \xi, \eta) = \eta,$$

но возможны и другие способы. Разумеется, $u(t, x, y)$, $v(t, x, y)$ в (6) считаются известными решениями уравнений газовой динамики.

Перейдем к выводу уравнений газовой динамики в форме Лагранжа, используя уравнения в форме Эйлера. Пусть известна функция эйлеровых координат $f(t, x, y)$, а x, y известны как функции лагранжевых координат t, ξ, η. Тем самым мы имеем f как функцию лагранжевых координат. Именно эта операция превращает функции u, v, ρ, $e(t, x, y)$ (решение уравнений газовой динамики в эйлеровых координатах) в функции $\tilde{u}$, $\tilde{v}$, $\tilde{\rho}$, $\tilde{e}(t, \xi, \eta)$, которые естественно считать решениями уравнений в лагранжевых координатах. Итак,

$$\tilde{f}(t, \xi, \eta) \equiv f(t, X(t, \xi, \eta), Y(t, \xi, \eta)).$$

Вычислим производную этой функции по t:

$$\tilde{f}_t = f_t + f_x X_t + f_y Y_t = f_t + u f_x + v f_y.$$

Такие выражения (субстанциальные производные) входят во все уравнения газовой динамики, которые можно переписать в форме

$$\begin{gathered} \tilde{u}_t + \tilde{\rho}^{-1} \tilde{p}_x = 0, \quad \tilde{v}_t + \tilde{\rho}^{-1} \tilde{p}_y = 0, \\ \tilde{\rho}_t + \tilde{\rho}(\tilde{u}_x + \tilde{v}_y) = 0, \quad \tilde{e}_t + \tilde{\rho}^{-1}\, \tilde{p}(\tilde{u}_x + \tilde{v}_y) = 0. \end{gathered} \tag{7}$$

Уравнения (7) содержат производные по x, y, а не по ξ, η, как хотелось бы, чтобы иметь замкнутую систему уравнений в переменных t, ξ, η. Система уравнений (7) дополняется уравнениями (6) для X и Y.

Теперь осталось выписать выражения для $\tilde{p}_x$, $\tilde{p}_y$, $\tilde{u}_x$, $\tilde{v}_y$ через производные $\tilde{p}$, $\tilde{u}$, $\tilde{v}$ по ξ, η. Продифференцируем $\tilde{f}$ по ξ, η:

$$\tilde{f}_\xi = f_x X_\xi + f_y Y_\xi, \quad \tilde{f}_\eta = f_x X_\eta + f_y Y_\eta.$$

Эту систему мы рассматриваем как систему линейных алгебраических уравнений относительно неизвестных f_x, f_y. Решая ее, получаем

$$\begin{aligned} f_x &= (\tilde{f}_\xi Y_\eta - \tilde{f}_\eta Y_\xi)/(X_\xi Y_\eta - X_\eta Y_\xi), \\ f_y &= (\tilde{f}_\eta X_\xi - \tilde{f}_\xi X_\eta)/(X_\xi Y_\eta - X_\eta Y_\xi). \end{aligned} \tag{8}$$

Формулы (8) определяют правила вычисления входящих в (7) производных по x, y через производные по ξ, η. Таким образом, уравнения в форме Лагранжа — это совокупность уравнений (6), (7) и формул (8).

Задачи, в которых удобны координаты Эйлера. Рассмотрим характерную прикладную задачу, в которой удобна и естественна эйлерова форма уравнений. Это — важная в разных областях прикладной аэродинамики задача обтекания. Пусть имеется некоторое тело, обтекаемое потоком газа. Нас интересует картина течения газа около тела и значения основных газодинамических переменных, так как ими определяются такие характеристики, как сопротивление, подъемная сила, температура, давление на поверхности тела и т.п.

Систему координат обычно выбирают связанную с телом. В этой задаче интересующие нас события разворачиваются в некоторой фиксированной в геометрическом пространстве области (рис. 38). Лагранжево представление здесь явно неудобно. Если мы выделим некоторую область в лагранжевых координатах, то она вместе с потоком газа пройдет мимо тела, удалится от него, и что в ней будет происходить, уже не очень интересно.

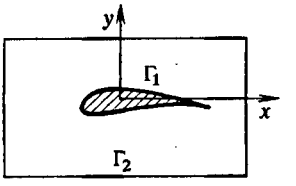


Рис. 38

Кроме задач, связанных с расчетом, например, аэродинамических характеристик крыльев, к этому классу относятся задачи расчета течений в соплах, задачи внешней баллистики, в том числе задачи о спуске космических кораблей, и т.п. Отметим, что в этих задачах есть проблема постановки краевых условий. Граница расчетной области состоит из двух частей. Первая часть границы есть граница тела Γ_1. Это — естественная граница, и на ней ставится физически очевидное условие непротекания: нормальная компонента скорости потока равна нулю, т.е. $un_x + vn_y = 0$, где n_x, n_y — вектор нормали к границе Γ_1.

Вторая часть границы Γ_2 вводится искусственно. По существу задача ставится в неограниченной плоскости, но реализация расчетных схем неизбежно требует ограничить область. Никаких «естественных», точных граничных условий на Γ_2 нет. Вычислители стараются отнести границу Γ_2 подальше от тела, чтобы искусственные граничные условия мало влияли на картину течения вблизи тела. Этот факт контролируется численными методами. Решив задачу один раз, повторяют расчет, отодвинув границу. Если основные интересующие нас характеристики изменились не очень сильно, считают их достаточно достоверными, несмотря на искусственность математической задачи.

Задачи, в которых удобны координаты Лагранжа. Типичный пример такой задачи — задача, связанная с проблемой лазерного термояда. Напомним в общих чертах суть дела. Сферическая мишень, состоящая из нескольких сферических слоев, выполненных

из разных веществ, подвергается мощному кратковременному облучению со всех сторон (рис. 39). На поверхности мишени быстро создается высокая температура и, следовательно, высокое давление, сжимающее мишень. Процесс носит сложный характер. Высокое давление на границе порождает тепловую и, возможно, ударную волны, сходящиеся к центру. В то же время поверхностные слои вещества начинают разлетаться от центра — идет так называемая волна разрежения. Будет ли в результате достигнут желаемый результат (создание в центре «термоядерных параметров», т.е. некоторой области с очень высокой температурой и достаточной плотностью), — на этот вопрос должен дать ответ расчет.

Для нас сейчас важны следующие обстоятельства. Рассматриваемая среда состоит из нескольких областей, в которых физические свойства газа существенно различаются. Рассчитываемый процесс сопровождается сильной деформацией первоначального расположения границ. Большая часть вещества сжимается в очень узкую зону.

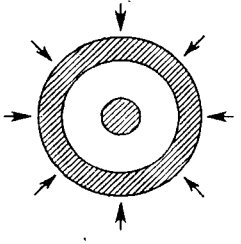


Рис. 39

Если бы мы пытались решать задачу методом конечных разностей в эйлеровых координатах, мы покрыли бы область, первоначально занимаемую газом, какой-то сеткой. В начале процесса в разных зонах имеется достаточно большое число счетных ячеек, что обеспечивает нужную точность разностной аппроксимации. По мере развития явления ситуация меняется. Почти все вещество сосредотачивается в очень узкой области, в которую попадает небольшое число ячеек сетки, и точность расчета, естественно, становится недопустимо низкой. Если же расчет ведется в лагранжевых координатах, узлы счетной сетки движутся вместе с веществом и число ячеек сетки в каждой зоне остается неизменным, как бы ни сжимались сами области.

Задачи, в которых неудобны как эйлеровы, так и лагранжевы координаты. Эйлеровы координаты оказываются неудобными в таких задачах, где рассматривается среда, состоящая из областей, заполненных веществами с разными физическими свойствами. Если в процессе течения границы таких областей передвигаются на заметные расстояния, т.е. контактная граница проходит последовательно через много счетных ячеек, происходит чисто вычислительное «размазывание» границы. Если не вводить в расчет эту границу явно (в виде отдельного математического объекта), трудно указать, где имеется вещество одного типа, а где — другого.

Приведем примеры содержательных задач, в которых нас как раз интересует достаточно точная картина эволюции контактных границ и в которых эти границы заметно перемещаются в пространстве.

Задача о волнах на поверхности. Рассмотрим течение, возникающее вследствие неустойчивости тангенциального разрыва. Пусть при $t=0$ линия $y=0$ является линией разрыва в начальных данных: при $y<0$ заданы постоянные значения $v=u=0$, p и ρ_1, при $y>0$ — значения $v=0$, $u>0$, p и $\rho_2 \ll \rho_1$ (при $y<0$ — покоящаяся вода, при $y>0$ — воздух с горизонтальным «ветром»). Такое течение является стационарным решением уравнений газовой динамики («чистый» тангенциальный разрыв). Но если поверхность раздела сред немного возмутить, разовьется сложное течение с сильной деформацией поверхности раздела. Лагранжевы координаты здесь неудобны: отображение $(t, x, y) \rightleftarrows (t, \xi, \eta)$ разрывно.

Задача о дифракции сильной ударной волны. Рассмотрим течение, особенности которого поясняет рис 40. По Г-образному каналу, заполненному газом, движется очень сильная ударная волна. В некоторый момент она выходит на границу твердого тела (заштрихованного на рис. 40*а*), возникает сложное течение, основными объектами которого являются прошедшая и отраженная ударные волны в газе, ударная волна в твердом теле. При этом возникает существенное искажение первоначальной контактной границы, связанное с течением типа мощной струи (рис. 40*б*).

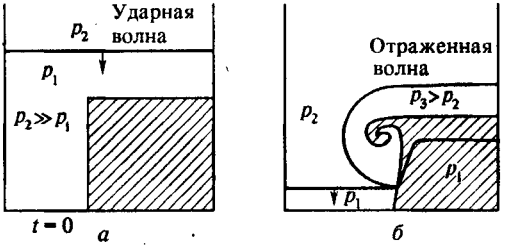


Рис. 40

Расчет подобных течений в координатах Эйлера затруднен тем, что определяющую роль в развитии явления играет именно форма поверхности, разделяющей разные газы. В эйлеровой системе эта поверхность «теряется». Трудности расчета в лагранжевых переменных связаны с существенными искажениями первоначальной геометрии лагранжевой сетки при расчете течений с сильными деформациями. Дело в том, что в таких течениях происходит, так сказать, перемешивание вещества. Частицы газа, бывшие в начале процесса близкими друг к другу, с течением времени расходятся на большие расстояния. Наоборот, далекие вначале частицы газа могут сблизиться. Физически на движение частицы оказывают влияние лишь те частицы газа, которые в данный момент непосредственно примыкают к ней. (Этот факт связан с тем, что уравнения газовой динамики — это дифференциальные уравнения в частных производных.)

В лагранжевых координатах близкими всегда считаются те частицы, которые были близки в начале процесса. Это приводит к тому, что разностные формулы в лагранжевых координатах с течени-

ем времени теряют точность до такой степени, что расчет не заслуживает никакого доверия. Первоначально прямоугольные ячейки постепенно деформируются, теряют даже форму параллелограммов, а в особо сложных случаях часто наблюдается «выворачивание» лагранжевой ячейки.

PIC-метод (метод частиц в ячейке). PIC-метод (Particle-In-Cell) был предложен и разработан Ф. Харлоу (Лос-Аламос) в 1955 г. и является типично американским. Этот метод требует большого объема оперативной памяти и большого числа арифметических операций, поэтому в нашей стране он нашел применение лишь в последние годы — после ввода в эксплуатацию мощных машин серии ЕС. Теоретическое обоснование метода, видимо, до настоящего времени еще не получило законченной формы, а первоначально метод имел чисто прагматическое оправдание. С его помощью были проведены расчеты очень сложных течений.

Хотя точность проведенных расчетов признается не всеми и большинство вычислителей согласны с тем, что она не очень высока, качественная картина, полученная в расчетах, выглядит убедительной и правдоподобной. А если учесть, что попытки проведения подобных расчетов другими методами приводят обычно к явно недостоверным результатам, легко понять популярность PIC-метода. Характерная ситуация, в которой PIC-метод демонстрирует свои преимущества перед иными, — это течение среды, первоначально разделенной на простые по форме области, заполненные разными веществами. В процессе течения эти области сильно деформируются и перемещаются в пространстве.

Для расчета подобных течений неудобны как эйлерова сетка (происходят большие перемещения контактных границ в геометрическом пространстве), так и лагранжева (происходит сильная деформация первоначальной лагранжевой сетки). PIC-метод — это попытка совместить достоинства эйлерова и лагранжева описаний течений сплошной среды.

Перейдем к описанию вычислительной схемы метода.

Эйлерова сетка. Область течения покрыта неподвижной в пространстве сеткой. Для простоты будем считать шаг сетки h постоянным, одинаковым по x, y. Ячейки сетки занумерованы парами индексов i, j. Величины, которые в дальнейшем помечаются индексами i, j, трактуются либо как относящиеся к ячейке в целом, либо (при разностной аппроксимации уравнений) как относящиеся к центру ячейки.

Основные счетные величины. В центрах счетных ячеек определены величины $u^n_{i,j}$, $v^n_{i,j}$ — компоненты скорости среды (индекс n показывает их принадлежность ко времени t_n). Совокупность чисел $t_0, t_1, \ldots, t_n, \ldots$ образует временную сетку, вообще говоря неравно-

мерную; $t_{n+1} = t_n + \tau_{n+1/2}$ (шаг $\tau_{n+1/2}$ выбирается на основе соображений точности и устойчивости в зависимости от состояния среды в момент t_n). Кроме того, в ячейках (i, j) определены величины $\{E^n_{\alpha, i, j}, M^n_{\alpha, i, j}\}$ $(\alpha = 1, 2, \ldots, A)$.

Поясним их смысл. Иногда ради удобства мы будем опускать индексы i, j, но не следует забывать, что эти величины — свои в каждой ячейке. Напомним, что в задаче изучается течение в области, заполненной в разных частях веществами с разными физическими свойствами (α — номер вещества). Вообще говоря, в данной ячейке может быть либо одно вещество, либо несколько разных. В каждой ячейке (i, j) в данный момент времени t_n индекс α пробегает свой (зависящий от i, j, n) набор значений. Проще будет считать, что индекс α пробегает все допустимые в данной задаче значения. Однако если вещества с номером α в данной ячейке нет, то соответствующие значения $E^n_\alpha = M^n_\alpha = 0$. Физический смысл этих величин таков: $E^n_{\alpha, i, j}$ — удельная внутренняя энергия вещества с номером α, в момент времени t_n находящегося в ячейке (i, j); $M^n_{\alpha, i, j}$ — масса этого вещества.

Кроме переменных u, v, E, M, в расчете участвует большое число «частиц». Будем нумеровать эти частицы индексом k. Число частиц должно быть много бо́льшим числа ячеек: на каждую ячейку в среднем должно приходится, как минимум, пять-десять частиц. Каждая k-я частица в момент времени t_n характеризуется следующими величинами: X^n_k, Y^n_k — координаты положения частицы; m_k — масса частицы (не зависящая от t); α_k — номер вещества, из которого «состоит» частица с номером k $(k = 1, 2, \ldots, K)$.

Это — основные счетные величины, полностью характеризующие (в принятой расчетной модели) состояние среды. Все остальные величины, которые появятся в дальнейшем, носят вспомогательный характер и выражаются через основные. Стандартный шаг интегрирования задачи состоит в переходе от величин $\{u, v, E_\alpha, M_\alpha\}^n_{i, j}$, $\{X, Y\}^n_k$ к величинам $\{u, v, E_\alpha, M_\alpha\}^{n+1}_{i, j}$, $\{X, Y\}^{n+1}_k$ (переход на следующий временной слой).

Расщепление уравнений газовой динамики. Математическая задача, которую предстоит решать, состоит в интегрировании уравнений газовой динамики, записанных в эйлеровой дивергентной форме (3). Обозначим плотность полной энергии $w = \rho\,[e + (u^2 + v^2)/2]$, где e — удельная внутренняя энергия, связанная с p и ρ уравнением состояния $p = P_\alpha(e, \rho)$, своим для каждого типа вещества.

PIC-метод применяется для расчета быстрых процессов, в которых диффузия не играет заметной роли. Поэтому перемешивания вещества не происходит. Считается, что на протяжении всего времени расчета сохраняются четкие границы, разделяющие разные вещества. Но форма этих границ претерпевает существенные изменения. Может измениться даже их «топология»: если при $t = 0$ какое-то вещество заполняло связную область, в дальнейшем оно может распасться на отдельные части, разделяемые веществом другого типа. Аккуратный расчет таких течений требует знания в каждый момент времени положения контактных границ. Это очень сложная вычислительная задача, и PIC-метод является попыткой решить ее относительно простыми средствами. Разумеется, эта простота оплачивается большим объемом памяти и машинных операций. Мы не будем обсуждать важных вопросов, связанных с граничными условиями, но скажем несколько слов о начальных данных.

Физическая постановка задачи обычно связана с заданием начальных данных в виде функций $u^0(x, y)$, $v^0(x, y)$, $\rho^0(x, y)$, $e^0(x, y)$ и границ, разделяющих разные вещества. В начальный момент времени эти границы, как правило, имеют простую геометрическую форму. Начальные данные для расчета $u^0_{i,j}$, $v^0_{i,j}$ связаны с $u^0(x, y)$, $v^0(x, y)$ — это просто значения в центрах ячеек. Если ячейка (i, j) целиком заполнена (допустим, веществом $\alpha = 1$), то E^0_1 в ней очевидным образом связано с $e^0(x, y)$ в центре. Остальные значения $E^0_\alpha = 0$. Если через ячейку проходит граница раздела двух веществ (т.е. в ней есть вещества двух типов), то и значения $E^0_{\alpha, i, j}$ задаются равными значениям в начальных данных (с точностью до шага сетки h). Напомним, что E — это удельная энергия вещества, а не количество энергии данного вещества в ячейке.

Начальные положения частиц (X^0_k, Y^0_k) задаются так, что, например, в каждую ячейку попадает равное число частиц (впрочем, это не обязательно, иногда полезно увеличить число частиц в тех ячейках, в которых ожидаются наиболее сложные события). Распределение m_k должно быть определенным образом согласовано с $\rho^0(x, y)$. Предполагается, что величина $M^n_{\alpha, i, j}$ согласована с положениями частиц, а именно:

$$M^n_{\alpha, i, j} = \sum m_k, \quad k: \{X^n_k, Y^n_k\} \in C_{i, j}, \quad \alpha = \alpha_k. \tag{9}$$

В дальнейшем нам часто придется иметь дело с суммами подобного рода.

Условный смысл суммы в (9) — это суммирование величин m_k для частиц с номером k, координаты которых (X^n_k, Y^n_k) в момент t_n

попали в ячейку (i, j) и которые имеют тип $\alpha_k = \alpha$. Если известно начальное распределение $\rho^0(x, y)$, то известна масса вещества в каждой ячейке. И если через ячейку проходит контактная граница, то известно, сколько вещества $M^0_{\alpha, i, j}$ каждого типа находится в данной ячейке (i, j). Задав положения частиц, нужно приписать им массы m_k так, чтобы выполнялось соотношение (9) при $n = 0$. Оно будет, как мы увидим, выполняться и в дальнейшем.

Уравнения (3) описывают изменения основных физических величин (масса, импульс, полная энергия) за счет процессов двух типов — работы сил давления и перетекания физических величин со скоростью потока (u, v). В PIC-методе переход за малое время τ от величин на временном слое n к величинам на слое $n + 1$ осуществляется в два этапа, на каждом из которых основные физические величины меняются за счет процесса только одного типа.

I. На первом этапе учитываются изменения основных величин только за счет работы сил давления (процессы перетекания пока исключены). Разностные формулы на этом этапе аппроксимируют (в привычном, наглядном смысле этого слова) следующие уравнения:

$$\begin{gathered} \rho_t = 0, \quad (\rho u)_t + p_x = 0, \\ (\rho v)_t + p_y = 0, \quad w_t + (pu)_x + (pv)_y = 0. \end{gathered} \tag{10}$$

II. На втором этапе используются уравнения, в которых, наоборот, оставлены только процессы перетекания:

$$\begin{gathered} \rho_t + (\rho u)_x + (\rho v)_y = 0, \quad (\rho u)_t + (\rho u u)_x + (\rho u v)_y = 0, \\ (\rho v)_t + (\rho u v)_x + (\rho v v)_y = 0, \quad w_t + (wu)_x + (wv)_y = 0. \end{gathered} \tag{11}$$

Разностная аппроксимация процессов переноса осуществляется с помощью частиц и имеет, как мы увидим, не очень привычный для метода конечных разностей характер. Скорее, здесь используются методы дискретного моделирования сплошной среды, апеллирующие к основным понятиям механики.

Численная реализация первого этапа. Исходная информация состоит в каждой ячейке (i, j) из величин u^n, v^n, M^n_α, E^n_α. Расчет начинается с вычисления давления $p^n_{i,j}$. Если в ячейке имеются частицы нескольких типов, то при расчете давления $p^n_{i,j}$ используются «физические» соображения о равенстве давлений на границе двух сред.

Введем величины σ_α — части объема h^2 ячейки, занимаемые веществом типа α. Очевидно, $\sum \sigma_\alpha = h^2$. Зная массу M^n_α вещества типа α, находим его плотность $\rho_\alpha = M^n_\alpha/\sigma_\alpha$, а зная его внутреннюю

энергию E_α^n вычисляем давление $p_\alpha = P_\alpha(E_\alpha^n, M_\alpha^n/\sigma)$. Приравнивая величины p_α друг другу, получаем систему уравнений типа

$$P_1(E_1^n, M_1^n/\sigma_1) = P_2(E_2^n, M_2^n/\sigma_2) = \ldots$$

Здесь число уравнений на единицу меньше числа веществ в ячейке.

Присоединяя к ним уравнение $\sum \sigma_\alpha = h^2$, приходим к полной системе уравнений относительно неизвестных σ_α. Она решается итерационным методом. Заметим, что в распространенном случае, когда уравнение состояния имеет вид $p = f_\alpha(e)\rho$, можно выписать явное решение. После того как σ_α найдены, определяется величина p, которую мы обозначим $p_{i,j}^n$, приписав ее центру ячейки. Теперь у нас есть все для того, чтобы рассчитать первый этап по стандартным разностным уравнениям, аппроксимирующим уравнения (10). Результатом будут величины $\tilde{u}_{i,j}$, $\tilde{v}_{i,j}$, $\tilde{M}_{\alpha,i,j}$, $\tilde{E}_{\alpha,i,j}$.

Уравнение $\rho_t = 0$ аппроксимируется просто:

$$\tilde{M}_{\alpha,i,j} = M_{\alpha,i,j}^n.$$

Определяем

$$p_{i+1/2,j}^n = 0.5(p_{i,j}^n + p_{i+1,j}^n), \qquad p_{i,j+1/2}^n = 0.5(p_{i,j}^n + p_{i,j+1}^n),$$

$$M_{i,j}^n = \sum_\alpha M_{\alpha,i,j}^n, \qquad \rho_{i,j}^n = M_{i,j}^n/h^2.$$

Так как $\rho_t = 0$, то используется аппроксимация

$$\rho_{i,j}^n \frac{\tilde{u}_{i,j} - u_{i,j}^n}{\tau} + \frac{1}{h}(p_{i+1/2,j}^n - p_{i-1/2,j}^n) = 0, \tag{12}$$

$$\rho_{i,j}^n \frac{\tilde{v}_{i,j} - v_{i,j}^n}{\tau} + \frac{1}{h}(p_{i,j+1/2}^n - p_{i,j-1/2}^n) = 0. \tag{13}$$

Из этих уравнений в явном виде находим $\tilde{u}_{i,j}$, $\tilde{v}_{i,j}$.

Уравнение изменения энергии аппроксимируется следующим образом:

$$\frac{\tilde{w}_{i,j} - w_{i,j}^n}{\tau} + \frac{1}{h}(p_{i+1/2,j}^n u_{i+1/2,j}^{n+1/2} - p_{i-1/2,j}^n u_{i-1/2,j}^{n+1/2}) + \\ + \frac{1}{h}(p_{i,j+1/2}^n v_{i,j+1/2}^{n+1/2} - p_{i,j-1/2}^n v_{i,j-1/2}^{n+1/2}) = 0. \tag{14}$$

Некоторые величины в этой формуле требуют пояснения:

$$u_{i+1/2,j}^{n+1/2} = (1/4)(\tilde{u}_{i,j} + u_{i,j}^n + \tilde{u}_{i+1,j} + u_{i+1,j}^n);$$

аналогично вычисляется $v_{i,j+1/2}^{n+1/2}$.

Сложнее обстоит дело с величиной $w^n_{i,j}$. Напомним, что плотность полной энергии $w = \rho\,[e + (u^2 + v^2)/2]$. Величина $h^2 w$ имеет смысл энергии в ячейке $h \times h$:

$$h^2 w^n_{i,j} = \left(h^2 \rho E + h^2 \rho \frac{u^2+v^2}{2}\right)^n_{i,j}.$$

Здесь $h^2\rho$ — масса ячейки, т.е. $M^n_{i,j} = \sum\limits_{\alpha} M^n_{\alpha,i,j}$. Итак,

$$\left(h^2 \rho \frac{u^2+v^2}{2}\right)^n_{i,j} = M^n_{i,j} \frac{(u^n_{i,j})^2 + (v^n_{i,j})^2}{2}$$

Вычислим $(h^2\rho E)^n_{i,j}$ — полную внутреннюю энергию в ячейке. Нам известны массы $M^n_{\alpha,i,j}$ и удельные внутренние энергии $E^n_{\alpha,i,j}$ веществ типа α. Естественно положить

$$(h^2\rho E)^n_{i,j} = \sum_{\alpha} M^n_{\alpha,i,j} E^n_{\alpha,i,j}.$$

Итак, вычислены $w^n_{i,j}$ и, следовательно, $\tilde{w}_{i,j}$ по формуле (14).

Теперь из выражения для полной энергии ячейки

$$h^2 \tilde{w}_{i,j} = h^2 \rho_{i,j} \tilde{E}_{i,j} + h^2 \rho_{i,j} \frac{(\tilde{u}_{i,j})^2 + (\tilde{v}_{i,j})^2}{2} \qquad (15)$$

можно найти величину $\tilde{E}_{i,j}$. Но это еще не все: ведь основными счетными величинами являются внутренние энергии $E_{\alpha,i,j}$ веществ разного типа. Введем изменения ΔE_α удельной внутренней энергии за шаг (точнее, за первый этап шага) по каждому веществу отдельно. Учитывая массу каждого вещества M_α, запишем полное приращение внутренней энергии в ячейке через ΔE_α и приравняем его известному нам полному приращению:

$$\sum_{\alpha} (M^n_{\alpha,i,j} \Delta E_{\alpha,i,j}) = M^n_{i,j} (\tilde{E}_{i,j} - E^n_{i,j}). \qquad (16)$$

Чтобы «поделить» полное приращение между разными веществами, нужно принять какие-то правдоподобные физические гипотезы. Например, можно считать, что все $\Delta E_{\alpha,i,j}$ одинаковы. Следовательно, $\Delta E_{\alpha,i,j} = \tilde{E}_{i,j} - E^n_{i,j}$. Теперь можно вычислить величины $\tilde{E}_{\alpha,i,j} = E^n_{\alpha,i,j} + \Delta E_{\alpha,i,j}$. Тем самым первый этап шага интегрирования завершен.

Численная реализация второго этапа. На втором этапе происходит учет перемещения частиц и величины с тильдой переходят в величины на $(n+1)$-м слое по времени.

1. Движение частиц. Прежде всего вычисляются новые положения частиц. Их координаты удовлетворяют уравнениям

$$\dot{X}_k = u(t, X_k, Y_k), \quad \dot{Y}_k = v(t, X_k, Y_k).$$

Расчеты ведутся по очевидным разностным аналогам этих уравнений:

$$X_k^{n+1} = X_k^n + \tau U_k, \quad Y_k^{n+1} = Y_k^n + \tau V_k.$$

Скорости U_k, V_k определяются некоторой интерполяцией величин $\tilde{u}$, $\tilde{v}$ в ячейках, окружающих точку (X_k^n, Y_k^n). Один из вопросов, который здесь возникает: почему используется интерполяция величин $\tilde{u}$, $\tilde{v}$, а не u^n, v^n или $(\tilde{u}+u^n)/2$, $(\tilde{v}+v^n)/2$? Ответ простой: использовавшие этот метод специалисты утверждают, что лучшие результаты дает именно интерполяция $\tilde{u}$, $\tilde{v}$, ссылаясь на опыт решения задач, в которых точность приближенного решения может быть проконтролирована.

Итак, мы знаем новые положения частиц X_k^{n+1}, Y_k^{n+1}. Теперь можно перейти к учету изменения основных физических величин за счет переноса.

2. Перенос массы и вычисление $M_{\alpha, i, j}^{n+1}$. Зная старые положения частиц X_k^n, Y_k^n и их новые положения X_k^{n+1}, Y_k^{n+1}, для каждой ячейки можно выделить три группы частиц.

а) Частицы, оставшиеся в пределах ячейки:

$$\{X_k^n, Y_k^n\} \in G_{i,j}, \quad \{X_k^{n+1}, Y_k^{n+1}\} \in C_{i,j}.$$

Эти частицы на данном шаге не вносят изменений в массу, импульс и энергию ячейки $C_{i,j}$.

б) Частицы, покинувшие ячейку $C_{i,j}$ и перешедшие в соседние:

$$\{X_k^n, Y_k^n\} \in C_{i,j}, \quad \{X_k^{n+1}, Y_k^{n+1}\} \notin C_{i,j}.$$

в) Частицы, пришедшие в $C_{i,j}$ из соседних ячеек:

$$\{X_k^n, Y_k^n\} \notin C_{i,j}, \quad \{X_k^{n+1}, Y_k^{n+1}\} \in C_{i,j}.$$

В расчете используется ограничение шага по t типа $\tau\sqrt{u^2+v^2} < h$, т.е. за один шаг частица может переместиться только в соседнюю ячейку. Каждая k-я частица, перешедшая на данном шаге из одной ячейки в соседнюю, переносит с собой свою массу m_k. Таким образом, величины $M_{\alpha, i, j}^{n+1}$ считаются по очевидным формулам: нужно

взять все частицы типа α, для которых $\{X_k^{n+1}, Y_k^{n+1}\} \in C_{i,j}$, и просуммировать их массы.

3. Перенос импульса. Промежуточное состояние (найденное на первом этапе расчета и отмеченное тильдой) характеризуется относящимися к каждой ячейке $C_{i,j}$ значениями импульса и полной энергии. Компоненты полного импульса могут быть вычислены по формулам $M_{i,j}^n \tilde{u}_{i,j}$, $M_{i,j}^n \tilde{v}_{i,j}$. Каждая k-я частица, покинувшая ячейку $C_{i,j}$, уносит с собой импульс $m_k \tilde{u}_{i,j}$, $m_k \tilde{v}_{i,j}$. (Впрочем, при желании импульсы можно вычислять и по формулам $m_k U_k$, $m_k V_k$.)

Вычислим изменение импульса в ячейке $C_{i,j}$ за один шаг:

$$\begin{aligned} M_{i,j}^{n+1} u_{i,j}^{n+1} &= M_{i,j}^n \tilde{u}_{i,j} - \sum\nolimits^1 m_k \tilde{u}_{i,j} + \sum\nolimits^2 m_k \tilde{u}_{i',j'}, \\ M_{i,j}^{n+1} v_{i,j}^{n+1} &= M_{i,j}^n \tilde{v}_{i,j} - \sum\nolimits^1 m_k \tilde{v}_{i,j} + \sum\nolimits^2 m_k \tilde{v}_{i',j'}. \end{aligned} \tag{17}$$

Здесь $\sum^1$ означает суммирование по k, соответствующим частицам, покинувшим ячейку $C_{i,j}$; $\sum^2$ соответствует частицам, пришедшим в $C_{i,j}$ из соседних $C_{i',j'}$. Из (17) вычисляются $u_{i,j}^{n+1}$, $v_{i,j}^{n+1}$, так как все остальные величины известны.

4. Перенос энергии. Если частица с номером k имеет тип α и переходит из одной ячейки $C_{i,j}$ в другую, она переносит с собой полную энергию

$$\Delta w_k = m_k \left(\tilde{E}_{\alpha,i,j} + \frac{(\tilde{u}_{i,j})^2 + (\tilde{v}_{i,j})^2}{2} \right).$$

Теперь можно вычислить полную энергию вещества типа α, находящегося в ячейке $C_{i,j}$. На промежуточном этапе эта величина есть

$$\tilde{w}_{\alpha,i,j} = \sum_{\alpha_k = \alpha} m_k \left(\tilde{E}_{\alpha,i,j} + \frac{(\tilde{u}_{i,j})^2 + (\tilde{v}_{i,j})^2}{2} \right)$$

(суммирование по всем частицам типа α, находящимся в момент t_n в ячейке $C_{i,j}$).

В момент времени t_{n+1} полная энергия изменится за счет переноса в соответствии с формулой

$$w_{\alpha,i,j}^{n+1} = \tilde{w}_{\alpha,i,j} - \sum_{\alpha_k = \alpha}^1 \Delta w_k + \sum_{\alpha_k = \alpha}^2 \Delta w_k,$$

где первая сумма берется по всем частицам, покинувшим ячейку $C_{i,j}$, а вторая — по всем частицам, пришедшим в ячейку $C_{i,j}$. Ра-

зумеется, учитываются только те частицы, у которых $\alpha_k = \alpha$. Завершается шаг вычислением величин $E^{n+1}_{\alpha, i, j}$ по формулам

$$E^{n+1}_{\alpha, i, j} = \frac{h^2 w^{n+1}_{\alpha, i, j}}{M^{n+1}_{\alpha, i, j}} - \frac{(u^{n+1}_{i, j})^2 + (v^{n+1}_{i, j})^2}{2}. \qquad (18)$$

Дивергентность PIC-метода. При конструировании разностных схем приближенного интегрирования уравнений газовой динамики, как правило, стремятся обеспечить дивергентность разностных уравнений. Другими словами, стараются получить дискретную модель среды, в которой выполняются простые и наглядные аналоги законов сохранения основных физических величин: массы, импульса и полной энергии. Их изменения внутри области (взаимодействия потоков с границами мы сейчас не рассматриваем) должны определяться только «перетеканием» из одной части пространства в другую. Не должно быть так называемых «разностных» источников (стоков) этих величин. Покажем, что PIC-метод удовлетворяет этим требованиям.

I. Сохранение массы. Проследим эволюцию массы в ячейке $C_{i, j}$ при переходе от момента времени t_n к t_{n+1}. Эта величина, как указывалось, вычисляется двумя способами, согласованными между собой:

$$M^n_{i, j} = \sum_{\alpha} M^n_{\alpha, i, j} = \sum m_k, \quad k: (X^n_k, Y^n_k) \in C_{i, j}, \quad \alpha = \alpha_k.$$

На первом этапе масса просто сохраняется: $\tilde{M}_{i, j} = M^n_{i, j}$.

На втором этапе изменение массы осуществляется за счет перемещения частиц. Можно ввести потоки

$$\Pi^{i', j'}_{\alpha, i, j} = -\sum\nolimits^{1} m_k + \sum\nolimits^{2} m_k,$$

где $\sum^1$ — сумма по тем k, для которых $((i', j') \neq (i, j))$

$$\alpha_k = \alpha, \quad (X^n_k, Y^n_k) \in C_{i, j}, \quad (X^{n+1}_k, Y^{n+1}_k) \in C_{i', j'};$$

$\sum^2$ — сумма по тем k, для которых

$$\alpha_k = \alpha, \quad (X^n_k, Y^n_k) \in C_{i', j'}, \quad (X^{n+1}_k, Y^{n+1}_k) \in C_{i, j}.$$

Итак, $\Pi^{i', j'}_{\alpha, i, j}$ есть количество вещества типа α, перенесенное потоком за время (t_n, t_{n+1}) в ячейку $C_{i, j}$ через «границу» между ячейками (i, j) и (i', j'). В терминах потоков изменение массы можно выразить так:

$$M^{n+1}_{\alpha, i, j} = M^n_{\alpha, i, j} + \sum_{l=-1}^{1} \sum_{n=-1}^{1} \Pi^{i+l, j+n}_{\alpha, i, j}.$$

Дивергентность этой формулы есть следствие очевидного соотношения $\Pi^{i',j'}_{\alpha,i,j} = -\Pi^{i,j}_{\alpha,i',j'}$. Таким образом, в расчетной схеме закон сохранения массы выполнен по всем веществам отдельно.

II. Сохранение импульса. Ограничимся анализом изменения только одной компоненты. На первом этапе импульс ячейки $C_{i,j}$ изменяется от $M^n_{i,j}u^n_{i,j}$ до $\tilde{M}^n_{i,j}\tilde{u}_{i,j}$ по формуле

$$\tilde{M}_{i,j}\tilde{u}_{i,j} = M^n_{i,j}u^n_{i,j} + \tilde{\Pi}^{i+1,j}_{i,j} + \tilde{\Pi}^{i-1,j}_{i,j}, \quad \tilde{\Pi}^{i\pm1,j}_{i,j} = \mp h\tau p_{i\pm1/2,j}.$$

Дивергентность связана с соотношением $\Pi^{i\pm1,j}_{i,j} = -\Pi^{i,j}_{i\pm1,j}$.

На втором этапе импульс изменяется по формуле

$$M^{n+1}_{i,j}u^{n+1}_{i,j} = \tilde{M}_{i,j}\tilde{u}_{i,j} + \sum_{l=-1}^{1}\sum_{m=-1}^{1}\Pi^{i+l,j+m}_{i,j}.$$

Выражения для потоков и свойство $\Pi^{i',j'}_{i,j} = -\Pi^{i,j}_{i',j'}$, предоставим вывести читателю (они в сущности очевидны). Полное изменение импульса за шаг есть

$$M^{n+1}_{i,j}u^{n+1}_{i,j} = M^n_{i,j}u^n_{i,j} + \sum_{l=-1}^{1}\sum_{n=-1}^{1}\Pi^{i+l,j+m}_{i,j}.$$

Здесь, конечно, значения двух потоков пересчитаны:

$$\Pi^{i\pm1,j}_{i,j} := \Pi^{i\pm1,j}_{i,j} + \tilde{\Pi}^{i\pm1,j}_{i,j}.$$

III. Сохранение энергии. Не будем выписывать потоков полной энергии из ячейки в ячейку и проверять их «кососимметричность». Это почти очевидно. В проверке нуждается дивергентность по времени. Напомним схему вычисления энергии. На первом этапе из величин $E^n_{\alpha,i,j}$ образуется полная энергия ячейки:

$$h^2w^n_{i,j} = \sum_{\alpha} M^n_{\alpha,i,j}E^n_{\alpha,i,j} + M^n_{i,j}\frac{(u^n_{i,j})^2 + (v^n_{i,j})^2}{2}. \qquad (19)$$

Для вычисления полной энергии используется дивергентная схема $h^2\tilde{w}_{i,j} = h^2w^n_{i,j} + \ldots$ Далее величина $\tilde{w}_{i,j}$ определяет значения $\tilde{E}_{\alpha,i,j}$.

На втором этапе из величин $\tilde{E}_{\alpha,i,j}$ образуются значения полной энергии вещества α в ячейке $\tilde{w}_{\alpha,i,j}$, и для каждого из них используется дивергентная схема $w^{n+1}_{\alpha,i,j} = \tilde{w}_{\alpha,i,j} + \ldots$ (потоков мы явно не выписываем, при последующих выкладках они остаются кососим-

метричными, дивергентность схемы по пространству очевидна). Суммируя по α, получаем дивергентную схему

$$\overline{w}^{n+1}_{i,j} = \overline{w}_{i,j} + \ldots, \qquad \overline{w}_{i,j} = \sum_{\alpha} \tilde{w}_{\alpha,i,j}, \qquad \overline{w}^{n+1}_{i,j} = \sum_{\alpha} w^{n+1}_{\alpha,i,j}.$$

Для того чтобы установить дивергентность по времени, нужно проверить равенство $\overline{\tilde{w}}_{i,j} = \tilde{w}_{i,j}$ и то, что $\overline{w}^{n+1}_{i,j}$ вычисляется по формуле типа (19). Проверим первое равенство, сравнивая выражения для $\tilde{w}_{i,j}$ и $\overline{w}_{i,j}$:

$$h^2\tilde{w}_{i,j} = \tilde{M}_{i,j}\tilde{E}_{i,j} + \tilde{M}_{i,j}\frac{(\tilde{u}_{i,j})^2 + (\tilde{v}_{i,j})^2}{2}, \tag{20}$$

$$h^2\overline{w}_{i,j} = \sum_{\alpha} \tilde{M}_{\alpha,i,j}\tilde{E}_{\alpha,i,j} + \tilde{M}_{i,j}\frac{(\tilde{u}_{i,j})^2 + (\tilde{v}_{i,j})^2}{2}. \tag{21}$$

Для величин $\tilde{E}_{\alpha,i,j}$ использована формула $\tilde{E}_{\alpha,i,j} = E^n_{\alpha,i,j} + \Delta_{i,j}$, где $\Delta_{i,j} = \tilde{E}_{i,j} - E^n_{i,j}$, а $E^n_{i,j}$ вычислялась через $w^n_{i,j}$ (см. формулу (19)), т.е. из соотношения $M^n_{i,j}E^n_{i,j} = \sum_{\alpha} M^n_{\alpha,i,j}E^n_{\alpha,i,j}$.

Вычислим входящую в (19) внутреннюю энергию:

$$\sum_{\alpha} \tilde{M}_{\alpha,i,j}\tilde{E}_{\alpha,i,j} = \sum_{\alpha} \tilde{M}_{\alpha,i,j}\{E^n_{\alpha,i,j} + (\tilde{E}_{i,j} - E^n_{i,j})\} =$$
$$= \sum_{\alpha} \tilde{M}_{\alpha,i,j}E^n_{\alpha,i,j} + \tilde{M}_{i,j}\tilde{E}_{i,j} - \tilde{M}_{i,j}E^n_{i,j} = \tilde{M}_{i,j}\tilde{E}_{i,j}.$$

Таким образом, установлено равенство $\tilde{w}_{i,j} = \overline{w}_{i,j}$.

Для того чтобы установить второе равенство, обратимся к формуле (18) для вычисления $E^{n+1}_{\alpha,i,j}$, переписав ее в виде

$$h^2 w^{n+1}_{\alpha,i,j} = M^{n+1}_{\alpha,i,j}E^{n+1}_{\alpha,i,j} + M^{n+1}_{\alpha,i,j}\frac{(u^{n+1}_{i,j})^2 + (v^{n+1}_{i,j})^2}{2}.$$

Суммируя по α, получаем

$$h^2 \sum_{\alpha} w^{n+1}_{\alpha,i,j} = h^2 w^{n+1}_{i,j} = \sum_{\alpha} M^{n+1}_{\alpha,i,j}E^{n+1}_{\alpha,i,j} + M^{n+1}_{i,j}\frac{(u^{n+1}_{i,j})^2 + (v^{n+1}_{i,j})^2}{2},$$

что совпадает с (19).

Основные недостатки PIC-метода. Укажем два момента, с которыми связана критика метода. Первый момент — дискретность плотности. В вышеописанной схеме, если имеется, допустим, в среднем по десять частиц на ячейку (это еще очень хорошо, часто их

меньше), плотность может принимать небольшое число дискретных значений. Особенно сильно это сказывается в областях разрежения, где плотность мала (по сравнению с первоначальной, например), т.е. на ячейку в среднем может приходиться одна-две частицы. В этом случае небольшое изменение положения частицы, находящейся близко около границы ячейки, приводит к ее переходу в другую ячейку. Плотность в соседних ячейках резко изменяется, что обычно приводит к соответствующему изменению давления. Возникает градиент давления, меняющий скорость, частица стремится вернуться назад и т.д. Так возникают колебания положений частицы, имеющие явно нефизический характер. Именно дефекты такого рода в первую очередь бросаются в глаза при анализе полученных PIC-методом приближенных решений.

Второй тонкий момент PIC-метода — реализация краевых условий. Здесь осложнения связаны с тем, что на каждом из этапов используется искаженная система уравнений. В частности, каждая из неполных систем уравнений имеет свою систему характеристик. Точнее, в двумерном случае надо говорить о характеристических конусах. Полная система уравнений имеет один вырожденный конус —линию с направлением $dt : dx : dy = 1 : u : v$ и наклонный звуковой конус, осью которого является вышеуказанная энтропийная характеристика, а раствором — скорость звука c. В соответствии с наклоном этих конусов относительно границы требуется на ней поставить то или иное число краевых условий. Неполные системы имеют иную картину характеристик. Так, система первого этапа (с исключенным переносом) имеет вертикальную энтропийную характеристику $dt : dx : dy = 1 : 0 : 0$ и звуковой конус вокруг нее. Система второго этапа имеет тройную вырожденную характеристику с направлением $dt : dx : dy = 1 : u : v$.

Таким образом, может оказаться, что на разных этапах стандартного шага по времени система дифференциальных уравнений на границе требует своего числа краевых условий, не совпадающего с тем, которое задано исходной постановкой задачи. Следует подчеркнуть, что реализация краевых условий — один из деликатных моментов схем расщепления, еще не получивший должной методологической разработки. В принципе, можно использовать процедуру исключения промежуточных (с тильдой) величин и получать разностные уравнения в терминах только величин n-го и $(n+1)$-го слоев. Можно ожидать, что это будет какая-то относительно стандартная схема, в которой можно будет так или иначе разобраться.

К сожалению, дело не так просто. Процедура исключения величин с тильдой приводит к «расползанию» шаблона. Аппроксимация входящих в уравнения газодинамики первых пространственных производных станет многоточечной, и такие разностные уравнения требуют значительно большего числа краевых условий, чем в исходной постановке задачи. Эти дополнительные краевые условия должны

быть определенным образом согласованы с уравнениями, чтобы не «подменить» настоящих краевых условий какими-то неявными. Вопрос еще более осложняется в приграничных узлах, когда начинают работать нестандартные аппроксимации. Реализация переноса с помощью частиц, разумеется, еще больше запутывает ситуацию. По этой причине мы воздержимся от изложения реализации краевых условий в PIC-методе. Здесь нет еще полной ясности, и тем, кто этим интересуется, придется обратиться к специальной литературе.

Выше мы нигде не включали в формулы искусственной вязкости. Вопрос о том, как обобщить, например, вязкость фон-Неймана, не так-то прост. Один из возможных рецептов состоит в том, что вязкость включается только на первом этапе, причем в членах p_x, $(pu)_x$ к p добавляется «вязкость по x», т.е. величина, пропорциональная $(u_x)^2$. Такой способ удобен тем, что u_x естественно вычисляется именно в нужных точках: $(u_x)_{i+1/2,\,j} = (u_{i+1,\,j} - u_{i,\,j})/h$. Точно так же в членах p_y, $(pv)_y$ добавляется «вязкость по y», пропорциональная $(v_y)^2$.

Корректнее добавить к p вязкую компоненту $q = \varepsilon\, d(d - |d|)$, где d — некоторая аппроксимация дивергенции скорости $u_x + v_y$. При использовании такой вязкости в уравнениях в форме Лагранжа вычислители сталкиваются с неприятным эффектом: вязкость равна нулю при деформациях счетных ячеек, сохраняющих площадь. Развитие таких деформаций приводит иногда к потере свойства выпуклости ячейки и к еще более неприятному «выворачиванию» ячейки, когда противоположные стороны квадратной в лагранжевых переменных ячейки в эйлеровых (т.е. в геометрических) переменных пересекаются.

Метод крупных частиц. Метод аппроксимации уравнений газовой динамики, описываемый ниже, часто трактуют как некоторое развитие PIC-метода, в котором исключены частицы и устранен один из главных дефектов — дискретность возможных значений плотности. Хотя, как будет показано, в методе крупных частиц действительно используется одна из существенных деталей PIC-метода — расщепление уравнений газовой динамики «по физическим процессам», суть дела все-таки в другом. Метод крупных частиц ориентирован совсем на другой класс газодинамических течений. Это в основном задачи обтекания тел потоком однородного газа, в которых нет проблемы контактных границ, сильно деформирующихся в процессе развития течения. Поэтому вся вычислительная схема носит иной характер: никаких частиц в ней нет. Схема строится достаточно традиционным способом с весьма прозрачными и наглядными рецептами замены производных конечными разностями. Слово «частицы» в названии метода отражает лишь историю возникновения расчетной схемы.

Уравнения. Исходной для аппроксимации выбирается эйлерова дивергентная форма уравнений газовой динамики (3). Система координат, естественно, связана с обтекаемым телом. На краевых условиях не останавливаемся (это — тема отдельного разговора).

Сетка. Область расчета (обычно, прямоугольник) покрывается равномерной (для простоты) сеткой, ячейки которой нумеруются парами индексов (i, j).

Счетные величины. Состояние среды описывается сеточными функциями $u^n_{i,j}$, $v^n_{i,j}$, $\rho^n_{i,j}$, $w^n_{i,j}$. Эти величины относятся к центрам ячеек и представляют собой приближенные значения компонент скорости u, v, плотности вещества ρ и плотности полной энергии $e + (u^2 + v^2)/2$. Уравнение состояния используется в виде $p = P(e, \rho)$, где $e = w - (u^2 + v^2)/2$. В дальнейшем мы будем использовать величины типа $p^n_{i,j}$, понимая под ними вспомогательные числа, полученные из уравнения состояния очевидным образом.

Шаги по x и y считаем, для простоты, равными и обозначаем h, шаг по времени — τ, хотя он, конечно, не фиксирован, а выбирается на каждом слое в зависимости от реализовавшихся значений u, v, ρ, w (из условий устойчивости и прочих). Схема метода крупных частиц явная. Как и в PIC-методе, стандартный шаг численного интегрирования состоит из двух этапов:

$$\text{I этап: } (u, v, \rho, w)^n_{i,j} \to (\tilde{u}, \tilde{v}, \tilde{\rho}, \tilde{w})_{i,j},$$

$$\text{II этап: } (\tilde{u}, \tilde{v}, \tilde{\rho}, \tilde{w})_{i,j} \to (u, v, \rho, w)^{n+1}_{i,j}.$$

На первом этапе учитываем силы давления, пренебрегая переносом. Используем простую аппроксимацию уравнений (10):

$$\frac{1}{\tau}(\tilde{\rho}_{i,j} - \rho^n_{i,j}) = 0,$$

$$\frac{1}{\tau}\rho^n_{i,j}(\tilde{u}_{i,j} - u^n_{i,j}) + \frac{1}{h}(p^n_{i+1/2,j} - p^n_{i-1/2,j}) = 0,$$

$$\frac{1}{\tau}\rho^n_{i,j}(\tilde{v}_{i,j} - v^n_{i,j}) + \frac{1}{h}(p^n_{i,j+1/2} - p^n_{i,j-1/2}) = 0,$$

где $p^n_{i+1/2,j}$, $p^n_{i,j+1/2}$ — полусуммы значений в центрах ячеек. Уравнение энергии имеет вид

$$\frac{1}{\tau}\rho^n_{i,j}(\tilde{w}_{i,j} - w^n_{i,j}) + \frac{1}{h}[(pu)^n_{i+1/2,j} - (pu)^n_{i-1/2,j}] + \\ + \frac{1}{h}[(pv)^n_{i,j+1/2} - (pv)^n_{i,j-1/2}] = 0.$$

Здесь возможны варианты: можно p, u, v порознь интерполировать с центров ячеек на их стороны, а можно интерполировать произведения (pu), (pv). Таким образом, первый этап очень прост и не содержит каких-либо нестандартных приемов аппроксимации.

Несколько сложнее и своеобразнее реализация второго этапа. На этом этапе учитываются процессы переноса: схема зависит от направления потока в данной точке и приобретает явно несимметричных характер. Второй этап начинается вычислением скоростей на сторонах ячейки: $u^*_{i+1/2,\,j} = \tilde{u}_{i,\,j} + \tilde{u}_{i+1,\,j}$, $v^*_{i,\,j+1/2}$. Они используются только для определения направления потока. Затем вычисляются скорости в серединах сторон ячейки на основе отрезка ряда Тейлора: $\tilde{u}(x \pm h/2) = \tilde{u}(x) \pm (h/2)\tilde{u}_x$.

Однако при вычислении $\tilde{u}_{i+1/2,\,j}$, например, можно использовать разложение как в точке (i, j), так и в точке $(i+1, j)$ — это определяется направлением потока. Предпочтение отдается тому направлению, откуда «приносится информация», т.е. откуда течет газ, попадающий в точку $(i+1/2, j)$. В результате мы получаем

$$\tilde{u}_{i+1/2,\,j} = \begin{cases} \tilde{u}_{i,\,j} + (\tilde{u}_{i+1,\,j} - \tilde{u}_{i-1,\,j})/4, & u^*_{i+1/2,\,j} > 0, \\ \tilde{u}_{i+1,\,j} - (\tilde{u}_{i+2,\,j} - \tilde{u}_{i,\,j})/4, & u^*_{i+1/2,\,j} < 0. \end{cases}$$

Таким образом, для аппроксимации u_x используется центральная разность (второй порядок точности) в точке разложения. Вышеизложенный принцип вычисления величин в точках $(i+1/2, j)$ и $(i, j+1/2)$ (в последнем случае, очевидно, играет роль знак $v^*_{i,\,j+1/2}$) используется для вычисления всех остальных величин, фигурирующих в формулах для потоков через соответствующую сторону ячейки. Вводится, однако, дополнительная корректировка: если знаки $u^*_{i+1/2,\,j}$ и $\tilde{u}_{i+1/2,\,j}$ противоположны, потоки всех величин (массы, импульса, энергии) через сторону ячейки $(i+1/2, j)$ считаются равными нулю (аналогично для потоков $\Pi_{i,\,j+1/2}$).

Теперь уравнение для ρ аппроксимируется следующим образом:

$$h^2(\rho^{n+1}_{i,\,j} - \tilde{\rho}_{i,\,j}) + (\Pi_{i+1/2,\,j} - \Pi_{i-1/2,\,j}) + (\Pi_{i,\,j+1/2} - \Pi_{i,\,j-1/2}) = 0,$$

где

$$\Pi_{i+1/2,\,j} = h\tau\,\tilde{u}_{i+1/2,\,j}\,\tilde{\rho}_{i+1/2,\,j}, \qquad \Pi_{i,\,j+1/2} = h\tau\,\tilde{v}_{i,\,j+1/2}\,\tilde{\rho}_{i,\,j+1/2}.$$

Здесь уравнение записано в форме, подчеркивающей связь с законом сохранения массы ячейки $h^2\rho$; величины Π имеют смысл потоков массы через границу ячейки за время шага τ. Подчеркнем, что каждый поток $\Pi_{i+1/2,\,j}$, например, вычисляется для разделяющей ячейки (i, j) и $(i+1, j)$ стороны независимо от того, является ли

она правой для одной ячейки или левой для другой. Это свойство обеспечивает дивергентность схемы.

Остальные уравнения (законов переноса импульса и полной энергии) имеют общую форму:

$$(\rho Q)_t + (\rho u Q)_x + (\rho v Q)_y = 0,$$

где Q принимает значения u, v, w соответственно. Эти уравнения аппроксимируются по одной и той же схеме. Мы уже имеем значения потоков массы $\tilde{\rho}\tilde{u}$ (на правой и левой границах ячейки) и $\tilde{\rho}\tilde{v}$ (на верхней и нижней границах ячейки) — это величины П, вычисление которых продемонстрировано выше.

Определим теперь правила вычисления величин Q на этих же границах. При этом используем величины Q в центрах ячеек, где они вычисляются естественным образом: $Q_{i,j} = \tilde{u}_{i,j}$, и т.п. Итак,

$$Q_{i+1/2,j} = \{\tilde{Q}_{i,j} \text{ при } \tilde{u}_{i+1/2,j} > 0; \quad \tilde{Q}_{i+1,j} \text{ при } \tilde{u}_{i+1/2,j} < 0\},$$

$$Q_{i,j+1/2} = \{\tilde{Q}_{i,j} \text{ при } \tilde{v}_{i,j+1/2} > 0; \quad \tilde{Q}_{i,j+1} \text{ при } \tilde{v}_{i,j+1/2} < 0\}.$$

После этого уравнение аппроксимируется просто (напомним, что $\rho_{i,j}^{n+1}$ уже известно):

$$h^2(\rho_{i,j}^{n+1}Q_{i,j}^{n+1} - \tilde{\rho}_{i,j}\tilde{Q}_{i,j}) + (\Pi_{i+1/2,j}\tilde{Q}_{i+1/2,j} - \Pi_{i-1/2,j}\tilde{Q}_{i-1/2,j}) + \\ + (\Pi_{i,j+1/2}\tilde{Q}_{i,j+1/2} - \Pi_{i,j-1/2}\tilde{Q}_{i,j-1/2}) = 0.$$

Из этого соотношения вычисляется величина $Q_{i,j}^{n+1}$.

Выше была описана одна из возможных реализаций метода крупных частиц. Многие детали могут быть оформлены иначе. В частности, естественно возникает вопрос: почему для вычисления $\tilde{\rho}_{i+1/2,j}$ применялось разложение (слева или справа, в зависимости от направления потока), а для величин, обозначенных Q, — «снос» по потоку на полшага? Теоретических обоснований такого способа, видимо, нет. Схемы, которые условно можно отнести к схемам типа крупных частиц, формировались под воздействием анализа результатов расчетов.

С причинами, определившими выбор той или иной расчетной формулы в какой-то мере можно познакомиться в специальной литературе, посвященной методу крупных частиц и практике его применения. Основные черты этой группы методов: расщепление системы уравнений (и связанный с ним «двухэтапный» счет) и наличие «односторонних» разностных аппроксимаций первых производных, ориентированных против направления потока. Эти особенности приводят к не очень высокой точности метода. В частности, в методе крупных частиц считается возможным не вводить искусственную вязкость. Функции сглаживания решения берет на себя «счетная вязкость», возникающая в таких «односторонних» схемах. Диверген-

тность разностной схемы метода крупных частиц также является его характерной чертой, которую обычно сохраняют при различных реализациях.

Проблемы геометрии. Одним из наиболее серьезных вопросов, решение которого существенно определяет расчетные схемы, является проблема достаточно аккуратного отражения геометрии течения, если она не слишком проста. Здесь есть два аспекта проблемы: внешняя геометрия течения и внутренняя. Поясним суть дела. К проблемам внешней геометрии мы отнесем те, которые обычно возникают при решении задач обтекания тел достаточно сложной формы или расчет течений в каналах сложного профиля. Характерным примером является, например, задача обтекания самолета (или даже его части). Если связать систему координат с обтекаемым телом, то расчет течения газа проводится в области, для которой поверхность тела является границей. На ней ставится достаточно простое «условие непротекания»: нормальная к поверхности тела компонента скорости равна нулю.

Реализация этого условия несложна, когда поверхность тела проходит по линиям счетной сетки, например $j = 1/2$. (Читатель без труда внесет необходимые дополнения в описанную выше схему для учета условия $v = 0$ на границе.) Все потоки $\Pi_{i,\,1/2} = 0$, и единственная проблема, которая возникает при использовании стандартных формул из-за отсутствия величин в узлах ниже границы, — это отсутствие в них давления. Во всех остальных случаях величины, формально зависящие от значений в узлах, не входящих в область определения сеточных функций, умножаются на нулевой поток массы на границе. Исключением является величина p, необходимая для аппроксимации члена p_y в уравнении для v. Однако уравнение

$$(\rho v)_t + (\rho uv)_x + (\rho v^2)_y + p_y = 0$$

на границе при $v(t, x, 0) = 0$ превращается в $p_y = 0$, что дает основание полагать p на границе равным значению p в центре ячейки.

Однако все это просто в случае, когда граница тела проходит по линиям координатной сетки. А если обтекаемое тело имеет сложную форму? Эта проблема возникла в начале шестидесятых годов, когда мощности ЭВМ уже позволяли приступать к решению двумерных задач обтекания тел. В то время выявились два направления. В одном направлении используется простая (декартова прямоугольная) система координат и так или иначе решаются проблемы построения аппроксимаций уравнений в нестандартных ситуациях около границы. В другом направлении строится специальная система координат, в которой граница тела является координатной линией. Построение таких сеток, называемых адаптирующимися (к форме тела), — не такое простое

дело. Ведь обычно граница тела не задается простой формулой, она может быть даже задана графически.

К тому же предъявляются определенные требования к координатной системе: переход от декартовых координат x, y к криволинейным ξ, η должен быть по возможности гладким, чему явно препятствует наличие угловых точек на контуре обтекаемого тела. Итак, первая проблема на этом пути — само построение адаптирующейся сетки. Далее, описание сетки в координатах ξ, η состоит в том, что для узлов (i, j) нужно вычислять и хранить в памяти декартовы координаты $x_{i,j}$, $y_{i,j}$. Они необходимы при построении аппроксимаций уравнений.

После перехода к уравнениям газовой динамики в переменных (t, ξ, η) вид уравнений резко усложняется: в них появляются выражения x_ξ, x_η, ... Использование такой формы уравнений требует запаса гладкости в отображении $(x, y) \rightleftarrows (\xi, \eta)$, что, как указывалось, трудно обеспечить при сложной форме контура тела. Эта гладкость нужна, в частности, для разностной аппроксимации производных x_ξ, x_η, ... Здесь возможен и часто используется другой путь, тоже не очень простой, — аппроксимация уравнений газовой динамики на неправильной и не очень регулярной сетке. Если ячейки сетки заметно отличаются от параллелограммов, стандартный и наглядный способ построения разностных схем (состоящий в замене входящих в уравнение производных простыми разностными отношениями) начинает отказывать. На смену ему приходит другой способ, к которому прибегают все чаще, так как возрастающие требования адаптации к геометрии рассчитываемого явления заставляют использовать сложно устроенные, нерегулярные сетки.

Несколько слов об этом способе, базирующемся на использовании интерполяционных полиномов, мы скажем ниже (в связи с изложением основных идей так называемого метода *свободных точек*). Важным достоинством метода адаптирующихся сеток является возможность учета априорной информации о гладкости решения. Эта информация имеет достаточно неопределенный характер и состоит в предположении о том, что рассчитываемое течение является кусочно-гладким, т.е. пространство (t, x, y) можно разбить на некоторое число частей достаточно гладкими поверхностями и внутри каждой части искомые функции достаточно гладкие.

Таким образом, вышеупомянутые разделяющие поверхности — это поверхности разрывов (сильных или слабых). К ним могут быть присоединены и поверхности разрывов (слабых) в отображении $(x, y) \rightleftarrows (\xi, \eta)$, которое, тем самым, является тоже кусочно-гладким. С точки зрения математический постановки задачи эти поверхности являются в некотором смысле «внутренними границами», на которых ставятся соответствующие граничные условия, связывающие значения искомых функций на разных сторонах поверхности

разрыва. Если такая поверхность является ударной волной, это — соотношения Гюгонио; в случае контактного разрыва, это — условия непрерывности давления и нормальной компоненты скорости при произвольных разрывах плотности и касательной к поверхности компоненты скорости, и т.п.

В каждой из выделенных частей обычно вводят свою систему координат таким образом, чтобы в этих координатах область стала прямоугольником, а сетка, как говорят, была топологически эквивалентна прямоугольной. Это весьма удобно для программирования вычислительного процесса, который организуется, как система вложенных циклов. Однако в таких системах координат точность разностных аппроксимаций зависит не только от гладкости искомого решения, но и от гладкости отображения $(x, y) \rightleftharpoons (\xi, \eta)$. Построенная сетка должна быть достаточно регулярной, ее ячейки (прямоугольники в координатах ξ, η) не должны слишком сильно отличаться от параллелограммов в пространстве x, y. Трудности возникают, если, например, ячейки оказываются сильно скошенными параллелограммами, и т.п. Построение хороших сеток, топологически эквивалентных прямоугольным, в областях даже не слишком вычурной формы — сложная задача, решение которой составляет специальный раздел вычислительной математики. Часто трудно даже обеспечить построение взаимно однозначного отображения $(x, y) \rightleftharpoons (\xi, \eta)$. Следует еще подчеркнуть, что число и топологическая структура выделяемых областей гладкости заранее не известны и определяются в процессе решения задачи, что заставляет использовать алгоритмы построения сетки в оперативном режиме — почти на каждом шаге интегрирования уравнения по времени.

Сказанного достаточно, чтобы понять, что реализация вышеизложенного подхода связана со значительными трудностями. Программы получаются очень сложными; они разрабатываются целыми коллективами в течение многих лет. В процессе эксплуатации мощного вычислительного аппарата происходит его постоянное развитие. Тем не менее такие программы созданы, и полученные с их помощью результаты считаются наиболее достоверными.

Наряду с этим направлением, естественно, возникла идея использовать самые простые сетки — прямоугольные в декартовых координатах и, преодолевая трудности аппроксимации уравнений около границы (с учетом краевых условий), получать достаточно простые программы для расчета течений. Слабым местом такого подхода является то обстоятельство, что часто рассчитываемое течение имеет разные характеристики гладкости в разных частях области. Адаптирующиеся сетки могут это учитывать (хотя и не без определенных трудностей). Расчет же на равномерной сетке, размер которой диктуется наиболее «узким» местом, требует слишком малого шага. Поэтому, хотя это направление начало развиваться в нашей стране с первых лет работы на ЭВМ, на некоторое время оно

было оставлено. Это было связано, видимо, с малым объемом оперативной памяти ЭВМ того времени.

До недавнего времени наиболее распространенной ЭВМ была БЭСМ-6, имевшая оперативную память в 32 000 слов. После введения в эксплуатацию новых ЭВМ с большими ресурсами оперативной памяти (порядка 10^6) интерес к этому направлению стал возрождаться. Конечно, число узлов равномерной сетки должно быть существенно больше числа узлов адаптирующейся, но расчетные формулы оказываются намного проще. По оценкам специалистов, программа трехмерной газовой динамики с адаптирующейся сеткой тратит порядка 10^4 операций на один узел (при расчете одного временного шага); программа же, основанная на прямоугольной декартовой сетке, около 10^2 операций. Смогут ли такие простые программы конкурировать с расчетами на адаптирующихся сетках — покажет будущее. Работа здесь в сущности только начинается. Ниже мы опишем некоторый способ учета криволинейной границы на прямоугольной декартовой сетке, который продемонстрирует, какого сорта проблемы здесь возникают.

Аппроксимация около криволинейной границы. Рассмотрим течение газа, описываемое уравнениями в форме Эйлера в декартовой системе координат. Граница, на которой поставлено условие непротекания (контур обтекаемого тела), является относительно гладкой и проходит более или менее произвольно относительно системы координат. Введем прямоугольную равномерную сетку с шагом h по обоим направлениям (для простоты). Вертикальные и горизонтальные линии сетки занумеруем целыми индексами i и j, а различные счетные величины будем вводить в разных точках сетки.

Определим в центрах ячеек сетки термодинамические неизвестные $\rho^n_{i+1/2,\,j+1/2}$, $p^n_{i+1/2,\,j+1/2}$. Определим компоненты скорости на серединах сторон ячейки $u^n_{i,\,j+1/2}$, $v^n_{i+1/2,\,j}$. Такая «шахматная» сетка удобна для аппроксимации оператора дивергенции в точках $(i+1/2,\ j+1/2)$ и grad p, причем x-компоненту удобно аппроксимировать в u-точке (где она только и нужна), а y-компоненту — соответственно в v-точке.

Среди введенных формально узлов квадратной сетки выделим *счетные узлы*, т.е. те, в которых соответствующая величина считается определенной. Узел сетки считаем счетным, если квадрат $h \times h$ с центром в этом узле целиком помещается в области течения (не пересекается с обтекаемым телом). Если такой квадрат частично находится в области потока, частично — внутри тела, он считается *фиктивным*: ему не соответствует никакая величина, входящая в основные массивы счетных величин. Однако такая величина в фиктивном узле может появиться как вспомогательная, вычисляемая через основные.

Рисунок 41 поясняет сказанное. На нем в части области показаны сетка, счетные (p, ρ)-узлы (темные кружки), счетные u-узлы (знаки «минус») и v-узлы (крестики). Указана граница и расположенные на ней нестандартные счетные узлы. Будем считать, что рассматриваемый участок границы задается кривой $Y(x)$, причем $|Y_x(x)| \leqslant 1$. (При $|Y_x(x)| > 1$ контур следует задавать функцией $X(y)$ и в последующем поменять роли x и y.)

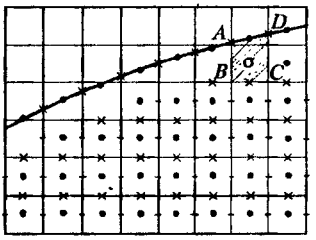


Рис. 41

Итак, на контуре вводятся (p, ρ)-узлы с координатами $x_{i+1/2} = (i + 1/2)h$, $Y(x_{i+1/2})$, в которых определены счетные величины $p^n_{i+1/2}$, $\rho^n_{i+1/2}$ (отсутствие второго индекса — признак принадлежности границе). Кроме того, на контуре вводятся w-узлы с координатами x_i, $Y(x_i)$, в которых определена счетная величина w^n_i, имеющая смысл касательной к контуру компоненты скорости.

Опишем структуру одного шага интегрирования по времени в явной схеме. Шаг состоит из трех основных операций.

1. По значениям величин в счетных узлах (внутренних и граничных) производится интерполяция соответствующих величин в фиктивные узлы. Тем самым создается ситуация, позволяющая во внутренних счетных узлах рассчитать величины на следующем шаге по стандартным формулам.

2. Во внутренних счетных узлах вычисляются значения на верхнем, $(n + 1)$-м, слое по стандартным формулам.

3. Расчет граничных значений $\rho^{n+1}_{i+1/2}$, $p^{n+1}_{i+1/2}$, w^{n+1}_i производится по специальным формулам.

Конкретизируем вид расчетных формул. Вычисление какой-либо величины в фиктивном узле производится с помощью линейной интерполяции по значениям соответствующей величины на той же вертикальной линии сетки. Используются последняя внутренняя точка и граничная. Если в граничной точке нет нужной величины, например $u^n_{i+1/2}$, она вычисляется как $0.5(w^n_i + w^n_{i+1}) \cos \alpha_{i+1/2}$, где $\alpha_{i+1/2}$ — угол наклона границы.

Уравнение для плотности ρ аппроксимируется так (за скобки выносятся общие индексы):

$$\tau^{-1}(\rho^{n+1} - \rho^n)_{i+1/2,\, j+1/2} + (ud_1\rho)^n_{i+1/2,\, j+1/2} + \\ + (vd_2\rho)^n_{i+1/2,\, j+1/2} + (\rho \operatorname{div})^n_{i+1/2,\, j+1/2} = 0.$$

Поясним смысл некоторых величин. Если в указанной точке отсутствует, например, величина $u^n_{i+1/2,\,j+1/2}$, она вычисляется линейной интерполяцией по ближайшим точкам:

$$u^n_{i+1/2,\,j+1/2} = (u^n_{i,\,j+1/2} + u^n_{i+1,\,j+1/2})/2.$$

Символом d_1 обозначена односторонняя разностная производная по x, ориентированная против потока. Так,

$$(d_1\rho)^n_{i+1/2,\,j+1/2} = h^{-1}(\rho^n_{i+1/2,\,j+1/2} - \rho^n_{i-1/2,\,j+1/2}),$$

если $u^n_{i+1/2,\,j+1/2} > 0$. Так же строится d_2 — аппроксимация производной по y.

Оператор дивергенции аппроксимируется естественно:

$$(\mathrm{div})^n_{i+1/2,\,j+1/2} = h^{-1}[(u^n_{i+1,\,j+1/2} - u^n_{i,\,j+1/2}) + (v^n_{i+1/2,\,j+1} - v^n_{i+1/2,\,j})].$$

Аналогично строится аппроксимация уравнения для давления p;:

$$p_t + up_x + vp_y + [\gamma p + (\gamma - 1)q](u_x + v_y) = 0$$

(в случае уравнения состояния идеального газа). Здесь величина $q = Ch^2\,\rho(u_x + v_y)^2$ — искусственная вязкость.

Формулы расчета граничных значений $p^{n+1}_{i+1/2}$, $\rho^{n+1}_{i+1/2}$ будут описаны отдельно. После вычисления всех p^{n+1}, ρ^{n+1} величины u^{n+1}, v^{n+1} находятся по формулам типа

$$\tau^{-1}(u^{n+1} - u^n)_{i,\,j+1/2} + (ud_1u)^n_{i,\,j+1/2} + (vd_2u)^n_{i,\,j+1/2} +$$

$$+ h^{-1}[(p^{n+1} + q^n)_{i+1/2,\,j+1/2} - (p^{n+1} + q^n)_{i-1/2,\,j+1/2}]/\rho^n_{i,\,j+1/2}.$$

Разностное уравнение для v строится по такому же принципу. Подчеркнем, что p берется уже с верхнего слоя.

Перейдем к аппроксимации уравнений на границе. С учетом условия непротекания они имеют вид

$$\rho_t + w\rho_s + \rho(\mathrm{div}) = 0,$$

где $\mathrm{div} = u_x + v_y$, s — длина дуги на контуре. Члены $\rho_t + w\rho_s$ аппроксимируются по тем же принципам, но с учетом знака w. Пояснения требует только вычисление $(\mathrm{div})_{i+1/2}$. Каждой (p, ρ)-точке ставится в соответствие свой элементарный объем (см. рис. 41). Если точка внутренняя, элементарный объем есть ячейка $h \times h$ с центром в точке $(i + 1/2,\ j + 1/2)$. Если точка граничная, это — четырехугольник $[x_i \leqslant x \leqslant x_{i+1}] \times [y_i \leqslant y \leqslant Y(x)]$, где y_i — уровень верхней границы элементарной ячейки, соответствующей последней на

линии $x = x_{i+1/2}$ внутренней счетной (p, ρ)-точке. Таким образом, элементарные объемы, соответствующие (p, ρ)-точкам (как внутренним, так и граничным), покрывают всю область течения без пустот и перекрытий.

Теперь для построения аппроксимации div используем известную формулу

$$\operatorname{div} \mathbf{w} = \lim_{\sigma \to 0} (1/\sigma) \oint (\mathbf{w}, \mathbf{n})\, ds, \quad \mathbf{w} = \{u, v\},$$

где σ — малая, стягивающаяся к точке (x, y) область, $\mathbf{n}$ — нормаль к ее границе (внешняя). Мы используем допредельный аналог этой формулы, беря в качестве σ элементарный объем, соответствующий данной граничной (p, ρ)-точке. Интеграл аппроксимируется почти очевидным образом: на сторонах AB и CD (см. рис. 41) по значениям w на границе и $u_{i(+1), j+1/2}$ интерполируется u, и эта величина интегрируется. На стороне BC известно $v_{i+1/2, j}$, и

$$\int_B^C (\mathbf{w}, \mathbf{n})\, ds \approx -h v_{i+1/2, j}.$$

Остальные детали аппроксимации не уточняем; они аналогичны тем, которые используются в стандартных счетных точках.

Опыт показал, что при вышеописанном способе формирования счетной сетки могут образоваться ячейки с маленькими линейными размерами. Наличие подобных ячеек вынуждает уменьшить шаг по времени τ по сравнению с тем, который обусловлен требованиями вычислительной устойчивости во внутренних счетных ячейках. В противном случае в области «уменьшенных» ячеек проявляется вычислительная неустойчивость. Чтобы избежать этого, следует вводить более крупные приграничные ячейки, присоединяя к ним примыкающие внутренние ячейки $h \times h$ и исключая соответствующие внутренние счетные точки.

Изложенная выше схема была испытана расчетом следующей задачи. Пусть начальные данные имеют вид (X_0 берется несколько левее острия клина; см. рис. 42)

$$p, \rho, u, v = \begin{cases} 0,\ 1,\ 0,\ \ 0 & \text{при } x > X_0, \\ 10,\ 4,\ 2.5,\ 0 & \text{при } x < X_0. \end{cases}$$

Они удовлетворяют (при $\gamma = 5/3$) условиям Гюгонио. Ударная волна, движущаяся вправо, падает на острый клин, возникает сложное течение, известное в газовой динамике под названием «тришок». Образуется точка, в которой «сходятся» три ударных волны: исходная ударная волна (ее фронт — вертикаль), «преломленная прямая волна» (ее фронт ортогонален линии клина) и отраженная ударная

волна. Эти три линии сильного разрыва сходятся в «тройной точке», движущейся по прямой линии под некоторым углом к линии клина.

На рис. 42 представлены линии уровня давления, которые соответствуют (справа—налево) значениям p, равным 2, 4, 6, 8, 10, 12, 13, 14, 15, 16, 17. Перед фронтом волны $p = 1$. Хорошо видна характерная конфигурация тришока, хотя число узлов на рис. 42 не очень велико для столь сложного течения. Анализ поля скорости показывает, что около твердой стенки поток параллелен ей с хорошей точностью, а скорость внутри области непрерывно сопрягается с граничной скоростью w_i.

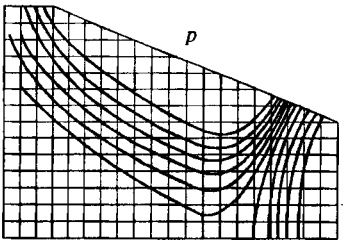


Рис. 42

Графики линий уровня плотности намного хуже, хотя «топология» тришока просматривается и там. Это не случайно. Графики плотности в сложных течениях часто получаются в расчетах очень «корявыми». Сложные, хотя локализованные в малых областях пространства-времени события в рассчитываемом явлении (пересечение ударных волн, отражение волны от препятствия и т.п.) оставляют на графиках плотности долго несглаживающиеся следы, получившие специальное название «энтропийные». В графиках давления такие следы не сохраняются, так как перепады давления вызывают изменение скорости и локальный максимум (или минимум) в давлении долго держаться не может. Плотность же является «пассивной» величиной: мы уже видели, что сколь угодно долго может существовать разрыв в плотности (контактный разрыв). Разумеется, неровности плотности согласованы с неровностями температуры, так что давление оказывается гладким.

Видимо, читатель догадался, что реализация вышеизложенной схемы в общем случае (и тем более в трехмерных задачах) достаточно сложна, так как нужно реализовать разветвленную логику программы, ориентирующейся в отношении сетки к заданной каким-то образом (обычно, таблично) границе тела. Поэтому большинство вычислителей предпочитает более простые алгоритмы «фиктивных» точек. В них перед очередным шагом интегрирования по t во все «внешние» узлы сетки (которые сами не являются счетными, но которые необходимы для реализации стандартного во внутренних узлах счета) засылаются некоторые значения, подбираемые так, что расчет во всех счетных точках по стандартным формулам учитывает краевые условия. Это сравнительно легко сделать, если граница все-таки проходит по узлам счетной сетки (например, граница проходит по диагонали сетки). Но как работает эта упрощенная технология в достаточно сложной геометрии, не очень ясно.

Метод свободных точек. Следующий подход к построению методов приближенного интегрирования уравнений газовой динамики ориентирован на тот же класс задач, который имелся в виду при разработке PIC-метода. Это — расчет течений с контактными границами, подвергающимися сильным деформациям. Можно говорить о задачах и без контактных границ, но с очень сильными деформациями первоначальной лагранжевой сетки. В некотором смысле (достаточно осторожно) можно иметь в виду течения, в процессе которых происходит сильное перемешивание вещества (но это все-таки далеко не турбулентность).

Основу метода составляет отказ от сетки как таковой. Основные счетные величины, описывающие состояние среды, здесь следующие.

1. Частицы нумеруются индексом k ($k = 1, 2, \ldots, K$). Каждая частица имеет номер α_k вещества, из которого она «сделана». Этот номер влияет только на выбор уравнения состояния, которым следует пользоваться при расчете давления в той точке, в которой в данный момент находится k-я частица. Положение частицы в момент времени t_n определяется ее координатами X_k^n, Y_k^n.

2. С каждой частицей связываются значения u_k^n, v_k^n, ρ_k^n, e_k^n. Они трактуются как значения скорости, плотности и относительной внутренней энергии в точке $\{t_n, X_k^n, Y_k^n\}$. Таким образом, точки $\{X_k^n, Y_k^n\}$ ($k = 1, 2, \ldots, K$) образуют подвижную сетку, в узлах которой определены основные величины.

Стандартный шаг интегрирования (переход от n-го слоя к $(n+1)$-му) начинается с того, что для каждой k-й точки формируется набор соседей — точек с номерами j_k^i ($i = 1, 2, \ldots, I_k^n$). В набор входят точки, находящиеся в момент t_n в некоторой малой окрестности точки $\{X_k^n, Y_k^n\}$. Именно по этому набору соседей и по значениям функций в них производится аппроксимация уравнений газовой динамики в k-й точке.

Технику аппроксимации поясним, начав с некоторой простейшей модели. Пусть некоторая функция $f(x, y)$ известна в точках (X_j^n, Y_j^n), где $j = j_k^i$ ($i = 0, 1, 2, \ldots$). Обозначим эти значения f_i (считаем, что $j_k^0 = k$). По этим значениям можно проинтерполировать функцию и получить непрерывную функцию $\tilde{f}(x, y)$, совпадающую в узлах с f_i ($i = 0, 1, 2, \ldots$). Теперь можно продифференцировать $\tilde{f}$ в точке (X_k^n, Y_k^n) и полученные выражения для производных подставить в уравнения. Таким образом получим аппроксимацию тех или иных дифференциальных операторов.

Выше изложена общая идея, которая сейчас постепенно начинает применяться при расчетах на неправильных сетках, узлы кото-

рых не образуют простой и гладкой структуры. Фактически схема расчета более сложна. Это объясняется тем, что решения уравнений газовой динамики не настолько гладки, чтобы использовать интерполирующий полином высокой степени (степень зависит от числа соседей I). Желательно использовать аппарат, работающий при минимальном запасе гладкости. Разъясним основные идеи, используемые при построении разностных формул стандартного шага.

Линеаризация. В окрестности некоторой точки (t_n, x^*, y^*) функции $f(t, x, y)$ представляются в виде

$$f(t, x, y) = f(t_n, x^*, y^*) + \delta f(t', x, y), \quad t' \in (0, \tau),$$

где δf — малое приращение f, τ — малый шаг численного интегрирования. Подставляя такие выражения в уравнение, разлагая нелинейные члены в ряд Тейлора и ограничиваясь линейными по δf членами, можно получить для них линейную систему уравнений с постоянными («замороженными» в точке (t_n, x^*, y^*)) коэффициентами.

Формула Герглотца—Петровского. Для уравнений газовой динамики линеаризация приводит к линейной гиперболической системе. Такие системы хорошо изучены. Решение задачи Коши (без краевых условий) для них может быть получено по упомянутой выше формуле. Если f — полный вектор газодинамических параметров u, v, p, ρ, то эта формула принимает вид

$$\delta f(\tau, x^*, y^*) = \int_0^{2\pi} d\varphi \int_0^{R(\varphi)} G(\varphi, r)\, \delta f\,(0_n, x^* + r\cos\varphi, y^* + r\sin\varphi)\, dr,$$

где $G(\varphi, r)$ — известное ядро; $R(\varphi)$ — граница пересечения плоскости $t' = 0$ с некоторым наклонным конусом, имеющим вершину в точке (τ, x^*, y^*). Это так называемый «характеристический конус», осью которого является «энтропийная характеристика» — линия, касающаяся направления $dt : dx : dy = 1 : u^* : v^*$ (u^*, v^* — скорости в точке (t_n, x^*, y^*)). Раствор этого конуса определяется скоростью звука c^*. Для вычисления интеграла от уже «известного» при $t = t_n$ решения нужно воспользоваться интерполяцией газодинамический параметров, заданных в узлах нерегулярной сетки $\{X_k^n, Y_n^k\}$ $(k = 1, 2, \ldots, K)$.

Интерполяция. Множество соседей точки (X_k^n, Y_k^n) формируется так, чтобы они как бы окружали k-ю точку, т.е. можно разбить окрестность k-й точки на некоторое число треугольников. В каждом треугольнике функции интерполируются линейно по x, y. Полученная

таким образом на n-м слое функция $\tilde{f}(x, y)$ (кусочно-линейная) интегрируется с весом G по основанию конуса. Условие устойчивости формул требует, чтобы эта область интегрирования целиком находилась внутри выделенной набором соседей окрестности. Этого всегда можно добиться выбором достаточно малого шага интегрирования τ. Однако чрезмерно сблизившиеся точки могут сделать такую окрестность слишком малой, что приведет к нерационально малому τ. В этом случае слишком близкая к k-й точка игнорируется.

Таковы основные методические моменты метода свободных точек. Они соответствуют одной из первых работ этого направления, выполненной В. Ф. Дьяченко. В дальнейшем эти идеи развивались, детали изменялись, они достаточно разнообразны. Все это привело к формированию специфического научного направления в вычислительной гидродинамике, получившего в западной литературе название «Free Lagrange Method» (FLM).

Следует отметить, что реализация подобных методов в первую очередь требует тщательной проработки алгоритма формирования соседей для каждой точки. Если оперировать только с той информацией, которая была введена выше (это минимально необходимая, но достаточная информация), придется в каждой k-й точке устраивать просмотр координат всех остальных точек. Это требует $O(K^2)$ операций на каждом шаге интегрирования по времени, что слишком дорого и ограничивает число K. Поэтому приходится разрабатывать систему сопровождающей расчет информации, которая грубо разделяет точки на близкие друг к другу и при выборе соседей k-й точки существенно сужает необходимый перебор.

За малый шаг τ координаты точек мало меняются, и сопутствующая информация, соответственно, корректируется. От удачного и остроумного решения этой достаточно сложной проблемы существенно зависит трудоемкость метода. Другое обстоятельство, стимулирующее совершенствование методов этого направления, — стремление обеспечить дивергентность схемы, т.е. наличие в дискретной системе прозрачных и естественных аналогов основных законов сохранения. В схеме, описанной выше, этого нет.

Вышеизложенная методика была использована для расчета сложных течений, сопровождавшихся сильной деформацией контактных границ, разделяющих газы с существенно различными физическими свойствами. Рисунок 40*а* дает представление об одной из таких задач. По газу идет мощная ударная волна, наталкивающаяся на твердое препятствие. (Хотя заштрихованная часть есть часть металлической конструкции, ее поведение описывается уравнениями газовой динамики в рассматриваемой задаче.) Рисунок 40*б* дает представление о последующем течении: показаны первичная ударная волна (которая в «твердом» теле движется медленнее, чем в газе), отраженная волна и сложная деформация контактной границы. Процесс сопровождается образованием мощной струи.

§ 24. Приближенное интегрирование уравнения Власова

Уравнение Власова описывает движение совокупности большого числа заряженных частиц (ионов и электронов, например) в условиях, когда можно пренебречь столкновениями частиц и их взаимодействие определяется только электрическими силами. Это уравнение описывает события, пространственный масштаб которых много меньше длины свободного пробега и характерное время много меньше времени свободного пробега. Такие ситуации достаточно распространены в теории сильно разреженной плазмы, а их практическое значение связано с изучением, например, космического пространства, взаимодействия космических аппаратов с очень высокими (и, следовательно, сильно разреженными) слоями атмосферы и некоторых других вопросов.

В этой модели состояние плазмы описывается двумя функциями: $f_e(r, v, t)$ и $f_i(r, v, t)$. Здесь независимые координаты имеют следующий смысл: $r = \{x, y, z\}$ — декартовы координаты точки пространства, трехмерная координата $v = \{v_x, v_y, v_z\}$ представляет координаты в импульсном пространстве, t — время, функции f_e, f_i — плотности электронов и ионов. Таким образом, например, если мы выделяем в пространстве маленький кубик $[r, r + \Delta r]$ и интересуемся числом частиц, содержащихся в этом кубике и имеющих скорости в диапазоне $[v, v + \Delta v]$, то оно выражается (в первом порядке) величиной $f\,\Delta r\,\Delta v$.

Область фазового пространства r, v, которая нас интересует, обычно не ограничена по скорости, но функции f быстро спадают при $|v| \to \infty$; можно ограничиться конечной областью $|v| \leqslant V$, поставив граничное условие $f|_{|v|=V} = 0$. В пространстве область ограничена размером $L = \{L_x, L_y, L_z\}$. Будем предполагать, что по пространственным переменным все функции периодичны с периодом L (в такой постановке решается большое число задач, связанных с теоретическим изучением процессов в плазме).

В дальнейшем, ради простоты изложения, будем рассматривать двумерные задачи: $r = \{x, y\}$, $v = \{v_x, v_y\}$. Уравнение Власова описывает эволюцию во времени функций f_e, f_i:

$$\begin{aligned} &\frac{\partial f_e}{\partial t} + v_x \frac{\partial f_e}{\partial x} + v_y \frac{\partial f_e}{\partial y} - \frac{q_e}{m_e}\left(\frac{\partial \varphi}{\partial x}\frac{\partial f_e}{\partial v_x} + \frac{\partial \varphi}{\partial y}\frac{\partial f_e}{\partial v_y}\right) = 0, \\ &\frac{\partial f_i}{\partial t} + v_x \frac{\partial f_i}{\partial x} + v_y \frac{\partial f_i}{\partial y} - \frac{q_i}{m_i}\left(\frac{\partial \varphi}{\partial x}\frac{\partial f_i}{\partial v_x} + \frac{\partial \varphi}{\partial y}\frac{\partial f_i}{\partial v_y}\right) = 0. \end{aligned} \tag{1}$$

Здесь $\varphi(x, y, t)$ — потенциал электрического поля; $(-\varphi_x, -\varphi_y)$ — компоненты напряженности электрического поля. Потенциал φ определяется уравнением Пуассона

$$\Delta\varphi = 4\pi\left\{q_e\int_{-\infty}^{+\infty} f_e(x, y, v, t)\,dv + q_i\int_{-\infty}^{+\infty} f_i(x, y, v, t)\,dv\right\}, \qquad (2)$$

где q_e, q_i — заряды электрона и иона; m_e, m_i — их массы. Система уравнений (1), (2) замкнута. Уравнение (2) называют уравнением самосогласованного электрического поля (в том смысле, что оно не задается расположением каких-то внешних зарядов, а создается участвующими в процессе частицами).

Существует альтернативная математическая модель, в которой рассматриваются отдельно все частицы. Движение каждой из них описывается уравнениями

$$\frac{dr_k}{dt} = v_k, \qquad \frac{dv_k}{dt} = \frac{q_k}{m_k}E(x_k, y_k, t), \quad k = 1, 2, \ldots, K. \qquad (3)$$

Здесь k — номер частицы; q_k, m_k — ее заряд и масса; E — напряженность электрического поля, связанная с частицами тем же уравнением Пуассона

$$E = -\mathrm{grad}\,\varphi, \qquad \Delta\varphi = 4\pi\rho(x, y, t);$$

ρ — плотность заряда. В соответствии с моделью частиц

$$\varphi(r, t) = \sum_k \frac{q_k}{|r - r_k(t)|}, \qquad (4)$$

где $q_k/|r - r_k|$ — потенциал, создаваемый k-м зарядом, находящимся в данный момент в точке $r_k(t)$.

При определенных условиях, когда нас не интересуют детали, имеющие пространственный размер порядка так называемого дебаевского радиуса (и меньше), можно считать, что поле определяется уравнением Пуассона, а плотность $\rho(x, y, t)$ определяется как «предел» при стремлении к нулю некоторого объема ω, окружающего точку $\{x, y\}$:

$$\rho(x, y, t) = \lim_{\omega\to 0}\sum_{\omega} q_k/|\omega|,$$

где $\sum_{\omega}$ — сумма зарядов частиц, попавших в этот объем.

Предел в вышеприведенной формуле надо понимать «физически» в следующем, например, смысле. Пусть ω — квадрат размером h, т.е. $|\omega| = h^2$. Тогда можно ввести величину $\rho_h = \sum_{\omega} q_k/h^2$. При

уменьшении h эта величина как-то меняется, приближаясь к некоторому пределу $\rho(x, y, t)$, и при $h_* < h < h^*$ (где $h_* \ll h^*$) более или менее совпадает с этим пределом. В дальнейшем (при $h \ll h_*$) величина ρ_h начинает беспорядочно и весьма ощутимо меняться — начинает сказываться дискретное строение вещества, в объеме ω оказывается уже слишком мало частиц, и, конечно, никакого математического предела не существует.

Дифференциальные уравнения пишутся именно для относительно устойчивых значений ρ_h, $h \in (h_*, h^*)$. Используя эти уравнения, мы неявно предполагаем, что интересующие нас явления описываются подобными функциями, причем они гладкие относительно шага $H \gg h^*$, т.е. мало меняются при изменении аргументов на H. Вообще, доверие к уравнениям, убеждение в том, что они описывают реальную действительность, держится не столько на строгости их вывода, сколько на опыте успешного их применения в анализе различных физических ситуаций.

Характеристики уравнения Власова. Постановка краевых задач. Две модели разреженной плазмы (модель (1), (2) в терминах полей f, φ, ρ, E и модель частиц (3), (4)) тесно связаны друг с другом и с чисто математической точки зрения. Покажем, что траектории частиц являются характеристиками уравнения Власова и что вдоль этих характеристик значения f_e, f_i сохраняются.

Пусть функции $f_e(r, v, t)$, $f_i(r, v, t)$, $\varphi(r, t)$, $E(r, t)$, $\rho(r, t)$ — решение уравнения Власова. Тогда на этом решении могут быть определены траектории системы (3): функции $R_e(t, R^0, V^0)$ и $V_e(t, R^0, V^0)$, где R^0, V^0 — данные Коши при $t = 0$. Определим на этой траектории функцию

$$\tilde{f}_e(t) \equiv f_e(R(t, R^0, V^0), V(t, R^0, V^0), t).$$

Индекс e указывает, что рассматривается система (3) с «электронными» параметрами q и m. Из той же точки (R^0, V^0) выходит и «ионная» характеристика, вдоль которой сохраняется значение f_i.

Вычислим производную $\tilde{f}_e$ по t:

$$\frac{d\tilde{f}_e}{dt} = \frac{\partial f_e}{\partial t} + \frac{\partial f_e}{\partial r}\dot{R} + \frac{\partial f_e}{\partial v}\dot{V} = \frac{\partial f_e}{\partial t} + \left(V, \frac{\partial f_e}{\partial r}\right) + \frac{q}{m}\left(E, \frac{\partial f_e}{\partial v}\right).$$

Производная $d\tilde{f}_e/dt = 0$, так как f_e удовлетворяет уравнению Власова. Итак,

$$f_e(R(t, R^0, V^0), V(t, R^0, V^0), t) = f_e(R^0, V^0, 0).$$

Перенос значений f вдоль характеристик имеет определенные неприятные последствия, которые обсуждаются ниже. Пока же мы используем картину характеристик для пояснения математической структуры тех краевых задач, постановка которых является корректной.

Граница области разбивается на две части в зависимости от знака скалярного произведения $(\mathbf{U}, \mathbf{N})$, где N — 6-мерный вектор внутренней нормали к границе, а U — 6-мерный вектор скорости характеристики в данной точке, т.е. $U = \{\mathbf{v}, q\mathbf{E}\}$ (это разбиение различно для электронов и ионов). Там, где $(\mathbf{U}, \mathbf{N}) > 0$, характеристики входят в область извне; на этой части границы нужно задавать краевое условие, например значение f, соответствующее входящей характеристике. Там, где $(\mathbf{U}, \mathbf{N}) < 0$, характеристика приходит на границу изнутри области, и краевое условие не ставится. Если, например, область является прямоугольной ($0 \leqslant x \leqslant L_x$, $0 \leqslant y \leqslant L_y$, $0 \leqslant z \leqslant L_z$), то часть границы $x = 0$ разбивается на две области: при $v_x > 0$ нужно задавать значения f_e, f_i, при $v_x < 0$ никакие краевые условия на ставятся. Что касается краевых условий для φ, то здесь мы имеем стандартную для оператора Лапласа ситуацию.

Отметим основные особенности уравнения Власова, определяющие трудности его численного решения.

1. Высокая размерность задачи. Если нас интересуют задачи двумерные с точки зрения геометрического пространства, т.е. задачи в (x, y), то, естественно, появляются две импульсные координаты (v_x, v_y) и события развиваются в области четырехмерного фазового пространства. Даже при миллионе узлов шаг сетки по каждому направлению будет порядка 1/30 линейного размера.

2. Наличие разных масштабов времени. На электроны и ионы действует одна и та же сила, но массы их существенно разные, соответственно различны их ускорения и характерные масштабы времени. Самый легкий ион (водорода) примерно в 1800 раз тяжелее электрона. В плазме явления носят обычно колебательный характер, и в ней имеются характерные частоты (или периоды):

$$\omega_e = \sqrt{4\pi n_e q_e^2/m_e}, \qquad \tau_e = 2\pi/\omega_e = \sqrt{\pi m_e/n_e q_e^2},$$
$$\omega_i = \sqrt{4\pi n_i q_i^2/m_i}, \qquad \tau_i = 2\pi/\omega_i = \sqrt{\pi m_i/n_i q_i^2}. \tag{5}$$

Здесь n_e, n_i — плотности частиц (числа частиц на единицу объема), они одинаковы; q_e, q_i — их заряды, они тоже одинаковы (по абсолютной величине); массы же частиц разные.

Если нас интересуют времена порядка $100\tau_i$, то считать приходится с шагом, существенно меньшим τ_e, например $dt = 0.1\tau_e$. Поскольку $\tau_e < \tau_i/40$, расчет требует 40 000 шагов. Это очень много. Поэтому в расчетах часто искусственно полагают $m_i \approx (10 \div 100)m_e$. Опре-

деляющим считается сам факт, что ионы намного тяжелее электронов и что за один ионный период электрон совершит «много» колебаний. А сколько именно, не так уж важно.

Дифференциальные свойства функций f (r, v, t). Заметим, что в уравнениях отсутствуют какие-либо диссипативные (сглаживающие) факторы и, наоборот, существуют факторы, приводящие к очень негладким функциям, особенно по переменным v. Это можно пояснить следующим образом. Как уже отмечалось, значения f сохраняются на характеристиках. Пусть при $t = 0$ заданы гладкие функции $f(r, v, 0)$. Фиксируем при каком-то t точку r_0 и рассмотрим $f(r_0, v, t)$ как функцию только v. В точках с близкими значениями v могут оказаться (и фактически оказываются) частицы (характеристики), пришедшие из разных начальных точек и принесшие с собой разные значения f. Поэтому график f приобретает после некоторого времени (большого, но еще не настолько, что расчет можно прекратить) «пилообразный» характер. Это — не неустойчивость (никакой катастрофы нет), а просто потеря гладкости решения, свойственная самим уравнениям.

Попытки решения уравнения Власова методом конечных разностей были не очень успешными (в сложных задачах), что, конечно, существенно связано и с тем, что сетки приходилось брать относительно скромные. Однако в задачах с учетом столкновений (в правую часть уравнения Власова (1) добавляются интегралы от f по импульсному пространству — так называемые интегралы столкновений) ситуация более благоприятная. Столкновения — это диссипативный процесс, приводящий к сглаживанию функций f. Но здесь, к сожалению, появляется другая трудность — вычисление интегралов столкновений.

Модель частиц. Первые попытки расчета явлений в бесстолкновительной плазме предпринимались на основе модели взаимодействующих частиц (3), (4), и они были достаточно успешными, пока можно было ограничиться небольшим числом частиц. И алгоритмов изобретать не приходилось: ведь модель (3), (4) — это просто задача Коши для системы обыкновенных дифференциальных уравнений только довольно большого порядка. Можно было пользоваться обычными методами интегрирования задачи Коши.

Однако приложения быстро потребовали существенного увеличения числа частиц. Тут возникли трудности и весьма существенные. Первая и, видимо, важнейшая из них — рост числа операций. В самом деле, мы имеем дело с большим числом K частиц (в современных расчетах K порядка 10^3, 10^4, 10^5), попарно взаимодействующих друг с другом. Следовательно, вычисление правых частей (сил) «стоит» $O(K^2)$ операций. Конечно, существенное влияние на движе-

ние данной частицы оказывают ближайшие к ней; действие остальных можно учитывать более грубо. Но как это сделать? Ведь даже само определение того, какая частица близка к данной, какая нет, требует (если не применять каких-то алгоритмических изобретений) такого же примерно объема вычислений, что и прямое вычисление всех сил.

Вторая трудность — малый шаг по времени. Он часто определяется постоянно происходящими сближениями небольшого числа частиц на столь малые расстояния, при которых силы их взаимодействия быстро меняются, движение приобретает сложный характер и требует слишком малого шага интегрирования. И этот шаг навязывается всей системе, хотя движение большинства частиц можно было бы интегрировать с гораздо большим шагом.

Указанные выше сложности построения расчетной схемы как на базе уравнений в частных производных (1), (2), так и на базе модели частиц (3), (4) привели к некоторому их синтезу, который и стал основой дальнейших конструкций, успешно применяемых для расчета сложных явлений в плазме.

Метод заряженных облаков. Наиболее успешным методом численного решения уравнения Власова является метод «облаков в ячейках», или метод «макрочастиц». Изложим его основную вычислительную схему. Состояние плазмы будем описывать следующими функциями.

1. Потенциал $\varphi^n_{l,m}$ определяется в узлах некоторой фиксированной сетки в пространстве x, y (n — временной индекс; l, m — пространственные индексы). Величину $\varphi^n_{l,m}$ мы трактуем как значение φ в точке $(x_l = lh,\ y_m = mh,\ t_n)$.

2. «Макрочастицы» нумеруются индексом k. Каждая частица характеризуется следующими величинами: X^n_k, Y^n_k — положение частицы; $U^{n+1/2}_k$, $V^{n+1/2}_k$ — компоненты ее скорости в момент $t_n + \tau/2$; m_k, q_k — масса и заряд частицы. Отметим, что скорости относятся не к моменту t_n, а к полуцелому моменту $t_{n+1/2}$.

Опишем стандартный шаг процесса интегрирования, при котором информация $\{\varphi^n, X^n, Y^n, U^{n+1/2}, V^{n+1/2}\}$ переходит в данные на следующем $(n+1)$-м слое $\{\varphi^{n+1}, X^{n+1}, Y^{n+1}, U^{n+3/2}, V^{n+3/2}\}$. Наиболее популярна схема «leap frog» (у нас ее называют «чехардой»). Расчет начинается с вычисления новых положений X^{n+1}_k, Y^{n+1}_k в соответствии с уравнениями характеристик (3):

$$(X^{n+1}_k - X^n_k)/\tau = U^{n+1/2}_k, \qquad (Y^{n+1}_k - Y^n_k)/\tau = V^{n+1/2}_k \tag{6}$$

(схема второго порядка точности).

Вычислим потенциал $\varphi_{k,m}^{n+1}$ в момент времени t_{n+1}. Он определяется из уравнения Пуассона, аппроксимированного по обычной разностной схеме:

$$(\varphi_{l+1,m} - 2\varphi_{l,m} + \varphi_{l-1,m}) + (\varphi_{l,m+1} - 2\varphi_{l,m} + \varphi_{l,m-1}) = 4\pi\rho_{l,m}^{n+1}h^2, \tag{7}$$

но сначала нужно с каждым узлом связать величину $\rho_{l,m}^{n+1}$ — плотность заряда. Напрашивается такой способ: возьмем ячейку размером $h \times h$ с центром в точке (x_l, y_m), посчитаем сумму зарядов частиц, находящихся в ней в момент t_{n+1}, и положим $\rho_{l,m}^{n+1} = \sum q_k/h^2$.

Первые же попытки расчетов показали существенный дефект такого подхода. При относительно небольшом числе частиц, приходящихся на каждую ячейку (10 частиц — это уже хорошо), плотность принимает лишь дискретные значения: $q/h^2, 2q/h^2, \ldots$ Кроме того, она разрывно зависит от положений частиц. Если частица пересекает границу ячейки, плотность в смежных с этой границей ячейках резко изменяется, следствием чего является изменение напряженности поля E. Движение частиц становится нерегулярным, проявляются колебания явно счетного, а не физического характера.

Выход был найден в том, что точку (X_k^n, Y_k^n) стали трактовать как положение центра некоторого заряженного облака малого размера, например $3h \times 3h$. В каждой из 9 ячеек $h \times h$ этого облака плотность заряда считалась постоянной. При вычислении $\rho_{l,m}$ определялась часть облака, попавшая в связанную с точкой (l, m) ячейку $h \times h$, и в эту точку «передавалась» соответствующая часть заряда облака. Кстати, кусочно-постоянная плотность в облаке использовалась для того, чтобы облегчить необходимый подсчет.

После вычисления $\rho_{l,m}^{n+1}$ решается уравнение (7). Это наиболее трудоемкий элемент алгоритма. Используются наиболее эффективные методы, в частности метод Фурье, реализованный в форме быстрого дискретного преобразования Фурье, являющегося одним из фундаментальных алгоритмов современного численного анализа. Его развитие связано с решением рассматриваемых здесь задач. Он будет подробно описан ниже.

Вычисление скоростей производится из очевидной аппроксимации уравнений (3):

$$\frac{U_k^{n+3/2} - U_k^{n+1/2}}{\tau} = \frac{q_k}{m_k}(E_x)_k^{n+1}, \qquad \frac{V_k^{n+3/2} - V_k^{n+1/2}}{\tau} = \frac{q_k}{m_k}(E_y)_k^{n+1}. \tag{8}$$

Здесь нужно еще определить силы, действующие на облако. Напомним, что $E = -\operatorname{grad}\varphi$.

Имея $\varphi_{l,m}^{n+1}$ в узлах сетки, можно определить в узлах значения

$$(E_x)_{l,m}^{n+1} = -\frac{\varphi_{l+1,m}^{n+1} - \varphi_{l-1,m}^{n+1}}{2h}, \quad (E_y)_{l,m}^{n+1} = -\frac{\varphi_{l,m+1}^{n+1} - \varphi_{l,m-1}^{n+1}}{2h}. \quad (9)$$

Опыт показал, что силу, действующую на k-е облако, нужно вычислять достаточно аккуратно. Например, с помощью интерполяции можно определить функцию $E(x, y)$. При вычислении правых частей уравнений (8) интегрируется функция $q(x, y)\,E(x, y)$ по облаку с учетом распределения заряда в нем. Это достаточно сложная процедура, но при упрощенном вычислении правых частей (8) проявляются нежелательные счетные эффекты. Анализ привел к пониманию вызывающих их причин и к упрощению техники конструирования расчетных схем, свободных от таких дефектов.

Отметим еще один момент, связанный с назначением для расчетов масс и зарядов макрочастиц. При моделировании явлений в плазме каждая макрочастица представляет $10^{10} \div 10^{15}$ реальных частиц, заряды и массы которых — хорошо известные физические константы. Какими же брать при моделировании массы и заряды макрочастиц? Естественный ответ «их нужно брать в $10^{10} \div 10^{15}$ раз бо́льшими реальных» оправдан еще и тем, что, как нетрудно проверить, в такой «макроплазме» значения характерных величин (периодов электронных и ионных колебаний) оказываются теми же, что и в реальной плазме.

Расчетная схема с законом сохранения импульса. Выше была описана схема, в которой преодолен один самый грубый недостаток, — флуктуации плотности, появляющиеся при примитивной технике расчета плотности в счетном узле сетки. Следующим шагом было построение схем с более тонким свойством — сохранением импульса. Поясним суть проблемы. Мы имеем дело как бы с набором твердых частиц, обладающих в некотором смысле пространственной структурой. Но эта структура сказывается только на формулах вычисления сил, действующих между частицами. Реализуется следующая цепочка зависимостей. Положения частиц определяют плотность заряда в узлах сетки; заряд определяет потенциал φ в узлах; потенциал определяет силу, действующую на частицы, находящиеся в каких-то точках пространства; в данной схеме центр частицы движется по тем же законам, что и «точка».

Итак, речь идет о системе точек, между которыми действуют силы, вызывающие их движение. А плотность заряда, потенциал — промежуточные объекты, облегчающие вычисление сил по сравнению с физической моделью плазмы, как совокупности заряженных частиц с обычным электростатическим потенциалом. В этой «чистой» модели взаимодействующих частиц действуют известные законы, в частности закон сохранения импульса, являющийся следстви-

ем известного закона Ньютона: если частица a действует на частицу b с силой F, то частица b действует на частицу a с силой $-F$. Следствием этого является отсутствие «самодействия»: если частица в системе только одна, на нее никакие силы не действуют.

В первых схемах метода облаков (или макрочастиц), составленных так, как было описано выше, был обнаружен недостаток: при единственной частице образовывалось поле, действующее на нее, и частица начинала движение, являющееся чисто счетным эффектом. Расчетная схема не обладала законом «действие равно противодействию», т.е. законом сохранения импульса, что порождало нефизические эффекты. Конечно, формально такие эффекты исчезающе малы при возрастании числа частиц, но реальные расчеты приходится проводить при не столь уж большом числе частиц. И если есть возможность избавиться от такого эффекта, это следует сделать. Изложим прежде всего способ анализа схемы метода макрочастиц с точки зрения закона Ньютона «действие равно противодействию».

Проследим и формально опишем процесс формирования сил в какой-то фиксированный момент времени. Движение мы не рассматриваем, поэтому речь идет о следующей цепочке вычислений.

Рассмотрим совокупность частиц, имеющих заряды q_k и координаты $\{X_k, Y_k\}$ $(k = 1, 2, \ldots, K)$.

Первая операция — расчет плотности $\rho_{l,m}$ в узлах эйлеровой сетки. Отвлечемся от конкретного способа, связанного с той или иной формулой распределения заряда в облаке. Ведь в любом случае заряд k-й частицы q_k распределяется по узлам сетки в соответствии с некоторой функцией $Q(l, m; X, Y)$ таким образом, что вклад k-й частицы в заряд в точке (l, m) есть $\rho^{(k)}_{l,m} = q_k\, Q(l, m; X_k, Y_k)$, а полный заряд в точке (l, m) есть, очевидно,

$$\rho_{l,m} = \sum_{k=1}^{K} q_k\, Q(l, m; X_k, Y_k). \tag{10}$$

В дальнейшем мы обсудим необходимые свойства функции Q.

Вторая операция — решение уравнения Пуассона (7). Уравнение дополняется условиями периодичности — это важное для дальнейшего обстоятельство. Оно означает, что все двумерное пространство заполнено частицами, положения и скорости которых периодически (по x, y) повторяются. Поэтому никаких других сил, кроме сил взаимного электростатического отталкивания (притяжения) частиц, в системе нет. Учитывать можно только взаимодействия частиц, расположенных в одном прямоугольнике в пространстве x, y, сторонами которого являются периоды по x, y соответственно. Действие всех остальных частиц учитывается периодичностью потенциала.

Если бы мы рассматривали уравнение Пуассона с какими-то другими условиями, например с заданными значениями φ на границе, это означало бы наличие внешних сил. В такой системе импульс не обязан сохраняться. Решение разностного уравнения Пуассона может быть выражено через разностную функцию Грина:

$$\varphi_{l,m} = \sum_{i,j} G(l, m; i, j)\, \rho_{i,j}. \tag{11}$$

Это очевидный факт. Функция $G(l, m; i, j)$ — просто матрица, обратная к матрице разностного оператора Лапласа. Здесь и в дальнейшем мы не будем аккуратно выписывать пределов суммирования: оно ведется по одной ячейке периодичности (или, если угодно, на торе). Мы обсудим свойства G ниже, пока же оставим запись преобразования ρ в φ в самой общей форме. Очевидно, потенциал всей системы частиц есть сумма потенциалов, порожденных каждой частицей:

$$\varphi^{(k)}_{l,m} = \sum_{i,j} G(l, m; i, j)\, \rho^{(k)}_{i,j}. \tag{12}$$

Третья операция — вычисление сил (градиента φ) в узлах сетки. Ограничимся только одной компонентой силы, так как для другой все будет точно так же. Итак, пусть сила $f_{i,j}$ есть $-\varphi_x(x_j, y_j)$. В разностной реализации любая аппроксимация может быть записана в общем виде:

$$f_{i,j} = \sum_{l,m} F(i, j; l, m)\, \varphi_{l,m}. \tag{13}$$

Четвертая операция — вычисление силы, действующей на частицу с зарядом q, расположенную в точке (x, y). Это достигается процедурой интерполяции сеточной функции $f_{i,j}$ в точку (x, y), которую можно записать в виде

$$S(x, y) = q \sum_{i,j} I(x, y; i, j)\, f_{i,j}. \tag{14}$$

Если нам нужно подсчитать силу, действующую на r-ю частицу со стороны k-й, необходимо провести следующую цепочку преобразований:

$$\begin{aligned} S(X_r, Y_r; X_k, Y_k) &= q_r q_k \sum_{i,j} I(X_r, Y_r; i, j) \times \\ &\times \sum_{l,m} F(i, j; l, m) \sum_{p,s} G(l, m; p, s)\, Q(p, s; X_k, Y_k) = \\ &= q_r q_k\, I(X_r, Y_r; \cdot)\, F(\cdot; *)\, G(*; \#)\, Q(\#; X_k, Y_k). \end{aligned} \tag{15}$$

Здесь $\cdot$, $*$, $\#$ — символы «немых» индексов, по которым произведена свертка. Сила $S(X_r, Y_r; X_k, Y_k)$ есть элемент (r, k) матрицы, являющейся произведением некоторых других матриц.

Введем C и D — пространства функций, определенных в точках $\{x, y\}$ и $\{l, m\}$ соответственно. Алгоритм интегрирования уравнения Власова определяется при конкретизации следующих операторов:

Q: $C \to D$ — оператор «раздачи заряда» (из точки (x, y) в узлы сетки);

G: $D \to D$ — разностный оператор Грина, обратный к оператору Лапласа с периодическими условиями;

F: $D \to D$ — оператор вычисления первой разностной производной;

I: $D \to C$ — оператор интерполяции силы с узлов сетки на произвольную точку.

Таким образом, S есть оператор типа $C \to C$, а действие k-й частицы на r-ю есть (множитель $q_r q_k$ опускаем)

$$S(X_r, Y_r; X_k, Y_k) = (IFGQ)_{r,k}.$$

Тогда действие r-й частицы на k-ю есть

$$S(X_k, Y_k; X_r, Y_r) = (IFGQ)_{k,r} = (IFGQ)^*_{r,k} = (Q^*G^*F^*I^*)_{r,k}.$$

Для того чтобы в системе частиц был справедлив закон «действие равно противодействию» и, как следствие, сохранялся импульс, нужно обеспечить соотношение

$$Q^*G^*F^*I^* = -IFGQ. \tag{16}$$

Используем следующие почти очевидные свойства операторов.

а) $G^* = G$ (следствие самосопряженности оператора Лапласа в классе периодических функций). Бóльшая часть разностных аппроксимаций оператора на равномерной сетке автоматически наследует это свойство, хотя при желании его можно и нарушить, не потеряв аппроксимации.

б) $F^* = -F$ (следствие кососимметричности оператора $\partial/\partial x$). Аппроксимация типа (9), как нетрудно проверить, сохраняет это свойство; аппроксимация $(\partial f/\partial x)_i = (f_i - f_{i-1})/h$ его нарушает: $(\partial f/\partial x)^*_i = -(f_{i+1} - f_i)/h$.

в) $FG = GF$ (следствие перестановочности операторов $\partial/\partial x$ и Δ). Это свойство наследуется использованными выше разностными аналогами операторов.

Таким образом, нужно обеспечить равенство

$$Q^*FGI^* = IFGQ. \tag{17}$$

Оно выполняется, если $Q = I^*$. В схеме с гарантированным сохранением импульса из операторов Q, I только один строится независи-

мо, например оператор интерполяции I. Оператор раздачи заряда Q после выбора I определяется автоматически. Поясним смысл сказанного. Пусть оператор $I(x, y; i, j)$ реализован как кусочно-билинейный. Это означает, что для вычисления

$$\tilde{f}(x, y) = \sum_{i,j} I(x, y; i, j)\, f_{i,j} \qquad (18)$$

нужно найти индексы l, m, при которых $x \in [lh, lh + 1)$, $y \in [mh, mh + h)$, и вычислить соответствующие коэффициенты интерполяции $w_{\alpha,\beta}$ $(\alpha, \beta = 0, 1)$. Тогда

$$\tilde{f}(x, y) = w_{0,0} f_{l,m} + w_{1,0} f_{l+1,m} + w_{0,1} f_{l,m+1} + w_{1,1} f_{l+1,m+1}, \qquad (19)$$

где $w_{0,0} = h^{-2}(lh + h - x)(mh + h - y)$ и т.д. Элемент «матрицы» $I^*(i, j; x, y)$, очевидно, вычисляется так: по точке (x, y) нужно найти индексы l, m и коэффициенты интерполяции, после чего определяются элементы матрицы I^*:

$$I^*(l + \alpha, m + \beta; x, y) = w_{\alpha,\beta}, \quad \alpha, \beta = 0, 1.$$

Все остальные элементы I^* равны нулю.

Итак, операция раздачи заряда q_k, находящегося в точке (X_k, Y_k) состоит в следующем. По значениям X_k, Y_k находится ячейка сетки (l, m), в которой эта точка находится, вычисляются соответствующие коэффициенты интерполяции $w_{\alpha,\beta}$, и доли заряда $q_k w_{\alpha,\beta}$ раздаются в узлы $(l + \alpha, m + \beta)$:

$$\rho_{l+\alpha, m+\beta} := \rho_{l+\alpha, m+\beta} + q_k w_{\alpha,\beta}.$$

Если бы использовалась более точная интерполяция, мы столкнулись бы с таким «нефизичным» явлением, как раздача отрицательной доли заряда в какой-то узел. Видимо, поэтому на практике ограничиваются простейшим кусочно-билинейным оператором интерполяции, когда все $w_{\alpha,\beta} \geqslant 0$ и $\sum w_{\alpha,\beta} = 1$.

Таким образом конструируются схемы интегрирования, обеспечивающие непрерывность плотности заряда $\rho_{l,m}$ по положениям частиц (X_k, Y_k) и сохранение импульса.

Быстрое дискретное преобразование Фурье. Имеется некоторая сеточная функция $f(k)$ $(k = 0, 1, \ldots, N - 1)$. Она может быть представлена разложением в сумму Фурье:

$$f(k) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} A(n)\, w^{kn}, \quad w = e^{2\pi i/N}, \quad k = 0, 1, \ldots, N - 1. \qquad (20)$$

Коэффициенты Фурье $A(n)$ вычисляются по аналогичным формулам:

$$A(n) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} f(k)\, w^{-kn}, \quad n = 0, 1, \ldots, N-1. \tag{21}$$

Формулы (21), (20) представляют собой прямое и обратное преобразования. Выполнение каждого из них требует $O(N^2)$ операций, если программировать непосредственно формулы (20), (21). Заметим, что нужно аккуратно вычислять степени w, чтобы избежать лишних операций. Ниже это будет сделано. Однако наиболее существенное сокращение объема вычислений связано с тем, что при вычислении сумм многие слагаемые и даже группы слагаемых вычисляются неоднократно, и этого повторения можно избежать. На этом основан алгоритм FFT (Fast Fourie Transform).

Рассмотрим реализацию (20), опуская, для простоты, множитель $1/\sqrt{N}$. Заметное сокращение операций происходит тогда, когда число N можно разложить на множители. Пусть $N = MN_1$. Представим k и n в виде

$$k = \varkappa N_1 + k_1, \qquad n = \nu + n_1 M. \tag{22}$$

Здесь $\varkappa = 0, 1, \ldots, M-1$; $k_1 = 0, 1, \ldots, N_1 - 1$; $\nu = 0, 1, \ldots, M-1$; $n_1 = 0, 1, \ldots, N_1 - 1$. Показатель степени kn можно записать так:

$$kn = k(\nu + n_1 M) = k\nu + kn_1 M = k\nu + n_1 M(\varkappa N_1 + k_1) =$$
$$= k\nu + n_1 \varkappa M N_1 + k_1 n_1 M.$$

Поскольку $w^{MN_1} = w^N = 1$, слагаемое $n_1 \varkappa M N_1$ можно из kn убрать и формула (20) примет вид

$$f(k) = \sum_{n=0}^{N-1} A(n)\, w^{k\nu}\, w^{k_1 n_1 M} = \sum_{n=0}^{N-1} A(n)\, w^{k\nu}\, w_1^{k_1 n_1}, \quad w_1 = w^M. \tag{20а}$$

Введем двухиндексную нумерацию, используя одно и то же имя для функций с разным числом целочисленных аргументов:

$$f(\varkappa, k_1) \equiv f(\varkappa N_1 + k_1), \qquad A(\nu, n_1) \equiv A(\nu + n_1 M). \tag{23}$$

Сумму по $n = 0, 1, \ldots, N-1$ представим как двойную:

$$f(\varkappa, k_1) = \sum_{\nu=0}^{M-1} \sum_{n_1=0}^{N_1-1} A(\nu, n_1)\, w^{k\nu}\, w^{k_1 n_1}, \tag{20б}$$

где $\varkappa = 0, 1, \ldots, M-1$; $k_1 = 0, 1, \ldots, N_1 - 1$. Внутренняя сумма в (20б) вычисляется при постоянных ν и k, поэтому множитель $w^{k\nu}$

можно вынести за скобки:

$$f(\varkappa, k_1) = \sum_{\nu=0}^{M-1} \left[\sum_{n_1=0}^{N_1-1} A(\nu, n_1)\ w_1^{k_1 n_1} \right] w^{k\nu}. \tag{20в}$$

Обозначим внутренние суммы, зависящие лишь от ν и k_1:

$$A_1(\nu, k_1) = \sum_{n_1=0}^{N_1-1} A(\nu, n_1)\ w_1^{k_1 n_1}, \tag{24}$$

$$\nu = 0, 1, \ldots, M-1, \quad k_1 = 0, 1, \ldots, N_1 - 1.$$

Тогда

$$f(\varkappa, k_1) = \sum_{\nu=0}^{M-1} A_1(\nu, k_1)\ w^{k\nu}, \tag{25}$$

$$\varkappa = 0, 1, \ldots, M-1, \quad k_1 = 0, 1, \ldots, N_1 - 1, \quad k = \varkappa N_1 + k_1.$$

Итак, преобразование (20) распалось на два последовательных. Сначала исходные величины A преобразуются в A_1 по формуле (24); это стоит CMN_1N_1 операций (постоянную C вычислим ниже: $C = O(1)$). Затем за $CMMN_1$ операций выполняется преобразование (25). Итого мы имеем $CN(M + N_1)$ операций вместо $CNMN_1$. При больших N и $M \approx N_1 \approx \sqrt{N}$ это уже заметный выигрыш.

Если число N_1 может быть разложено на множители, этот прием может быть использован снова. Будем понимать (24) как M преобразований типа (20). В самом деле, единственное свойство основания w в (20), которое было использовано, есть соотношение $w^N = 1$. Но и $w_1^{N_1} = w^{MN_1} = 1$. Пусть $M = M_1$ — простое число, а $N_1 = M_2N_2$. Тогда при каждом $\nu = 0, 1, \ldots, M_1 - 1$ нужно вычислить N_1 коэффициентов $A_1(\nu, k_1)$, а это можно сделать за $CN_1(M_2 + N_2)$ операций. Таким образом, мы имеем $CN_1M_1(M_2 + N_2)$ операций. Вместе с $CM_1M_1N_1$ операциями на вычисление (25) получаем $CN(M_1 + M_2 + N_2)$ операций. И т.д. Наибольший эффект достигается в случае, когда N разлагается на множители малой величины (2, 3, например).

Наиболее популярен алгоритм FFT при $N = 2^p$. В этом случае число операций сводится к $CN(2p) = 2CN \log_2 N$. Для оценки постоянной C приведем текст программы, реализующей, напри-

мер, преобразования (25). Используем нестрогую версию языка FORTRAN:

$v_0 = w^{N_1}$ \$ $v_1 = 1$
do 1 $\varkappa = 0, M-1$ \$ $v_2 = 1$
do 2 $k_1 = 0, N_1 - 1$ \$ $v_3 = 1$
do 3 $\nu = 0, M-1$ \$ $f(\varkappa, k_1) = f(\varkappa, k_1) + A_1(\nu, k_1)*v_3$
3 $v_3 = v_3*v_2*v_1$
2 $v_2 = v_2*w$
1 $v_1 = v_1*v_0$

Заметим, что $w^{k\nu} = w^{(\varkappa N_1 + k_1)\nu}$, а величины v_1, v_2, v_3 в циклах по $\varkappa$, k_1, ν принимают значения

$$v_1 = v_0^{\varkappa} = w^{\varkappa N_1}, \qquad v_2 = w^{k_1}, \qquad v_3 = (v_1*v_2)^{\nu} = w^{(k_1 + \varkappa N_1)\nu}.$$

Решение уравнения Пуассона методом Фурье. Разностное уравнение Пуассона (7) легко решается методом Фурье. Обозначим $f_{l,m} = 4\pi\rho_{l,m}^{n+1}h^2$. Сначала делаем преобразование по первому индексу. Получаем представление $f_{l,m}$ в виде

$$f_{l,m} = \sum_{\lambda} F_m^{\lambda} w^{l\lambda}.$$

Это N быстрых преобразований (при каждом m отдельно), которые выполняются за $CN^2 \log_2 N$ операций (при $N = 2^p$). Затем при каждом λ выполняем преобразование по индексу μ. Получаем выражение

$$f_{l,m} = \alpha \sum_{\lambda} \sum_{\mu} F^{\lambda,\mu} w^{l\lambda + m\mu}, \quad \alpha = 1/\sqrt{N}.$$

Это стоит еще $CN^2 \log_2 N$ операций.

В § 14 было показано, что сеточные функции $w^{l\lambda+m\mu}$ являются собственными функциями разностного оператора Лапласа. Соответствующие собственные значения суть

$$\Lambda_{\lambda,\mu} = -\frac{4}{h^2}\left(\sin^2\frac{\lambda\pi}{2N} + \sin^2\frac{\mu\pi}{2N}\right).$$

Из (7) получаем коэффициенты Фурье искомой функции φ: $\Phi^{\lambda,\mu} = F^{\lambda,\mu}/\Lambda_{\lambda,\mu}$. После этого дважды делаем обратное преобразование и находим значения в узлах сетки:

$$\varphi_{l,m} = \alpha \sum_{\lambda} \sum_{\mu} \Phi^{\lambda,\mu} w^{l\lambda + m\mu}.$$

Итак, решение (7) стоит $4CN^2 \log_2 N$ операций.

§ 25. Некорректные задачи и их приближенное решение

Почти все задачи, с которыми мы имеем дело, можно записать в общей форме:

$$R(u) = f, \tag{1}$$

где u — искомая функция из некоторого функционального пространства U, f — известная «правая часть» уравнения, принадлежащая пространству F, а R — оператор (вообще говоря, нелинейный) из U в F. Обычно правая часть известна нам не точно, а с некоторой погрешностью δ, т.е. «реальной действительности» соответствует «истинная» функция $\tilde{f}$, а решается задача с правой частью f, причем $\|\tilde{f} - f\|_F \leqslant \delta$ (где $\|\cdot\|_F$ — норма в пространстве F). Естественно, возникает вопрос: в какой мере погрешность в f сказывается на решении, т.е. каково отличие u от $\tilde{u}$, где $\tilde{u}$ — решение «точной» задачи

$$R(\tilde{u}) = \tilde{f}\,?$$

Определение 1. Задача (1) называется поставленной корректно, если из $\|\tilde{f} - f\|_F \leqslant \delta$ следует $\|\tilde{u} - u\|_U \leqslant \varepsilon$, где ε может быть сделано сколь угодно малым за счет достаточно малого δ.

Обычно считается, что физические задачи приводят к корректно поставленным математическим и только последние подлежат приближенному решению. Наоборот, задача считается поставленной некорректно, если при сколь угодно малом δ и сколь угодно большом Δ найдется функция f, такая, что при $\|f - \tilde{f}\| \leqslant \delta$ оказывается $\|\tilde{u} - u\| > \Delta$. Есть ли смысл находить решение некорректной задачи, если сколь угодно малые погрешности в правой части (а они всегда есть) приводят к большим погрешностям в решении?

Заметим, что характеристика задачи как корректной или некорректной зависит от выбора норм $\|\cdot\|_U$ и $\|\cdot\|_F$, и задача, некорректная при одном выборе норм, может оказаться корректной при другом. Однако этот формальный способ «исправления» задачи обычно не проходит, так как выбор норм не является совершенно произвольным и должен правильно отражать суть дела в содержательной постановке задачи. В дальнейшем мы поясним эти вещи на конкретном примере.

Рассмотрим классический пример (Адамара) некорректной задачи — так называемую обратную задачу теплопроводности. Ищется функция $u(t, x)$, удовлетворяющая уравнению

$$\frac{\partial u}{\partial t} = -\frac{\partial^2 u}{\partial x^2} \tag{2}$$

в прямоугольной области $0 \leqslant x \leqslant 1$, $0 \leqslant t \leqslant T$, с начальными данными $u(0, x) = \tilde{f}(x)$ и краевыми условиями $u(t, 0) = u(t, 1) = 0$. Пусть $\tilde{u}(t, x)$ — решение этой задачи. Возмутим «правую часть» (входную информацию), т.е. рассмотрим ту же задачу, но с начальными данными $u(0, x) = \tilde{f}(x) + \delta \sin(k\pi x)$, где δ — очень малое число. Выпишем явно решение новой задачи:

$$u(t, x) = \tilde{u}(t, x) + \delta\, e^{k^2\pi^2 t} \sin(k\pi x).$$

При достаточно большой частоте k возмущение решения может стать сколь угодно большим при сколь угодно малом возмущении начальных данных.

Покажем, как сделать эту задачу корректной, выбирая нормы подходящим образом. Идея почти очевидна: нужно так определить норму $\|\cdot\|_F$, чтобы величина $\delta \| \sin(k\pi \cdot)\|_F$ была очень большой, тем большей, чем больше k. Проще всего это сделать, определяя норму элемента $f \in F$ через его коэффициенты Фурье. Пусть $f(x) = \sum f_k \sin(k\pi x)$. Определим, например,

$$\|f\|_F = \left(\sum_k |f_k|^2 e^{2k^2\pi^2 T} \right)^{1/2}, \qquad \|u(\cdot)\|_U = \left(\int_0^1 u^2(x)\, dx \right)^{1/2}. \tag{3}$$

Нетрудно проверить, что $\|\cdot\|_F$ удовлетворяет требованиям, предъявляемым к нормам.

Теперь обратная задача теплопроводности стала корректной. В самом деле, если две функции $f(x)$ и $\tilde{f}(x)$ отличаются друг от друга на величину $\delta(x) = f(x) - \tilde{f}(x)$ и $\|\delta(\cdot)\|_F \leqslant \delta$, то соответствующие решения обратных задач $u(T, x)$ и $\tilde{u}(T, x)$ отличаются друг от друга на функцию

$$u(T, x) - \tilde{u}(T, x) = \sum_k \delta_k\, e^{k^2\pi^2 T} \sin(k\pi x)$$

и, очевидно, $\|u(T, \cdot) - \tilde{u}(T, \cdot)\|_U \leqslant \delta$, т.е. если правые части мало отличаются в норме (3), то так же мало отличаются соответствующие им решения в обычной норме $\|\cdot\|_{L_2}$. Конечно, обратная задача теплопроводности сама по себе от введения нормы (3) не изменилась. Просто предположение о том, что f мало отличается от $\tilde{f}$ в норме (3) включает в себя очень сильные ограничения: все производные f мало отличаются от соответствующих производных $\tilde{f}$, и это отличие тем меньше, чем выше порядок производной.

Можно проверить, что, определив норму формулой

$$\|f\|_F \equiv \sum_{n=0}^{N} c_n \left\| \frac{d^n f}{dx^n} \right\|_{L_2}, \tag{4}$$

нельзя сделать обратную задачу теплопроводности корректной. Это следует из того, что $e^{k^2\pi^2 T}$ растет быстрее k^N при любом N. Нормы (4) при небольших $N = 0, 1, 2$ будем называть «обычными», «естественными». Норму (3) будем называть «сильной». Пространства, в которых используются сильные нормы типа (4) при $N = \infty$, в математике известны. Они связаны с изучением аналитических и, в частности, целых функций.

Если мы предположим, что обратная задача теплопроводности ставится так, что правая часть должна быть элементом пространства очень гладких (аналитических) функций, то она окажется достаточно благополучной задачей. Содержательно обратная задача теплопроводности связана, например, с попыткой по известному в настоящий момент распределению температуры тела восстановить его температуру в прошлом. Это распределение температуры $f(x)$ известно не очень точно, с достаточно грубыми погрешностями. Если мы просто решим обратную задачу теплопроводности, то получим решение явно нефизического характера: в нем будут огромные пики отрицательных и положительных температур, которых в природе не бывает.

Предположим, что из общих качественных физических соображений с достаточными основаниями можно утверждать, что искомая температура $u(T, x)$ была не чрезмерно большой и достаточно простой функцией. Функция $f(t, x)$ связана с $u(T, x)$ прямой задачей теплопроводности, т.е. если решить обычную задачу

$$v_t = v_{xx}, \quad v(t, 0) = v(t, 1) = 0,$$

с начальными данными $v(0, x) = u(T, x)$, то $v(T, x) = f(x)$. Как известно, решение прямого уравнения теплопроводности есть очень гладкая, аналитическая функция при любых начальных данных и при всех $t > 0$, причем степень гладкости повышается с ростом времени t.

Таким образом, сделанное выше предложение превратить обратную задачу теплопроводности в корректную, взяв в качестве пространства F очень узкое пространство функций, не так искусственно, как это могло показаться на первый взгляд. Оно оказывается достаточно естественным и, как мы увидим в дальнейшем, составляет основу методов решения некорректных задач. Но беда в том, что реальная функция $f(x)$, которая получается физическими измерениями, включает в себя погрешности, которые выводят ее из пространства F.

Вот типичная картина в некорректных задачах. Для них исходная информация (правая часть f) состоит из двух компонент: $f = f_0 + \delta f$, причем $f_0 \in F$, где F — узкое пространство очень гладких функций с нормой $\|\cdot\|_F$, в которой задача корректна. Погрешность измерения (или способа задания f) $\delta f \notin F$, она мала в естественной, обычной норме, например L_2, т.е. $\|\delta f\|_{L_2} \leqslant \delta$, где δ —

малое число, но $\|\delta f\|_F = \infty$. Таким образом, при решении некорректной задачи нужно каким-то образом «отфильтровать» влияние δf.

Следующий характерный пример некорректной задачи — интегральное уравнение первого рода:

$$\int_0^1 K(x, s)\, u(s)\, ds = f(x), \quad x \in [0, 1],$$

где $K(x, s)$ — известное гладкое ядро. К таким уравнениям приводят задачи восстановления сигнала по некоторым измерениям. В этом случае искомая функция $u(s)$ представляет собой первичный сигнал, $f(x)$ — показание прибора, $K(x, s)$ — так называемая «аппаратная функция», т.е. непосредственно в опыте измеряется не сам сигнал $u(s)$, а некоторое его преобразование.

Характерным фактором, определяющим сложность такой задачи, является именно гладкость ядра $K(x, s)$. Прямое преобразование $u \to f$, т.е. $\int_0^1 K(x, s)\, u(s)\, ds$, обладает типичным свойством — это «сглаживающее» преобразование: оно преобразует негладкую функцию $u(s)$ в гладкую. В этом можно убедиться следующим образом. Сравним между собой преобразования функций $u(x)$ и $u(x) + C \sin(k\pi x)$. Очевидно, они отличаются друг от друга на функцию

$$C \int_0^1 K(x, s) \sin(k\pi s)\, ds = C\varkappa_k(x),$$

где $\varkappa_k(x)$ — k-й коэффициент Фурье функции $K(x, \cdot)$. Известно, что если функция $K(x, s)$ гладкая, то ее коэффициенты Фурье убывают с ростом k тем быстрее, чем более гладкая функция $K(x, s)$.

Таким образом, интегральное преобразование с гладким ядром отображает широкое функциональное пространство (негладких функций) в очень узкое пространство (гладких функций). Это преобразование мы запишем в форме $f = Ku$. Но нас интересует обратное преобразование $u = K^{-1}f$. Оно обладает противоположными свойствами, отображая узкое пространство в очень широкое, и в этом, в сущности, причина некорректности задачи. Р самом деле, функции $u(x)$ и $\tilde{u}(x) = u(x) + C \sin(k\pi x)$, сколь угодно сильно отличающиеся друг от друга, являются решениями уравнений

$$\int_0^1 K(x, s)\, u(s)\, ds = f(x), \qquad \int_0^1 K(x, s)\, \tilde{u}(s)\, ds = f(x) + C\varkappa_k(x)$$

со сколь угодно малыми (при достаточно большом k) отличиями в правых частях. Это и есть некорректность, так как малые погрешности в правых частях приводят к большим различиям в решениях. И здесь некорректность проявляется именно на высокочастотных возмущениях правых частей.

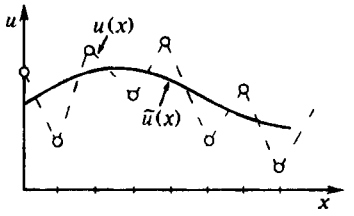


Рис. 43

Что же произойдет, если мы попробуем наивно решать интегральное уравнение первого рода? Один из возможных методов состоит в следующем. Введем на $[0, 1]$ сетку $x_n = nh$ $(n = 0, 1, \ldots, N)$ и заменим интегральное уравнение сеточным, аппроксимируя интеграл конечной суммой по правилу трапеций, например. Получаем систему линейных алгебраических уравнений $(n = 0, 1, \ldots, N)$

$$h\left[0.5K(x_n, s_0)\, u_0 + \sum_{j=1}^{N-1} K(x_n, s_j)\, u_j + 0.5K(x_n, s_N)\, U_N\right] = f_n.$$

Здесь u_j, f_n $(j, n \in [0, N])$ — сеточные аппроксимации искомой функции и правой части.

Мы имеем систему $N+1$ линейных уравнений с $N+1$ неизвестными. Результат решения такой системы показан (качественно) на рис. 43. При увеличении N картина становится все хуже и хуже — пилообразные пики растут. И это понятно: ведь чем больше число узлов N (размерность конечномерного сеточного пространства, аппроксимирующего функциональное), тем более высокие гармоники появляются в пространстве сеточных функций и тем сильнее проявляется некорректность.

Метод квазирешений. Перейдем к изложению основных идей, используемых при решении некорректных задач. Для определенности будем иметь в виду обратную задачу теплопроводности (2). Трудности решения были бы (в принципе) преодолены, если бы правая часть f выбиралась из очень узкого пространства гладких функций, ограниченных в сильной норме (3). К сожалению, эта рекомендация в прикладных задачах совершенно неприемлема.

Предложенный А. Н. Тихоновым подход к решению некорректных задач основан на следующем предположении. Искомое решение u существует. Более того, мы располагаем априорной информацией о его свойствах. Такая информация обычно имеет доста-

точно неопределенную форму. Например, задаются ограничения функции и ее производных такого типа:

$$u(x) \geqslant 0, \quad u(x) \leqslant C_1, \quad \forall\, x,$$

$$\int_0^1 [u_x(x)]^2\, dx \leqslant c_2, \quad |u_x(x)| \leqslant C_3,$$

$$\operatorname{var} u(\cdot) \leqslant C_4, \quad \text{т.е.} \int_0^1 |u_x(x)|\, dx \leqslant C_4, \quad \text{и т.д.}$$

Совокупность этих условий выделяет в пространстве U некоторое относительно узкое множество — компакт M. Компакт, напомним, — это ограниченное замкнутое множество, такое, что из всякой бесконечной последовательности $u_j \in M$ можно выбрать подпоследовательность, сходящуюся к $u^* \in M$. Сходимость предполагает какую-то норму — типа C или L_2, так что термин «компакт» нужно понимать в смысле этой нормы.

Рассмотрим множество $N = KM$, т.е. совокупность функций Ku, $\forall\, u \in M$. Это — узкое множество очень гладких функций: если M состоит из «хороших» функций, то N — из еще «лучших». Если бы мы рассматривали задачу

$$Ku = f, \quad f \in N,$$

то это была бы корректная задача. К сожалению, $f \notin N$. Будем исходить из предположения, что функция f состоит из двух слагаемых: $f = f_т + \delta f$, где $f_т = Ku_т$ ($u_т$ — точное решение задачи), причем $u_т \in M$ и, следовательно, $f_т \in N$. Что касается δf, то это плохая, негладкая функция. Она возникает, например, вследствие погрешностей измерения или математического представления. Поэтому $\|\delta f\| \leqslant \delta$, где δ — малое число, норма $\|\cdot\|$ — обычная (типа C или L_2).

К сожалению, достаточно объективно и однозначно разделить f на $f_т$ и δf, т.е. «отфильтровать» погрешности задания f, не удается. Одним из основоположников теории решения некорректных задач В. К. Ивановым было предложено искать вместо $u_т$ так называемое «квазирешение». Имеется в виду следующее. В компакте M (а это множество нам задано более или менее «явно» — системой неравенств) находится квазирешение u_q, такое, что расстояние между Ku_q и f минимально.

Другими словами, определение квазирешения сводится к задаче на условный экстремум:

$$\min_{u \in M} \|Ku - f\|. \tag{5}$$

Эта задача получила в настоящее время наименование «задача математического программирования». Методы ее численного решения достаточно хорошо развиты (см. § 26), хотя она и не относится к числу очень уж простых. Иногда употребляется обозначение

$$u_q = \arg\min_{u \in M} \|Ku - f\|. \tag{6}$$

Прежде всего следует установить существование квазирешения. Это почти очевидный факт: все определяется предположением о том, что M — компакт, и ограниченностью (непрерывностью) оператора K.

Теорема 1. Квазирешение существует.

Доказательство. Введем функцию $\Phi(u) \equiv \|Ku - f\|$. Непрерывная ограниченная снизу функция на компакте достигает своего наименьшего значения. Нужно только доказать непрерывность $\Phi(u)$. Оценим

$$\Phi(u + \delta u) = \|Ku + K\delta u - f\| \leqslant \|Ku - f\| + \|K\,\delta u\|.$$

В результате имеем

$$\Phi(u + \delta u) \leqslant \Phi(u) + \|K\|\,\|\delta u\|.$$

С другой стороны, аналогично,

$$\Phi(u) = \Phi(u + \delta u - \delta u) \leqslant \Phi(u + \delta u) + \|K\|\,\|\delta u\|,$$

т.е. $\Phi(u) - \Phi(u + \delta u) \leqslant \|K\|\,\|\delta u\|$. Следовательно,

$$|\Phi(u + \delta u) - \Phi(u)| \leqslant \|K\|\,\|\delta u\|.$$

Для того чтобы квазирешение представляло интерес с прикладной точки зрения, оно должно обладать важным свойством — непрерывной зависимостью от правой части f. Это свойство обеспечивается при дополнительных требованиях к компакту M.

Определение 2. Компакт M называется множеством корректности в смысле Тихонова, если существует такая функция скалярного аргумента $\omega(\xi)$, что:

а) $\lim\limits_{\xi \to 0} \omega(\xi) = 0$;

б) для любых двух элементов $u' \in M$ и $u'' \in M$ имеет место соотношение

$$\|u'' - u'\| \leqslant \omega(\|Ku'' - Ku'\|). \tag{7}$$

Определение 3. Задача $Ku = f$, $u \in M$ называется корректной в смысле Тихонова, если:

а) априори известно, что существует ее единственное решение;

б) компакт M является множеством корректности этой задачи.

Поясним смысл этих определений. Прежде всего подчеркнем, что в (7) используются обычные нормы, например нормы в L_2. Из определений следует, что если правые части f', f'' брать из множества $N = KM$, то соответствующие им решения u', u'' мало отличаются друг от друга при малом отличии f', f'': $\|u'' - u'\| \leqslant \omega(\|f'' - f'\|)$. Другими словами, «сужение» задачи на множество N делает ее корректной. Однако решению подлежит задача, в которой $f \notin N$.

Теорема 2. Пусть задача $Ku = f$ корректна в смысле Тихонова и M есть ее множество корректности. Тогда квазирешение непрерывно зависит от правой части f.

Уточним суть дела. Предполагается, что существует точное решение $u_т \in M$. Ему соответствует точная правая часть $f_т = Ku_т \in N$. Известна правая часть f, близкая к $f_т$ в обычной норме: $\|f - f_т\| \leqslant \delta$, где δ — малое число. Однако в сильной норме типа (3) погрешность в правой части бесконечна. Пусть найдено квазирешение u_q. Теорема утверждает, что $\|u_q - u_т\| \to 0$ при $\delta \to 0$. Перейдем к доказательству.

Доказательство. Введем $f_q = Ku_q \in N$ и оценим

$$\|f_q - f_т\| \leqslant \|f_q - f\| + \|f_т - f\|.$$

Второе слагаемое в правой части оценивается величиной δ. Оценим первое:

$$\|f_q - f\| = \|Ku_q - f\| = \min_{u \in M} \|Ku - f\| \leqslant \|Ku_т - f\| = \|f_т - f\| \leqslant \delta.$$

Итак, $\|f_q - f_т\| \leqslant 2\delta$ и в силу (7) имеем $\|u_q - u_т\| \leqslant \omega(2\delta)$, так как $f_q \in N$ и $f_т \in N$. Теорема доказана.

Множество корректности в задаче Адамара. Поясним технику построения множеств корректности в конкретной задаче (2). Удобно обозначить $\lambda_k = e^{k^2\pi^2 T}$. Это собственные значения оператора K^{-1}, если задачу (2) представить в виде $Ku = f$. Сначала покажем, что множество M функций $|u(x)| \leqslant 1$, $x \in [0, 1]$, множеством корректности не является. (Напомним, что искомая функция $u(x)$ служит начальными данными для обычной задачи теплопроводности, решение которой в момент времени T должно совпасть с заданной функцией $f(x)$.)

Возьмем $u'(x) \equiv 0$, $u''(x) = \sin(k\pi x)$. Им соответствуют функции из $N = KM$, $f' = 0$, $f'' = \lambda_k^{-1} \sin(k\pi x)$. Итак, $\|u'' - u'\| = 1$, $\|f'' - f'\| = e^{-k^2\pi^2 T}$. Нельзя построить никакой требуемой в определении 2 функции $\omega(\xi)$, такой, что

$$1 = \|u'' - u'\| \leqslant \omega(\|f'' - f'\|) = \omega(e^{-k^2\pi^2 T}).$$

Определим теперь множество M условием $\|u_x(x)\| \leqslant 1$ (кроме того, есть еще и условия $u(0) = u(1) = 0$). Рассмотрим функции f' и f'' из N. Пусть a_k, b_k — их коэффициенты Фурье. Обозначим расстояние между ними δ (ради простоты, ниже мы будем иметь дело с квадратами расстояний):

$$\sum_{k=1}^{\infty} (b_k - a_k)^2 = \delta^2. \tag{8}$$

Принадлежность f' и f'' множеству KM (т.е. $K^{-1}f \in M$) означает выполнение неравенств

$$\pi^2 \sum_{k=1}^{\infty} \lambda_k^2 k^2 a_k^2 \leqslant 1, \qquad \pi^2 \sum_{k=1}^{\infty} \lambda_k^2 k^2 b_k^2 \leqslant 1. \tag{9}$$

Функциям f' и f'' соответствуют элементы u', u'', коэффициентами Фурье которых являются числа $\lambda_k a_k$ и $\lambda_k b_k$. Расстояние между ними есть

$$\|u'' - u'\|^2 = \sum_{k=1}^{\infty} \lambda_k^2 (b_k - a_k)^2. \tag{10}$$

Коэффициенты Фурье производных u_x', u_x'' суть $k\lambda_k a_k$, $k\lambda_k b_k$. Неравенства (9) выражают принадлежность f', f'' множеству $N = KM$.

Оценим (10), используя (8), (9). Выберем некоторое число m и воспользуемся следующей из (9) оценкой

$$\lambda_k^2 a_k^2 \leqslant 1/(\pi^2 k^2), \qquad \lambda_k^2 b_k^2 \leqslant 1/(\pi^2 k^2).$$

Имеем

$$\sum_{k=1}^{\infty} \lambda_k^2 (b_k - a_k)^2 \leqslant \sum_{k=1}^{m} \lambda_k^2 (b_k - a_k)^2 + 2 \sum_{k=m+1}^{\infty} \lambda_k^2 (a_k^2 + b_k^2). \tag{11}$$

Второе слагаемое в правой части (11) оценим так:

$$\sum_{k=m+1}^{\infty} \lambda_k^2 (a_k^2 + b_k^2) \leqslant \frac{2}{\pi^2} \sum_{k=m+1}^{\infty} \frac{1}{k^2} \leqslant \frac{2}{\pi^2} \frac{1}{m}.$$

Первое слагаемое оценим, используя (8):

$$\sum_{k=1}^{\infty} \lambda_k^2 (b_k - a_k)^2 \leqslant \lambda_m^2 \sum_{k=1}^{m} (b_k - a_k)^2 \leqslant \lambda_m^2 \delta^2. \tag{12}$$

Получаем $\|u'' - u'\|^2 \leqslant e^{2m^2\pi^2 T}\delta^2 + \frac{4}{\pi^2}\frac{1}{m}$.

Теперь нужно распорядиться числом m так, например, чтобы оба слагаемых в оценке были равны. Логарифмируя выражение $e^{2m^2\pi^2T}\delta^2 = 4/(\pi^2 m)$, приходим к уравнению для m:

$$\varphi(m) \equiv \alpha m^2 + \ln(\delta^2\pi^2/4) + \ln m = 0, \quad \alpha = 2\pi^2 T. \tag{13}$$

Будем ориентироваться на задачу с $T = 0.01$, $\delta \approx 10^{-2}$. В этом случае $\alpha \approx 0.2$, $\ln\{\delta^2\pi^2/4\}) \approx -7.5$. В первом приближении можно отбросить в формуле (13) для φ третий член, после чего уравнение решается: $m_1 = [-\alpha^{-1}\ln(\delta^2\pi^2/4)]^{1/2}$ (в примере $m_1 \approx 6$).

Однако это слишком грубый результат. При таком выборе m первый член в оценке (12) для $\|u'' - u'\|$ есть $O(1)$. Полученная оценка не дает права утверждать, что M есть множество корректности. Уточним корень уравнения $\varphi(m) = 0$ одной итерацией по Ньютону: $m_2 = m_1 - \varphi(m_1)/\varphi'(m_1)$. Очевидно,

$$\varphi(m_1) = \ln m_1, \qquad \varphi'(m_1) = 2\alpha m_1 + 1/m_1.$$

Так как мы рассматриваем все-таки значения $\delta \ll 1$, то $m_1 \gg 1$ и можно упростить выкладки, полагая $m_2 = m_1 - \ln m_1/(2\alpha m_1)$.

Оценим первое слагаемое в (12), используя приближение m_2:

$$\exp(\alpha m_2^2) = \exp\left[\alpha\left(m_1^2 - \frac{1}{2\alpha}\ln m_1 + \beta\right)\right],$$

где $\beta = [\ln m_1/(2\alpha m_1)]^2$ — величина, пренебрежимо малая. Мы не будем доводить оценки до абсолютной строгости — это дело техники, не очень сложной, но громоздкой. Итак, имеем

$$\exp(\alpha m_2^2)\,\delta^2 \approx \exp(\alpha m_1^2)\,\delta^2 \exp\left(-\frac{1}{2\alpha}\ln m_1\right) =$$

$$= \frac{4}{\delta^2\pi^2}\,\delta^2 \exp\left[-\frac{1}{2\alpha}\ln\left(-\frac{1}{\alpha}\ln\frac{\delta^2\pi^2}{4}\right)\right].$$

Таким образом, опуская несущественные детали, мы получили оценку типа

$$\|u'' - u'\|^2 \approx (1/\ln\delta^{-1})^{1/2\alpha}.$$

Тем самым доказано, что условие $\|u_x(\cdot)\| \leqslant C$ определяет множество корректности для обратной задачи теплопроводности. Однако очень медленное стремление к нулю соответствующей функции $\omega(\xi)$ при $\xi \to 0$ (см. определение 2) служит предостережением тем, кто на этом основании счел бы исследование задачи (2) законченным.

Совершенно ясно, что множество функций $u(x) = \sum_{k=1}^{m} a_k \sin(k\pi x)$ при условии $\sum_{k=1}^{m} a_k^2 \leqslant A$ и любом заданном m является множеством корректности для задачи (2). Более того, соответствующая функция $\omega(\xi) = C\xi$, где $C = e^{\pi^2 m^2 T}$. Это настолько большая величина, что реально можно использовать соответствующее множество M лишь при очень малых $m = 2, 3, 4$ и в том случае, когда есть уверенность, что искомое точное решение может быть аппроксимировано с нужной точностью тремя-четырьмя гармониками.

Приближенное решение обратной задачи теплопроводности. Дальнейшее знакомство с некорректными задачами удобно провести в более конкретной форме — в виде комментария к процессу приближенного решения модельной задачи обратной теплопроводности. Она конструируется просто. Возьмем относительно простую функцию $u_\tau(x)$ и решим прямую задачу теплопроводности $u_t = u_{xx}$ при краевых условиях $u(t, 0) = u(t, 1) = 0$ с начальными данными $u(0, x) = u_\tau(x)$. Полученную (численно) функцию $u(T, x)$ используем как начальные данные для обратной задачи. Полезно еще возмутить ее малой случайной погрешностью. Итак, построим функцию

$$f(x) = u(x, T) + \delta(x), \quad |\delta(x)| \leqslant \delta.$$

Теперь попытаемся решить обратную задачу. При построении модели надо достаточно разумно выбрать два числа: T и δ (уровень погрешностей). Значение δ выбирается из таких соображений. Если найдена функция $u(T, x)$, то в качестве δ можно взять, например, 0.01 среднего значения u. В дальнейшем, говоря о решении уравнения теплопроводности, мы имеем в виду приближенное решение, получаемое методом сеток с шагом $h = 0.01$ (по x) и с шагом $\tau \approx h^2$ (по t) при использовании самой простой, например явной, схемы.

Что касается T, то обсуждаемые ниже расчеты проводились при $T = 0.01$. Этот выбор может удивить читателя, но для рассматриваемой задачи время 0.01 не такое уж малое. Оценим, какие события могут произойти в задаче за это время. Если разложить начальные данные прямой задачи в ряд Фурье: $u_\tau(x) = \sum c_k \sin(k\pi x)$, то решение в момент времени $t = T$ есть

$$u(x, T) = \sum c_k e^{-k^2\pi^2 T} \sin(k\pi x).$$

В табл. 13 приведены значения $e^{-k^2\pi^2 T}$ для разных T и k. Видно, что время $T = 0.1$ является почти асимптотически большим. За это время из всех гармоник, входивших в начальные данные, «выжива-

Таблица 13

T \ k	1	2	3	4	5	6	7	8
0.01	0.9	0.7	0.4	0.2	0.08	0.03	0.007	0.002
0.02	0.8	0.5	0.2	0.04	0.006	$9\cdot10^{-4}$	$5\cdot10^{-5}$	$3\cdot10^{-6}$
0.1	0.4	0.02	10^{-4}	10^{-7}	10^{-11}	10^{-16}		

ют» только две первых. За время $T = 0.01$ в системе, описываемой уравнением теплопроводности, происходят достаточно сложные события. Нетрудно сообразить, что функции $u(0, x)$ и $u(T, x)$ достаточно сильно отличаются друг от друга. К обсуждению приведенных в табл. 13 значений мы вернемся чуть позже. На рис. 44 показаны функции $u(0, x)$ и $u(T, x)$. Возмущенная функция $f(x) = u(T, x) + \delta(x)$ при $\delta = 0.015$ в таком масштабе не отличается от $u(T, x)$.

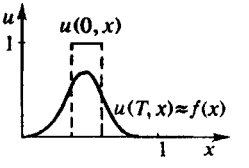


Рис. 44

Итак, нам задана функция $f(x)$ и мы предполагаем, что для обратной задачи $f(x)$ отличается от точных «начальных данных» на величину, не превосходящую 0.015. Решение обратной задачи теплопроводности будем искать как решение задачи математического программирования. Требуется найти функцию $v(x)$ («начальные данные прямой задачи»), такую, чтобы:

а) значение $\int_0^1 q(x)v(x)\,dx$ было минимальным;

б) $\max_x |v(T, x) - f(x)| \leqslant \delta$;

в) $\operatorname{var} v(\cdot) \leqslant W$.

Здесь $v(T, x)$ — решение прямой задачи теплопроводности с начальными данными $v(x)$. Речь идет о том, чтобы «подобрать» начальные данные прямой задачи так, чтобы ее решение в момент времени T попало в «коридор» шириной δ около заданной функции $f(x)$.

Мы знаем (по постановке задачи), что искомый ответ $u_т(x)$ порождает решение, попадающее в тот же коридор, и не собираемся извлекать из $f(x)$ более точной информации. Мы отступили от рекомендаций метода квазирешений, согласно которому следовало бы выбирать функцию $v(x)$ такой, чтобы минимизировать $\|v(T, \cdot) - f(\cdot)\|$. Оба

подхода имеют свои основания. Если оценка погрешности реальна, то, видимо, полезной информацией является упоминавшийся выше коридор. Если же (а такое тоже часто бывает) оценка погрешности сильно завышена и фактическая погрешность может быть существенно меньшей, функция $f(x)$ несет бо́льшую информацию, чем δ-коридор около нее, и рекомендации метода квазирешений будут предпочтительнее.

Перейдем к обсуждению условия в), в котором величина W — это априорная оценка вариации искомого ответа. Если мы ограничимся только условием б), решение будет принципиально неединственным и эта неединственность будет очень сильной. Прообраз «коридора» в отображении начальных данных прямой задачи в решение при $t = T$ состоит из множества функций, сильно отличающихся друг от друга. Можно взять какую-то функцию, отображающуюся в коридор, добавить к ней $A \sin(k\pi x)$ с большим значением A при, соответственно, большом k. Эти новые начальные данные отображаются, в сущности, в тот же коридор. Методы решения таких обратных некорректных задач основаны на том, что в постановку задачи вводится качественная информация об искомом решении, которая, так сказать, отсекает высокие гармоники.

Другими словами, мы рассматриваем пересечение прообраза коридора с множеством достаточно хороших функций, таких, каким, по имеющимся априорным сведениям, решение могло бы быть (множество корректности по Тихонову). В данном случае мы задаем это множество ограничением вариации искомой функции — числом W. При этом, как нетрудно понять, отсекаются функции с очень частыми колебаниями (при заметной их амплитуде), но не отсекаются разрывные функции. И если последние по каким-то причинам нужно оставить (априорные данные о решении не исключают разрывных функций), то такой способ очень удобен. Правда, с вычислительной точки зрения он достаточно сложен: приходится использовать непростые алгоритмы решения задачи математического программирования. Конечно, в реальных задачах точное значение W едва ли известно (ниже мы еще вернемся к обсуждению роли этой величины в процессе решения). Обозначим: $M(W)$ — множество функций с вариацией, не превосходящей W, D — полный прообраз коридора. Итак, пока на роль решения может претендовать любая функция из множества $\pi = D \cap M(W)$.

Обсудим теперь роль условия а), где $q(x)$ — «произвольная» функция. Формально это условие выбирает из возможных претендентов на роль решения какое-то одно, определяемое заданием $q(x)$. Само множество π может оказаться настолько широким, что решение задачи с «точностью до π» не всегда имеет смысл. При решении обратной некорректной задачи обычно имеется (хотя часто явно не фигурирует в постановке задачи) некоторое представление о том, с какой погрешностью ε (в подходящей

норме) необходим ответ. Если все функции, входящие в π, различаются не более чем на ε, любой элемент этого множества может считаться решением задачи. Если же в нем имеются элементы, различающиеся на величину, существенно бо́льшую ε, следует признать априорную информацию, используемую в решении, недостаточной и отказаться (при таком уровне априорной информированности) от решения задачи. Однако множество претендентов на звание «решения» описано не очень эффективно: его не так-то просто просмотреть и оценить. Для того чтобы иметь хоть какую-то информацию о нем, и вводится условие а).

Решая задачу с разными q (например, $q_1(x) = 1$, $q_2(x) = -1$ и т.п.), мы будем получать различные крайние точки множества π, что позволит составить хоть какое-то представление о его размерах. Полезным при этом является предъявление «физику» различных функций, каждая из которых может быть решением. Конкретные примеры возможных решений задачи часто содержат внешние

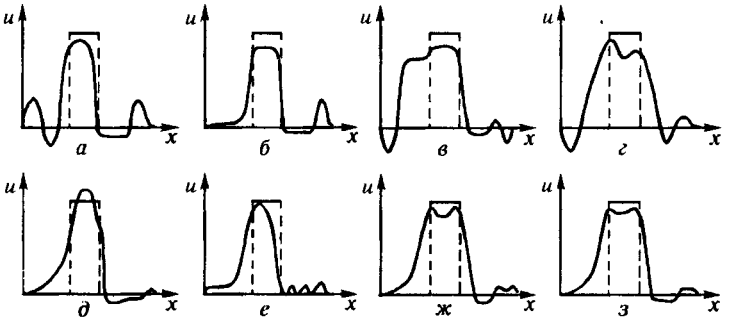


Рис. 45

дефекты, которых, по мнению физика, настоящее решение не должно иметь. В этом случае могут быть сформулированы дополнительные требования к решению, которые включаются в постановку задачи, решается новая, более сложная задача, и т.д. Опыт показывает, что такое извлечение априорной информации из «физика» успешнее проходит при предъявлении ему «решений», удовлетворяющих всем сформулированным им требованиям, но отвергаемого по каким-то интуитивным соображениям.

Перейдем к обсуждению результатов решения задачи. На рис. 45 представлены восемь функций, полученных с помощью достаточно сложного алгоритма (см. § 28). Они соответствуют разным значениям параметров, входящих в определение множества M, т.е. в постановку задачи математического программирования. Таких параметров четыре: $q(x)$, W, δ и V, входящий в условие

$v(x) \geqslant V$. Таблица 14 содержит значения параметров, соответствующие решениям, представленным на рис. 45.

Вышеприведенный вычислительный эксперимент преследует несколько целей. Многие параметры, входящие в априорную информа-

Таблица 14

	а	*б*	*в*	*г*	*д*	*е*	*ж*	*з*
$q(x)$	1	1	−1	−1	−1	−1	−1	−1
W	3.2	2.2	3.0	3.0	3.0	2.2	1.8	1.8
δ	0.015	0.015	0.015	0.0075	0.015	0.015	0.015	0.015
V	$-\infty$	$-\infty$	$-\infty$	$-\infty$	0	0	0	$-\infty$

цию, не имеют точных значений. Полезно знать, как их изменение влияет на квазирешение. Можно предположить, что каждое «решение» содержит как объективную информацию, так и случайную, связанную с неопределенностью описания M. Вторая компонента численного решения, видимо, наиболее чувствительна к вариации параметров. Рисунок 45 это предположение, кажется, подтверждает. На рис. 46 показаны четыре первых решения одновременно. Такое представление позволяет оценить, что в разных решениях является устойчивым, а что — случайным. Наконец, подобные расчеты позволяют более объективно оценить точность решения некорректной задачи при той априорной информации, которая была использована.

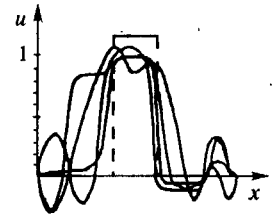


Рис. 46

Метод квазиобращения. Французским математиком Ж. Л. Лионсом был предложен гораздо более простой способ решения обратной задачи теплопроводности. Предлагается решать задачу Коши

$$-\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} - \varepsilon \frac{\partial^4 u}{\partial x^4}, \quad u(0, x) = f(x), \tag{5}$$

и решение $u(T, x)$ считать теми начальными данными, которые при решении прямой задачи через время T дают результат, близкий к заданной функции $f(x)$. Механизм, обеспечивающий «устойчивость» этой процедуры, т.е. отсечение высоких гармоник, почти очевиден.

Разлагая $f(x)$ в ряд Фурье, выпишем решение (5):

$$u(t, x) = \sum c_k e^{k^2\pi^2 t - \varepsilon k^4 \pi^4 t} \sin(k\pi x). \tag{6}$$

Достаточно высокие гармоники (при $\varepsilon k^2\pi^2 \gg 1$) оказываются уже сильно затухающими, решение обратной задачи получается фор-

мально гладким. Но ε — малый параметр, влияние которого имеет противоречивый характер. При слишком малом ε решение содержит быстро растущие компоненты и некорректность практически сохраняется. При слишком большом ε решение сильно искажается. Что можно получить на этом пути, заранее сказать нельзя, надо оценивать соотношение этих факторов.

Проведем такую оценку в первом приближении, заметив, что безусловно сильной стороной этого подхода является его простота. Задача сводится к решению (методом сеток, например) уравнения, немногим более сложного, чем уравнение теплопроводности. Вычислительная цена такого решения неизмеримо ниже цены тех вычислений, которые были обсуждены выше. Но это достоинство ничего не стоит, если результаты окажутся неудовлетворительными. Приступим к оценкам.

Имеем функцию $u_т(x)$ (точное, искомое решение). Разложим ее в ряд Фурье:

$$u_т = \sum c_k \varphi_k, \quad \varphi_k \equiv \sin(k\pi x).$$

С начальными данными $u_т(x)$ решим прямую задачу теплопроводности. При $t = T$ имеем

$$f_т(x) \equiv u(T, x) = \sum_k c_k e^{-k^2\pi T} \varphi_k.$$

Внесем в $f_т$ «погрешности измерения» и превратим ее в функцию

$$f(x) = \sum (c_k e^{-k^2\pi^2 T} + \delta_k) \varphi_k.$$

Решая регуляризованную обратную задачу теплопроводности (5) с начальными данными $f(x)$ и параметром ε, получаем

$$u^*(x) = \sum (c_k e^{-k^2\pi^2 T} + \delta_k) e^{\pi^2 k^2 (1 - \varepsilon\pi^2 k^2) T} \varphi_k =$$

$$= \sum c_k \varphi_k - \sum (1 - e^{-\varepsilon\pi^4 k^4 T}) c_k \varphi_k + \sum \delta_k e^{\pi^2 k^2 (1 - \varepsilon\pi^2 k^2) T} \varphi_k.$$

В решении задачи u^* выделены три основные компоненты: точное решение $u_т(x)$, погрешность регуляризации (вторая сумма) и «погрешность некорректности» (третья сумма). Положим, для определенности, $\|u_т\| = 1$, $T = 0.01$, $\delta = 0.01$ и вычислим при разных ε значения

$$r_k = 1 - e^{-\varepsilon\pi^4 k^4 T}, \qquad q_k = 0.01 e^{\pi^2 k^2 (1 - \varepsilon k^2 \pi^2) T}.$$

Они имеют простой содержательный смысл. Если, например, $u_т = \varphi_k$, то в функции $u^*(x)$ погрешность регуляризации достигает величины r_k. Если функция $f_т(x)$ возмущена только членом $\delta_k \varphi_k$, то он дает вклад в погрешность восстановления порядка q_k.

В табл. 15 представлены значения этих величин. При $k \to \infty$ величина $r_k \to 1$, при $k \approx 1/\sqrt{2\pi^2\varepsilon}$ величина q_k достигает максимального значения $\delta e^{T/4\varepsilon}$.

Прокомментируем табл. 15, приняв не очень высокие требования к точности восстановления $u_т$ (будем считать допустимой погрешность порядка 10 %).

При $\varepsilon = 5 \cdot 10^{-3}$ погрешность некорректности q_k достаточно мала, но погрешность регуляризации r_k такова, что удовлетворительный

Таблица 15

ε	k	1	2	3	4	5	6	7	8	9	10	11
$5 \cdot 10^{-3}$	r	0.005	0.074	0.32	0.71	0.95						
	q	0.01	0.014	0.016	0.014	0.006						
10^{-3}	r	0.001	0.015	0.076	0.22	0.46	0.71	0.90				
	q	0.01	0.015	0.022	0.038	0.064	0.10	0.12	0.10			
$7.5 \cdot 10^{-4}$	r	$7 \cdot 10^{-4}$	0.01	0.06	0.17	0.36	0.61	0.82				
	q	0.01	0.015	0.023	0.04	0.075	0.14	0.22	0.28	0.25		
$5 \cdot 10^{-4}$	r	$4 \cdot 10^{-4}$	0.008	0.038	0.11	0.26	0.46	0.69	0.86	0.96		
	q	0.01	0.015	0.023	0.042	0.09	0.19	0.39	0.75	1.21	1.5	1.2

результат получается лишь в случае, когда функция $u_т$ состоит из двух первых гармоник.

При $\varepsilon = 10^{-3}$ погрешность некорректности q_k достигает 10 %, функция $u_т$ может состоять из двух-трех первых гармоник.

При $\varepsilon = 7.5 \cdot 10^{-4}$ функция $u_т$ может состоять из трех первых гармоник, но погрешность некорректности q_k может достигать 25 %.

При $\varepsilon = 5 \cdot 10^{-4}$ функция $u_т$ может состоять из четырех первых гармоник, но погрешность некорректности q_k может все испортить.

И т.д.

Заметим, что вышеприведенные результаты существенно связаны со значением $\delta = 0.01$. При меньших δ можно в принципе использовать и меньшие ε, однако, например, при $\varepsilon = 10^{-4}$ максимальное значение $\max q_k \approx q_{23} \approx \delta e^{25} \approx \delta \cdot 10^{11}$, так что далеко здесь не продвинешься. Правда, есть еще один резерв регуляризации — проведение расчетов в пространстве, базис в котором составляют несколько первых гармоник ($k = 10$, например). Этот резерв

неявно используется при решении задачи методом сеток, когда берутся сетки из $10 \div 20$ узлов. Вывод из проведенного обсуждения почти очевиден: возможности метода квазиобращения, видимо, достаточно ограничены, и пользоваться им следует осторожно.

§ 26. Поиск минимума

Приведем общие сведения о методах решения так называемой *задачи математического программирования*. Это название в современной литературе присвоено задаче на условный экстремум. Общая ее формулировка такова. Требуется найти точку x (из R^N), минимизирующую значение функции f^0:

$$\min_x f^0(x), \tag{1}$$

при условиях

$$\text{а)}\ X_n^- \leqslant x_n \leqslant X_n^+, \quad n = 1, 2, \ldots, N, \tag{2}$$

$$\text{б)}\ F_i^- \leqslant f^i(x) \leqslant F_i^+, \quad i = 1, 2, \ldots, I.$$

Здесь $f^i(x)$ — заданные функции, которые, если не оговорено иное, предполагаются гладкими (например, имеющими вторые производные); X_n^-, X_n^+, F_i^-, F_i^+ — заданные числа.

Начнем обзор методов решения с простейших вариантов этой общей задачи.

Поиск безусловного минимума. Имеется в виду задача (1). Никаких условий и ограничений на диапазон изменения x нет. Конструкции алгоритмов решения этой задачи основаны на идеях, которые, соответственно усложняясь и модифицируясь, используются и при решении общей задачи. Основная идея заключается в построении минимизирующей последовательности точек x^j. Начиная с некоторой заданной точки x^0 (начального приближения) строят последовательность точек x^1, x^2, x^3, ... таким образом, чтобы значение f^0 монотонно понижалось:

$$f^0(x^{j+1}) < f^0(x^j), \quad \lim_{j\to\infty} f^0(x^j) = \min f^0(x).$$

Эта общая идея конкретизируется построением алгоритма «улучшения» текущей точки x^j: если она не является точкой минимума, в ее окрестности должна найтись другая точка x^{j+1}, в которой $f^0(x^{j+1}) < f^0(x^j)$. Есть несколько способов найти такую точку.

Метод покоординатного спуска. Точка x^{j+1} ищется в виде $x^j + se_k$, где e_k — k-й орт в пространстве R^N. Скалярный параметр s («шаг спуска») определяется задачей того же типа:

$$\min_s f^0(x^j + se_k).$$

Ее решение (так называемый *линейный поиск*) осуществляется специальными алгоритмами (разумеется, приближенно), они описаны в специальной литературе. Что касается индекса k, то он меняется на каждом шаге j, циклически пробегая значения 1, 2, ..., N.

Алгоритм достаточно прост, но возникает вопрос: к чему он сходится, действительно ли он позволяет отыскивать точку минимума? Мы не будем рассматривать ситуации, когда точки x^j уходят в бесконечность, когда $f(x^j) \to -\infty$ и т.п. Предположим, что все x^j остаются в ограниченной области пространства R^N. Следовательно, имеется предел $\lim x^j = x^*$ (либо предел какой-то подпоследовательности x^{j_i}). Что можно сказать о такой точке x^*? Очевидно, она является стационарной точкой метода, т.е. если задать x^* в качестве x^0, то попытки переместиться из нее по любому из ортов e_k ни к чему не приведут.

Очевидно, стационарными для метода покоординатного спуска являются точки, в которых

$$\partial f^0(x^*)/\partial x_k = 0, \quad \partial^2 f(x^*)/\partial x_k^2 > 0, \quad k = 1, 2, \ldots, N.$$

Однако точка x^* может и не быть точкой минимума, даже локального; она может быть точкой перегиба. Если метод приведет в такую точку, процесс изменения x^j прекратится. Однако вероятность попасть в подобную точку не очень велика, так как в ее окрестности есть точки со значениями $f(x) < f(x^*)$, и если хоть одна точка x^j именно такова, то в дальнейшем точки x^{j+1}, ... не приблизятся к x^*.

Наиболее вероятным результатом описанного процесса является сходимость последовательности x^j к точке *локального минимума* $f^0(x)$. Подчеркнем — именно локального, а не точного, «глобального» минимума. Если функция $f^0(x)$ имеет несколько точек локального минимума, результат, естественно, зависит от выбора стартовой точки x^0.

Каждая точка локального минимума x^* имеет свою «область притяжения» — совокупность точек x^0, начиная с которых процесс спуска приводит именно к точке x^*. Это относится и ко всем остальным методам построения минимизирующих последовательностей. Они отличаются друг от друга в первую очередь способом построе-

ния направлений спуска. Легко понять, что для таких методов области притяжения к той или иной точке локального минимума практически одинаковы.

Метод спуска по градиенту. Более эффективным является метод, отличающийся от описанного выше только выбором направления спуска. В каждой точке x^j вычисляется градиент $f_x^0(x^j)$, и следующая точка ищется как точка минимума функции f^0 на луче $x(s) = x^j - sf_x^0(x^j)$. Очевидно, множество стационарных точек процесса здесь шире — это все точки, в которых $f_x^0(x) = 0$. Однако наиболее вероятным исходом является сходимость x^j к точке локального минимума.

Метод спуска по градиенту можно получить, применяя к задаче одну из самых плодотворных в вычислительной математике конструкций — линеаризацию в окрестности текущего приближения и решение последовательности линейных задач (вспомним в связи с этим метод Ньютона). Кстати, мы не доказываем теорем о сходимости методов спуска, так как они дословно повторяли бы доказательство сходимости модифицированного метода Ньютона (см. § 1). Итак, в точке x^j найдем $x^{j+1} = x^j + \delta x$, где поправка δx является решением линеаризованной задачи

$$\min_{\delta x} \{f^0(x^j) + f_x^0(x^j)\,\delta x\}, \quad \|\delta x\| \leqslant \varepsilon. \tag{3}$$

Ограничение $\|\delta x\|$ необходимо, чтобы избежать бесконечного решения.

Решение легко находится методом множителей Лагранжа. Формируем функцию Лагранжа

$$\mathscr{L}(\delta x, \lambda) = f^0(x^j) + f_x^0(x^j)\,\delta x + 0.5\,\lambda(\delta x, \delta x)$$

с неопределенным пока множителем λ и ищем точку ее минимума по δx. Задача решается просто. Приравнивая нулю производную по δx, находим $\delta x(\lambda) = -(1/\lambda) f_x^0(x^j)$. Множитель λ определяется условием $(\delta x(\lambda), \delta x(\lambda)) = \varepsilon^2$. Теперь можно использовать δx двумя способами: либо считать δx направлением спуска и определять $x^{j+1} = x^j + s\delta x$ после решения задачи минимизации по s, либо определять $x^{j+1} = x^j + \delta x$.

В первом случае величина ε, очевидно, никакой роли не играет. Этот способ надежный, но требует нескольких дополнительных вычислений f^0 для определения s. Второй способ более экономный, но величину ε надо назначать очень ответственно: она должна быть достаточ-

но мала, чтобы линейная аппроксимация $f^0(x+\delta x) \approx f^0(x) + f^0_x\,\delta x$ была достаточно точной. Однако ε не должно быть слишком малой величиной, чтобы «движение» x^j проходило не слишком малыми шагами. В своей работе автор обычно использовал второй способ (в сложных задачах вычисление f^0 часто оказывается одной из наиболее дорогих операций).

Для определения ε используется алгоритм адаптации. Сначала ε назначается из каких-то грубых соображений. После очередного шага сравнивается приращение $\Delta f^0 = f^0(x^{j+1}) - f^0(x^j)$ с вариацией $\delta f = f_x(x^j)\,\delta x$. Если они совпадают с высокой точностью, значение ε, соответственно, увеличивается; если совпадение плохое, — уменьшается. (Обычно увеличение и уменьшение ε осуществляются умножением на числа, не сильно отличающиеся от единицы. В дальнейшем мы подробнее обсудим эти вопросы в более сложной ситуации.) Если $\Delta f^0 > 0$, происходит «возврат» в точку x^j и δx вычисляется заново после пересчета $\varepsilon := \varepsilon/2$, например.

Метод случайного спуска. Он отличается от описанных выше тем, что в качестве направления движения выбирается «случайное» направление, т.е. единичный вектор e, генерируемый каким-либо датчиком случайных векторов, равномерно распределенных на единичной сфере в R^N (такие датчики входят в состав математического обеспечения современных ЭВМ). «Почти любое» направление e является направлением «спуска», если, конечно, рассматривать как положительные, так и отрицательные значения s. Стационарными точками процесса построения минимизирующей последовательности являются только точки локального минимума f^0.

Эффективность методов спуска. «Овраги». Задача поиска минимума гладкой функции с общематематической точки зрения является одной из простейших. Основная идея решения (построение минимизирующей последовательности) очевидна, да и конструктивная ее реализация не очень сложна. Проблема существования решения и сходимости процесса решается тоже не очень сложными средствами. Ответ дают такие простые факторы, как непрерывность и принадлежность всех точек x^j некоторому компакту. Поэтому сведение какой-либо задачи к поиску $\min f^0(x)$ на компакте справедливо считают почти исчерпывающим ее решением.

Многие сложные задачи естествознания и техники стремятся оформить именно как вариационные задачи. Однако внешняя простота решения обманчива. Дело в том, что задача проста, если она решается «в принципе»: на уровне доказательства сходимости. Но пока не обсуждался важнейший фактор — эффективность процесса поиска минимума, количество вычислений функции f^0, которое по-

надобится для определения $\min f^0(x)$ с какой-то (часто не очень высокой) степенью точности.

Продолжающееся до сих пор конструирование алгоритмов поиска минимума имеет основной целью повышение их эффективности и надежности. Такая работа должна опираться на достаточно четкую теоретическую концепцию, объясняющую причины возможной крайне низкой скорости убывания $f^0(x^j)$. Конечно, одной из причин этого может быть существенная фактическая негладкость f^0, даже если формально она имеет сколько угодно непрерывных производных. С точки зрения вычислителя гладкость — это не число существующих производных, а константы, ограничивающие их значения. Если эти константы просто «конечны», то нет особой разницы между классами гладких и негладких функций. Эту причину (существенную негладкость f^0) оставим пока в стороне. Ведь отношение вычислителя к тем или иным методам определяется не столько их способностью решать задачу данного типа в ее общей формулировке, сколько эффективностью метода в классе тех задач, которые нуждаются в фактическом решении.

Итак, в каком случае методы спуска оказываются эффективными, а в каком нет? Этот вопрос сейчас изучен достаточно полно. Основной моделью, на которой получаются точные результаты, является класс квадратичных функций $f^0(x) = (a, x) + 0.5(Ax, x)$ с положительно-определенной самосопряженной матрицей A. В окрестности точки минимума (слово «локальный» будем для краткости опускать) гладкая функция f^0 хорошо аппроксимируется именно квадратичной функцией. Матрица A в данном случае аналогична матрице $\partial^2 f^0/\partial x_i \partial x_j$, именуемой *гессианом*. Существенным фактором, определяющим эффективность метода спуска по градиенту, является обусловленность матрицы A, т.е. отношение $\eta = l/L$, где L и l — максимальное и минимальное собственные числа A.

Расстояние $\|x^j - x^*\|$ убывает, как q^j, где $q = (L - l)/(L + l)$. При малых значениях η имеем $q \approx 1 - 2\eta$. Чем меньше η, тем медленнее осуществляется поиск минимума. Число η имеет простую геометрическую интерпретацию: линии уровня квадратичной функции (при $A > 0$) суть «эллипсоиды», отношение экстремальных полуосей которых как раз и есть η. Таким образом, «трудная» функция $f^0(x)$ — это функция, график которой похож на «овраг» с крутыми «склонами» и очень длинным пологим «дном», вдоль которого нужно очень долго идти до точки минимума.

Первые шаги процессов поиска приводят к быстрому спуску со «склона» на дно оврага, после чего начинается длительное «зигзагообразное» движение вдоль «дна» с очень медленным темпом убывания f^0 за шаг. При $\eta = 1$ (линии уровня — сферы) метод спуска по

градиенту приводит к минимуму за один шаг. Скорость сходимости (число q) покоординатного спуска в этом случае легко оценить. Предоставим это поучительное упражнение читателю (здесь интересна зависимость q от размерности пространства).

Несколько сложнее оценивается математическое ожидание убывания f^0 (при $\eta = 1$) за один шаг спуска в случайном направлении. Здесь существенна размерность, причем сказывается следующее неприятное обстоятельство: почти вся площадь сферы в R^N сосредоточена в узком поясе около экватора (это свойство сфер проявляется тем резче, чем больше N). Поэтому «почти любое» случайное направление «почти ортогонально» направлению градиента, т.е. направлению в точку минимума такой простой функции, как $\sum x_n^2$. (Предоставим читателю эти вычисления. Они не составят труда для того, кто знает формулу площади многомерной сферы.)

Эффективность процесса поиска минимума можно существенно повысить линейным преобразованием пространства, т.е. используя замену $x = By$. Легко понять, какой должна быть матрица B: квадратичная форма после замены переменных перейдет в $(ABy, By) = (B^*ABy, y)$. Чтобы в переменных y получить просто сумму квадратов, следует найти B из уравнения $B^*AB = E$, например. В качестве B можно взять $A^{-1/2}$. Это, конечно, рецепт чисто теоретический. Он только указывает направление, в котором следует искать B: ведь можно брать матрицы B, близкие к идеальной, но более доступные на практике. Приведенные выше соображения можно трактовать несколько иначе. В линеаризованной задаче ограничение δx можно сформулировать в какой-то другой метрике, например $(B\,\delta x, \delta x) \leqslant \varepsilon^2$ с положительной самосопряженной матрицей B. В этом случае методом Лагранжа найдем $\delta x \approx B^{-1} f_x^0(x^j)$.

Квадратичная модель подсказывает идеальный выбор B: это должна быть матрица, близкая к гессиану A. Практическая реализация такой подсказки возможна двумя способами. Самый очевидный — использовать метод, основанный на квадратичной аппроксимации:

$$f^0(x + \delta x) \approx f^0(x) + f_x^0\,\delta x + 0.5(f_{xx}^0\,\delta x, \delta x).$$

В очередной точке x^j следует вычислить $f_x^0(x^j)$ и гессиан $f_{xx}^0(x^j)$, решить задачу минимизации квадратичной функции. Например, если размерность пространства N не очень велика, можно решить систему линейных уравнений $f_x^0(x^j) + f_{xx}^0(x^j)\,\delta x = 0$. Такой алгоритм иногда называют методом Ньютона, так как его можно трактовать как решение системы нелинейных уравнений $f_x^0(x) = 0$.

Применение вышеприведенной схемы вычислений сталкивается с двумя препятствиями. В начальной точке x^0 гессиан f_{xx}^0 может не

быть положительно-определенным. Тогда решение задачи минимизации квадратичной формы (если мы ее действительно минимизируем) уводит нас в бесконечность. Если же решается система линейных уравнений (необходимое условие экстремума квадратичной формы), то мы уже не отличаем минимума от максимума и от другого типа стационарных точек. Поэтому такая техника применяется после некоторого числа шагов спуска по градиенту, которые проходят достаточно эффективно и выводят точку x^j в область положительности гессиана.

Более серьезным препятствием является необходимость вычисления вторых производных. В пространствах не очень малой размерности это очень дорогая операция. В семидесятых годах был найден удачный компромисс, приведший к созданию так называемых квазиньютоновских процедур. Они основаны на следующем соображении. В методе спуска по градиенту, располагая значениями градиента в разных точках $f_x(x_j)$, мы получаем некоторую ограниченную на каждом шаге информацию и о f_{xx}. В самом деле, если смещение $\|x^{j+1} - x^j\|$ не очень велико,

$$f_x(x^{j+1}) - f_x(x^j) \approx f_{xx}(x^{j+1} - x^j),$$

т.е. если пренебречь величинами $O(\|x^{j+1} - x^j\|^2)$, мы знаем величины N линейных комбинаций из элементов $N \to N$ матрицы f_{xx}. Накапливая такую информацию на нескольких подряд идущих шагах, можно с какой-то точностью восстановить и гессиан.

Практическая реализация вышеизложенных соображений приводит к процессу следующего типа. Кроме точки x^j, имеем еще и положительно-определенную самосопряженную матрицу H^j. Функцию $f(x^j + \delta x)$ аппроксимируем разложением

$$f(x^j + \delta x) \approx f(x^j) + f_x(x^j)\,\delta x + 0.5(H^j \delta x, \delta x).$$

Минимизируя правую часть, определяем δx, т.е. $\delta x = -(H^j)^{-1} f_x$. После нахождения точки $x^{j+1} = x^j + \delta x$ определяем $f_x(x^{j+1})$ и пересчитываем матрицу H, вычисляя H^{j+1} таким образом, чтобы выполнялись N вышеупомянутых соотношений между $N(N+1)/2$ элементами гессиана.

Элементы H этими соотношениями, конечно, однозначно не определяются. Поэтому нужно привлечь какие-то дополнительные эвристические соображения, например минимизацию отличия H^{j+1} от H^j или что-либо в этом роде. Не будем доводить дело до конкретных расчетных формул (все это описано в обширной литературе по математическому программированию); ограничимся лишь изложением основных идей. В настоящее время квазиньютоновские ме-

тоды составляют основу наиболее эффективных алгоритмов безусловной минимизации. Правда, их высокая эффективность проявилась пока в задачах сравнительно невысокой размерности.

Поиск глобального минимума. Это характерный пример проблемы, которая с одной точки зрения тривиальна, с другой — в сущности неразрешима. В самом деле, вот ее тривиальное решение. Введем в R^N куб $|x_n| \leqslant X$ и сетку с шагом h, покрывающую куб. Вычислим $f^0(x)$ в узлах сетки (это потребует конечного числа операций) и найдем точку сетки, в которой достигается минимальное значение f^0. Затем удвоим размер куба ($X := 2X$), уменьшим шаг h вдвое и повторим вышеописанную операцию.

Нетрудно доказать, что для любой непрерывной функции f^0 можно получить последовательность точек, сходящихся к точке глобального (абсолютного) минимума. Описанная выше операция носит название *сканирования*. Единственное возражение против ее использования — число $(X/h)^N$ вычислений f^0 на каждом этапе. Сканирование используется в пространствах невысокой размерности ($N = 1, 2, 3$) на достаточно грубых сетках, но серьезного практического значения эта универсальная процедура поиска не имеет.

Более реалистичной и достаточно часто используемой является процедура случайного поиска. При поиске $\min f^0(x)$ в кубе $|x| \leqslant X$ (величина X задает априорную информацию о расположении минимума) значение f^0 вычисляется в последовательности точек x^j, генерируемых датчиком случайных чисел, равномерно распределенных в кубе. Из этих точек выбирается точка с минимальным значением f^0. Доказательство того, что найденная точка находится на расстоянии ε_1 от точки минимума с вероятностью $1 - \varepsilon_2$, где ε_1, ε_2 могут быть сделаны сколь угодно малыми за счет соответственно большого числа «испытаний», является упражнением по теории вероятностей студенческого уровня и особого интереса не представляет. Ответ почти очевиден.

Действительно трудной и интересной алгоритмической задачей является конструирование программных датчиков так называемых «псевдослучайных» чисел с равномерным распределением. Хороший датчик генерирует такую последовательность точек, что каждый ее отрезок достаточно хорошо имитирует равномерное распределение: точки отрезка не должны «сбиваться в кучу», с одной стороны, и не должны оставлять «пустот» в кубе — с другой. Имея такие генераторы точек, можно (при приемлемом числе вычислений f^0) «прощупать» ее значение в кубе и найти не слишком уж грубую оценку $\min f^0(x)$. Число необходимых испытаний (вычислений f^0) существенным образом зависит от размерности куба,

требуемой точности и гладкости $f^0(x)$. Ограничимся этими общими и почти очевидными сведениями. Случайный поиск давно оформился в самостоятельную дисциплину, и по этому вопросу есть богатая специальная литература.

Метод исчерпывания. Пусть известно, что f^0 удовлетворяет условию Липшица с константой C. Генерируется последовательность точек куба, выбирается точка с минимальным значением f^0. Обозначим через f_j^* минимальное значение функции после j-го испытания. Вычислим значение в очередной точке x^{j+1}. Если $f^0(x^{j+1}) > f_j^*$, то около точки x^{j+1} можно «вырезать» сферу радиусом $(f^0(x^{j+1}) - f_j^*)/C$, в которой значение f^0 заведомо не меньше f_j^* и в которой в дальнейшем вычислять значения f^0 не имеет смысла.

Таким образом, после каждого испытания накапливается информация о тех частях куба, в которых минимум заведомо не находится. Реализация столь простого соображения связана с большими алгоритмическими сложностями, касающимися в сущности двух проблем: хранения накапливающейся информации (и, возможно, ее коррекции; если после очередного испытания значение f^* изменилось, исключаемые из просмотра части куба могут быть, соответственно, расширены) и ее использования (не так-то просто генерировать разумным образом распределенные точки в оставшемся «дырявом» множестве). Так что и эта конструкция имеет достаточно ограниченные возможности практической реализации.

Одним из наиболее часто применяемых способов хоть какой-то борьбы с «опасностью локального минимума» является поиск локального минимума при разных выборах стартовых точек x^0; а для выбора x^0 используются соображения, например, случайного поиска. Итак, простейшая задача (1) послужила поводом познакомить читателя с основными понятиями этой темы. Перейдем к более сложным задачам.

Поиск условного минимума. Начнем постепенное усложнение постановки задачи. Рассмотрим задачу (1) при условии (2а). В этом случае применимы предыдущие алгоритмы спуска по различным направлениям с небольшим алгоритмически несложным дополнением — «проецированием» точек на прямоугольник в R^N: $\Pi = [X^-, X^+]$. Пусть в точке x^j найдено направление e (например, $e = -f_x^0(x^j)$) и образуется «линия» (в общем случае ломаная) $x(s) = P(x^j + se)$, где Pz — операция проецирования точки

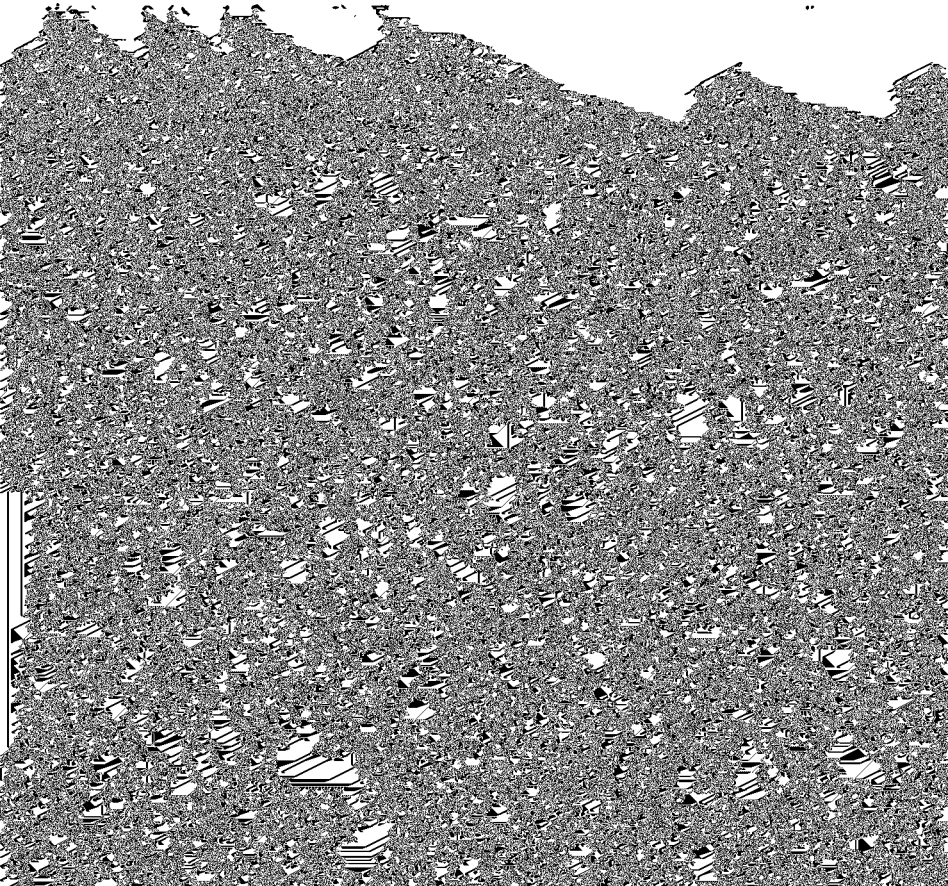
z на π. Проецирование z сводится к нахождению в π точки, ближайшей к z.

Решение этой задачи элементарно и состоит в покомпонентной «срезке»:

$$(Pz)_n = \{X_n^-,\ z_n < X_n^-;\ \ z_n,\ \ X_n^- \leqslant z_n \leqslant X_n^+;\ \ X_n^+,\ \ z_n > X_n^+\}.$$

Параметр s находится решением задачи $\min\limits_s f^0(P(x^j + se))$. Надо, правда, иметь в виду, что в некоторых случаях при гладкой функции f^0 функция $f^0(P(x + se))$ может иметь разрывы производных. Вместо условий $X^- \leqslant x \leqslant X^+$ можно ввести общее условие $x \in \mathscr{X}$, где $\mathscr{X}$ — некоторое замкнутое множество. Но, конечно, с усложнением геометрии $\mathscr{X}$ операция проецирования на $\mathscr{X}$ усложняется.

Метод множителей Лагранжа. Следующий класс задач был исследован очень давно. Это — задачи поиска $\min f^0(x)$ при условиях



Вышеприведенное рассуждение необходимо дополнить не очень сложным процессом коррекции на величины $O(\|\delta x\|^2)$, с тем чтобы доказать существование таких малых $\Delta x = \delta x + O(\|\delta x\|^2)$, что с учетом нелинейности $f^0(x + \Delta x) < f^0(x)$ и $f^i(x + \Delta x) = f^i(x + \Delta x)$.

С этим результатом связаны два возможных алгоритма решения задачи. Первый следует классическому рецепту Лагранжа. Образуется функция Лагранжа и ищется точка ее безусловного минимума

$$x(\lambda) = \arg \min_x \mathscr{L}(x, \lambda) \quad \mathscr{L}(x, \lambda) \equiv (\lambda, f(x)), \quad |\lambda_0| = 1. \quad (4)$$

Множители λ здесь пока не определены.

Разумеется, при произвольных множителях λ условия будут нарушены и для их определения ставится естественная задача

$$\partial \mathscr{L}/\partial \lambda_i = f^i(x(\lambda)) = 0, \quad i = 1, 2, \ldots, I.$$

Это — система I нелинейных уравнений с I неизвестными. Ее следует решать подходящим методом, например методом Ньютона. Здесь есть, конечно, осложнения. Зависимость (4) для $x(\lambda)$ реализуется решением задачи поиска безусловного минимума, а ее придется решать много раз при разных значениях λ. Положение несколько облегчается тем, что при вычислении $x(\lambda)$ предыдущие значения x могут служить хорошим начальным приближением.

Сложным является и вычисление производных $\partial f^i(x(\lambda))/\partial \lambda$, т.е. дифференцирование функций $x(\lambda)$, определенных не совсем обычным образом. Численное дифференцирование в принципе решает проблему, но это требует дополнительных вычислений $x(\lambda)$. К тому же не хотелось бы вычислять $x(\lambda)$ слишком уж точно: это требует большого объема вычислительной работы.

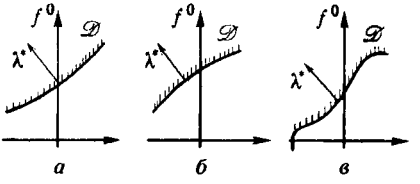


Рис. 47

Теперь разъясним самое важное обстоятельство — сходимость предложенной процедуры требует (и это по существу дела) предположения о выпуклости так называемой *области достижимости*. Так называют область $\mathscr{D}$ в пространстве R^{I+1} точек $\{f^0(x), f^1(x), \ldots, f^I(x)\}$, которые могут быть получены при всех допустимых x.

Не будем пока давать строгих определений некоторых понятий (выпуклость, строгая выпуклость и т.п.), апеллируя к простым геометрическим образам. На рис. 47 показаны типичные ситуации. Ось абсцисс представляет I-мерное пространство. Точка $x(\lambda)$ является самой низкой точкой области $\mathscr{D}$ в направлении λ. Вектор λ является

нормалью к $\partial\mathscr{D}$ в точке $x(\lambda)$, если в этой точке граница $\partial\mathscr{D}$ гладкая (если $x(\lambda)$ является угловой точкой границы, λ принадлежит множеству «опорных векторов»).

Рассмотрим характерные ситуации, представленные на рис. 47.

а) Нижняя граница $\mathscr{D}$ строго выпукла вниз. Описанный выше метод имеет шансы на успех, так связь между λ и $x(\lambda)$ однозначна: каждому λ соответствует единственная точка $x(\lambda)$.

б) Нижняя граница $\mathscr{D}$ вогнута. Метод не будет работать, так как даже при точном значении λ^* поиск минимума $\mathscr{L}(x, \lambda^*)$ приведет к далекой от решения точке $x(\lambda)$.

в) Нижняя граница $\mathscr{D}$ строго выпукла вниз только в окрестности оси ординат. Метод имеет шансы на успех при хорошем начальном приближении. (Напомним, что, решая задачу (4), находят именно локальный минимум.)

Свойства выпуклости области $\mathscr{D}$ обычно неизвестны. Поэтому метод множителей Лагранжа в той форме, в какой он описан выше, применяется в вычислениях редко. Перейдем к описанию второго алгоритма.

Метод условного градиента. Более употребительна другая форма использования идей Лагранжа, к которой можно прийти разными путями. (В зависимости от этого одинаковые по существу методы получают разные названия: метод линеаризации, метод приведенного градиента и т.п.) Мы предпочтем вывести основную конструкцию на основе линеаризации. Итак, пусть точка x допустима в смысле выполнения всех условий $f^i = 0$ $(i = 1, 2, \ldots, I)$. Ищем малую поправку δx, линеаризуя задачу и добавляя ограничение на δx.

Смещение δx определяется задачей (3) при условиях

$$f^i(x) + f^i_x(x)\,\delta x = 0, \quad i = 1, 2, \ldots, I. \tag{5}$$

Составляя для этой задачи функцию Лагранжа:

$$\mathscr{L}(\delta x, \lambda, \mu) = f^0_x\,\delta x + \sum \lambda_i (f^i + f^i_x\,\delta x) - \frac{\mu}{2}(\delta x, \delta x),$$

находим минимум из уравнений $\partial\mathscr{L}/\partial\delta x = 0$. В результате имеем систему уравнений для δx, λ, μ:

$$\delta x = \frac{1}{\mu}\left(f^0_x + \sum \lambda_i f^i_x\right).$$

Подставляя это выражение в (5), получаем для λ систему уравнений с параметром μ:

$$\mu f^i + (f^0_x, f^i_x) + \sum_{j=1}^{I} \lambda_j (f^i_x, f^j_x) = 0, \quad i = 1, 2, \ldots, I.$$

Решая систему уравнений дважды (один раз с правыми частями (f_x^0, f_x^i), второй раз с f^i), находим общее решение вида $\lambda_i' + \mu\lambda_i''$, после чего можно определить μ из условия $(\delta x, \delta x) = \varepsilon^2$. Здесь не случайно в задачу включены значения $f^i \neq 0$. Ниже нам понадобится именно такая более общая конструкция. Заметим только, что при вычислении μ следует выбрать решение, дающее минимум, а не максимум функции Лагранжа $\mathscr{L}$.

Итак, пусть в точке x выполнены условия $f^i(x) = 0$ и определено возмущение δx. Теперь есть два способа использовать этот результат.

Первый способ: формируется «линия спуска» $x + s\,\delta x$. Тогда (если в точке x выполнены условия $f^i = 0$) можно утверждать, что

$$\frac{d}{ds} f^i(x + s\,\delta x)\,|_{s=0} = f_x^i(x)\,\delta x,$$

т.е. смещение по s сопровождается линейным (по s) убыванием f^0 и соблюдением (в первом порядке) условий $f^i = 0$.

Однако мы должны выбрать какое-то конечное значение s. Работающие по вышеприведенной схеме алгоритмы различаются выбором того критерия, по которому при увеличении s падение f^0 считается еще выгодным, хотя оно и сопровождается нарушением условий. На следующей итерации приходится варьировать x с учетом уже имеющихся малых нарушений в условиях $f^i = 0$, и проблема выбора s осложняется.

Второй способ: δx считается уже готовой вариацией, т.е. нужно переходить к точке $x + \delta x$. Здесь возникают те же, в сущности, проблемы, но они должны решаться в терминах достаточно ответственного назначения величины ε (в первом способе, очевидно, величина ε никакой роли не играет).

Выпуклое программирование. Перейдем к описанию некоторой общей идеи, приведшей в конечном счете к одной (из двух, в сущности) фундаментальных конструкций эффективных алгоритмов математического программирования. Предположим, что область достижимости $\mathscr{D}$ есть строго выпуклое множество.

Определение 1. Множество $\mathscr{D}$ называют строго выпуклым, если для любых двух его граничных точек z' и z'' все точки соединяющего их интервала $z(s) = sz' + (1 - s)z''$ $(s \in (0, 1))$ лежат строго внутри $\mathscr{D}$.

Можно характеризовать свойство строгой выпуклости и по-другому. Для любого вектора $\lambda = \{1, \lambda_1, \ldots, \lambda_I\}$ задача $\min\limits_{z \in \mathscr{D}} (z, \lambda)$ имеет единственное решение $z(\lambda) \in \partial\mathscr{D}$. Заметим, что граница строго

выпуклого множества может содержать угловые точки. При этом разные λ могут дать одну и ту же точку $z(\lambda)$. Все такие λ называют опорными к $\mathscr{D}$ в точке $z(\lambda)$. Если граница в точке $z_0 \in \partial\mathscr{D}$ является гладкой, то значение λ_0, для которого $z(\lambda_0) = z_0$, является внутренней нормалью к $\partial\mathscr{D}$ в точке z_0. Докажем важную теорему.

Теорема 1. Если область достижимости $\mathscr{D}$ является строго выпуклой, то задача на условный экстремум (1), (2) эквивалентна суперпозиции задач на безусловный экстремум

$$\max_{\lambda} \{ \min_{X^- \leqslant x \leqslant X^+} \mathscr{L}(x, \lambda) \},$$

где $\mathscr{L}(x, \lambda) = (\mathbf{f}(x), \boldsymbol{\lambda})$.

Эту задачу можно сформулировать в несколько иной редакции:

$$\max_{\lambda} F(\lambda), \quad \text{где} \quad F(\lambda) = \min_{X^- \leqslant x \leqslant X^+} (\mathbf{f}(x), \boldsymbol{\lambda}).$$

Таким образом, речь идет просто о нахождении максимума функции $F(\lambda)$ при достаточно сложном ее определении: эта функция задана алгоритмом поиска минимума.

Доказательство. Пусть x^* — точка, решающая исходную задачу на условный экстремум (1), (2). Существование такой точки следует из общих теорем, в которых используется непрерывность функций f, а также, например, ограниченность и замкнутость $\mathscr{D}$. Пусть λ^* — опорный вектор к $\partial\mathscr{D}$ в точке $z^* = f(x^*)$. Покажем, что для всех $\boldsymbol{\lambda}$ имеет место $F(\lambda) \leqslant F(\lambda^*)$.

В самом деле, пусть $\lambda \neq \lambda^*$. Вычислим $z(\lambda)$ и $F(\lambda) = (z(\lambda), \lambda)$. Тогда вся область $\mathscr{D}$ и, следовательно, точка z^* лежат выше гиперплоскости, проходящей через $z(\lambda)$ ортогонально $\boldsymbol{\lambda}$:

$$(z^* - z(\lambda), \lambda) \geqslant 0, \quad \text{т.е.} \quad (z^*, \lambda) \geqslant F(\lambda).$$

Но $z^* = \{z_0^*, 0, \ldots, 0\}$, т.е. $(z^*, \lambda) = z_0^*$. Точно так же величина $F(\lambda^*) = (z^*, \lambda^*) = z_0^*$. Итак, $F(\lambda) \leqslant (z^*, \lambda) = (z^*, \lambda^*) = F(\lambda^*)$. Теорема доказана.

Для того чтобы на основе этой теоремы построить достаточно эффективные алгоритмы, нужно указать способ вычисления градиента $F_\lambda(\lambda)$. Этот вопрос (и это очень важное обстоятельство) допускает совсем простое решение.

Теорема 2. Пусть $\mathscr{D}$ — строго выпуклая ограниченная область. Тогда функция $F(\lambda)$ дифференцируема и ее градиент вычисляется по формуле

$$\frac{\partial F(\lambda)}{\partial \lambda} = z(\lambda) = \arg \min_{z \in \mathscr{D}} (z, \lambda). \tag{6}$$

Не проводя полного доказательства, укажем его основные моменты. Начнем с формального вычисления производной:

$$\frac{\partial F(\lambda)}{\partial \lambda} = \frac{\partial}{\partial \lambda}(z(\lambda), \lambda) = \frac{\partial z(\lambda)}{\partial \lambda}\lambda + z(\lambda).$$

Строками матрицы z_λ являются векторы $\partial z(\lambda)/\partial \lambda_i$. Если граница области $\mathscr{D}$ в точке $z(\lambda)$ имеет касательную гиперплоскость, все эти векторы лежат в ней, так как при малом изменении λ точка $z(\lambda)$ непрерывно перемещается по $\partial\mathscr{D}$. А так как в этом случае λ есть нормаль к касательной гиперплоскости, то $z_\lambda \lambda = 0$. Если $z(\lambda)$ есть угловая точка границы $\partial\mathscr{D}$, то при малом изменении λ точка $z(\lambda)$ не смещается и $z_\lambda(\lambda) = 0$. Мы опускаем анализ смешанных ситуаций, когда граница $\partial\mathscr{D}$ в точке $z(\lambda)$ имеет характер «ребра».

Перейдем к описанию алгоритма решения задачи (1), (2) на основе доказанных теорем. Задаем начальные значения x и λ. Фиксируя λ, начинаем поиск $\min\limits_x \mathscr{L}(x, \lambda)$ из заданной точки x. Получаем новую точку $x(\lambda)$. Тогда $F(\lambda) = \mathscr{L}(x(\lambda), \lambda)$, а $\partial F(\lambda)/\partial \lambda_i = f^i(x(\lambda))$. В общем случае $f^i \neq 0$, поэтому делаем пересчет множителей Лагранжа с целью максимизации $\mathscr{L}$:

$$\lambda_i := \lambda_i + s f^i(x(\lambda)), \quad i = 1, 2, \ldots, I. \tag{7}$$

Реализация этого алгоритма связана с двумя достаточно сложными вопросами. Первый вопрос: с какой точностью находить $x(\lambda)$? Решая эту задачу слишком точно, мы затратим, видимо, без особой нужды слишком большое машинное время. Решая задачу слишком грубо, мы рискуем снизить эффективность процесса в целом. Второй вопрос — выбор шага s в (7).

Ограничимся постановкой этих вопросов, адресуя читателя, интересующегося их решением (решения могут быть различными; их основа — не строгие теоремы, а различного рода эвристические соображения) к специальной литературе. Алгоритмы такого сорта оказываются достаточно эффективными, но подчеркнем еще раз, что существенной предпосылкой их применимости является строгая выпуклость множества достижимости. А это свойство нельзя считать типичным в прикладных задачах, оно встречается достаточно редко.

Метод штрафных функций. Следующая простая конструкция одно время была очень популярна среди теоретиков оптимизации. Решение задачи на условный экстремум (1), (2) «сводится» к решению задачи на безусловный экстремум для функции

$$F(x, A) = f^0(x) + A\sum_{i=1}^{I}\{[f^i(x) - F_i^+]_+^2 + [F_i^- - f^i(x)]_+^2\}. \tag{8}$$

Здесь $[a]_+ = a$, если $a > 0$, $[a]_+ = 0$ в противном случае; A — большой коэффициент «штрафа» за нарушение условий.

Функцию $F(x, A)$ называют «штрафной функцией». Сведение задачи (1), (2) к минимизации $F(x, A)$, разумеется, приближенное, но тем более точное, чем больше A. К сожалению, практика применения метода оказалась не очень успешной. Введение в конструкцию (8) большого параметра A приводит к тому, что в точках x, где $f^i(x) \approx 0$ $(i = 1, 2, 3, \ldots)$, функция F оказывается очень негладкой, имеет сложное «овражное» строение и очень трудно минимизируется.

Скорость сходимости методов поиска минимума оказывается столь медленной, что точки «минимизирующей» последовательности x^j практически «стоят на месте», и это часто принимают за достижение минимума. Таким образом, метод дает неверные результаты, которые нередко принимаются за решение сложных задач. Полезно еще раз отметить, что эффективность методов приближенных вычислений существенным образом связана с гладкостью используемых функциональных зависимостей. Метод штрафных функций достигает своих целей (замена задачи на условный экстремум задачей безусловной оптимизации) ценой именно этого важнейшего фактора, делая из «хороших» функций $f^i(x)$ «плохую» функцию F (см. (8)).

Недостатки метода пытаются преодолеть следующим образом. Параметр A сначала берется не очень большим, так что свойства гладкости F еще немногим хуже свойств f. Найденное решение задачи $x(A) = \arg\min F(x, A)$, конечно, сильно нарушает условия $f^i = 0$ $(i = 1, 2, 3, \ldots)$, и его используют как начальное приближение в задаче с несколько увеличенным значением A, и т.д. Однако и эта идея не привела к очень уж большим успехам.

Метод модифицированной функции Лагранжа. Перейдем к одной из наиболее удачных вычислительных конструкций. Исходная задача заменяется следующей:

$$\min_x F(x, A) \quad \text{при условиях} \quad f^i(x) = 0, \quad i = 1, 2, \ldots, I. \qquad (9)$$

Суть дела в том, что теперь параметр A принимает умеренные значения. Его цель — сделать границу области достижимости выпуклой вниз (хотя бы локально, в окрестности искомого решения; см. рис. 47). Применим к задаче (9) алгоритм выпуклого программирования.

К методу модифицированной функции Лагранжа можно прийти и из других соображений. Минимизируем функцию $F(x, A)$ со слабым штрафом. Пусть $x(A) = \arg\min F(x, A)$. Тогда, как уже отмечалось, $f^i(x(A)) = r_i \neq 0$. Введем «невязки» r_i в конструкцию (8) заранее. Используем простое соображение. Если введение штрафа A смещает

точку минимума на расстояния r_i, попробуем решать с этим же штрафом A задачу с условиями $f^i(x) + r_i$, надеясь, что новые условия будут выполнены с теми же погрешностями r_i в точке $x(A)$:

$$f^i(x) + r_i = r_i, \quad \text{т.е.} \quad f^i(x(A)) = 0, \quad i = 1, 2, 3, \ldots$$

Итак, новая штрафная функция будет такой:

$$F(x, A) = f^0(x) + A \sum_{i=1}^{I} |f^i(x) + r_i|^2 =$$
$$= f^0(x) + A \sum |f^i(x)|^2 + \sum A r_i f^i(x) + A \sum r_i^2.$$

Значения r_i берутся с предыдущей итерации, Ar_i играют роль множителей Лагранжа λ_i. Последнее слагаемое, очевидно, можно опустить.

Метод линеаризации. Вторая основная вычислительная конструкция, которую удалось довести до сравнительно эффективных алгоритмов, связана с одной из фундаментальных идей вычислительной математики. По традиции ее связывают с именем Ньютона. Задача линеаризуется в окрестности некоторого уже найденного приближения к решению, и следующее, лучшее, приближение находится решением линеаризованной задачи.

Пусть x — некоторое текущее приближение к решению задачи поиска условного экстремума. Следующее приближение ищется в виде $x + \delta x$, где δx — «малая» поправка, которая решает линеаризованную задачу

$$\min_{\delta x} \{f^0(x) + f_x^0(x)\, \delta x\} \tag{10}$$

при условиях

$$\begin{aligned} &\text{а)}\ F_i^- \leqslant f^i(x) + f_x^i(x)\, \delta x \leqslant F_i^+, \quad i = 1, 2, \ldots, I, \\ &\text{б)}\ s_n^- \leqslant \delta x_n \leqslant s_n^+, \quad n = 1, 2, \ldots, N. \end{aligned} \tag{11}$$

Числа s_n^-, s_n^+ определяют «шаг» процесса. Задача (10), (11) является хорошо изученной задачей линейного программирования, для решения которой давно разработаны достаточно надежные алгоритмы (так называемый симплекс-метод). Эффективность конкретного алгоритма существенно определяется тактикой подбора шага (s^-, s^+), которая, конечно же, должна опираться на информацию, получаемую в процессе решения задачи.

Опишем основные технологические элементы метода линеаризации, разработанного автором в начале шестидесятых годов и применявшегося в существенно более сложной ситуации (подробнее об

этом см. в § 28). При организации вычислений нужно избежать двух опасностей:

при слишком малом значении шага процесс протекает надежно, но медленно;

при слишком большом значении шага пренебрежение квадратичными членами становится необоснованным и процесс перехода от x к $x+\delta x$ становится бессмысленным.

Итак, шаг должен быть максимально возможной величиной, при которой линеаризация функций обладает достаточной точностью. Это качественное соображение предстоит превратить в алгоритм подбора шага в зависимости от фактического хода вычислений. Введем параметры, управляющие процессом решения.

а) Числа ε_i $(i=1, 2, \ldots, I)$ определяют заданную постановкой задачи точность выполнения условий (2б).

б) Число S (шаг процесса) входит в алгоритм генерации чисел $s_n^- = O(S)$, $s_n^+ = O(S)$. Учет условий (2а) очевиден.

в) Число C в начале процесса — достаточно большое число $(C = 10^3 \div 10^5)$. В процессе решения задачи это число автоматически меняется и стремится к единице. На данном этапе поиска решения в условиях (2б) допускается погрешность $C\varepsilon_i$.

Стандартный шаг алгоритма состоит из следующих операций.

1. Пусть уже получена точка x, выработались какие-то числа S и C.

2. Задача линеаризуется, т.е. вычисляются значения $f^i(x)$, производные $f_x^i(x)$ и числа s_n^-, s_n^+. Таким образом, задача (10), (11) сформирована.

3. Решается задача (10), (11). При значительных нарушениях в точке x условий (2б) задача может и не иметь решения. Алгоритм обнаруживает этот факт и переходит к определению δx с целью минимизации

$$r(\delta x) = \sum_{i=1}^{I} [f^i(x) + f_x^i(x)\,\delta x]_*^2,$$

где

$$[a^i]_* = \{|a^i - F_i^-|,\ a^i < F_i^-;\ |a^i - F_i^+|,\ a^i > F_i^+;\ \text{иначе } 0\}.$$

Другими словами, игнорируя значение f^0, мы стремимся получить так называемую допустимую точку x, в которой с требуемой точностью выполнены все условия. В любом случае получается вариация δx.

4. Вычисляются вариации функций $\delta f^i = f_x^i\,\delta x$ и (после вычисления значений $f^i(x+\delta x)$) их приращения $\Delta f^i = f^i(x+\delta x) - f^i(x)$.

5. После этого начинается логически наиболее сложная операция — анализ и принятие решений об изменении управляющих параметров. Выделяются характерные ситуации, в каждой из которых реализуется свое целесообразное поведение.

5.1. Пусть точка x недопустима, т.е. $[f^i(x)]_* > C\varepsilon_i$ хотя бы при одном i. В этом случае вычисляется погрешность линейного приближения $\eta = \max_{i \geqslant 1} |(\delta f^i - \Delta f^i)/\Delta f^i|$. При этом максимум вычисляется лишь для тех индексов, для которых условие нарушено и в новой точке $x + \delta x$, т.е. $[f^i(x + \delta x)]_* > C\varepsilon_i$. В зависимости от значения η изменяется шаг S для следующей итерации. Если $\eta \approx 0$, шаг S увеличивается; если $\eta \approx 1$, шаг S уменьшается. При определенных обстоятельствах ($r(\delta x) > r(0)$) итерация «отменяется», шаг S уменьшается (например, вдвое) и задача линейного программирования решается заново.

Описанная выше ситуация типична для начала решения задачи, когда взятая из каких-то соображений точка начального приближения x грубо нарушает условия (2б). Какое-то число первых итераций процесса тратится на определение допустимой точки x. Значение f^0 на этом этапе обычно растет.

5.2. Пусть точка x допустима и точка $x + \delta x$ тоже удовлетворяет всем условиям с погрешностью $C\varepsilon$. При этом погрешность η вычисляется только по индексу $i = 0$. Здесь могут представиться разные ситуации.

а) $\Delta f^0 < 0$, т.е. при выполнении условий с погрешностью $C\varepsilon_i$ происходит уменьшение f^0. В этом случае пересчитывается шаг S (так же, как это было описано выше), точка x заменяется на $x + \delta x$ и процесс продолжается с операции 5.1 (конечно, вычисления f^i в точке $x + \delta x$ не повторяются, они уже известны).

б) $\delta f^0 < 0$, но $\Delta f^0 > 0$. Это означает, что шаг S слишком велик. Поэтому итерация «отменяется», шаг S уменьшается (например, вдвое) и задача линейного программирования решается заново.

в) $\delta f^0 > 0$ и точность линейного приближения удовлетворительна ($\eta \approx 0.1$, например). Решение задачи линейного программирования приводит к росту f^0. Такая ситуация обычно связана не с погрешностью линеаризации, а с тем, что условия (2б) в точке $x + \delta x$ выполнены с большей точностью, чем условия в точке x. Улучшение точки $x + \delta x$ получено ценой некоторого увеличения f^0, хотя обе точки удовлетворяли условиям (2б) с погрешностью $C\varepsilon$. В этом случае величина C уменьшается настолько, что точка x становится «недопустимой». Задача линейного программирования заново не решается, но анализ проводится заново, при новых требованиях к точности выполнения условий (2б).

Вышеописанное решение об изменении C основано на следующем простом соображении. Не следует с самого начала требовать очень точного выполнения условий (2б) (для всех промежуточных результатов): это приводит к слишком малому шагу. Точность вы-

полнения условий (2б) должна быть «согласована» с желаемым ходом вычислительного процесса. Если при колебаниях значений f^i в пределах $C\varepsilon$-погрешности происходит монотонное понижение f^0, все «идет так, как надо».

Конечно, мы могли бы с самого начала задать слишком малое значение C и завысить требования к точности выполнения условий. Эту ситуацию можно распознать по слишком малой величине η (шаг S слишком мал, точность линейного приближения излишне высока). В этом случае значение C несколько увеличивается. Ограничимся этим не очень строгим и не исчерпывающе полным описанием. Стремление к точному описанию алгоритма затруднило бы понимание основных идей. Здесь же мы ставим целью не безупречность описания, а подготовку читателя к знакомству с более точным изложением.

Для иллюстрации сказанного выше приведем пример решения не очень сложной модельной задачи. Она входит в набор тестов, принятых для оценки фактической эффективности алгоритмов. Задача имеет следующий вид ($i = 1, 2, \ldots, 5$):

$$f^0(x) = -\sum_{i=1}^{10} b_i x_i + \sum_{i=1}^{5}\sum_{j=1}^{5} c_{ij} x_{10+i} x_{10+j} + 2\sum_{i=1}^{5} d_i x_{10+i}^3,$$

$$f^i(x) = \sum_{j=1}^{10} a_{ij} x_j - 2\sum_{j=1}^{5} c_{ij} x_{10+j} - 3 d_i x_{10+i}^2 - e_i.$$

Кроме того, поставлены условия $x_i \geqslant 0$ ($i = 1, 2, \ldots, 15$).

Таким образом, мы имеем в задаче 15 неизвестных и 5 условий-равенств. В качестве начального вектора берется $x_7 = 60$, остальные $x_i = 0$. Требуемая точность выполнения условий определялась числами $\varepsilon_i = 10^{-5}$, $C = 10^6$, $S = 1.3$. Значения параметров, входящих в выражения для функций, можно найти в известной монографии Д. Химмельблау «Прикладное нелинейное программирование» (см. задачу 18).

Процесс решения задачи иллюстрирует табл. 16. Поясним обозначения: ν — номер шага (итерации); $r = \max |f^i(x^\nu)|$ — характеристика погрешности выполнения условий ($i = 1, 2, \ldots, 5$); S — шаг варьирования; η — характеристика погрешности линейного приближения; K — число вычислений функций f^i (вычисление f^i для каждого отдельного i увеличивает счетчик K на единицу).

Каждый шаг процесса стоит примерно шести вычислений f^i, f^i_x. На некоторых шагах приходится повторять вариацию меньшим шагом; поэтому $K_\nu \approx 1.35\,(I+1)\nu$. При $\eta = 0$ условия (2б) выполнены с допустимой на данном этапе погрешностью. Эта погрешность есть 10 при $\nu \leqslant 32$. После 32-й итерации допустимая погрешность

есть 0.45, после 41-й — 0.02, после 61-й — $5 \cdot 10^{-4}$. Фактическая точность заметно выше. Расчет требует около минуты работы

Таблица 16

ν	f^0	r	S	η	δf^0	Δf^0	K
0	2400.	48	1	0.7	−20.3	−20.2	18
1	2380.	43	0.19	0.4	−3.3	4.6	24
2	2384.	44	0.14	0.14	−15.8	−13.4	30
3	2371.	32	0.17	0.19	−17.5	−16.5	36
4	2355.	35	0.21	0.13	−25.5	−23.2	42
9	1757.6	7.2	24.0	0	−86.6	−86.4	96
10	1671.2	3	1.7	0	−74.1	−68.5	108
11	1602.8	3.8	0.99	0	−98.3	−92.7	114
19	829.3	7.4	1.1	0	−135.7	−125.3	168
20	704.0	5	1.3	0	−150.6	−146.5	174
21	557.48	2.4	1.5	0	−186.7	−180.6	180
24	108.90	5	1	0	−74.5	−72.6	204
25	36.29522	1.2	1.2	0	−2.7	−2.8	216
26	35.51306	3.4	0.3	0	−2.73	−1.72	222
30	32.62674	0.06	0.18	0	−0.36	−0.20	246
31	32.42416	0.6	0.15	0	−0.39	−0.35	252
32	32.07438	0.9	0.15	0	0.35	0.38	264
33	32.45504	0.17	0.6	0	−0.084	−0.044	282
41	32.36572	0.01	0.07	0	−0.011	−0.0023	366
42	32.36345	6E−3	0.017	0	−6.9E−3	−6.2E−3	372
43	32.35725	4E−4	0.017	0	−3.8E−3	−3.1E−3	378
51	32.34916	2E−5	5E−3	0	−1.9E−5	−12E−2	426
52	32.34894	4E−5	4E−3	0	−83E−5	−70E−5	432
61	32.34879	1E−6	9E−4	0	−6E−6	−5E−6	510
65	32.34869	1E−7	6E−4	0			

БЭСМ-6, причем половина этого времени тратится на вычисление f и их производных, половина — на решение задач линейного программирования. Решением является точка

$x = \{0, 0, 5.17278, 0, 3.06138, 11.83698, 0, 0, 0.10337, 0,$

$0.30007, 0.33342, 0.40013, 0.42823, 0.22397\}.$

Задачи негладкой оптимизации. Выше были рассмотрены алгоритмы, ориентированные на гладкие задачи в том смысле, что все функции f^i предполагались достаточно простыми, гладкими. Кстати, что это значит? Ответ не так прост, как может показаться. Мы используем аппроксимацию функций $f^i(x)$ линейными или квадратичными выражениями. Эффективность алгоритмов существенно зависит от того, как соотносятся между собой два числа: S — расстояние, на котором используемая аппроксимация описывает изменение функций с некоторой (10 %-ной, допустим) погрешностью, и L — расстояние от начального приближения x^0 до искомой точки минимума x^*. Грубо говоря, число шагов можно ориентировочно оценивать величиной L/S. Это отношение есть естественная мера гладкости функции. Читателю не должно казаться странным, что эта характеристика функции зависит от начального приближения, т.е. от априорной информации о расположении искомой точки x^*.

Если говорить просто о классе негладких функций, то он включает в себя необозримое множество слишком разных функций. Универсальные алгоритмы поиска минимума таких функций строить можно, но они крайне неэффективны и особого интереса для практики не представляют. К счастью, обычно в практической работе возникают негладкие функции специфического типа, для которых можно разрабатывать специальные достаточно эффективные алгоритмы минимизации. В частности, одним из наиболее важных источников задач негладкой оптимизации является задача следующего типа (иногда ее называют чебышевской задачей):

$$\min_x f^0(x), \quad \text{где} \quad f^0(x) \equiv \max_j f^{0,j}(x), \quad x \in R^N. \tag{12}$$

В задачу могут входить условия (2), и каждая из функций $f^i(x)$ может иметь форму (12). Важно подчеркнуть, что каждую из функций $f^{0,j}$ мы будем предполагать гладкой в том смысле, который выше мы придали этому термину. Функция $f^0(x)$ вида (12) принадлежит важному классу функций, дифференцируемых лишь по направлениям.

Определение 2. Функция $f(x)$ называется дифференцируемой в точке x по направлениям, если для любого направления e ($\|e\| = 1$) существует предел

$$D(x, e) = \lim_{s \to +0} \frac{f(x+se) - f(x)}{s}.$$

(Здесь s — скалярный параметр, всегда положительный!) Величина $D(x, e)$ называется производной f в точке x по направлению e. При гладких $f^{0,j}$ функция (12) почти во всех точках является просто дифференцируемой. Вычисляется эта производная очень просто.

Пусть $j(x)$ — тот, пока единственный индекс, на котором достигается максимум в (12), т.е. $f^0(x) = f^{0,j(x)}(x)$, и это соотношение (в си-

лу непрерывности) остается справедливым в некоторой окрестности точки x. Тогда $f^0(x)$ дифференцируема и $D(x, e) = (f_x^{0, j(x)}(x), e)$. Осложнения возникают в том случае, когда максимум в (12) достигается при двух (и более) индексах. Обозначим множество таких индексов $j(x) = \arg\max_j f^{0, j}(x)$. В этом случае

$$D(x, e) = \max_{j \in j(x)} (f_x^{0, j}(x), e).$$

Конечно, множество точек x, в которых в $j(x)$ входит больше одного индекса, имеет «нулевую меру», но при решении задач минимизации таких функций приходится иметь дело именно с точками из этого множества. Более того, если J больше размерности x, то типичной является ситуация, когда в точке минимума $f^0(x)$ в $j(x^*)$ входит $N + 1$ индексов. Для построения минимизирующей последовательности точек x важно уметь находить направление e, вдоль которого $f^0(x)$ убывает. Это направление должно быть направлением убывания для всех функций $f^{0, j}$ $(j \in j(x))$, т.е. речь идет о выпуклом конусе, образованном пересечением подпространств:

$$(f_x^{0, j}(x), e) < 0, \quad j \in j(x). \tag{13}$$

По мере роста числа индексов в $j(x)$ конус (13) становится все уже. В общем положении (т.е. если нет каких-то случайных совпадений) этот конус не пуст до тех пор, пока число индексов в $j(x)$ не больше N. Но как только это число станет хотя бы на единицу больше, конус (13) оказывается пустым.

Сложность задачи негладкой оптимизации связана со следующим обстоятельством. Когда функция $f^0(x)$ является гладкой, перемещение точки x по лучу $x(s) = x + se$ (где s — скалярный параметр, e — почти любой вектор) приводит к убыванию $f(x + se)$, если не при росте s, то при его убывании. В случае функции f^0 вида (12) ситуация иная. Для большинства векторов e такое движение по лучу сопровождается ростом f^0: функция $f^0(x + se)$ аналогична $|s|$. Исключение составляют лишь векторы конуса (13). Обобщение метода линеаризации на задачи с f^0 типа (12) проходит почти автоматически: несколько изменяется лишь форма задачи линейного программирования для определения вариации δx, а именно, вместо (10) имеем, очевидно,

$$\min_{\delta x} \{ \max_{j \in j(x)} [f^{0, j}(x) + f_x^{0, j}(x)\, \delta x] \}. \tag{14}$$

Замена (10) на (14) не создает никаких проблем для хороших программ симплекс-метода.

В алгоритме следует пользоваться другим определением множества индексов $j(x)$: $j \in j(x)$, если $f^{0, j}(x) > f^0(x) - \varepsilon$. Положитель-

ное число ε должно быть таким, чтобы при переходе от точки x к $x+\delta x$ не достигался $\max_j f^{0,j}(x+\delta x)$ при значении j, не входившем в $j(x)$. В противном случае вариация δx может быть выбрана такой, что $\max_{j\in j(x)} f^{0,j}(x+\delta x) < f^0(x)$, но $f^0(x+\delta x) > f^0(x)$, и алгоритм не обеспечивает монотонного понижения $f^0(x)$. Выбор ε особых трудностей не содержит, так как значительное увеличение ε часто приводит лишь к включению в $j(x)$ нескольких лишних индексов, что делает задачу линейного программирования несколько сложнее, чем она могла бы быть.

Более трудные задачи (так называемые минимаксные задачи) возникают при определении $f^0(x)$ формулой

$$f^0(x) = \max_{y\in Y} f^0(x, y). \tag{15}$$

Здесь существенные сложности связаны с тем, что нужно найти все точки множества

$$y(x) = \arg\max_y f^0(x, y). \tag{16}$$

Стандартная ситуация в задачах такого типа такова. Имеется множество $y^*(x)$, состоящее из всех точек локальных максимумов $f^0(x, y)$ по y. Это множество состоит из конечного числа точек. Их число обычно того же порядка, что и размерность x. Мы ограничимся случаем, когда Y — какая-то простая область (шар или прямоугольник, например). Часть этих точек входит в множество $y(x)$, и производная по направлению e от функции $f^0(x)$ вычисляется просто:

$$D(x, e) = \max_{y(x)} (f^0_x(x, y), e). \tag{17}$$

Формула (17) почти очевидна. Единственное обстоятельство, требующее анализа, связано со следующим. При изменении x на малую величину εe значение $f^0(x)$ изменяется не только за счет изменения значений $f^0(x, y)$ на $\varepsilon(f^0_x(x, y), e)$ для $y \in y(x)$ (эти изменения учтены в (17)), но и за счет смещения точек множества $y(x)$. Однако эти смещения изменяют величину $f^0(x+\varepsilon e)$ на $o(\varepsilon)$, поэтому на первую производную они не влияют. Причина здесь та же, что и в теореме 2.

Алгоритм поиска минимума функции $f^0(x)$ типа (15) отличается от алгоритма решения задачи (12) (когда Y есть конечное множество) тем, что требуется достаточно надежно отслеживать множество $y^*(x)$, выбирая из него на каждой итерации подмножество $y(x)$. Оно используется при формировании задачи типа (14). При переходе от точки x к близкой точке $x+\delta x$ множество $y^*(x+\delta x)$ обычно слегка смеща-

ется относительно множества $y^*(x)$. Однако, в принципе, при этом могут «рождаться» новые точки множества $y^*(x)$ и нужно считаться с тем, что до данного этапа процесса программа работала с неполным множеством $y^*(x)$ и значение f^0 вычислялось неверно.

Проиллюстрируем сказанное выше примером решения следующей задачи (она имеет прикладное происхождение, но мы будем рассматривать задачу как пример):

$$f^0(x, y) = \sum_{j=1}^{J} \varphi(x_j^1 - y^1)\, \varphi(x_j^2 - y^2),$$

где

$$\varphi(z) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-h}^{h} \exp\left(-\frac{(z+t)^2}{2\sigma^2}\right) dt.$$

Здесь $y = \{y^1, y^2\} \in R^2$, $x = \{x_j^1, x_j^2\}_{j=1}^J \in R^{2J}$; область Y — единичный квадрат. В рассматриваемом ниже примере $J = 16$, $h = 0.1$. Кроме того, решалась задача $\max\limits_x \min\limits_y f^0(x, y)$. В качестве начального приближения брались точки $x_j \in R^2$, расположенные равномерно на окружности радиусом 0.1 с центром в центре единичного квадрата.

Расчет начинался с того, что при фиксированном x генерировались случайные точки $y \in Y$, каждая из которых была стартовой точкой поиска $\min\limits_y f^0(x, y)$. Таким образом формировались множества $y^*(x)$ и $y(x)$. После каждого перехода от x к $x + \delta x$ уточнялись положения точек $y(x + \delta x)$. Для этого каждая точка $y \in y^*(x)$ бралась в качестве стартовой в процессе поиска $\max\limits_y f^0(x + \delta x, y)$, делалось небольшое число шагов подъема по градиенту, т.е. корректировалось множество $y^*(x + \delta x)$.

Однако множество $y^*(x + \delta x)$ может быть принципиально неполным. В нем может отсутствовать какая-то еще не обнаруженная точка локального максимума $f^0(x, y)$ по y. Поиск таких точек должен продолжаться. Для этого берется некоторое число случайных точек в Y, каждая из них используется в качестве стартовой при решении задачи $\max\limits_y f^0(x + \delta x, y)$. Получающиеся после некоторого числа шагов подъема по градиенту точки анализируются. Некоторые из них могут оказаться близкими к точкам, уже входящим в $y^*(x + \delta x)$, они, естественно, игнорируются. Но некоторые могут оказаться новыми, и тогда они включаются в $y^*(x + \delta x)$, расширяя его.

Хотя мы ограничились выше общим описанием, не уточняя деталей, читатель не ошибется, если сочтет алгоритм не абсолютно надежным. Это действительно так, и для таких задач практически

неизбежен некоторый риск. Мы можем повышать надежность различных элементов алгоритма, но лишь ценой существенного увеличения вычислительной работы, причем полной надежности никогда не достигнем (при конечном числе операций).

Итак, перейдем к описанию результатов вычислений, представленных в табл. 17. Поясним обозначения: ν — номер шага по x; m — число точек в множестве $y(x)$; F — значение $f^0 \cdot 10^5$, K — число вычислений функции f^0 (число вычислений производных f^0 по x или y составляет примерно $0.35K$). Заметим, что весь расчет занял 10 минут на БЭСМ-6.

Обратим внимание на 45-ю итерацию. В этот момент была найдена новая существенная точка множества $y(x)$, значение f^0 резко упало, затем ситуация выправилась. Разумеется, нет гарантии того, что задача решена очень точно. Однако стабилизация значений F и m в какой-то мере свидетельствует об этом. Во всяком случае, результаты

Таблица 17

ν	m	F	K	ν	m	F	K
1	11	3321	442	30	18	34413	7213
3	11	9158	667	33	18	34482	7773
6	12	16180	2196	36	18	34515	8151
9	14	24177	3061	39	18	34545	8731
12	16	29312	3685	42	18	34557	9676
15	17	31893	4468	45	19	29342	10260
18	18	33320	5187	48	18	31615	10812
21	18	33687	5724	51	17	33223	11332
24	18	34117	6275	54	18	34774	12146
27	18	34340	6650				

создают впечатление, что продолжение вычислений едва ли будет оправдано: либо задача решена, либо метод перестал работать.

Большую роль при этом играет репутация метода. Она создается решением большого числа задач, в которых результат удается так или иначе проконтролировать. Кстати, описываемую выше задачу автор заимствовал в одной из работ, в которой она решалась методом штрафных функций. Автор, скептически относясь к его возможностям, проконтролировал эти расчеты с помощью метода линеаризации и без труда обнаружил грубость полученных методом штрафных функций результатов (в некоторых случаях такую, что едва ли можно было говорить о приближенном решении задачи). Конечно, нельзя исключать и того, что кто-то таким же образом обнаружит ошибочность приближенного решения, найденного автором. Но пока этого не произошло.

§ 27. Дифференцирование функционалов

В самых различных задачах возникает необходимость использовать функциональные производные. Основным источником таких задач являются вариационные принципы, широко используемые в разных областях естествознания. Но есть и другие задачи, методы решения которых связаны с использованием функциональных производных, например нелинейные функциональные уравнения. В настоящее время сложилась достаточно общая формальная задача, которую иногда называют задачей оптимального управления, хотя это название не столько отражает существо дела, сколько является исторически сложившимся. Рассмотрим ее в общей форме

Имеется уравнение

$$R(x, u) = 0, \tag{1}$$

связывающее состояние некоторого объекта x с «управлением» u, т.е. с совокупностью функций и параметров, входящих в уравнение. Например, R может быть обозначением краевой задачи для уравнений в частных производных относительно x, а u в этой ситуации может обозначать функции и параметры, входящие в краевые и начальные условия или коэффициенты уравнения. Важным является следующее свойство уравнения (1), которое в абстрактной формулировке является, конечно, предположением. При любом «управлении» u уравнение (1) имеет решение и оно единственно. Более того, это решение $\mathscr{X}(u)$ зависит от u достаточно гладко, например непрерывно дифференцируемо по u.

Пусть по тем или иным причинам нас интересует, как изменяется решение x при малом изменении u. Точнее, нас не интересует полная картина изменения решения. Достаточно более грубой информации об изменении некоторых общих («усредненных») характеристик решения, или, проще говоря, некоторых функционалов от решения. Итак, пусть задана некоторая конкретная формула $\Phi(x, u)$, позволяющая вычислить значение Φ через x и u. Так как x однозначно определяется заданием u, можно ввести обозначение

$$F(u) \equiv \Phi(x, u), \quad \Phi \in R^1.$$

Здесь левая часть — абстрактный символ, означающий, что, коль скоро задан элемент u, можно вычислить число F. Правая часть расшифровывает способ вычисления: зная u, нужно решить уравнение (1), найти x и вычислить Φ, т.е. $F(u) = \Phi(\mathscr{X}(u), u)$.

Продифференцируем F, т.е. вычислим (в первом порядке) изменение F при малом изменении u на δu:

$$F(u + \delta u) \approx F(u) + \Phi_u(\mathscr{X}(u), u)\,\delta u + \Phi_x \mathscr{X}_u\,\delta u. \tag{2}$$

Таким образом, речь идет о дифференцировании суперпозиции функций. Дело осложняется тем, что зависимости $\mathscr{X}(u)$ мы явно не

имеем. Уравнение (1) обычно носит настолько сложный характер, что можно рассчитывать лишь на приближенное его решение при заданном u. Поэтому формула (2) неэффективна, ее следует заменить некоторыми выполнимыми операциями.

Первый шаг используемой в этих ситуациях техники — это прямое варьирование Ф. Считая, что малая вариация u приведет к малому возмущению состояния δx, запишем предварительную формулу, в которой производные Φ_x, Φ_u известны:

$$F(u+\delta u) \approx \Phi(x, u) + \Phi_x(x, u)\,\delta x + \Phi_u(x, u)\,\delta u. \tag{3}$$

Заменим линейный функционал $\Phi_x\,\delta x$ равным ему функционалом от δu, используя то, что δu однозначно определяет δx посредством так называемого *уравнения в вариациях*. Оно получается формальным варьированием уравнения (1):

$$R(x+\delta x, u+\delta u) \approx R(x, u) + R_x(x, u)\,\delta x + R_u(x, u)\,\delta u.$$

Отсюда, так как $R(x, u) = 0$, приходим к уравнению

$$R_x(x, u)\,\delta x + R_u(x, u)\,\delta u = 0. \tag{4}$$

Оно линейно относительно δx и δu и определено в той точке (x, u), в которой производится вычисление производной. Конечно, предполагается, что (4) однозначно разрешимо относительно δx при заданном δu.

Следующий шаг носит несколько искусственный характер. Используем тождество Лагранжа, являющееся в сущности определением сопряженного оператора:

$$(R_x\delta x, \psi) = (\delta x, R_x^*\psi), \quad \forall\,\psi. \tag{5}$$

Здесь ψ пока произвольно. При подходящем выборе ψ эта формула позволяет выражать линейный функционал от δx в виде линейного функционала от δu. Заметим, что нас интересует выражение $\Phi_x\delta x$, которое, конечно же, точнее следует записывать в виде скалярного произведения $(\Phi_x, \delta x)$. (Производной в смысле Фреше функционала $\Phi(x, u)$ по x, если она существует, является элемент пространства, двойственного к пространству элементов δx.) В качестве ψ возьмем решение «сопряженного» уравнения

$$R_x^*(x, u)\,\psi = \Phi_x(x, u). \tag{6}$$

Нетрудно сообразить, что в левой части (5) следует заменить $R_x\delta x$ на $-R_u\delta u$ в силу уравнения в вариациях (4). Объединяя эти преобразования, получаем

$$(\Phi_x, \delta x) = -(R_u\delta u, \psi) = -(R_u^*\psi, \delta u). \tag{7}$$

Подставляя (7) в (3), мы имеем окончательную формулу для вычисления функциональной производной:

$$(F_u(u), \delta u) = (\Phi_u(x, u) - R_u^*\psi, \delta u).$$

Итак,

$$F_u(u) = \Phi_u(x, u) - R_u^*\psi.$$

Подведем итог, перечислив вычисления, которые дают функциональную производную F_u в точке u. Имея u, можно решить уравнение (1) и получить x; имея x, u, можно сформировать уравнение (6). При этом мы неявно предполагаем, что операции дифференцирования по x и u оператора R и функционала Φ являются элементарными. Во многих достаточно сложных задачах это действительно очень простые операции, но встречаются и более сложные ситуации, в которых не так-то просто разобраться. Решая уравнение (7), находим ψ и вычисляем функциональную производную F_u.

Выше была приведена общая схема дифференцирования функционалов, определенных на решениях функционального уравнения. В изложении мы опустили многочисленные тонкости строгого математического оформления схемы, выделяя содержательно существенные моменты. По этой схеме ниже мы рассмотрим более аккуратно характерные конкретные примеры.

Дифференцирование функционалов от решений обыкновенных дифференциальных уравнений. Рассмотрим ситуацию, которая связана с задачами оптимального управления в первоначальном смысле этого слова (см. § 28). Изучается система дифференциальных уравнений, «управляемая» выбором функции $u(\cdot)$ и параметров p:

$$\dot{x} = f(x, u, p), \quad x(0) = \mathscr{X}_0(p), \quad 0 \leqslant t \leqslant T. \tag{8}$$

Траектория системы (8) полностью определена заданием управления $\{u(\cdot), p\}$. Пусть управление подверглось малому возмущению: $u(\cdot) \to u(\cdot) + \delta u(\cdot)$, $p \to p + \delta p$.

Возникает вопрос: что значит «малое возмущение $\delta u(\cdot)$»? Пока ограничимся самым простым случаем, считая, что $\max_t \|\delta u(t)\| = O(\varepsilon)$, $\|\delta p\| = O(\varepsilon)$. (Если вид нормы не конкретизирован, можно считать, что $\|\cdot\|$ — любая из употребляемых в конечномерных пространствах норм.) В нижеследующих выкладках используется тривиальная теория малых возмущений первого порядка (см. § 19). Прежде всего необходимо установить, что малое возмущение управления порождает, соответственно, малое возмущение траектории. Обозначим через $x(t, u(\cdot), p)$ решение задачи Коши (8), определяемое управлением

$\{u(\cdot), p\}$, через $\Delta x(t)$ — приращение $x(t)$, вызванное возмущением управления:

$$\Delta x(t) = x(t, u(\cdot) + \delta u(\cdot), p + \delta p) - x(t, u(\cdot), p).$$

Оценка $\|\Delta x(t)\| = O(\varepsilon)$ устанавливается аналогично тому, как исследовался ряд Пуассона в § 19.

Пусть определен функционал от $\{u(\cdot), p\}$:

$$F[u(\cdot), p] \equiv \int_0^T \Phi(x(t), u(t), p)\, dp, \tag{9}$$

где Φ — заданная гладкая функция. Мы используем символ Δ в идентификаторах, присваиваемых точным приращениям величин, символ δ — в идентификаторах вариаций этих величин (Δ от δ отличаются в следующем по ε порядке). Символ $u(\cdot)$ означает функцию, взятую в целом как аргумент функционала; $u(t)$ есть точка конечномерного пространства (сечение $u(\cdot)$ в точке t).

Вычислим δF прямым варьированием формулы (9). Подставляя в правую часть $x + \delta x$, $u + \delta u$, $p + \delta p$ и используя первые члены ряда Тейлора, имеем

$$\delta F[\delta u(\cdot), \delta p] = \int_0^T \{\Phi_x[t]\, \delta x(t) + \Phi_u[t]\, \delta u(t) + \Phi_p[t]\, \delta p\}\, dt.$$

Здесь $\Phi_x[t]$ обозначает $\Phi_x[x(t), u(t), p]$ — зависящую от t матрицу, определенную в той точке $\{u(\cdot), p\}$, в которой вычисляется производная. Итак, получена формула типа (3). Выпишем уравнение в вариациях (4) таким же формальным варьированием уравнения (8):

$$\delta\dot{x} = f_x[t]\, \delta x + f_u[t]\, \delta u + f_p[t]\, \delta p, \quad \delta x(0) = \mathscr{X}_p(p)\, \delta p.$$

Это и есть уравнение в вариациях. В нем δx можно заменить на Δx, добавив к правой части выражение $o(\varepsilon)$.

Используем элемент общей схемы — тождество Лагранжа:

$$\int_0^T \left(\psi, \left[\frac{d}{dt} - f_x\right] \delta x\right) dt - \int_0^T \left(\left[\frac{d}{dt} - f_x\right]^*, \delta x\right) dt = (\delta x, \psi)\Big|_0^T. \tag{10}$$

Вывод (10) сводится к интегрированию по частям, определению f_x^* и к соотношению $(d/dt)^* = -d/dt$. Заключительный шаг преобразований требует нехитрого угадывания вида правой части сопряженного уравнения, с тем чтобы выражение $(-\dot{\psi} - f_x^*\psi, \delta x)$ превратилось в $(\Phi_x, \delta x)$. Очевидно, в качестве ψ следует взять решение уравнения

$$-\dot{\psi} = f_x^*[t]\, \psi + \Phi_x[t]. \tag{11}$$

Преобразуем выражение $(\delta x(T), \psi(T)) - (\delta x(0), \psi(0))$, заменяя $\delta x(0) = \mathscr{X}_p(p)\,\delta p$ и уничтожая лишнее слагаемое выбором значения $\psi(T) = 0$. Это то краевое условие, которое однозначно определяет $\psi(t)$ как решение задачи Коши.

Итак, имеем окончательный результат:

$$\delta F[\delta u(\cdot), \delta p] = \int_0^T (w(t), \delta u(t))\,dt + (W, \delta p),$$

где функциональные производные суть

$$\begin{aligned} w(t) &= \frac{\partial F[u(\cdot), p]}{\partial u(t)} = \Phi_u[t] + (f_u^*[t], \psi(t)), \\ W &= \frac{\partial F[u(\cdot), p]}{\partial p} = \mathscr{X}_p^*(p)\,\psi(0) + \int_0^T (f_p^*, \psi)\,dt + \int_0^T \Phi_p[t]\,dt. \end{aligned} \tag{12}$$

Проведенные выше выкладки по существу содержат в себе доказательство дифференцируемости функционала (9) по Фреше. В качестве упражнения рекомендуем читателю проделать аналогичные вычисления в следующих, мало отличающихся от рассмотренной ситуациях:

а) Пусть $F[u(\cdot), p] \equiv \Phi[x(t^*), p]$, где Φ и t^* заданы. (Отдельно необходимо рассмотреть часто встречающийся в приложениях случай $t^* = T$.)

б) Пусть вместо начальных данных Коши $x(0) = \mathscr{X}(p)$ заданы общие краевые условия $\mathscr{X}(x(0), x(T), p) = 0$. Существование, единственность решения такой краевой задачи, и гладкую его зависимость от управления следует предположить (доказательство этих свойств — отдельная наука, которой мы здесь не касаемся).

Конечные вариации u на множествах малой меры. Важным элементом современного вариационного исчисления является следующий класс «малых» возмущений управления, также приводящих к малому (в обычном смысле слова) возмущению траектории и функционалов. Пусть задано невозмущенное управление $u(\cdot)$ и соответствующая ему траектория $x(t)$ (параметры p, ради простоты, опустим). Рассмотрим другое, в некотором смысле близкое, управление:

$$\tilde{u}(t) = \begin{cases} v(t), & t \in \mu, \\ u(t), & t \notin \mu. \end{cases} \tag{13}$$

Здесь $v(\cdot)$ — некоторая функция того же типа, что и u, причем $\|u(t) - v(t)\| = O(1)$, μ — некоторое множество малой меры: $\operatorname{mes} \mu = \varepsilon$. Конструкции типа (13) называются конечными возмущениями управления на множествах малой меры (рассматриваются всевозможные множества μ и функции $v(\cdot)$).

Пусть управлению $\tilde{u}(t)$ соответствует возмущенная траектория $\tilde{x}(t)$. Тогда (для задачи (8), во всяком случае) можно получить оценку $\Delta x(t) = O(\varepsilon)$. Вычисление вариации функционала (теперь уже лучше не говорить о производной — она в данном случае не определена) проводится по той же схеме, что и раньше, но некоторые детали следует уточнить.

Итак, сначала вычислим вариацию функционала

$$\Delta F = \int_0^T \Phi(x + \Delta x, \tilde{u})\, dt - \int_0^T \Phi(x, u)\, dt \approx$$
$$\approx \int_0^T \Phi_x[t]\, \Delta x(t)\, dt + \int_\mu \{\Phi(x(t), v(t)) - \Phi(x(t), u(t))\}\, dt. \quad (14)$$

Кроме обычных формул Тейлора, в (1′) использовано следующее:

а) в члене $\Phi_x \Delta x$ производная Φ_x везде вычисляется в точке $u(t)$; это неверно при $t \in \mu$, но мера μ мала и связанная с этим погрешность есть, очевидно, $O(\varepsilon^2)$;

б) по тем же причинам в последнем интеграле в (14) $x + \Delta x$ заменено на x.

Перейдем к уравнению в вариациях. Имеем уравнения

$$\dot{x} = f(x, u), \qquad \dot{\tilde{x}} = f(\tilde{x}, \tilde{u}), \qquad x(0) = \tilde{x}(0).$$

Беря их разность, проделаем простые преобразования:

$$\Delta\dot{x} = f(x + \Delta x, \tilde{u}) - f(x, u) =$$
$$= f_x[t]\, \Delta x(t) + \{f(x(t), \tilde{u}(t)) - f(x(t), u(t))\} + R(t). \quad (15)$$

Здесь $R(t)$ — разность между точным значением $\Delta\dot{x}$ и первыми тремя членами последнего в (15) выражения. По тем же соображениям, которые использовались в преобразованиях (14), можно показать, что $R(t) = O(\varepsilon^2)$ при $t \notin \mu$ и $R(t) = O(\varepsilon)$ при $t \in \mu$. Поэтому для вариации $\delta x(t) = \Delta x(t) + O(\varepsilon^2)$ можно использовать уравнение в вариациях в форме

$$\delta\dot{x} = f_x[t]\, \delta x + \{f(x(t), \tilde{u}(t)) - f(x(t), u(t))\}, \quad \delta x(0) = 0. \quad (16)$$

Выражение в фигурных скобках есть, очевидно, величина $O(1)$ при $t \in \mu$; оно обращается в нуль при $t \notin \mu$.

Тождество Лагранжа (10) берется в той же форме (10); уравнение для ψ такое же, как (11). В результате получаем окончательную формулу для вариации функционала:

$$\delta F[\mu, v(\cdot)] = \int_\mu \{\Phi[x(t), v(t)] - \Phi[x(t), u(t)]\}\, dt +$$
$$+ \int_\mu (\psi(t), f(x(t), v(t)) - f(x(t), u(t)))\, dt. \quad (17)$$

В вариационном исчислении различают *слабый относительный минимум* — это точка $u(\cdot)$, которая не может быть «улучшена» при возмущениях управления вариациями $\delta u(\cdot)$, малыми в обычной метрике (типа C), и *сильный относительный минимум* — это точка $u(\cdot)$, которую нельзя улучшить, используя конечные вариации на множествах малой меры. Современные теории, если это удается, строят именно как теории сильного экстремума.

Дифференцирование спектра. В приложениях часто встречаются задачи следующего типа. Состояние системы $x(t)$ определяется как та или иная (например, главная) собственная функция линейного дифференциального оператора, зависящего от «управления» u:

$$L(u)\, x = \lambda x. \tag{18}$$

В таком компактном виде записываются как дифференциальное уравнение, так и краевые условия, которые обычно оформляются указанием на принадлежность x некоторому линейному пространству функций. Это пространство определяется числом необходимых производных и краевыми условиями. Уравнение (18) следует дополнить четким указанием о том, какая именно точка спектра имеется в виду в данной задаче и как она нормируется.

Итак, мы считаем, что задание u однозначно определяет как x, так и λ. Рассмотрим функционал и его прямую вариацию:

$$F(u) \equiv \lambda, \qquad \delta F[\delta u] = \delta\lambda. \tag{19}$$

Предположим, для простоты, что спектр вещественный, дискретный и непрерывно зависит от u. (Этот факт нужно доказывать и это делается в соответствующих разделах теории. Мы будем действовать формально.) Выпишем уравнение в вариациях. Оно получается теми же операциями — подстановкой в (18) $u + \delta u$, $x + \delta x$, $\lambda + \delta\lambda$, использованием первых членов ряда Тейлора и группировкой членов одного порядка малости:

$$L(u)\, \delta x + M(u, x)\, \delta u = \lambda\, \delta x + \delta\lambda\, x. \tag{20}$$

Здесь u — заданное управление, x и λ — соответствующие ему собственные функция и число. Таким образом, относительно δx мы имеем линейное дифференциальное уравнение с переменными коэффициентами ($M = L_u x$ — матрица, зависящая от x, u).

Представим (20) в другой форме:

$$(L - \lambda E)\, \delta x = -M\, \delta u + \delta\lambda\, x.$$

Относительно δx это есть вырожденная задача (задача «на спектре»). Как известно, она имеет решение только в случае, когда правая часть ортогональна собственной функции сопряженного к L оператора, соответствующей той же точке спектра λ (или $\bar{\lambda}$, если оператор L несамосопряженный). Обозначим эту функцию ψ, т.е.

$L^*(u)\psi = \bar{\lambda}\psi$. Тогда условие существования решения (20) есть $(-M\,\delta u + \delta\lambda\, x, \psi) = 0$. Отсюда получаем формулу

$$\delta\lambda(\delta u) = (M^*(u, x)\psi, \delta u)/(x, \psi). \tag{21}$$

Заметим, что в (18) x есть функция, заданная в некоторой области Ω, u может быть функцией, заданной в Ω, а может быть определена только на ее границе $\partial\Omega$ (если u есть коэффициент, входящий в линейные однородные краевые условия). Возможен и такой случай, когда u есть комплекс, одна компонента которого определена в Ω, другая — на $\partial\Omega$. Поэтому в (21) скалярное произведение (x, ψ) есть интеграл по Ω, $(M^*\psi, \delta u)$ может состоять из интеграла по Ω и интеграла по $\partial\Omega$. В этом случае M^* отображает функцию, определенную в Ω, в комплекс, одна компонента которого есть функция, определенная в Ω, другая — на $\partial\Omega$. Очевидно, функция M линейно зависит от x. Поэтому правая часть (21) не зависит от способа нормировки x, ψ.

Формулы типа (21) используются, например, при решении вариационных задач для математических моделей ядерных реакторов (их состояние определяется главной собственной функцией некоторой краевой задачи для системы уравнений в частных производных),при оптимизации некоторых конструкций (например, мембран, важные технические характеристики которых выражаются через частоты собственных колебаний) и т.п.

Варьирование слабого разрыва. Рассмотрим задачу, в которой траектория имеет точку слабого разрыва, причем сама эта точка при варьировании управления меняет свое положение. С такими ситуациями имеют дело в случае, когда правая часть уравнения меняется при пересечении траекторией некоторой заданной поверхности в фазовом пространстве. Итак, рассматривается обычная задача для управляемой системы обыкновенных дифференциальных уравнений вида $(x(0) = \mathscr{X},\ 0 \leqslant t \leqslant T)$

$$\dot{x} = \begin{cases} f(x, u), & G(x(t)) < 0, \\ \tilde{f}(x, u), & G(x(t)) > 0. \end{cases}$$

Ради простоты предположим, что исследуемая траектория $x(t)$ только один раз пересекает поверхность $G(x) = 0$, причем пересекает, как говорят, версально, без касания. Другими словами, требуется (при всех рассматриваемых значениях u) выполнение неравенств

$$(f(x, u), G_x(x)) > 0, \quad (\tilde{f}(x, u), G_x(x)) > 0,$$

где x — точка, в которой анализируемая траектория пересекает поверхность $G = 0$. Это условие существенно. Если оно не выполняется, перестает работать теория малых возмущений: малое возмущение управления может привести к конечному ($O(1)$) изменению траектории.

Рисунок 48 иллюстрирует сказанное. На нем показана линия G, на которой рвется поле направлений рассматриваемой системы уравнений, и две траектории. Одна из них пересекает поверхность $G = 0$ версально. Близкая к ней в области $G < 0$ траектория после пересечения поверхности разрыва остается близкой. Мы будем анализировать только этот случай. Другая траектория пересекает поверхность, касаясь ее. Близкая к ней в области $G < 0$, траектория не пересекает поверхности $G = 0$, и такие траектории расходятся на расстояние $O(1)$ как бы ни были они близки до приближения к поверхности разрыва. По существу в этом случае не работает теорема о единственности решения задачи Коши.

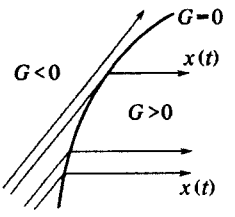


Рис. 48

Ниже мы ограничимся только тем основным моментом, которым эта задача отличается от стандартной. Рассмотрим вывод формулы

$$\int_0^T (Y(t), \delta x(t))\, dt = \int_0^T (w(t), \delta u(t))\, dt. \tag{22}$$

Здесь Y — заданная функция, w — функция, подлежащая вычислению. Пусть исследуемая траектория $x(t)$, порождаемая управлением $u(\cdot)$, пересекает поверхность $G = 0$ в момент t^*, а траектория, порожденная возмущенным управлением $u(\cdot) + \delta u(\cdot)$, пересекает эту поверхность в момент $t^* + \delta$. Для определенности, будем считать $\delta > 0$ (случай $\delta < 0$ приводит к тем же формулам); очевидно, $\delta = O(\|\delta u\|)$. Уравнение в вариациях имеет вид

$$\delta\dot{x} - f_x[t]\,\delta x = f_u[t]\,\delta u, \qquad 0 \leqslant t \leqslant t^*,$$

$$\delta\dot{x} - \tilde{f}_x[t]\,\delta x = \tilde{f}_u[t]\,\delta u, \quad t^* + \delta \leqslant t \leqslant T.$$

Тождество Лагранжа записывается очевидным образом:

$$\int_0^{t^*} [(\psi, \delta\dot{x} - f_x\,\delta x) + (\delta x, \dot{\psi} + f_x^*\,\psi)]\, dt +$$

$$+ \int_{t^*+\delta}^{T} [(\psi, \delta\dot{x} - \tilde{f}_x\,\delta x) + (\delta x, \dot{\psi} + \tilde{f}_x^*\,\psi)]\, dt + (\psi, \delta x)\,|_0^T - (\psi, \delta x)\,|_{t^*}^{t^*+\delta}. \tag{23}$$

Так как в левой части соотношения (22) можно пренебречь величиной $\int_{t^*}^{t^*+\delta} (Y, \delta x)\, dt = O(\|\delta u\|^2)$, то превращение (23) в (22) осуществляется стандартным подбором правой части для сопряженного

уравнения. Мешает только одно слагаемое в (23): $-(\psi, \delta x)\,|_{t^*}^{t^*+\delta}$. Мы уберем его за счет разрыва ψ в точке t^*.

Введем для возмущенной траектории обозначение $y(t)$, $v(t)$. Тогда с точностью до величин $O(\|\delta u\|^2)$ имеем

$$x(t^* + \delta) = x^* + \delta \tilde{f}(x^*, u^*), \quad x^* = x(t^*), \quad u^* = u(t^*),$$

$$y(t^* + \delta) = y^* + \delta f(y^*, v^*), \quad y^* = y(t^*), \quad v^* = v(t^*).$$

Вычитая, получаем

$$\delta x(t^* + \delta) = \delta x(t^*) + \delta \cdot [f(y^*, v^*) - \tilde{f}(x^*, u^*)]. \tag{24}$$

Здесь мы неявно предполагали непрерывность управлений u, v в точке t^*.

Используем связь δ с $\delta x(t^*)$:

$$G(y(t^* + \delta)) = 0 = G(y^* + \delta \cdot f(y^*, v^*)) = G(x^* + \delta x^* + \delta \cdot f(y^*, v^*)) =$$
$$= G(x^*) + G_x(x^*)\, \delta x^* + \delta \cdot G_x(x^*)\, f(y^*, v^*).$$

Так как $G(x^*) = 0$, то

$$\delta = -\frac{(G_x(x^*), \delta x^*)}{(G_x(x^*), f(y^*, v^*))} = -\frac{(G_x(x^*), \delta x^*)}{(G_x(x^*), f(x^*, u^*))}.$$

Второе равенство, конечно, неточное, но мы пренебрегаем малыми второго порядка, возникающими при замене y^* на x^* и v^* на u^*.

Подставляя найденное значение δ в (24), имеем

$$\delta x(t^* + \delta) = \delta x(t^*) - \frac{(G_x, \delta x^*)}{(G_x, f_1)} (f_1 - f_2),$$

где $f_1 = f(x^*, u^*)$, $f_2 = \tilde{f}(x^*, u^*)$. Легко подобрать такой скачок между величинами $\psi^- = \psi(t^*)$ и $\psi^+ = \psi(t^* + \delta)$, чтобы в первом порядке можно было уничтожить мешающие нам слагаемые в (23):

$$(\psi^+, \delta x(t^* + \delta)) - (\psi^-, \delta x(t^*)) = (\psi^+ - \psi^-, \delta x^*) -$$
$$- \frac{(G_x, \delta x^*)}{(G_x, f_1)} (f_1 - f_2, \psi^+) = \left(\psi^+ - \psi^- - \frac{(f_1 - f_2, \psi^+)}{(G_x, f_1)} G_x, \delta x^*\right).$$

Из полученного выражения вытекает требуемый результат. Функцию $\psi(t)$ следует взять как решение сопряженного уравнения

$$\dot{\psi} + \tilde{f}_x[t]\psi = -Y(t), \quad t \in (t^*, T),$$

$$\dot{\psi} + f_x[t]\psi = -Y(t), \quad t \in (0, t^*),$$

с начальными данными $\psi(T) = 0$ и условием скачка

$$\psi(t^* - 0) = \psi(t^* + 0) - \frac{(f_2 - f_1, \psi(t^* + 0))}{(G_x(x^*), f_1)} G_x(x^*). \qquad (25)$$

Здесь мы провели еще одно обобщение: заменили $\psi^+ = \psi(t^* + \delta)$ на $\psi(t^* + 0)$. Предоставим читателю несложную проверку того, что эта операция допустима в первом порядке теории возмущений. Если читатель повторит приведенный выше вывод для случая $\delta < 0$, он получит условие скачка в несколько иной форме:

$$\psi^+ = \psi^- - \frac{(f_2 - f_1, \psi^-)}{(G_x, f_2)} G_x(x^*).$$

Можно показать, что эти условия равносильны.

Подчеркнем, что относительная сложность вычислений связана с тем, что точка разрыва производной t^* варьируется при изменении управления. Если речь идет о разрыве правой части уравнения в фиксированный момент времени, стандартная техника вычисления производных не требует никаких изменений.

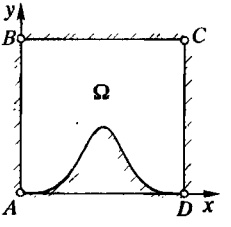


Рис. 49

Дифференцирование по границе области. Рассмотрим задачу, в которой состояние некоторого объекта определяется решением краевой задачи в некоторой области. «Качество» этого состояния оценивается функционалом от решения. Пусть форма области не фиксирована, но может в тех или иных пределах меняться. Это изменение следует производить с целью улучшения качества объекта. Перейдем к более конкретной постановке задачи. Рассмотрим модельную задачу, в которой, однако, присутствуют те моменты техники дифференцирования, которые мы хотим разъяснить.

Предположим, что состояние объекта описывается функцией $\mathscr{Z}(x, y)$, определенной в области Ω (рис. 49) и являющейся решением краевой задачи для уравнения Пуассона (Δ — оператор Лапласа)

$$\Delta\mathscr{Z} = f(x, y), \quad (x, y) \in \Omega, \qquad (26)$$

с краевым условием, для определенности, первого рода:

$$\mathscr{Z}(x, y)|_{\partial\Omega} = \varphi(s(x, y)). \qquad (27)$$

Здесь f, φ — заданные функции, форма области не фиксирована. Чтобы избежать чисто технических усложнений, будем считать, что граница $ABCD$ фиксирована и только ее часть AD может варьироваться. Пусть качество состояния оценивается функционалом

$$\int_0^1 \Phi(\mathscr{Z}_y(x, 1))\, dx, \qquad (28)$$

где Ф — заданная функция. Таким образом, (28) есть функционал от формы области, вычисляемый по очевидной схеме: если область задана, решается краевая задача (26), (27), затем вычисляется интеграл (28).

Для того чтобы продифференцировать (28), нужно выбрать форму задания границы. Допустим, что граница AD задана параметрически координатами $\xi(t)$, $\eta(t)$ $(0 \leqslant t \leqslant 1)$, а сами эти функции являются решением краевой задачи:

$$\ddot{\xi} = u_1(t), \quad \ddot{\eta} = u_2(t), \quad \xi(0) = \eta(0) = \eta(1) = 0, \ \xi(1) = 1. \tag{29}$$

Такая (или аналогичная) форма задания бывает удобна, когда возникает необходимость ограничить геометрические характеристики границы (кривизну, например). Итак, основным независимым аргументом в задаче является вектор-функция $u(\cdot) = \{u_1(\cdot), u_2(\cdot)\}$, по традиции называемая управлением. Интеграл (28) можно обозначить $F[u(\cdot)]$. Алгоритм его вычисления начинается с решения краевой задачи (29), далее — как было описано выше.

Вычисление производной начинается с прямой вариации функционала. Эта операция дает очевидную формулу

$$\delta F[\delta u(\cdot)] = \int_0^1 \Phi'(\mathscr{X}_y)\, \delta\mathscr{X}_y(x, 1)\, dx, \tag{30}$$

которую нужно преобразовать в выражение

$$\delta F[\delta u(\cdot)] = \int_0^1 (w_1(t)\, \delta u_1(t) + w_2(t)\, \delta u_2(t))\, dt.$$

Ниже описывается, как вычисляются w_1, w_2. Схема рассуждений обычная: вариация $u(\cdot)$ малыми величинами $\delta u(\cdot)$ приводит к малому изменению дуги AD, это влечет малое изменение состояния $(\mathscr{X} \to \mathscr{X} + \delta\mathscr{X})$, после чего по (30) вычисляется малое изменение функционала.

Основной элемент вычисления производной в данной ситуации — это правильная формулировка уравнения в вариациях; сложность здесь в том, что варьируется область. Выпишем уравнение для возмущенного состояния:

$$\Delta\tilde{\mathscr{X}} = f(x, y), \quad (x, y) \in \tilde{\Omega}. \tag{31}$$

Некоторые сложности связаны и с краевым условием для $\tilde{\mathscr{X}}$ на дуге AD, причем дело не в том, что нужно аккуратно разобраться в этом вопросе, а в том, что нужно уточнить постановку задачи. Пусть формально эти условия записаны в виде

$$\tilde{\mathscr{X}}(\tilde{\xi}(t), \tilde{\eta}(t)) = \tilde{\varphi}(t) = \varphi(t) + \delta\varphi(t), \quad 0 \leqslant t \leqslant 1;$$

при этом $\delta\varphi(t)$ не вычисляется, а задается постановкой задачи.

Рассмотрим вариацию $\delta\mathscr{Z}(x, y)$ и сформируем для нее краевую задачу. Учтем, что $\mathscr{Z}$ и $\tilde{\mathscr{Z}}$ определены в разных, хотя и мало отличающихся областях. На рис. 50 показана часть области (границы Ω и $\tilde{\Omega}$). Введем вспомогательную функцию $\overline{\mathscr{Z}}$, определенную в невозмущенной области Ω и мало отличающуюся от $\mathscr{Z}$ там, где последняя имеет смысл:

$$|\overline{\mathscr{Z}}(x, y) - \tilde{\mathscr{Z}}(x, y)| = O(\|\delta u\|^2), \quad (x, y) \in \Omega \cap \tilde{\Omega}.$$

Введя обозначение

$$\delta\mathscr{Z}(x, y) = \overline{\mathscr{Z}}(x, y) - \mathscr{Z}(x, y), \quad (x, y) \in \Omega,$$

получим для этой функции уравнение в вариациях.

Возмущенную границу удобно описывать с помощью скалярных функций $\alpha(t)$ и $\tau(t)$:

$$\begin{aligned} \tilde{\xi}(t + \tau(t)) &= \xi(t) + \alpha(t) n_1(t), \\ \tilde{\eta}(t + \tau(t)) &= \eta(t) + \alpha(t) n_2(t), \end{aligned} \tag{32}$$

где α, $\tau = O(\|\delta u\|)$, $n = \{n_1, n_2\}$ — внешняя нормаль к $\partial\Omega$ в точке t, $\alpha(t)$ — смещение $\partial\tilde{\Omega}$ относительно $\partial\Omega$ по нормали, $\tau(t)$ — малое возмущение параметра t. Можно обойтись и без этого смещения, изменив параметризацию $\partial\tilde{\Omega}$ таким образом. Припишем значение t точке пересечения $\partial\tilde{\Omega}$ с прямой $\{\xi(t) + \alpha n_1(t), \eta(t) + \alpha n_2(t)\}$. При этом надо пересчитать $\delta\varphi(t)$, увеличив ее, очевидно, на $\tau(t)\varphi_t(t)$. Будем считать эту операцию проделанной и уберем τ из (32). Величина $\alpha(t)$, конечно, функционально зависит от $\delta u(\cdot)$, и это в дальнейшем будет учтено. Величины $\delta\varphi$, α, τ суть малые первого порядка; только этот порядок и будет учитываться в дальнейших выкладках.

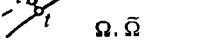


Рис. 50

Определим $\overline{\mathscr{Z}}(x, y)$ в Ω решением уравнения

$$\Delta\overline{\mathscr{Z}} = f(x, y), \quad (x, y) \in \Omega.$$

Краевые условия на границе $ABCD$ — те же, что и для $\mathscr{Z}$ (и $\tilde{\mathscr{Z}}$). На невозмущенной дуге AD поставим краевое условие

$$\overline{\mathscr{Z}}(\xi(t), \eta(t)) + \alpha(t) \frac{\partial\overline{\mathscr{Z}}}{\partial n} = \varphi(t) + \delta\varphi(t).$$

Смысл этого условия очевиден: функция $\overline{\mathscr{Z}}(x, y)$ на возмущенной дуге $\partial\tilde{\Omega}$ с точностью до малых второго порядка совпадает с $\tilde{\mathscr{Z}}$. Если какая-то часть $\partial\tilde{\Omega}$ лежит вне Ω, речь идет об экстраполяции значений $\overline{\mathscr{Z}}$ в малой окрестности $\partial\Omega$. Такая операция корректна, если

граница $\partial\Omega$ достаточно гладкая, что обеспечивается определением границы уравнениями (29) при некоторых ограничениях на величину $\|u(t)\|$, которые мы неявно считаем выполненными.

Таким образом, можно считать, что $\overline{\mathscr{X}}(x, y)$ определена и удовлетворяет в $\tilde{\Omega}$ тому же уравнению (стало быть, предполагается, что $f(x, y)$ определена в окрестности Ω и является достаточно гладкой) и совпадает с $\tilde{\mathscr{X}}$ на границе $\tilde{\Omega}$ с точностью до малых второго порядка. Следовательно, $\left|\overline{\mathscr{X}}(x, y) - \tilde{\mathscr{X}}(x, y)\right|$ есть величина второго порядка. Этой разницей мы пренебрегаем, т.е. вариацию $\delta\mathscr{X}$, определенную как $\overline{\mathscr{X}} - \mathscr{X}$, можно использовать в дальнейшем и в смысле $\delta\mathscr{X} = \tilde{\mathscr{X}} - \mathscr{X}$. Вычитая из (31) невозмущенное уравнение (26), получаем уравнение в вариациях:

$$\Delta\delta\mathscr{X} = 0, \quad (x, y) \in \Omega, \tag{33}$$

с краевыми условиями $\delta\mathscr{X}(x, y) = 0$ на границе $ABCD$. На границе AD краевое условие имеет вид

$$\delta\mathscr{X} + \alpha(t)\,\mathscr{X}_n = \delta\varphi(t), \quad t \in [0, 1]. \tag{34}$$

Здесь мы заменили $\overline{\mathscr{X}}_n$ на $\mathscr{X}_n$, отбросив возникающую при этом погрешность второго порядка.

Теперь используем тождество Лагранжа:

$$\iint\limits_{\Omega} (\Psi\,\Delta\delta\mathscr{X} - \delta\mathscr{X}\,\Delta\Psi)\,dx\,dy = \oint\limits_{\partial\Omega} \Psi\,\delta\mathscr{X}_n\,dl - \oint\limits_{\partial\Omega} \delta\mathscr{X}\,\Psi_n\,dl, \tag{35}$$

где l — длина дуги на $\partial\Omega$. Простой подбор краевых условий для Ψ позволяет получить из (35) выражение для δF через интеграл от возмущений на AD. В самом деле, принимая для $\Psi(x, y)$ уравнение $\Delta\Psi = 0$ и учитывая (33), обращаем в нуль левую часть (35). В силу краевых условий $\delta\mathscr{X} = 0$ на $ABCD$ второй интеграл в правой части превращается в $\int\limits_A^D \delta\mathscr{X}\,\Psi_n\,dl$. Первый же интеграл правой части превратим в δF, определив краевые условия для Ψ следующим образом:

$$\Psi = 0 \text{ на } BADC, \qquad \Psi = \Phi'[\mathscr{X}_y(x, 1)] \text{ на } BC. \tag{36}$$

Таким образом, краевая задача для Ψ полностью сформулирована. Заменяя $\delta\mathscr{X}$ на AD из краевого условия (34), получаем

$$\delta F = \int\limits_A^D \Psi_n\,\delta\mathscr{X}\,dl = \int\limits_0^1 \Psi_n[t]\{\delta\varphi(t) - \alpha(t)\,\mathscr{X}_n[t]\}\,\frac{dl}{dt}\,dt. \tag{37}$$

Здесь мы используем обозначение типа $\Psi_n[t] \equiv \Psi_n(\xi(t), \eta(t))$. Линейный функционал от $\alpha(\cdot)$ следует преобразовать в функционал от $\delta u(\cdot)$, что достигается сравнительно стандартными выкладками. Сначала находим внешнюю нормаль к AD в точке t:

$$n(t) = \{\dot\eta, \ -\dot\xi\}/\sqrt{\dot\xi^2 + \dot\eta^2}, \qquad dl = \sqrt{\dot\xi^2 + \dot\eta^2}\, dt.$$

Вычисляем $\alpha(t)$, выписывая условия пересечения нормали с возмущенной границей:

$$\xi(t + \tau) + \delta\xi(t + \tau) = \xi(t) + \alpha n_1(t),$$

$$\eta(t + \tau) + \delta\eta(t + \tau) = \eta(t) + \alpha n_2(t).$$

Пренебрегая малыми второго порядка, получаем систему линейных уравнений относительно α и смещения параметра τ:

$$\dot\xi\tau + \delta\xi(t) = \alpha n_1, \qquad \dot\eta\tau + \delta\eta(t) = \alpha n_2,$$

откуда

$$\alpha(t) = (\dot\eta(t)\ \delta\xi(t) - \dot\xi(t)\ \delta\eta(t))/\sqrt{\dot\xi^2 + \dot\eta^2}.$$

Используя полученные выражения, преобразуем формулу (37):

$$\delta F = \int_0^1 \Psi_n[t]\ \delta\varphi(t)\, dt - \int_0^1 \Psi_n[t]\ \mathscr{K}_n[t]\ (\dot\eta\ \delta\xi - \dot\xi\ \delta\eta)\, dt.$$

В дальнейшем мы будем преобразовывать в функционал от $\delta u(\cdot)$ только второй интеграл правой части. Такие же преобразования должны быть проделаны и над первым интегралом, но сначала (в зависимости от точной трактовки краевого условия на любой допустимой дуге AD) этот интеграл должен быть преобразован в функционал от $\delta\xi(\cdot)$, $\delta\eta(\cdot)$.

Выпишем очевидное уравнение в вариациях:

$$\delta\ddot\xi = \delta u_1, \qquad \delta\ddot\eta = \delta u_2, \qquad \delta\xi(0) = \delta\xi(1) = \delta\eta(0) = \delta\eta(1) = 0,$$

и соответствующие тождества Лагранжа:

$$\int_0^1 (\psi_1\ \delta\ddot\xi - \delta\xi\ \ddot\psi_1)\, dt = [\psi_1\ \delta\dot\xi - \delta\xi\ \dot\psi_1]_0^1,$$

$$\int_0^1 (\psi_2\ \delta\ddot\eta - \delta\eta\ \ddot\psi_2)\, dt = [\psi_2\ \delta\dot\eta - \delta\eta\ \dot\psi_2]_0^1.$$

Взяв в качестве ψ_1, ψ_2 решения краевых задач

$$\begin{aligned} \ddot{\psi}_1 &= -\Psi_n[t]\,\mathscr{Z}_n[t]\,\dot{\eta}(t), \quad \psi_1(0)=\psi_1(1)=0, \\ \ddot{\psi}_2 &= \Psi_n[t]\,\mathscr{Z}_n[t]\,\dot{\xi}(t), \qquad \psi_2(0)=\psi_2(1)=0, \end{aligned} \tag{38}$$

мы, очевидно, получим

$$\int_0^1 \Psi_n[t]\,\mathscr{Z}_n[t]\,(\dot{\eta}\,\delta\xi - \dot{\xi}\,\delta\eta)\,dt = \int_0^1 (\psi_1(t)\,\delta u_1(t) + \psi_2(t)\,\delta u_2(t))\,dt,$$

т.е. требуемый результат.

Подведем итог, перечислив последовательность операций, выполняемых при вычислении производной функционала:

1) задано невозмущенное управление $u_1(t)$, $u_2(t)$ $(t \in [0, 1])$;

2) решая систему (29), определяем $\xi(t)$, $\eta(t)$ и, тем самым, область Ω;

3) решая «прямую» краевую задачу (26), (27), вычисляем $\mathscr{Z}(x, y)$ и функционал $F[u(\cdot)]$;

4) решая задачу $\Delta\Psi = 0$ с краевыми условиями (36), находим $\Psi(x, y)$;

5) решая краевые задачи (38), находим ψ_1, ψ_2, являющиеся производными функционала:

$$\psi_1(t) = \frac{\partial F[u(\cdot)]}{\partial u_1(t)}, \qquad \psi_2(t) = \frac{\partial F[u(\cdot)]}{\partial u_2(t)}.$$

Дифференцирование по коэффициенту диффузии. Рассмотрим задачу, в которой состояние объекта $\mathscr{Z}(x, y)$ определяется решением эллиптического уравнения «с управлением» $u(x, y)$:

$$\operatorname{div}[u\ \operatorname{grad}\mathscr{Z}] = 0. \tag{39}$$

Это уравнение рассматривается в заданной области Ω с границей Γ, на которой поставлено краевое условие

$$\mathscr{Z}|_\Gamma = f. \tag{40}$$

Коэффициент диффузии $u(x, y)$ может как-то меняться и является в данном случае тем ресурсом, распоряжаясь которым можно влиять на состояние объекта в нужном направлении.

Предположим, что качество состояния $\mathscr{Z}$ оценивается функционалом $F[u(\cdot)]$, для которого, ради определенности, примем формулу

$$F[u(\cdot)] \equiv \oint_\Gamma \Phi(\mathscr{Z}_n)\,ds. \tag{41}$$

Здесь Φ — заданная функция, $\mathscr{Z}_n$ — нормальная производная. Покажем, что при дифференцировании функционала (41) в некоторых

(достаточно распространенных) ситуациях слишком наивное и прямое применение описанной выше техники вычисления функциональной производной может привести к грубой ошибке. Нужно достаточно внимательно относиться к некоторым чисто математическим тонкостям. Ситуация (простейший ее вариант) такая: в области Ω имеется внутренняя подобласть ω с границей γ (рис. 51). Пусть невозмущенный коэффициент диффузии $u(x, y)$ имеет разрыв на γ, будучи гладкой функцией в ω и $\Omega\backslash\omega$.

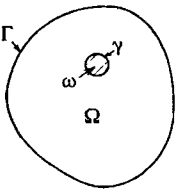


Рис. 51

Рассмотрим два варианта теории возмущений:

а) малое ($O(\varepsilon)$) возмущение u во всей области Ω (малое в метрике C); в этом случае $\operatorname{mes}\omega = O(1)$;

б) мала мера ω ($\operatorname{mes}\omega = \varepsilon$); в этом случае возмущение u есть $O(1)$ в ω и нуль (или, если угодно, $O(\varepsilon)$) в остальной части.

В обоих случаях соответствующее возмущение состояния $\delta\mathscr{Z} = O(\varepsilon)$ (такие теоремы для (39) доказаны) и вариация функционала вычисляется по формуле

$$\delta F = \int_\Gamma \Phi'[s]\,\delta\mathscr{Z}_n\,ds, \qquad \Phi'[s] = \Phi'(\mathscr{Z}_n(x(s), y(s))). \tag{42}$$

Здесь $\Phi'[s]$ — известная на Γ функция, вычисленная по известному невозмущенному состоянию.

Для преобразования (42) в линейный функционал от $\delta u(\cdot)$ (мы пока ограничимся более простой ситуацией малых возмущений на всей области Ω) выпишем наивное уравнение в вариациях:

$$\operatorname{div}[u\ \operatorname{grad}\delta\mathscr{Z}] + \operatorname{div}[\delta u\ \operatorname{grad}\mathscr{Z}] = 0, \tag{43}$$

и тождество Лагранжа:

$$\iint_\Omega \Psi \operatorname{div}(u\ \operatorname{grad}\delta\mathscr{Z})\,dx\,dy - \iint_\Omega \delta\mathscr{Z} \operatorname{div}(u\ \operatorname{grad}\Psi)\,dx\,dy =$$

$$= \oint_\Gamma \{\Psi u\,\delta\mathscr{Z}_n - \delta\mathscr{Z}\,u\Psi_n\}\,ds. \tag{44}$$

Определяя Ψ решением уравнения $\operatorname{div}(u\ \operatorname{grad}\Psi) = 0$ с краевыми условиями $u\Psi|_\Gamma = \Phi'[s]$, учитывая $\delta\mathscr{Z}|_\Gamma = 0$ и уравнение (43), из (44) получаем

$$\delta F[\delta u(\cdot)] = -\iint_\Omega \Psi(x, y) \operatorname{div}(\delta u\ \operatorname{grad}\mathscr{Z})\,dx\,dy. \tag{45}$$

Ошибка этой прямолинейной выкладки состоит в том, что при разрыве u на γ гладкой функцией является «поток» $u\mathscr{Z}_n$, где $\mathscr{Z} = (\operatorname{grad}\mathscr{Z}, n)$, n есть нормаль к γ. Функции u и $\mathscr{Z}_n$ на γ рвутся

и уравнение (43), будучи верным всюду вне γ, на этой линии теряет смысл. Видимо, в ситуации можно разобраться, используя теорию обобщенных функций, но мы предпочтем более прозрачный классический анализ. Итак, уравнением (43) и тождеством (44) можно пользоваться отдельно в ω и $\Omega\backslash\omega$. На разделяющей их кривой γ выполняются условия согласования

$$[\mathscr{Z}]_\gamma = 0, \quad [u\mathscr{Z}_n]_\gamma = 0,$$

т.е. разрывы решения и потока на γ равны нулю. Эти условия уже можно проварьировать обычным образом:

$$[\delta\mathscr{Z}]_\gamma = 0, \quad [u\,\delta\mathscr{Z}_n]_\gamma + [\delta u\,\mathscr{Z}_n]_\gamma = 0. \tag{46}$$

Используя уравнение (43) и тождество (44) отдельно в ω и $\Omega\backslash\omega$ (в этом случае в (44), очевидно, добавляется контурный интеграл по γ), складывая оба выражения типа (44), получаем правильное тождество:

$$\iint\limits_{\Omega\backslash\gamma} \{\Psi \operatorname{div}(u \operatorname{grad} \delta\mathscr{Z}) - \delta\mathscr{Z} \operatorname{div}(u \operatorname{grad} \Psi)\}\,dx\,dy =$$

$$= \oint\limits_\Gamma \{\Psi u\,\delta\mathscr{Z}_n - \delta\mathscr{Z}\,u\Psi_n\}\,ds + \oint\limits_\gamma \{[\Psi u\,\delta\mathscr{Z}_n]_\gamma + [\delta\mathscr{Z}\,u\Psi_n]_\gamma\}\,ds. \tag{47}$$

Интеграл по $\Omega\backslash\gamma$ означает просто сумму интегралов по ω и $\Omega\backslash\omega$. Теперь уже можно действовать стандартным способом.

Определим Ψ решением той же самой задачи со стандартным условием на γ:

$$[\Psi]_\gamma = 0, \quad [u\Psi_n]_\gamma = 0.$$

Используя непрерывность Ψ и $\delta\mathscr{Z}$ на γ, имеем

$$[\delta\mathscr{Z}\,u\Psi_n]_\gamma = 0,$$

а в силу (46)

$$[\Psi u\,\delta\mathscr{Z}_n]_\gamma = \Psi\cdot[u\,\delta\mathscr{Z}_n]_\gamma = -\Psi\cdot[\delta u\,\mathscr{Z}_n]_\gamma.$$

Теперь остается исправить формулу (45):

$$\delta F[\delta u(\cdot)] = -\iint\limits_{\Omega\backslash\gamma} \Psi \operatorname{div}(\delta u \operatorname{grad}\mathscr{Z})\,dx\,dy + \oint\limits_\gamma \Psi[\mathscr{Z}_n\,\delta u]_\gamma\,ds. \tag{48}$$

Рассмотрим второй случай — конечное возмущение управления на множестве малой меры. Возмущенное управление

$$\tilde{u}(x, y) = \begin{cases} u(x, y), & (x, y) \in \Omega\backslash\omega, \\ u(x, y) + v(x, y), & (x, y) \in \omega. \end{cases}$$

Начнем с уравнения в вариациях, выписав возмущенное и невозмущенное уравнения в ω и $\Omega\backslash\omega$:

$$\operatorname{div}(u \operatorname{grad} \tilde{\mathscr{X}}) = 0, \quad \operatorname{div}((u+v) \operatorname{grad} \tilde{\mathscr{X}}) = 0,$$

$$\operatorname{div}(u \operatorname{grad} \mathscr{X}) = 0, \quad \operatorname{div}((u)\operatorname{grad} \mathscr{X}) = 0.$$

Вычитая, получаем уравнения для $\delta\mathscr{X} \equiv \tilde{\mathscr{X}} - \mathscr{X}$:

$$\begin{gathered} \operatorname{div}(u \operatorname{grad} \delta\mathscr{X}) = 0 \quad \text{в } \Omega\backslash\omega, \\ \operatorname{div}(v \operatorname{grad} \mathscr{X}) + \operatorname{div}(u \operatorname{grad} \delta\mathscr{X}) + \operatorname{div}(v \operatorname{grad} \delta\mathscr{X}) = 0 \quad \text{в } \omega. \end{gathered} \tag{49}$$

В последнем уравнении пренебрежем третьим слагаемым, так как оно имеет величину порядка $O(\varepsilon)$ на множестве ω меры ε.

Что касается условий на γ, то их можно записать в форме

$$[\tilde{\mathscr{X}}]_\gamma = 0, \quad \{(u+v)\,\tilde{\mathscr{X}}_n\}_- = \{u\tilde{\mathscr{X}}_n\}_+,$$

$$[\mathscr{X}]_\gamma = 0, \quad \{u\mathscr{X}_n\}_- = \{u\mathscr{X}_n\}_+.$$

Здесь индексами «$-$» и «$+$» отмечены предельные значения величин на γ со стороны ω и $\Omega\backslash\omega$ соответственно. Вычитая, получаем соотношения для уравнения в вариациях:

$$[\delta\mathscr{X}]_\gamma = 0, \quad \{u\,\delta\mathscr{X}_n\}_- + \{v\,\delta\mathscr{X}_n\}_- + \{v\mathscr{X}_n\}_- = \{u\,\delta\mathscr{X}_n\}. \tag{50}$$

Используем тождество Лагранжа в форме (47). Дальнейшие преобразования носят стандартный характер, отличаясь от того, что было раньше, только другой формой уравнений в вариациях (49), (50).

В силу (49) и (50) левая часть (47) приобретает вид

$$-\iint_\omega \Psi \operatorname{div}(v \operatorname{grad} \mathscr{X})\,dx\,dy.$$

Интеграл по Γ в правой части (47) с учетом краевого условия для Ψ и $\delta\mathscr{X}|_\Gamma = 0$ превращается в δF. Второе слагаемое интеграла по γ обращается в нуль (в силу непрерывности $\delta\mathscr{X}$ на γ и условия $[u\Psi_n] = 0$). Первое слагаемое этого интеграла преобразуется с учетом непрерывности Ψ на γ и уравнений (50) следующим образом:

$$[\Psi u\,\delta\mathscr{X}_n]_\gamma = \Psi\cdot[u\,\delta\mathscr{X}_n]_\gamma = \Psi\{(v\,\delta\mathscr{X}_n)_- + (v\mathscr{X}_n)_-\} \approx \Psi\cdot(v\mathscr{X}_n)_-.$$

Пренебрегая первым членом как малой более высокого порядка по сравнению со вторым, получаем

$$\delta F[\delta u(\cdot)] = -\iint_\omega \Psi \operatorname{div}(v \operatorname{grad} \mathscr{X})\,dx\,dy + \oint_\gamma \Psi(v\mathscr{X}_n)_-\,ds.$$

§ 28. Задачи оптимального управления

Математическая теория оптимального управления начала бурно развиваться с начала шестидесятых годов. Истоки этой дисциплины лежат в классическом вариационном исчислении, но процесс математизации разнообразных прикладных наук привел к постановке задач, вариационных по своей сути, но не укладывавшихся в старые рамки. Особую роль в становлении теории оптимального управления сыграли ракетостроение и теория автоматического регулирования (как источники новых типов задач) и работы математиков под руководством Л. С. Понтрягина, в которых была выделена общая постановка задачи и получен основной теоретический результат — принцип максимума.

В настоящее время имеет смысл рассматривать задачи оптимального управления как задачи математического программирования в функциональном пространстве. И с этой точки зрения постановка задачи не отличается от рассмотренной в § 26. Требуется найти функцию u, обеспечивающую

$$\min_u F_0[u] \tag{1}$$

при условиях

$$F_i[u] \leqslant 0, \quad i = 1, 2, \ldots, m, \quad u \in U. \tag{2}$$

(Каждое условие может иметь и форму $F_i = 0$.) То обстоятельство, что u — это элемент функционального пространства, приводит к включению в форму (1), (2) различных задач, имеющих свои особенности. Начнем с конкретного примера.

Задача о подъеме ракеты. Движение ракеты описывается тремя функциями: $m(t)$ — масса, $h(t)$ — высота, $v(t)$ — скорость. Изменение этих величин определяется системой дифференциальных уравнений ($0 \leqslant t \leqslant T$)

$$\dot{m} = -u(t), \quad \dot{h} = v, \quad \dot{v} = -g + [Vu - Qe^{-\alpha h}(v)^2]/m, \tag{3}$$

дополненных данными Коши $m(0) = 1$, $h(0) = 0$, $v(0) = 0$. Величины g, V, Q, α, входящие в систему (3), — некоторые заданные постоянные. Функция $u(t)$ задает режим горения топлива. Ее нужно найти, с тем чтобы наилучшим образом выполнить стоящую перед управляемой системой (ракетой) задачу.

Ракета — управляемый объект, возможности управления которым ограничены выбором функции $u(t)$. Естественно, возникают ограничения на возможности выбора. Обычно их обозначают общей формой $u \in U$. В данном случае эта абстрактная форма принимает вид $0 \leqslant u(t) \leqslant u^+$, $\forall\, t$, где u^+ — техническое ограничение скорости расхода топлива. Цель управления — получить $\max h(T)$ при усло-

вии $m(T) = m_0$ (T, m_0 заданы). Итак, ставится задача достижения в момент T наибольшей высоты при заданном запасе топлива.

Общая задача оптимального управления. Приведем обобщенную формулировку вариационных задач подобного типа. Имеется управляемая система, состояние которой в момент времени t описывается фазовым вектором $x(t)$ (размерности p). Эволюция состояния системы во времени описывается системой обыкновенных дифференциальных уравнений

$$\dot{x} = f(t, x(t), u(t)), \quad 0 \leqslant t \leqslant T, \quad x(0) = x_0, \tag{5}$$

в правую часть которой входит искомая вектор-функция $u(t)$.

Управление системой состоит в выборе функции $u(t)$, ограниченной условиями

$$u(t) \in U, \quad \forall\, t \in [0, T], \tag{6}$$

где U — заданная область в r-мерном пространстве. В большинстве приложений она замкнутая, ограниченная и не очень сложной формы. Широко распространен простейший вид области U — прямоугольник:

$$u_i^- \leqslant u_i \leqslant u_i^+, \quad i = 1, 2, \ldots, r$$

(u_i^-, u_i^+ — заданные границы изменения u_i).

Пусть задано управление $u(\cdot)$. Так обозначается точка функционального пространства, т.е. функция, взятая «целиком». Под $u(t)$ мы теперь будем понимать значение этой функции в момент t, а под u — просто точку из r-мерного пространства. В теории управления имеют дело со всеми тремя объектами одновременно и их нужно четко различать, используя разные обозначения. Тогда задача Коши (5) интегрируется (мы считаем, что все условия существования и единственность для этой задачи выполнены). Таким образом, задание управления однозначно определяет состояние системы в любой момент времени.

Пусть определены функционалы от $F_i[u(\cdot)]$ ($i = 0, 1, \ldots, m$). Цель управления состоит в том, чтобы выполнить условия

$$F_i[u(\cdot)] \leqslant 0, \quad i = 1, 2, \ldots, m.$$

Этой цели нужно добиться самым экономным способом, т.е. нужно при этом получить

$$\min F_0[u(\cdot)].$$

Функционал $F_i[u(\cdot)]$ — это абстрактное обозначение алгоритма, позволяющего, коль скоро задано управление $u(\cdot)$, вычислить число F_i. Конкретные формулы вычисления F_i могут быть самым разнообразными.

Ограничимся пока несколькими наиболее часто встречающимися конструкциями функционалов:

$$F[u(\cdot)] \equiv \int_0^T \Phi[t, x(t), u(t)]\, dt, \tag{7}$$

$$F[u(\cdot)] \equiv \Phi[x(t^*)], \tag{8}$$

где t^* — заданная точка из $[0, T]$. Функции Φ, входящие в (7), (8), — заданные гладкие функции своих аргументов. Эти две конструкции представляют широкой класс функционалов, дифференцируемых в смысле Фреше (вопрос о вычислении производных Фреше был обсужден достаточно подробно в § 27).

Заметим, что любая гладкая функция от функционалов типа (7), (8) также приводит к дифференцируемому по Фреше функционалу, который может быть использован при формулировке вариационной задачи. Обозначение выражений в правой части (7), (8) через $F[u(\cdot)]$ оправдано тем, что именно $u(\cdot)$ является тем аргументом, задание которого позволяет (в принципе) вычислить значение F. Для того чтобы это фактически выполнить, следует при заданном $u(\cdot)$ проинтегрировать задачу Коши (5) (в общем случае только численно) и, получив $x(t)$, выполнить, например, интегрирование (тоже используя подходящий алгоритм приближенного вычисления квадратуры).

Перейдем к следующим двум важным конструкциям:

$$F[u(\cdot)] \equiv \max_t \Phi[t, x(t)], \tag{9}$$

$$F[u(\cdot)] \equiv \max_t \Phi[t, x(t), u(t)]. \tag{10}$$

Эти два функционала в общем случае не имеют производных Фреше, они (при сколь угодно гладких Φ) дифференцируемы лишь по направлениям в функциональном пространстве. Дифференцируемость функционала (9) зависит не от гладкости Φ, а от множества точек, на котором достигается максимум. Оно обозначается как

$$\arg\max_t \Phi[t, x(t)].$$

Если это множество состоит из одной точки, функционал (9), как правило, дифференцируем по Фреше; если их хотя бы две, производной Фреше не существует. Для (10) ситуация осложняется тем, что значения $u(t)$ на множестве меры нуль несущественны.

В терминах таких функционалов оформляются так называемые ограничения в фазовом пространстве. Пусть выбор управления $u(t)$ стеснен еще и требованием $x(t) \in G \subset R^p$, $\forall\, t \in [0, T]$, где G — заданная область в R^p. Наиболее распространенным способом

описания области G являются системы неравенств $\Phi^j(x) \leqslant 0$ $(j = 1, 2, \ldots, J)$. Каждое такое неравенство может быть оформлено как ограничение значения функционала типа (9). Функционалы типа (10) появляются таким же образом из требований $\{x(t), u(t)\} \in G \subset R^{p+r}$, $\forall\, t \in [0, T]$. Существуют причины, оправдывающие выделение (9) из более общей конструкции (10). Мы их сейчас обсудим.

«Измеримое» управление. Для того чтобы задача оптимального управления была поставлена достаточно четко, нужно указать то функциональное пространство, в котором разрешается искать $u(\cdot)$. При этом следует учесть чисто теоретические аспекты с одной стороны (это пространство должно быть достаточно широким, чтобы в нем существовало решение задачи), и интересы практики — с другой (найденное оптимальное управление должно быть достаточно простой функцией, чтобы его можно было использовать при управлении данной технической системой, например ракетой). Удобным оказался класс измеримых функций. При этом не возникает трудностей при интегрировании системы (5). Теория, как известно, требует от $f(t, x, u(t))$ выполнения условия Липшица по x и довольствуется произвольной, в сущности, зависимостью от t.

Конечно, класс измеримых (т.е. произвольных) функций слишком широк, техническая реализация такого управления кажется нереальной. К счастью, ситуация здесь оказалась достаточно благоприятной: при решении прикладных задач оптимальное управление, как привило, оказывается не очень сложно устроенной кусочно-гладкой функцией. Поэтому термин «измеримая функция» практически означает, что никаких требований гладкости функции $u(t)$ мы не ставим. В большинстве случаев достаточным был бы класс функций, имеющих конечное число разрывов. Между точками разрывов управление можно считать достаточно гладким. Правда, ни положения точек разрыва, ни их число заранее не известны.

Приближенное решение. Алгоритмы приближенного решения задач оптимального управления формально мало отличаются от алгоритмов решения задач математического программирования. Но здесь есть своя специфика, и некоторые алгоритмы практически оказываются почти нереализуемыми. Первый специфический момент — это вычисление производных (мы пока ограничимся задачами, в которых все функционалы дифференцируемы по Фреше).

Основу алгоритмов составляет формула первого члена ряда Тейлора. При малом изменении управления $u(\cdot)$ функцией $\delta u(\cdot)$ происходит малое изменение функционала:

$$F[u(\cdot) + \delta u(\cdot)] = F[u(\cdot)] + \frac{\partial F[u(\cdot)]}{\partial u(\cdot)}\, \delta u(\cdot).$$

Эта абстрактная формула должна быть конкретизирована:

$$\frac{\partial F[u(\cdot)]}{\partial u(\cdot)}\,\delta u(\cdot) = \int_0^T (w(t),\, \delta u(t))\, dt. \qquad (11)$$

Вектор-функция $w(t)$ (размерности r) называется производной Фреше функционала $F[u(\cdot)]$ в точке $u(\cdot)$. Функциональные производные в современных исследованиях используются достаточно часто (см. § 27).

Вычисление функции w требует интегрирования определенного в точке $u(\cdot)$ так называемого *сопряженного уравнения*

$$-\dot{\psi} = f_x^*[t, x(t), u(t)]\,\psi(t) + Y(t), \quad \psi(T) = 0, \qquad (12)$$

где $x(t)$ — траектория, соответствующая $u(\cdot)$. Функция $Y(t)$ для данного функционала легко вычисляется. Например, для F вида (7) имеем

$$Y(t) = \Phi_x[t, x(t), u(t)];$$

для F вида (8)

$$Y(t) = \Phi_x[x(t^*)]\,\delta(t - t^*).$$

Здесь $\delta(t - t^*)$ — функция Дирака с полюсом в точке t^*. После решения уравнения (12) функциональная производная вычисляется по формуле, полученной в § 27:

$$w(t) = f_u^*[t, x(t), u(t)]\,\psi(t) + \Phi_u[t, x(t), u(t)]. \qquad (13)$$

Реализация вычислительной схемы требует конечномерной аппроксимации всех объектов. Опишем возможный вариант.

Сетка и управление. Введем на $[0, T]$ сетку

$$0 = t_0 < t_1 < \ldots < t_N = T$$

и будем рассматривать кусочно-постоянные управления

$$u(t) = u_{n+1/2}, \quad t \in (t_n, t_{n+1}).$$

Обычно в расчетах $N \sim 10^2$. Вариация $\delta u(t)$ ищется в том же классе функций.

Траектория $x(t)$. Интегрируя (численно) задачу Коши (5), запомним значения x в узлах сетки t_n; обозначим их x_n $(n = 0, 1, \ldots, N)$. Так как каждое значение t_n является возможной точкой разрыва $u(t)$, следует быть осторожным, используя методы интегрирования высокого порядка точности. Эта точность реализуется лишь при достаточной гладкости f, в том числе и по t. Следую-

щее почти очевидное условие позволяет сохранить эту точность. Используя, например, метод типа Рунге—Кутты, необходимо брать шаг численного интегрирования таким, чтобы все точки сетки t_n входили в число узлов численного интегрирования.

Отметим, что сетка t_n не является сеткой численного интегрирования системы (5). Последняя обычно существенно гуще и в явном виде не присутствует. Сетка, однако, должна быть достаточной для представления траектории $x(t)$ и для ее восстановления (например, линейной интерполяцией значения x_n) с необходимой для дальнейшего точностью (не очень, в сущности, высокой).

Линеаризация задачи. Сопряженное уравнение (12) интегрируется многократно ($m+1$ раз; для каждого дифференцируемого функционала F_i требуется свое интегрирование). Уравнение (12) линейное с переменной матрицей $f_x^*[t, x(t), u(t)]$, определенной на варьируемой траектории $\{x(t), u(t)\}$. Реализуется это, например, аппроксимацией матрицы f_x кусочно-постоянной на той же сетке, т.е. вычисляются матрицы $f_x[n+1/2] \equiv f_x[t_{n+1/2}, x_{n+1/2}, u_{n+1/2}]$. Аналогично вычисляются матрицы $f_u[n+1/2]$ и векторы $Y[n+1/2]$, $\Phi_u[n+1/2]$. Теперь уже интегрирование системы (12) осуществляется без труда. Для дальнейшего нам нужны не $\psi(t)$, а интегралы

$$h^i_{n+1/2} = f^*_u[n+1/2] \int_{t_n}^{t_{n+1}} \psi^i(t)\, dt + \Phi^i_u[n+1/2], \quad i = 0, 1, \ldots, m.$$

Имея $h^i_{n+1/2}$, можно вычислить последствия возмущения управления величинами $\delta u_{n+1/2}$ ($n = 0, 1, \ldots, N-1$):

$$F[u(\cdot) + \delta u(\cdot)] \approx F[u(\cdot)] + \sum_{n=0}^{N-1} h_{n+1/2}\, \delta u_{n+1/2}. \qquad (14)$$

Здесь $h_{n+1/2}$ матрица $r \to m+1$. Формула (14), разумеется, приближенная. Ее погрешность связана как с пренебрежением величинами $O(\|\delta u\|^2)$, так и с погрешностями описанных выше аппроксимаций, из которых наибольшие последствия, видимо, имеет переход к кусочно-постоянным матрицам $f_x[n+1/2]$.

Располагая формулами (14), после вычисления всех $h_{n+1/2}$ можно осуществить выбор вариации $\{\delta u_{n+1/2}\}_{n=0}^{N-1}$ решением задачи линейного программирования. Процесс решения задачи поиска условного экстремума организуется так, как это было описано в § 26. Однако стоит отметить некоторые важные детали. Они связаны с тем, что формулы (14) получены аппроксимацией континуальных фор-

мул (11). Поэтому «горизонтальный» размер задачи линейного программирования (т.е. число неизвестных $\delta u_{n+1/2}$, равное Nr) обычно много больше ее «вертикального» размера $m+1$. Существенно еще и то, что эта задача сильно «почти вырождена»: компоненты $h_{n+1/2}$ для близких значений индексов очень близки друг к другу, как сеточное представление некоторых гладких функций. Эффективное решение таких задач линейного программирования требует специализированных алгоритмов. Попытки использования обычных стандартных программ линейного программирования часто оказываются в этих ситуациях неудачными.

Реализация методов квазиньютоновского типа, в которых появляются матрицы, аппроксимирующие гессиан функционала, здесь также встречается с трудностями. Это, прежде всего, — трудности больших размерностей: ведь такая матрица должна иметь размер $Nr \times Nr$. Да и перспективы построения хорошей аппроксимации гессиана процессом постепенного уточнения при высокой размерности пространства не очень ясны. Во всяком случае, этот путь еще не разведан вычислителями, и мы не знаем, с чем встретимся на этом пути. С этими оговорками, располагая формулами типа (14), можно реализовать любой из описанных в § 26 алгоритмов решения общей задачи математического программирования. Кстати, информация, содержащаяся в матрицах $h_{n+1/2}$, позволяет проверять приближенное выполнение необходимого условия оптимальности — принципа максимума. Здесь появляются объекты, полезные и в теории, и в практических вычислениях.

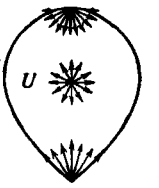


Рис. 52

Конус возможных вариаций K_u. Множество всех вариаций управления $\delta u(t)$, совместимых с условием $u(t) + \delta u(t) \in U$, обычно является выпуклым конусом K_u. Построение этого конуса (в функциональном пространстве) не очень сложно, если геометрия области U не слишком сложна. Нужно построить конусы $K(t)$ в каждой точке t отдельно, после чего конус K_u есть просто «топологическое произведение» конусов. Это означает, что $\delta u(\cdot) \in K_u$ эквивалентно $\delta u(t) \in K(t)$, $\forall\, t$. Построение конусов $K(t)$ при разных положениях $u(t)$ в U показано на рис. 52, которым мы и ограничимся, полагая, что читатель без труда обобщит эти простые соображения на общий случай.

Конус вариаций K_F. Рассмотрим точку функционального пространства $u(\cdot)$, которой соответствует точка $F[u(\cdot)]$ в $(m+1)$-мерном пространстве ($F = \{F_0, F_1, \ldots, F_m\}$). При возмущении управления $u(\cdot)$ малой функцией $\delta u(\cdot) \in K_u$ точка $F[u(\cdot)]$ переходит в

точку $F[u(\cdot)+\delta u(\cdot)]$, которую в первом приближении можно представить в виде

$$F[u(\cdot)+\delta u(\cdot)] = F[u(\cdot)] + \int_0^T W(t)\,\delta u(t)\,dt.$$

Матрица-функция $W(t)$ (физики называют ее *функцией влияния*, но это всего лишь функциональная производная) определяет линейное отображение K_u в K_F, и коль скоро K_u есть выпуклый конус, то и его образ есть выпуклый конус.

Конус запрещенных вариаций K_Z. Пусть вариационная задача поставлена в форме $F_i \leqslant 0$ $(i=1, 2, \ldots, m)$. Рассмотрим точку $u(\cdot)$, в которой $F_i[u(\cdot)]=0$. Нас интересует, можно ли за счет вариации $\delta u(\cdot) \in K_u$ сместить точку F таким образом, чтобы $\delta F_0 < 0$, $\delta F_i \leqslant 0$ $(i=1, 2, \ldots, m)$. Множество таких направлений δF образует простой выпуклый конус — «отрицательный квадрант» в $(m+1)$-мерном пространстве. Этот конус называют конусом запрещенных вариаций K_Z. Если точка $u(\cdot)$ — решение вариационной задачи, ни одно из направлений $\delta F \in K_F$ не должно попадать в K_Z. Если существует $\delta u(\cdot) \in K_u$, такая, что ей соответствует $\delta F \in K_Z$, точка $u(\cdot)$ не является оптимальной, ее можно «улучшить» (понизить значение F_0, не нарушая поставленных условий). Если такой $\delta u(\cdot)$ не существует, точка $u(\cdot)$ может быть оптимальной (здесь ситуация такая же, как и в обычной теории экстремума: если производная в какой-то точке равна нулю, эта точка может оказаться точкой экстремума).

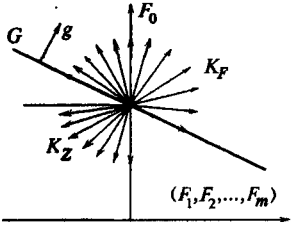


Рис. 53

Принцип максимума. Теперь перейдем к выводу основного уравнения теории оптимального уравнения — принципа максимума, являющегося необходимым условием оптимальности точки $u(\cdot)$.

Начнем с простого факта. Если $u(\cdot)$ — экстремум, конусы K_F и K_Z не должны пересекаться:

$$K_F \cap K_Z = \varnothing. \tag{15}$$

Расшифруем формулу (15). Если два выпуклых конуса не пересекаются, они могут быть разделены некоторой гиперплоскостью G (рис. 53). Пусть $g=\{1, g_1, \ldots, g_m\}$ — нормаль к G. Тогда (15) эквивалентно условиям

$$(g, \delta F) \leqslant 0 \quad \text{для всех } \delta F \in K_Z.$$

Так как векторы

$$\{0, -1, 0, \ldots, 0\}, \quad \{0, 0, -1, \ldots, 0\}, \quad \ldots, \quad \{0, 0, \ldots, 0, -1\}$$

лежат в K_Z, то мы получаем информацию о знаках g_i: $g_i \geqslant 0$.

Более сложная информация содержится в другом следствии из (15):

$$(g, \delta F) \geqslant 0 \quad \text{для всех} \quad \delta F \in K_F.$$

Все $\delta F \in K_F$ могут быть получены по формуле

$$\int_0^T W(t)\, \delta u(t)\, dt, \quad \delta u(\cdot) \in K_u.$$

Следовательно,

$$(g, \delta F) = \left(g, \int_0^T W(t)\, \delta u(t)\, dt\right) = \int_0^T (g, W\, \delta u)\, dt = \int_0^T (W^*(t)g, \delta u(t))\, dt \geqslant 0.$$

Это неравенство должно выполняться для всех $\delta u(\cdot) \in K_u$. Отсюда вытекает, что при всех t (точнее, при почти всех t) должно быть

$$(W^*(t)g, \delta u) \geqslant 0, \quad \forall\, \delta u \in K(t), \ \forall\, t.$$

Полученный результат можно преобразовать, вспомнив формулу (13). Строками матрицы $W(t)$ являются векторы

$$w_i(t) = f_u^*[t, x(t), u(t)]\, \psi^i(t) + \Phi_u^i[t, x(t), u(t)].$$

В результате

$$W^*(t)g = \sum_{i=0}^m g_i\, w_i(t) =$$

$$= f_u^*[t, x(t), u(t)] \sum_{i=0}^m g_i\, \psi^i(t) + \frac{\partial}{\partial u} \sum_{i=0}^m g_i\, \Phi^i[t, x(t), u(t)] =$$

$$= \frac{\partial}{\partial u} \left\{ (f[t, x(t), u(t)], \sum_{i=0}^m g_i\, \psi^i(t)) + \sum_{i=0}^m g_i\, \Phi^i[t, x(t), u(t)] \right\}.$$

Заметим, что каждое $\psi^i(t)$ — решение линейного дифференциального уравнения (12) со своей правой частью $Y^i(t)$. Таким образом, $\psi(t) = \sum_{i=0}^m g_i\, \psi^i(t)$ есть решение уравнения, содержащего m неопределенных параметров:

$$-\dot{\psi} = f_x^*[t, x(t), u(t)]\, \psi + \sum g_i\, Y^i(t), \quad \psi(T) = 0. \qquad (16)$$

Определим «функцию Гамильтона»:

$$H[t, x, u, g] \equiv (f(t, x, u), \psi) + \sum_{i=0}^{m} g_i \Phi^i(t, x, u).$$

Здесь $\psi(t)$ — решение уравнения (16). Теперь условие (15) примет вид

$$\frac{\partial}{\partial u} H[t, x(t), u(t), \psi(t), g]\, \delta u \geqslant 0, \quad \forall\, t, \ \forall\, \delta u \in K_t. \qquad (17)$$

Оно означает, что функция $H[t, x(t), u, \psi(t), g]$, рассматриваемая как функция u в области U, в точке $u(t)$ достигает локального минимума (максимума, если бы мы использовали для разделения конусов K_F и K_Z вектор $g' = -g$). Это и есть простейший вариант принципа максимума. Он утверждает, что если траектория $\{x(\cdot), u(\cdot)\}$ оптимальна ($x(t)$ — решение задачи Коши $\dot{x} = f$, соответствующее управлению $u(\cdot)$), то существует вектор g, такой, что выполняется условие (17) экстремума H по u в области U.

Конечно, в приведенном выше выводе мы опустили некоторые элементы математической аккуратности, но основные содержательные соображения сохранены. Теория, основанная на использовании конечных вариаций управления на множествах малой меры (см. § 27), позволяет утверждать, что функция Гамильтона достигает не локального, а точного минимума (максимума) по $u \in U$ именно в точке $u(t)$. Неопределенные коэффициенты g_i, входящие в H, играют роль множителей Лагранжа.

Некоторые обобщения задачи. Выше был рассмотрен относительно простой вариант задачи оптимального управления. В дальнейшем мы рассмотрим и задачи, существенно от нее отличающиеся. Здесь же мы ограничимся простым, но полезным обобщением. Расширим управление, включив в него набор параметров $p = \{p_1, p_2, \ldots, p_k\}$, которые должны быть определены из тех же соображений, что и $u(\cdot)$. Будем рассматривать систему уравнений вида

$$\dot{x} = f(t, x(t), u(t), p), \quad 0 \leqslant t \leqslant T, \quad x(0) = X_0(p).$$

В функции Ф, входящие в описание стандартных конструкций функционалов, наряду с указанными ранее аргументами, может входить и вектор p.

Будем считать, что, задав обобщенное управление $\{u(\cdot), p\}$, можно определить траекторию $x(t)$ и вычислить значения всех функционалов, которые теперь следует обозначать как $F[u(\cdot), p]$. Формула для вариации функционала при малом возмущении управ-

ления $\{u(\cdot), p\} \to \{u(\cdot) + \delta u(\cdot), p + \delta p\}$ очевидным образом обобщается:

$$F[u(\cdot) + \delta u(\cdot), p + \delta p] \approx F[u(\cdot), p] + \int_0^T (w, \delta u)\, dt + (a, \delta p).$$

Вычисление производной $a = \partial F[u(\cdot), p]/\partial p$ не требует новых сложных вычислений и осуществляется одновременно с вычислением $w(t)$ (производной по $u(\cdot)$); см. § 27.

Обратим внимание на то, что теперь задачу можно рассматривать на стандартном интервале времени $0 \leqslant t \leqslant 1$. В тех случаях, когда время процесса управления T не фиксировано и является вместе с $u(\cdot)$ «ресурсом оптимизации», можно перейти к системе $\dot{x} = Tf$, включив T в качестве одной из компонент в вектор параметров. Если $F_0[u(\cdot), p] \equiv T$, задача называется задачей оптимального быстродействия, так как целью является выполнение системой поставленной задачи за минимальное время.

Задачи с фазовыми ограничениями. Особенно сложным является приближенное решение задач оптимального управления, если среди требований к управлению поставлено условие невыхода траектории $x(t)$ из некоторой заданной области. Рассмотрим простейший пример.

Пусть поставлено условие $G[x(t)] \leqslant 0$, $\forall\, t$, где G — скалярная гладкая функция. Как уже было сказано, это условие можно оформить в терминах функционала: $F[u(\cdot)] \equiv \max\limits_t G[x(t)]$. Если, как это часто случается,

$$\mu = \arg\max_t G[x(t)]$$

есть не точка, а несколько точек или даже целый интервал, функционал оказывается недифференцируемым по Фреше. Однако он оказывается дифференцируемым по направлениям в функциональном пространстве, и можно написать почти очевидную (пока предварительную) формулу:

$$\delta F = \max_{\tau \in \mu} G_x[x(\tau)]\, \delta x(\tau).$$

Здесь $G_x[x(\tau)]\, \delta x(\tau)$ есть линейный функционал от вариации управления $\delta u(\cdot)$, и мы знаем, как он вычисляется:

$$G_x[x(\tau)]\, \delta x(\tau) = \int_0^T (w(t, \tau), \delta u(t))\, dt.$$

Причины, приведшие к появлению в w еще одного аргумента τ, понятны. Теперь мы имеем

$$\delta F[\delta u(\cdot)] = \max_{\tau \in \mu} \int_0^T (w(t, \tau), \delta u(t))\, dt.$$

Формально можно обобщить задачу поиска улучшающей вариации управления $\delta u(\cdot)$, включив в нее еще и условия

$$G[x(\tau)] + \int_0^T (w(t, \tau), \delta u(t))\, dt \leqslant 0, \quad \forall\, \tau \in \mu.$$

Но это не так просто, ведь этих условий очень много (континуум, если μ — отрезок, например).

Вычисление $w(t, \tau)$ для каждого $\tau \in \mu$ требует своего отдельного интегрирования сопряженной системы. Однако здесь есть некоторые возможности облегчения ситуации: $x(t)$ есть гладкая функция ($\dot{x} = f$, f ограничена при всех u; следовательно, $x(t)$ — непрерывная функция, с ограниченной кусочно-непрерывной производной). Такая функция не может очень сильно изгибаться. Поэтому если потребовать выполнения условия $G[x(t)] \leqslant 0$ не во всех $t \in [0, T]$, а только в узлах некоторой сетки τ^j ($j = 1, 2, \ldots, J$), то в остальных точках t оно будет, видимо, нарушено не очень сильно. Это соображение можно развивать и дальше, с тем чтобы число J было не слишком большим.

Пусть при каком-то $u(\cdot)$ найдена траектория $x(t)$ и вычислена функция $G[x(t)]$. Выделим на $[0, T]$ множество μ условием

$$G[x(\tau)] \geqslant -\varepsilon, \quad \varepsilon > 0, \quad \tau \in \mu.$$

На этом множестве разместим небольшое число точек τ^j по следующему, например, правилу. Предположим, для простоты, что μ есть просто отрезок. Разобьем его на заданное число J равных частей, и на каждой части найдем точку τ^j с наибольшим на этой части значением $G[x(\tau)]$. Конечно, мы не можем сказать заранее, сколько таких «контрольных» точек надо брать. Это зависит от структуры траектории, от меры множества μ, от оценок $|\dot{x}| = |f|$ на данной траектории и прочих трудно контролируемых факторов. Поэтому эти соображения дополняются алгоритмами, регулирующими изменение числа J в зависимости от хода процесса поиска экстремума.

Отметим, что мы не случайно не рассматриваем в таком же стиле близкую по форме конструкцию ограничения $G[x(t), u(t)] \leqslant 0$. Дело, конечно, в свойствах гладкости функции $G[x(t), u(t)]$. Так как $u(t)$ — произвольная («измеримая») функция, то и $G[x(t), u(t)]$, как функция t, — тоже произвольная функция. И даже контролируя условие $G \leqslant 0$ на всюду плотном множестве меры нуль, мы на самом де-

ле не обеспечиваем выполнения условия $G \leqslant 0$ при всех t. Однако реально задача решается в классе кусочно-постоянных $u(t)$, к которым термин «измеримость», кажется, никакого отношения не имеет. Это так: предложенная выше конструкция применима и в данном случае с конечным числом J. Все дело в том, каким будет это число J. Если u явно входит в G, то, скорее всего, число контрольных точек τ^j будет сравнимо с числом интервалов постоянства u, т.е. с числом узлов сетки N. Это делает задачу определения δu (и подготовки необходимой информации) слишком громоздкой и дорогой. Как показал опыт, часто условие $G(x) \leqslant 0$ с хорошей точностью можно обеспечить при небольшом числе J ($3 \div 5$, например).

Что касается условий $G(x, u) \leqslant 0$, то они могут быть учтены с помощью несложного искусственного приема. Явно входящие в G компоненты оформляются как дополнительные фазовые переменные, а управлением становятся их производные, т.е. делается замена переменных $\dot{u} = v$, $u(0) = p$. Теперь v — компонента нового управления, p — неизвестный параметр, тоже входящий в обобщенное управление. Этот прием имеет отрицательные последствия: сравнительно простые ограничения u (типа $0 \leqslant u \leqslant 1$) становятся ограничениями в фазовом пространстве. Кроме того, если $u(t)$ — разрывная функция, процессом малых вариаций $v(\cdot) \to v(\cdot) + \delta v(\cdot)$ приходится получать в $v(t)$ аналог δ-функции. Тем не менее этот прием с успехом применяется на практике. В дальнейшем мы познакомим читателя и с более прогрессивной идеей учета таких условий в методах приближенного решения.

Пример решения задачи. Выше были изложены основные идеи методов приближенного решения задач оптимального управления. Их реализация связана с необходимостью конкретизировать большое число деталей, кажущихся мелкими на первый взгляд, но оказывающих довольно большое влияние на эффективность алгоритма. Мы не можем здесь уделить внимание этим деталям, с ними читатель, если ему понадобится, может познакомиться по специальной литературе. Для иллюстрации приведем пример решения одной прикладной задачи — об оптимальном развороте самолета.

Система уравнений движения для такой задачи имеет вид ($0 \leqslant t \leqslant 1$)

$$\dot{x}^1 = x^4 \cos x^5 \cos x^6, \quad \dot{x}^2 = x^4 \sin x^5, \quad \dot{x}^3 = -x^4 \cos x^5 \sin x^6,$$

$$\dot{x}^4 = g\,[(x^3 p_0 \cos\alpha - C_\alpha q S)/x^7 - \sin x^5],$$

$$\dot{x}^5 = g\,(x^9 \cos x^{11} - \cos x^5)/x^4, \quad \dot{x}^6 = -g\,(x^9 \cos x^{11})/(x^4 \cos x^5),$$

$$\dot{x}^7 = -C_s, \quad \dot{x}^8 = u_1, \quad \dot{x}^9 = u_2, \quad \dot{x}^{10} = u_3, \quad \dot{x}^{11} = u_4.$$

Здесь $\dot{x} = p_1^{-1}\, dx/dt$, где p_1 — параметр, входящий в обобщенное управление $\{u(\cdot), p\}$. Он имеет смысл времени выполнения маневра, которое не задано заранее, а является ресурсом оптимизации наряду

с $u(\cdot)$. В монографии Ю. Г. Евтушенко (см. список литературы) подробно разъяснено содержание задачи, указан конкретный вид функций и значения параметров, входящих в систему уравнений

Таблица 18

i	1	2	3	4	5
$\tilde{\Phi}^i$	$\dot{x}^2-3000$	x^5	$x^6+\pi$	x^9-1	$x^{11}-1$

(p_0, C_x, S, q, α, g). Вектор фазовых координат $x=\{x^1, x^2, \ldots, x^{11}\}$, при этом четыре его компоненты x^8, x^9, x^{10}, x^{11} являются управлением в исходной постановке задачи. Они превращены в компоненты фазового вектора. Управлением стали их производные, а компонентами управляющего вектора p стали начальные данные.

Таблица 19

i	6	7	8	9	10
Φ	$x^5-\pi/2$	$-x^5-\pi/2$	$0.05-x^8$	x^8-1	x^9-8
i	11	12		13	14
Φ	$4.6\alpha(x)-1$	$x^7\cdot x^9-1.5\cdot 10^4$		$-x^{10}$	$x^{10}-1$

Опишем в общих чертах процесс решения этой задачи методом линеаризации (он подробно описан в § 26). На каждой итерации алгоритма вычисляются производные всех функционалов и решается задача линейного программирования. Некоторые детали метода описаны выше (например, дискретизация задачи, аппроксимация функционалов, недифференцируемых в смысле Фреше).

Вариационная задача ставится в терминах функционалов

$$F_0[u(\cdot), p] \equiv p_1 \quad \text{(задача быстродействия)}.$$

На правом конце траектории ставятся пять условий, определяющих требование попадания правого конца траектории $x(1)$ на некоторую гиперплоскость. Они имеют стандартную форму:

$$F_i[u(\cdot), p] \equiv \tilde{\Phi}^i[x(1)] = 0, \quad i=1, 2, \ldots, 5.$$

Функции $\tilde{\Phi}^i$ приведены в табл. 18. Эти пять функционалов дифференцируемы по Фреше. Таблица 19 содержит функции $\Phi^i(x)$ для функционалов типа

$$F_i[u(\cdot), p] \equiv \max_t \Phi^i[x(t)] \leqslant 0, \quad i=6, 7, \ldots, 14.$$

Как это часто бывает, многие из приведенных условий ставятся на всякий случай. Заранее не ясно, нужно ли оптимальной траектории нарушать поставленные ограничения. Если в том или ином условии $F_i[u(\cdot), p] \leqslant 0$ реализуется строгое неравенство $F_i < 0$, то условие называют «пассивным». Оно может быть выброшено из постановки задачи без изменения существа дела. К сожалению, до решения задачи мы не можем выделить такие условия. Иногда условие в процессе поиска экстремума бывает то активным, то пассивным.

Процесс решения задачи показан в табл. 20, в которой приведены номер шага (итерации) ν и значения некоторых функционалов F_i. Из недифференцируемых функционалов показаны значения только четырех, оказавшихся активными (т.е. существенными). На нескольких первых итерациях активным было условие

Таблица 20

ν	F_0	F_1	F_2	F_3	F_4	F_5	F_8	F_{11}	F_{12}	F_{14}	K
0	21	3204	.57	1.03	2.62	−.047	−.9	−.46	−25500	−1.0	0,0,0,0
5	20.45	377	−.45	1.00	.65	.13	−.31	−.45	−45200	.49	0,0,0,2
10	20.86	441	−.39	0.65	.13	.014	.016	−.33	35300	.22	1,0,0,2
15	21.02	23	−.011	−.008	.004	.0009	.001	−.19	79	.065	1,0,1,3
20	18.69	56	−.005	.02	.005	.007	.008	−.07	−81	.11	1,0,1,3
25	17.74	23	−.011	.008	.007	.006	.011	−.08	204	.03	3,1,1,3
30	16.95	28	−.010	.015	.007	.008	.008	−.05	1530	.07	3,1,1,4
35	15.88	31	−.012	.005	.007	.007	.026	−.003	1040	.09	3,1,2,4
40	14.75	32	−.012	.05	.006	.006	.019	.002	2160	.09	3,1,3,4
45	14.64	23	−.12	.004	.003	.007	.005	−.010	690	.05	3,1,3,4
50	14.46	23	−.013	.017	.006	−.0006	.023	−.016	1230	.105	3,1,3,4
55	14.59	7	−.006	.014	−.002	$4\cdot10^{-5}$	$4\cdot10^{-5}$	−.03	197	.05	6,0,2,5
60	14.49	10	−.012	.07	.003	.0009	.0015	−.04	18	.09	7,0,5,6
61	14.48	12	−.012	.12	.002	.0006	.0001	−.04	116	.03	7,0,6,8
62	14.53	8	−.006	.036	−.001	.0001	$3\cdot10^{-4}$	−.04	55	.009	7,0,6,8
64	14.57	0	−.0002	$2\cdot10^{-4}$	$6\cdot10^{-4}$	$6\cdot10^{-5}$	.0018	−.04	113	.007	6,0,6,8

$F_9 \leqslant 0$, потом оно прочно перешло в разряд пассивных. Условие $F_{13} \leqslant 0$ стало активным только после 49-й итерации, в конце процесса выполнено условие $F_{13} \leqslant 0.006$. Последний столбец табл. 20 содержит четыре целых числа — это величины J_i, показывающие, сколько точек J_i используется для аппроксимации условия $F_i \leqslant 0$ (для $i = 8, 11, 12, 14$).

Естественно возникает вопрос: можно ли при $F_{12} \approx 100$, например, считать, что условие $F_{12} \leqslant 0$ выполняется с достаточной точностью? Это, разумеется, зависит от того, какие значения для F_{12} считаются «средними», характерными. Обычно в содержательной постановке задачи условия формулируются в виде $F_i \leqslant C_i$, где заданные значения C_i определяют, как правило, характерные значения для F_i: малое значение для F_i — это значение, существенно меньшее значения C_i. Ради стандартизации постановки задачи все функционалы заменяются на $F_i - C_i$. О значениях C_i можно судить по значениям F_i в начальном приближении ($\nu = 0$). Видно, что для F_2, F_3, F_4, F_8 характерными являются значения порядка единицы, а для F_{12} — порядка 10^5. С учетом этих значений и следует оценивать точность выполнения условий $F_i \leqslant 0$ для разных i.

О совместных ограничениях u и x. Как было сказано выше, для повышения гладкости функций, входящих в выражения для недифференцируемых функционалов, используется искусственный прием. Первоначальные управления объявляются фазовыми координатами, а новыми управлениями становятся их производные. Это делается для того, чтобы от функционала (10) (с явным вхождением управления в Ф) перейти к функционалу типа (9). Есть и другой способ, предложенный В. Г. Болтянским. Пусть в задаче поставлено условие $\Phi[x(t), u(t)] \leqslant 0$, $\forall\, t$. При построении вариации управления мы должны использовать линеаризованное условие

$$\Phi[x(t), u(t)] + \Phi_x\, \delta x(t) + \Phi_u\, \delta u(t) \leqslant 0, \quad \forall\, t \tag{18}$$

(разумеется, на самом деле это условие нужно использовать не при всех значениях t, а лишь при тех, где $\Phi[x(t), u(t)] \geqslant -\varepsilon$ ($\varepsilon \approx \Phi_u\, \delta u \approx \Phi_x\, \delta x$).

Условие (18) очень неудобно с вычислительной точки зрения, так как $\delta x(t)$ зависит не от $\delta u(t)$, а от $\delta u(\cdot)$ (от всех значений $\delta u(t')$ для $t' < t$). Возникает идея как-то избавиться от $\delta x(t)$ и трактовать условие (18) независимо для каждого t (так же, как трактуются условия типа $u(t) \in U$, наиболее простые в данной задаче). Это достигается следующим образом. Будем искать вариацию управления в виде

$$\delta u(t) = \delta \tilde{u}(t) + C(t)\, \delta x(t), \tag{19}$$

где $\delta x(t)$ — вариация фазы, вызванная полной вариацией управления $\delta u(\cdot)$, $C(t)$ — некоторая подходящим образом построенная матрица.

Подставляя (19) в (18), получаем

$$\Phi[t] + \Phi_x[t]\ \delta x(t) + \Phi_u[t]\ \delta\tilde{u}(t) + \Phi_u[t]\ C(t)\ \delta x(t) \leqslant 0.$$

Здесь $\Phi[t] \equiv \Phi[x(t), u(t)]$ и т.п. Очевидно, поставленная цель будет достигнута, если в качестве $C(t)$ взять решение матричного уравнения

$$\Phi_x[t] + \Phi_u[t]\ C(t) = 0. \tag{20}$$

В этом случае условия (18) превратятся в «локальные» (независимые при разных t) условия для $\delta\tilde{u}(t)$:

$$\Phi[t] + \Phi_u[t]\ \delta\tilde{u}(t) \leqslant 0, \quad \forall\, t.$$

Разумеется, надо внести соответствующие изменения во все элементы техники вычисления функциональных производных (дифференцирование по $\delta\tilde{u}$ с учетом связи (19)). В частности, уравнение в вариациях преобразуется так:

$$\frac{d}{dt}\,\delta x = f_x[t]\ \delta x + f_u[t]\ \delta u = \{f_x[t] + f_u[t]\ C(t)\}\ \delta x + f_u[t]\ \delta\tilde{u}(t).$$

Остальное преобразуется таким же образом.

Что касается уравнения (20), то искомая матрица C есть матрица типа $\dim u \to \dim x$, т.е. она содержит $\dim u \cdot \dim x$ неизвестных элементов. Само же уравнение (20) есть (так как Φ — скаляр) $\dim x$ скалярных уравнений. Стало быть, это есть переопределенная система. Нас устраивает любое ее решение. Несколько сложнее случай, когда условие $\Phi \leqslant 0$ векторное. Тогда стандартной является ситуация, при которой в каждый момент времени t из всех условий $\Phi[x(t), u(t)] \leqslant 0$ не более $\dim u$ являются активными (они реализуются в виде равенства, остальные — в виде строгого неравенства, их можно игнорировать) и уравнений в (19) становится не больше, чем неизвестных.

§ 29. Вариационные задачи механики с недифференцируемыми функционалами

Очень многие задачи механики имеют вариационную формулировку. Это связано с такими фундаментальными в естествознании идеями, как принцип наименьшего действия, особое значение состояний с минимальной энергией и т.п. Таким образом получаются задачи, с математической точки зрения имеющие вариационный характер. Определено некоторое пространство U и на его элементах u — функционал $F(u)$. Требуется определить элемент u^*, решая задачу

$$\min_{u \in U} F(u). \tag{1}$$

Иногда то же самое записывают в виде $u^* = \arg\min F(u)$.

В классической механике такие задачи возникали весьма часто. При этом формулировка $u \in U$ включала в себя указание о числе производных у допустимых функций, для которых определено вычисление F, и о краевых условиях, которым они должны удовлетворять. В наше время все чаще возникают задачи типа (1), в которых в формулировку $u \in U$ включается, например, условие положительности функции u. В абстрактном представлении это оформляется как требование $u \in K$, где K является не линейным пространством, а, например, выпуклым замкнутым конусом. Разница между линейным пространством и конусом состоит в том, что если два элемента u_1 и u_2 принадлежат пространству, то ему же принадлежит и любая их линейная комбинация $\alpha u_1 + \beta u_2$ (α, β — скаляры). Конусу такая комбинация принадлежит только при неотрицательных α, β. В частности, множество положительных функций образует выпуклый конус («положительный квадрант» в бесконечномерном пространстве).

Кроме того, в классической механике обычно функционал $F(u)$ был дифференцируемым в смысле Фреше, т.е. при малом возмущении элемента u имеет место формула

$$F(u + \delta u) = F(u) + F_u(u)\,\delta u + O(\|\delta u\|^2).$$

Линейный функционал $F_u(u)$ есть производная Фреше от F в точке u (мы не останавливаемся на вопросе о том, в какой норме мало возмущение δu). В этом случае задачу можно решить не только в вариационной форме (1), но и используя необходимое условие экстремума $F_u(u) = 0$. Это уравнение обычно называют *уравнением Эйлера* для вариационной задачи (1). Например, известная задача Дирихле допускает две формулировки:

$$1)\ \min_{u \in U} \iint_G (u_x^2 + u_y^2)\,dx\,dy; \qquad 2)\ \Delta u = 0,\ u \in U. \tag{2}$$

Здесь U — пространство функций, удовлетворяющих следующим условиям:

а) u принимает заданные значения на ∂G;

б) u непрерывна и имеет первые производные, ограниченные в норме L_2; вторая формулировка задачи предполагает существование вторых производных.

В современной науке все чаще возникают задачи (1), в которых функционал $F(u)$ не имеет производной Фреше. Он дифференцируем в более слабом смысле Гато, т.е. лишь по направлениям в функциональном пространстве. Другими словами, для любого возмущения εv, такого, что $u + \varepsilon v \in U$, $\forall\, \varepsilon \geq 0$, имеем

$$F(u + \varepsilon v) = F(u) + \varepsilon F'(u, v) + o(\varepsilon).$$

При этом U считают конусом, а $F'(u, v)$ называют производной F в точке u по направлению v. Введем конус V, включающий такие элементы, для которых $u + \varepsilon v \in U$ при достаточно малом $\varepsilon > 0$. Этот конус V может быть своим для каждой точки u, т.е. его следует обозначать $V(u)$, и необходимое условие экстремума принимает форму так называемого *вариационного неравенства*: функция u является решением задачи, если

$$F(u + v) \geqslant F(u) \quad \text{для всех} \quad v \in V(u),$$

или

$$F'(u, v) \geqslant 0, \quad \forall\, v \in V(u).$$

Приближенные методы решения задач, сформулированных как вариационные с недифференцируемым по Фреше фукционалом или в терминах вариационного неравенства, в настоящее время делают первые шаги. В этой области открывается широкое поле для создания эффективных вычислительных методов. Однако это достаточно трудная область, она требует использования неклассических методов линейной алгебры, в частности алгоритмов линейного программирования (алгоритмов решения задач типа (26.1), линейных, но содержащих условия-неравенства).

Заметим, наконец, что часто функционалы $F(u)$ «почти всюду» имеют обычную производную Фреше: они дифференцируемы только в смысле Гато в очень редких точках u. К сожалению, именно таковыми являются искомые решения (и близкие к ним точки). Простейший пример $F(u) = |u|$ поясняет это замечание. Обратимся к некоторым характерным конкретным задачам.

Задача Бингама. В заданной двумерной области G ищется функция $u^*(x, y)$, минимизирующая функционал

$$F[u(\cdot, \cdot)] \equiv \iint\limits_G \left\{\tfrac{1}{2}\,(u_x^2 + u_y^2) + \sqrt{u_x^2 + u_y^2} - \alpha u\right\} dx\, dy, \quad u|_{\partial G} = 0. \tag{3}$$

Физически $u(x, y)$ есть продольная скорость движения в трубе сечением G так называемого вязкопластичного вещества, т.е. вещества, подчиняющегося обычному закону Ньютона (ускорение пропорционально силе), только если сила превосходит некоторый порог. Поэтому в этом случае говорят о стационарном движении неньютоновской среды. Такую среду образуют, например, пульпа, колбасный фарш, некоторые виды ракетного топлива и т.п. Параметр α связан с перепадом давления, G — область не очень сложной формы (круг, прямоугольник).

Что же можно сказать о дифференцируемости функционала (3)? Он недифференцируем в смысле Фреше в том случае, когда функция $u(x, y)$ тождественно равна постоянной в некоторой области $g \in G$, имеющей ненулевую плоскую меру. Это, к сожалению, типичная си-

туация. Некоторые части среды образуют как бы твердое тело: $u(x, y) = \text{const}$, $u_x = u_y = 0$ при $(x, y) \in g$. Если область g находится внутри G, ее называют «ядром» течения. Если она примыкает к границе G, ее называют «зоной застоя», так как в этой зоне $u = 0$ (зоны застоя часто образуются вблизи угловых точек, если G — прямоугольник). Наличие таких областей — характерное явление, если перепад давления α не очень велик. При достаточно большом α таких областей нет, при достаточно малом α все сечение G образует зону застоя и «жидкость» не движется. Функционал (3) называют функционалом Бингама, а описываемую им среду — средой Бингама.

Задача Ильюшина. Эта задача связана с течением той же самой вязкопластичной среды, только речь идет не о продольном ее движении, а о «вращении» в сечении G. Оно описывается функцией тока $u(x, y)$ (через которую скорость движения в плоскости сечения выражается известными формулами $\{-u_y, u_x\}$).

Рассматриваются функции u, удовлетворяющие краевому условию «прилипания» к границе: $u = 0$, $\partial u/\partial n = 0$ на ∂G. Вводится квадратичная форма $I(x, y) = (u_{xx} - u_{yy})^2 + 4u_{xy}^2$ и ставится задача минимизации функционала:

$$F[u(\cdot, \cdot)] \equiv \iint\limits_G \left\{ \tfrac{1}{2} I(x, y) + \sqrt{I(x, y)} + Rxu_x \right\} dx\, dy. \qquad (4)$$

(Этот функционал называют функционалом Ильюшина.) Здесь характерными являются течения, в которых образуются «ядра течения» — области $g \in G$, в которых $I(x, y) \equiv 0$. Очевидно, в этом случае $u_{xy} = 0$, $u_{xx} = u_{yy}$. Это, как нетрудно проверить, означает, что в g среда вращается, как твердое тело вокруг некоторого центра.

Дифференцируемость функционала (4) по направлениям подробно проверять не будем, ограничившись указанием на основной фактор такого анализа на примере функционала Бингама. Пусть точка $u(\cdot, \cdot)$ содержит ядро, т.е. в некоторой точке $(x, y) \in G$ значения $u_x = u_y = 0$. Пусть u возмущается на εv, причем $v(x, y)$ — дифференцируемая функция. Тогда подынтегральное выражение в (3) обычным образом разлагается в ряд по ε:

$$0.5\{(u_x^2 + u_y^2) + 2\varepsilon(u_x v_x + u_y v_y) + O(\varepsilon^2)\} +$$
$$+ \{(u_x^2 + u_y^2) + 2\varepsilon(u_x v_x + u_y v_y) + \varepsilon^2(v_x^2 + v_y^2)\}^{1/2} - \alpha u - \alpha\varepsilon v =$$
$$= O(\varepsilon^2) + |\varepsilon|\{v_x^2 + v_y^2\}^{1/2} - \alpha u - \alpha\varepsilon v.$$

Таким образом, главный член приращения интегранта в этой точке есть $|\varepsilon|\sqrt{v_x^2 + v_y^2} - \alpha\varepsilon v$. Однако дальнейшие, стандартные в вариационном исчислении выкладки (интегрирование по частям), имеющие целью избавиться от производных v и получить выраже-

ние в терминах только v, здесь принципиально невыполнимы. Это обстоятельство имеет весьма важное следствие, существенно осложняющее создание алгоритмов приближенного решения: необходимое условие оптимальности в этих задачах имеет принципиально нелокальный характер.

Имеется в виду следующее. Если взять классическую задачу с функционалом Дирихле (2), то функцию $u(x, y)$, подозреваемую в том, что она и есть точка минимума, можно проверить в каждой точке (x, y) отдельно: надо вычислить в этой точке Δu. Если эта величина всюду равна нулю, все в порядке, если хотя бы в одной точке $\Delta u \neq 0$, это не решение. В такой точке $u(\cdot, \cdot)$ значение функционала можно понизить. Это прямо вытекает из того, что вариация функционала Дирихле после интегрирования по частям преобразуется к виду

$$F[u(\cdot,\cdot) + \delta u(\cdot,\cdot)] = F[u(\cdot,\cdot)] - \iint \Delta u\, \delta u\, dx\, dy + O(\|\delta u\|^2).$$

Итак, если в какой-то точке (x, y) (а по непрерывности — и в ее окрестности) $\Delta u > 0$, то можно взять в качестве $\delta u(x, y)$ гладкую финитную функцию, положительную там, где $\Delta u > 0$, и равную нулю в остальной части G. Для такого возмущения $\delta F < 0$. Если функционал нельзя улучшить финитными возмущениями точки $u(\cdot, \cdot)$, то она является экстремумом. Это и имеется в виду, когда говорится, что необходимое условие в классической вариационной задаче имеет локальный характер.

Иное дело в неклассической задаче, хотя бы в задаче Бингама. Здесь необходимо испытывать функцию $u(x, y)$, подозреваемую в качестве экстремальной, специальными нелокальными возмущениями. Пусть, например, исследуется функция $u(x, y)$, содержащая ядро течения в форме круга радиусом ρ, в котором $u(x, y) = \text{const}$. В остальной части G (где $u_x^2 + u_y^2 \neq 0$) выполнено стандартное локальное условие экстремума, которое имеет форму дифференциального уравнения (уравнения Эйлера):

$$\Delta u + \left(u_x/\sqrt{u_x^2 + u_y^2}\right)_x + \left(u_y/\sqrt{u_x^2 + u_y^2}\right)_y + \alpha = 0.$$

Проварьируем функцию u в области ядра следующим образом. В окружности, концентрической с ядром радиуса $\rho - \delta$, положим возмущение равным ε. Пусть вне ядра возмущение будет нулевым, а в «поясе» шириной δ оно линейно по радиусу переходит от нуля до ε. Обозначим возмущенную функцию $\tilde{u}(x, y)$, область ядра g. Вычислим приращение функционала, опуская некоторые заведомо несущественные малые величины:

$$F[\tilde{u}(\cdot,\cdot)] - F[u(\cdot,\cdot)] = -\iint\limits_g \alpha\varepsilon\, dx\, dy + \iint\limits_q \{\sqrt{\tilde{u}_x^2 + \tilde{u}_y^2} + 0.5(\tilde{u}_x^2 + \tilde{u}_y^2)\}\, dx\, dy.$$

Здесь q — вышеупомянутый пояс. Легко понять, что в этом поясе $(\tilde{u}_x^2 + \tilde{u}_y^2)^{1/2} = |\varepsilon/\delta|$, так как величина $u_x^2 + u_y^2$ инвариантна при поворотах системы координат и при ее оценке удобно перейти в локальную систему координат, оси которой совпадают с касательной и нормалью к контуру g.

Для вариации функционала имеем

$$\delta F = -\alpha\varepsilon\pi\rho^2 + 2\pi\rho\delta(|\varepsilon/\delta| + 0.5|\varepsilon/\delta|^2).$$

Пусть возмущение $\varepsilon \ll \delta$. Тогда главная часть δF при $\varepsilon > 0$ ($\varepsilon < 0$ можно не рассматривать, так как в этом случае заведомо $\delta F > 0$) есть $\delta F = -\varepsilon(\alpha\pi\rho^2 - 2\pi\rho)$. Очевидно, если $\alpha\pi\rho^2 > 2\pi\rho$, точка u может быть улучшена. Итак, необходимым для оптимальности u является условие $\alpha\pi\rho^2 \leqslant 2\pi\rho$.

Предоставим читателю обобщить эту конструкцию на область ядра произвольной формы. Легко понять, что $\pi\rho^2$ надо заменить на S (площадь g), а $2\pi\rho$ — на длину контура L. Тем самым мы получаем общее необходимое условие на ядро течения в терминах его площади и длины контура границы: $\alpha S \leqslant L$. Более тонкий анализ показывает, что условие $\alpha = L/S$ является достаточным для того, чтобы точка $u(\cdot, \cdot)$ была решением задачи Бингама (это установлено П. П. Мосоловым и В. П. Мясниковым в 1965 г.).

Проверим, что в определенных, заведомо неоптимальных ситуациях попытки «улучшить» некоторое проверяемое «решение» с помощью финитных возмущений окажутся безуспешными, т.е. функция, явно не являющаяся точкой минимума функционала, оказывается «минимумом» относительно класса финитных возмущений. Пусть $u(x, y) \equiv 0$. Все финитные возмущения исследовать довольно сложно, но простое их множество поддается оценке и хорошо проясняет суть дела.

Итак, возьмем в качестве возмущений функции $\delta u(x, y)$ в форме конуса высотой ε и радиусом ρ, равные нулю вне круга радиусом ρ. Тогда всюду в круге радиусом ρ, очевидно, $(\delta u_x^2 + \delta u_y^2)^{1/2} = |\varepsilon/\delta|$ и приращение функционала на таком возмущении легко подсчитать:

$$\delta F = \pi\rho^2\{0.5(\varepsilon/\rho)^2 + |\varepsilon/\rho|\} - (1/3)\alpha\pi\rho^2\varepsilon =$$
$$= \pi\rho^2\{0.5(\varepsilon/\rho)^2 + |\varepsilon/\rho| - (1/3)\alpha\varepsilon\}.$$

Таким образом, как бы ни был велик параметр α, при достаточно малом ρ будет $\delta F > 0$, т.е. такая вариация только ухудшает функцию. В то же время легко строится нелокальная вариация δu того типа, который был описан выше, и для нее $\delta F < 0$.

Все это имеет прямое отношение к одной из распространенных схем приближенного решения вариационных задач.

Метод покоординатного спуска. Нелокальность условий экстремума («уравнения Эйлера») в неклассической вариационной задаче имеет серьезные последствия с точки зрения вычислителя. Рассмотрим универсальный метод построения минимизирующей последовательности. На этой основе естественно пытаться строить приближенные методы.

Здесь есть чисто технический вопрос — конечномерная аппроксимация вариационной задачи. Введем сетку с шагом h (по x и y) и узлами (k, m) и сеточную функцию $u_{k,m}$. Заменим функционал функцией конечного числа переменных. Обозначим

$$I_{k+1/2,\, m+1/2} = (u_{k+1,\, m} - u_{k,\, m})^2/h^2 + (u_{k,\, m+1} - u_{k,\, m})^2/h^2$$

и аппроксимируем функционал $F(u)$ так:

$$F(u) = h^2 \sum_{k,\, m} \{0.5 I_{k+1/2,\, m+1/2} + \sqrt{I_{k+1/2,\, m+1/2}} - \alpha u_{k+1/2,\, m+1/2}\} \qquad (5)$$

($u_{k+1/2,\, m+1/2}$ — среднее из четырех значений в узлах сетки).

Метод покоординатного спуска минимизации функционала (5) состоит в том, что поочередно меняются значения $u_{k,m}$ в одном узле с целью понизить значение F. Очевидно, при вариации значения $u_{k,m}$ в сумме (5) изменятся только три слагаемых (соответствующих узлам $(k-1, m)$, $(k, m-1)$ и (k, m)). Этот способ решения вариационных задач известен уже около ста лет под названием «релаксационный». Он является одним из наиболее медленно сходящихся, но в классических вариационных задачах в принципе приводит к успеху. Если в каждом узле (k, m) попытка понизить значение F оказывается безуспешной, минимум функционала (точнее, его конечномерной аппроксимации (5)) найден. Основу этого метода, очевидно, составляет множество финитных сеточных пробных функций.

Метод легко обобщается и на неклассические задачи. В частности, некоторый его вариант под именем «метод локальных вариаций» был одним из первых, предложенных для приближенного решения задачи Бингама. Однако из сказанного выше следует, что такой метод принципиально неадекватен природе задачи: здесь нужны более сложные и тонкие алгоритмы. Действительно, применение метода локальных вариаций, популярного благодаря его алгоритмической простоте, привело к публикации «решений» (задач Бингама, Ильюшина и некоторых других), опровергнутых последующими расчетами.

Задача качения. Следующий пример неклассической вариационной задачи связан с задачей качения шарика по плоскости с учетом сухого трения. Под действием силы, направленной ортогонально плоскости качения, материалы шарика и основания дефор-

мируются и образуется двумерная область контакта G (процесс считается стационарным). Область G задана (ее форма определяется решением другой вариационной задачи, которую мы не обсуждаем). В этой области определены искомые двумерные вектор-функции $s(x, y) = \{s_1, s_2\}$ и $\tau(x, y) = \{\tau_1, \tau_2\}$. Вектор $s(x, y)$ имеет смысл относительного проскальзывания — смещения контактирующих точек шарика относительно их положения в отсутствие движения. (Заметим, кстати, что задача рассматривается в подвижной системе координат, в которой вся картина стационарна.) Вектор $\tau(x, y)$ имеет смысл силы трения.

Перейдем к математической формулировке задачи. Итак, следует минимизировать функционал

$$F[\tau(\cdot)] \equiv \iint\limits_G \{f(x, y)\|s(x, y)\| - (\tau(x, y), s(x, y))\}\, dx\, dy \qquad (6)$$

при ограничении

$$\|\tau(x, y)\| \leqslant f(x, y), \quad \forall\, (x, y) \in G, \qquad (7)$$

и связи между s и τ в виде

$$s(x, y) = v(x, y) - \iint\limits_G B(x - x', y - y')\, \tau(x', y')\, dx'\, dy'. \qquad (8)$$

Здесь все, кроме τ и s, задано, f имеет смысл нормального давления, v — скорость движения точки (x, y) в подвижной системе координат, $B(x, y)$ — некоторая $(2 \to 2)$ матрица-функция. Функционал обозначен $F[\tau]$, так как s явно выражается через τ. Эта функция, таким образом, является единственным независимым аргументом.

Задача хорошо исследована. О ее решении известно следующее.

1. Решение существует и единственно.
2. Минимальное значение F есть нуль.
3. Область G разбивается на две части: область сцепления G_0, в которой $s(x, y) = 0$, $\|\tau(x, y)\| < f(x, y)$, и область проскальзывания $G \backslash G_0$, в которой $\|s(x, y)\| \neq 0$, $\tau(x, y) = f(x, y) s(x, y)/\|s(x, y)\|$. Это, в сущности, хорошо известные законы сухого трения. Если сила $\|\tau\|$ меньше некоторого порога, пропорционального силе нормального давления, скольжения нет ($s = 0$). Если сила достигает этого порога, начинается скольжение.

Мы сталкиваемся с недифференцируемостью функционала (6) в тех точках $\tau(\cdot)$, в которых уже имеется непустая область сцепления G_0.

Приближенное решение задачи качения. Опишем в общих чертах метод приближенного решения задачи. Первый элемент метода — конечномерная аппроксимация. В области G вводятся квад-

ратная сетка с шагом h и узлами (k, m) и сеточные функции $\tau_{k,m}$, $s_{k,m}$ и т.п. Функционал аппроксимируется суммой

$$F(\tau) = h^2 \sum_k \sum_m \{f_{k,m}\|s_{k,m}\| - (\tau_{k,m}, s_{k,m})\},$$

а связь между s и τ записывается в виде

$$s_{k,m} = v_{k,m} - h^2 \sum_i \sum_j B_{i-k,j-m}\tau_{i,j}.$$

Проблемы вычисления элементов матрицы B обсуждаются в § 30. Основная трудность состоит в построении алгоритма минимизации недифференцируемого функционала. Дело в том, что, когда образуется область сцепления, зависимость F от τ становится, вообще говоря, аналогичной зависимости типа $(a, \tau) + \|\tau\|$. График этой функции есть «наклоненный конус». Множество направлений ее убывания (если оно не пусто) есть конус, тем более узкий, чем ближе ситуация к экстремуму (т.е. чем ближе $\|a\|$ к единице). Найти хотя бы одно направление в таком конусе в пространстве очень высокой размерности (тем более высокой, чем меньше шаг сетки h) — сложная вычислительная задача. Она осложняется еще и тем, что интересы эффективности процесса минимизации требуют не просто какого-то направления убывания F, но, по возможности, направления наиболее быстрого убывания. Конечно, наличие ограничений $\|\tau\| \leqslant f$ вносит дополнительные осложнения и сокращает возможности выбора.

Метод численного решения, реализующийся в виде процесса построения минимизирующей последовательности, основан на анализе формулы для первой вариации функционала. Пусть текущая, уже найденная точка τ подвергается малому возмущению, т.е. переходит в $\tau + \delta\tau$. Как изменится при этом значение F? Для упрощения изложение будем вести в терминах функций и интегралов. Перевод полученных формул в сеточный вид достигается заменой аргументов x, y на индексы k, m, интегралов — на суммы. Кроме того, используем полезное свойство преобразования B в (8):

$$\iint_G (B\tau, \tau)\, dx\, dy = 0, \quad \forall\, \tau.$$

Это позволит упростить выражение для функционала, заменив в (6) интеграл от $(s, \tau) = (\tau, v - B\tau)$ на интеграл от (τ, v).

Первоначальное выражение (6) для F полезно в том отношении, что позволяет контролировать качество приближенного решения. В точном решении, как это следует из указанных выше сведений о нем, подынтегральное выражение

$$f(x, y)\|s(x, y)\| - (\tau(x, y), s(x, y)) \equiv 0, \quad \forall\, (x, y).$$

Будем проводить вычисления в некоторой точке $\tau(\cdot)$, для которой определена область сцепления G_ε: $\|s(x, y)\| \leqslant \varepsilon(x, y)$ (роль ε разъ-

ясняется ниже). В соответствии с этим функционал можно разбить на две части — на дифференцируемую (по Фреше) $F_д$ и недифференцируемую $F_н$:

$$F[\tau(\cdot)] = F_д + F_н = \iint\limits_{G\setminus G_\varepsilon} \{\ldots\}\, dx\, dy + \iint\limits_{G_\varepsilon} \{\ldots\}\, dx\, dy.$$

Пусть τ возмущено малой функцией $\delta\tau$. Тогда первая вариация (дифференциал) $F_д$ есть линейный функционал от $\delta\tau$, т.е. с точностью до $O(\|\delta\tau\|^2)$ имеем

$$F_д[\tau(\cdot) + \delta\tau(\cdot)] = F_д[\tau(\cdot)] + \iint\limits_G (D(x, y), \delta\tau(x, y))\, dx\, dy.$$

Здесь $D(x, y)$ — производная Фреше от $F_д$, которая вычисляется по формуле

$$D(x, y) = \iint\limits_{G\setminus G_\varepsilon} \frac{f(x', y')}{\|s(x', y')\|} B(x - x', y - y')\, s(x', y')\, dx'\, dy' - v(x, y). \tag{9}$$

Не станем выводить этой формулы, укажем лишь основные операции ее вывода.

Подставляем в (6) вместо τ и s соответственно $\tau + \delta\tau$ и $s + \delta s$. Пользуясь тем, что $\|s\| > \varepsilon$ в $G\setminus G_\varepsilon$, разлагаем подынтегральное выражение в ряд Тейлора с учетом первого порядка малых величин. Заменяя δs на $-\iint B\,\delta\tau$ (в соответствии с (8)), получаем выражение для $\delta F_д[\delta\tau(\cdot)]$ в виде четырехкратного интеграла. В этом интеграле меняем очередность интегрирования (переобозначая x', y' через x, y и наоборот). «Внутренний» интеграл и есть $D(x, y)$. Что касается приращения недифференцируемой части, то тут никаких особых упрощений нет. Итак, приращение функционала при вариации τ может быть записано в форме

$$F[\tau(\cdot) + \delta\tau(\cdot)] - F[\tau(\cdot)] = \iint\limits_G (D(x, y), \delta\tau(x, y))\, dx\, dy +$$
$$+ \iint\limits_{G_\varepsilon} f\left\| s(x, y) - \iint\limits_G B(x - x', y - y')\, \delta\tau(x', y')\, dx'\, dy' \right\| dx\, dy -$$
$$- \iint\limits_{G_\varepsilon} f(x, y)\, \|s(x, y)\|\, dx\, dy + O(\|\delta\tau\|^2). \tag{10}$$

Пренебрегая $O(\|\delta\tau\|^2)$, определим процедуру вычисления вариации $\delta\tau$, обеспечивающей убывание F. Обычно решение таких сложных задач начинают с относительно простых алгоритмов. Совсем не обязательно в конкретных расчетах должны появиться все неприятности, которые в принципе возможны. Начнем поиск минимума с начального приближения $\tau(x, y) \equiv 0$. Пока $G_0 = \varnothing$ и функционал

дифференцируем, работает метод спуска по градиенту. На каждом шаге τ пересчитывается по формуле $\tilde{\tau} := P(\tau - SD)$, где P — оператор проецирования, работающий локально, в каждой точке (x, y) независимо от других, S — шаг процесса.

Оператор P введен с целью учета поточечного ограничения $\|\tau(x, y)\| \leqslant f(x, y)$ и реализуется просто. Определяем $\tau^* = \tau - SD$ и, если $\|\tau^*(x, y)\| \geqslant f$, полагаем $\tilde{\tau}(x, y) = \tau^*(x, y) f(x, y)/\|\tau^*(x, y)\|$. Вычисляем фактическую вариацию $\delta\tau(x, y) = \tilde{\tau}(x, y) - \tau(x, y)$ и предсказанную вариацию функционала

$$\delta F = \iint_G (D, \delta\tau)\, dx\, dy.$$

Выбор шага процесса S играет большую роль, если нас интересует не только факт сходимости, но и скорость процесса минимизации. Затем вычисляем новое значение функционала $F[\tilde{\tau}(\cdot)]$ и фактическое приращение $\Delta F = F[\tilde{\tau}(\cdot)] - F[\tau(\cdot)]$. Если $\Delta F > 0$, итерация считается неудачной, шаг S уменьшается вдвое и с той же производной повторяется вариация τ. Если $\Delta F < 0$, итерация выполняется, т.е. τ заменяется на $\tilde{\tau}$, а шаг S корректируется в зависимости от точности линейного приближения — величины $\eta = 2|\delta F - \Delta F|/|\delta F + \Delta F|$. Если эта величина мала, шаг S увеличивается, если велика, — уменьшается.

Заметим, что трудоемкость итерации велика: она определяется необходимостью вычисления функционала F и его производной D. Обе операции стоят $O(h^{-4})$ операций (в каждой точке двумерной сетки нужно вычислить двумерный интеграл). Поэтому здесь не применяется надежный способ выбора шага, связанный с решением задачи $\min F[P(\tau - SD)]$ по S. Даже не очень точное ее решение требует нескольких вычислений F. Расчеты по этой простой схеме показали, что сначала функционал достаточно быстро убывает, затем образуется небольшая область сцепления G_ε, которая растет. По мере ее роста все большую роль в ΔF начинает играть недифференцируемая составляющая, шаг S катастрофически уменьшается и алгоритм «застревает» в заведомо неоптимальный точке $\tau(\cdot)$.

Метод регуляризации. Наиболее простой и дешевый способ продвинуться дальше, почти не усложняя алгоритма, состоит в регуляризации задачи, т.е. в данном случае в аппроксимации недифференцируемой функции $\|\cdot\|$ дифференцируемой. Практически это означает замену $\sqrt{s_1^2 + s_2^2}$ на $\sqrt{s_1^2 + s_2^2 + \varepsilon}$ $(\varepsilon > 0)$. Величину ε можно затем, по мере достижения минимума (для данного ε), постепенно уменьшать, используя найденное ранее решение как начальное приближение при новом значении ε.

Методы регуляризации задач весьма популярны, но, к сожалению, не очень эффективны. Дело в том, что нужно не только ап-

проксимировать недифференцируемую функцию дифференцируемой (эта цель легко достигается в данном случае), но и получить функцию, с хорошей точностью аппроксимируемую своей касательной на таких расстояниях от точки линеаризации, которые следует использовать в эффективном алгоритме построения минимизирующей последовательности. При замене $|s|$ на $\sqrt{s^2+\varepsilon}$ возникает конфликт между точностью аппроксимации и гладкостью регуляризованной функции.

В описываемом алгоритме разумный компромисс между точностью и гладкостью аппроксимации достигался следующим образом. Наряду с основными счетными массивами $\tau_{k,m}$, $s_{k,m}$, $D_{k,m}$ использовался массив $\varepsilon_{k,m}$, и вместо $\|s_{k,m}\|$ в формулы входила величина $(\|s_{k,m}\|^2+\varepsilon_{k,m})^{1/2}$. После осуществления вариации ($\tau\to\tau+\delta\tau$, $s\to s+\delta s$) величины $\varepsilon_{k,m}$ пересчитывались. При этом предполагалось, что на следующей итерации значение $\delta s_{k,m}$ будет примерно таким же. Для хорошей линеаризации нужно, чтобы значение $\|s_{k,m}\|^2+\varepsilon_{k,m}$ было раз в пять больше ожидаемой вариации $\delta s_{k,m}$. Таким образом, величины $\varepsilon_{k,m}$ автоматически убывали в процессе расчета при уменьшении шага S.

Таблица 21 дает представление о том, как протекал процесс минимизации. В ней представлены: номер шага ν (звездочкой отмечены неудачные итерации с $\Delta F>0$), F, δF, ΔF и шаг спуска S. Видно, что неудачные шаги сравнительно редки. Обратим внимание на то, что величина δF (в принципе пропорциональная $S\|D\|^2$) убывает намного быстрее, чем S. Это связано с убыванием производной D, т.е. с приближением к минимуму. Возникает вопрос: насколько рациональна вырабатывающаяся в ходе расчета величина S? Свидетельством в пользу этого алгоритма служит хорошо видный из таблицы факт: обычно резкое уменьшение шага S сопровождается ростом фактического убывания функционала ΔF.

Экспериментальные попытки волевым образом увеличить S не приводили к успеху: возникала ситуация $\Delta F>0$, шаг последовательно дробился несколько раз подряд и приходил к старому значению. Вышеприведенный расчет носил методический характер, поэтому число итераций относительно велико (впрочем, размерность конечномерного пространства, в котором решалась дискретная задача на минимум, здесь была около 1500). Видно, что после 15-й итерации расчет практически «стоит на месте» и продолжение его бесполезно. Каковы же полученные при этом результаты?

Анализ показал, что почти всюду в области G (это был эллипс) значение функции $I=|f\|s\|-(\tau,s)|$ было очень мало. Более точно это означает следующее. В начальном приближении среднее значение $I_{\text{ср}}\approx 2$ максимальное значение $I_{\text{макс}}\approx 5$. В конце расчета (при $F\approx 0.18$ в большей части области для $I\approx 0.0002\div 0.002$)

$I_{ср} \approx 0.027$, $I_{макс} = 2.2$. Имеется небольшая подобласть в G, в которой значение функции $I(x, y)$ достаточно велико, причем вместо требуемой в точном решении коллинеарности τ и s наблюдалась почти антиколлинеарность этих векторов. В этот момент G_0 занима-

Таблица 21

ν	F	$-\delta F$	$-\Delta F$	S
0	13.83	10.8	8.8	22
1	4.83	6.0	1.36	27
2	3.46	4.2	1.69	22
3*	1.772	3.6	−0.46	22
3	1.772	1.8	0.78	11
4	0.988	1.1	0.13	11
5*	0.864	1.3	−0.002	8.6
5	0.864	0.67	0.39	4.3
6	0.477	0.35	0.064	4.3
7*	0.413	0.43	−0.018	3.5
7	0.413	0.22	0.11	1.7
8	0.307	0.11	0.039	1.7
9*	0.268	0.15	−0.025	1.7
9	0.268	0.075	0.026	0.86
10	0.241	0.058	0.004	0.86
11	0.237	0.072	0.008	0.69
12	0.229	0.052	0.016	0.55
13*	0.212	0.045	−0.0014	0.55
13	0.212	0.023	0.011	0.28
14	0.202	0.013	0.005	0.28
15*	0.197	0.017	−0.001	0.28
15	0.197	0.009	0.004	0.14
16	0.193	0.0056	0.002	0.14
17	0.191	0.0064	0.001	0.14
21	0.186	0.0029	0.0006	0.07
25	0.184	0.0017	0.0005	0.036
29	0.182	0.0009	0.0002	0.029
33	0.181	0.0005	0.0002	0.023

ла значительную часть G и никакие ухищрения в рамках описанной выше методики не приводили к улучшению решения. (Впрочем, как показали дальнейшие расчеты, для грубых выводов часто бывает достаточно и полученного таким образом решения.)

Для того чтобы получить более аккуратные и достоверные результаты, пришлось существенно усложнить метод.

Метод линейного программирования. Этот сложный алгоритм мы опишем в самых общих чертах. Он основан на некоторой аппроксимации приращения недифференцируемого слагаемого F_n более простым, но тоже недифференцируемым. Используется возможность аппроксимировать круговой конус, как поверхность в трехмерном пространстве $\{\xi, \eta, \sqrt{\xi^2+\eta^2}\}$, шестигранным конусом $\{\xi, \eta, \Phi(\xi, \eta)\}$, где $\Phi(\xi, \eta)$ определяется решением следующей задачи линейного программирования:

$$\Phi(\xi, \eta) = 0.54 \min_{\alpha', \ldots, \gamma''} (\alpha' + \alpha'' + \beta' + \beta'' + \gamma' + \gamma'')$$

при условиях

$$(\alpha' - \alpha'')/\sqrt{3} + (2/\sqrt{3})(\beta' - \beta'') = \xi, \quad -(\beta' - \beta'') + (\gamma' - \gamma'') = \eta,$$

$$(\alpha' - \alpha'') + (\beta' - \beta'') + (\gamma' - \gamma'') = 0, \quad \alpha', \alpha'', \beta', \beta'', \gamma', \gamma'' \geqslant 0.$$

Не будем доказывать этого. Читатель, желающий понять, в чем тут дело, пусть начнет с вопроса о том, почему $|\xi| = \min(\alpha' + \alpha'')$ при $\alpha' - \alpha'' = \xi$, $\alpha' \geqslant 0$, $\alpha'' \geqslant 0$.

Используем введенную аппроксимацию конуса. Введем, кроме $\delta\tau(x, y)$, новые вспомогательные переменные $\alpha'(x, y), \ldots, \gamma''(x, y)$, $(x, y) \in G_\varepsilon$. В терминах этих переменных задача выбора направления спуска для недифференцируемого функционала ставится следующим образом. Требуется найти $\delta\tau(x, y), \alpha'(x, y), \ldots, \gamma''(x, y)$, обеспечивающие

$$\min \Big\{ \iint\limits_G (D(x, y), \delta\tau(x, y))\, dx\, dy + 0.54 \iint\limits_{G_\varepsilon} f(x, y)\, \{\alpha'(x, y) + \alpha''(x, y) + \ldots + \gamma''(x, y)\}\, dx\, dy \Big\}$$

при условиях

$$s_1(x', y') - \iint\limits_G (b_1(x - x', y - y'), \delta\tau(x, y))\, dx\, dy - \{(\alpha' - \alpha'')/\sqrt{3} + (2/\sqrt{3})(\beta' - \beta'')\}_{x', y'} = 0,$$

$$s_2(x', y') - \iint\limits_G (b_2(x - x', y - y'), \delta\tau(x, y))\, dx\, dy + \{(\beta' - \beta'') - (\gamma' - \gamma'')\}_{x', y'} = 0,$$

$$\{(\alpha' - \alpha'') + (\beta' - \beta'') + (\gamma' - \gamma'')\}_{x', y'} = 0, \quad \forall\, (x', y') \in G_\varepsilon.$$

Здесь b_1, b_2 — первая и вторая строки матрицы B, $\{\alpha\}_{x, y} = \alpha(x, y)$.

Выше были опущены некоторые несложные технические детали, обеспечивающие условия $\|\tau(x, y) + \delta\tau(x, y)\| \leqslant f(x, y)$. Все это превращается в конечномерную задачу линейного программирования, матрица которой схематически изображена на рис. 54. Поясним ее: N есть число узлов в области G (в расчете $N \approx 800$), N_0 — число узлов в G_ε (на разных этапах расчета $N_0 \approx 300 \div 600$), $b_{i,j}$ — элементы матрицы B, рассматриваемые здесь как четырехиндексные матрицы (первая пара индексов из G, вторая — из G_ε). В правой части матрицы стоит разреженный блок $6N_0 \times 3N_0$, состоящий из коротких (по шесть элементов) строк, расположение которых показано на рисунке.

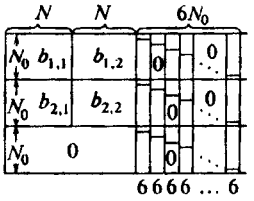


Рис. 54

Ясно, что такая задача сама по себе практически непосильна для БЭСМ-6, а ее предстоит решать много раз: на каждой итерации. Поэтому был использован прием, известный под названием «агрегирование неизвестных». Он состоит в том, что ячейки сетки $h \times h$ объединялись в блоки $H \times H$, где $H \approx (3 \div 4)h$, например, и все переменные считались постоянными в каждом блоке. Иными словами, процесс проводился не на основной h-сетке, а на более грубой H-сетке. Это приводит к существенному сокращению размеров задачи (до 360 неизвестных и 120 условий).

Вышеприведенная громоздкая методика использовалась попеременно с методом регуляризации. Когда последний переставал работать, делалась одна итерация с применением линейного программирования, которая «сдвигала» точку $\tau(\cdot)$ со стационарной для метода регуляризации ситуации. Снова применялся спуск по градиенту с регуляризацией, и т.д. В конечном счете было получено приближенное решение с такими характеристиками: $F = 0.026$, $I_{\text{макс}} \approx 0.15$, причем только в семи узлах (из 800) значение I попадает в интервал $[0.1, 0.15]$. Что касается затрат машинного времени (на БЭСМ-6), то первый этап стоил 25 мин, а в целом расчет занял 2 ч 15 мин. Это был один из первых расчетов, в процессе которого отрабатывалась «стратегия» проведения вычислений. В дальнейшем время подобных расчетов несколько сократилось.

То, что было описано выше, представляет вторую, видимо, попытку решения задач подобного рода (она осуществлялась автором), если, конечно, не считать типичных в механике приближенных решений, основанных на тех или иных упрощающих предположениях, априорных гипотезах о решении и т.п. Подобные решения оказываются удачными в той мере, в какой оправдываются такие гипотезы. Первый опыт решения задачи о качении в ма-

тематически замкнутой форме был предпринят И. И. Калькером. Он использовал несколько иную форму минимизируемого функционала и, соответственно, другие алгоритмы. Результаты его расчетов были проконтролированы решением по вышеизложенной методике.

На рис. 55 показаны данные сравнения решений одной задачи двумя разными методами. Показаны графики функций $\tau_2(x, 0)$, $s_2(x, 0)$. Линия $y = 0$ является линией симметрии, в силу которой $\tau_1(x, 0) = s_1(x, 0) = 0$. Решение Калькера изображено штриховой линией, решение по вышеописанной методике — сплошной. Хотя в целом, качественно, картины близки, можно отметить явные, видные даже на глаз дефекты «штрихового» решения. Там, где $s_2 \neq 0$, должно быть $\tau_2 = fs/\|s\|$. Сплошная линия точно следует этому правилу: знак τ_2 меняется точно в том месте, где s_2 проходит через нуль. Штриховая линия явно нарушает это правило точного решения. Нарушено условие задачи $\|\tau\| \leqslant f$ (примерно на 11 %). Нарушен и второй «закон»: там, где $\|\tau\| < f$, обязательно $s = 0$.

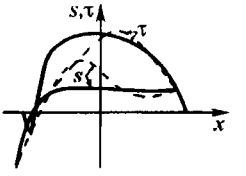


Рис. 55

Вычислительные методы математического (в частности, линейного) программирования возникли, развивались и применялись в первую очередь в связи с внедрением математических методов в экономические теории. Поэтому многие считают их некой «экзотикой», интересующей весьма узкий круг специалистов. Большинство физиков, механиков и представителей других естественных наук этих методов не знают. Между тем и в этих областях в последние годы все чаще возникают задачи, настоятельно требующие применения именно таких нетрадиционных алгоритмов.

Метод двойственности. Опишем основные моменты другого подхода к решению неклассических вариационных задач, в котором используется так называемая двойственная формулировка задачи. Вычислительные алгоритмы такого рода разрабатываются группой математиков, возглавляемых Ж. Лионсом. Одним из первых приложений этих методов было решение задачи Бингама. Применительно к ней мы и будем вести изложение. Итак, требуется минимизировать функционал $F[u(\cdot)]$ вида (3).

Основной момент нижеследующего состоит в замене недифференцируемой функции $\|\xi\|$ (ξ — некоторый вектор) решением специальной задачи на экстремум, сформулированной в терминах только гладких функций. Нетрудно проверить, что

$$\|\xi\| = \max_{\|p\| \leqslant 1} (p, \xi). \tag{11}$$

Очевидно, максимум достигается при $p = \xi/\|\xi\|$. Используя (11), задачу (3) можно переформулировать следующим образом:

$$\min_{u(\cdot)} \iint_G \left\{ \frac{u_x^2 + u_y^2}{2} + \max_{\|p\| \leqslant 1} (p, w(x, y)) - \alpha u(x, y) \right\} dx\, dy,$$

$$p(x, y) = \{p_1, p_2\}, \qquad w = \{u_x, u_y\}.$$

Первый шаг в дальнейших преобразованиях имеет целью вынести операцию взятия максимума по p за пределы интеграла, т.е. осуществить преобразование

$$\iint_G \max_p (p(x, y), w(x, y))\, dx\, dy = \max_{p(\cdot)} \iint_G (p(x, y), w(x, y))\, dx\, dy. \tag{12}$$

Справедливость его почти очевидна: обе части (12) достигают максимума при $p(x, y) = w(x, y)/\|w(x, y)\|$ и оба значения интегралов в (12) при этом совпадают.

Следующий шаг — перемена порядка операций взятия минимума и максимума. Определим функционал F от двух аргументов:

$$F[u(\cdot), p(\cdot)] \equiv \iint_G \left\{ \frac{1}{2} \|w\|^2 + (p, w) - \alpha u \right\} dx\, dy. \tag{13}$$

Итак, надо использовать преобразование

$$\min_{u(\cdot)} \max_{p(\cdot)} F[u(\cdot), p(\cdot)] = \max_{p(\cdot)} \min_{u(\cdot)} F[u(\cdot), p(\cdot)]. \tag{14}$$

В общем случае min max $\geqslant$ max min. Перестановочность этих операций возможна только при специальных свойствах $F[u, p]$. В нашем случае F, очевидно, линеен по p и выпукл (вниз) по u. Этих свойств достаточно для справедливости (14).

Теперь определим функционал

$$\Phi[p(\cdot)] \equiv \min_{u(\cdot)} F[u(\cdot), p(\cdot)].$$

В терминах Φ исходная задача сводится к задаче

$$\max_{p(\cdot)} \Phi[p(\cdot)],$$

где Φ, однако, определен не явным выражением, а каким-то алгоритмом решения «внутренней» задачи на min по $u(\cdot)$. Так как в дальнейшем предполагается искать max Φ методом подъема по градиенту, то нужно уметь вычислять не только $\Phi[p(\cdot)]$, но и его градиент.

Начнем с вычисления $\Phi[p(\cdot)]$. Если в (14) убрать max по p (считая $p(x, y)$ заданной функцией), то минимизация функционала по

$u(\cdot)$ есть достаточно хорошо изученная вариационная задача, обобщающая задачу Дирихле. Ее решение сводится к решению относительно простого уравнения (уравнения Эйлера для вариационной формулировки)

$$\Delta u + \operatorname{div} p + \alpha = 0, \quad u|_{\partial G} = 0. \tag{15}$$

Это уравнение получается стандартным способом. Подставляя в (13) вместо u возмущенное $u + \delta u$, разлагая подынтегральное выражение в ряд по δu (отбрасываем члены второго порядка), интегрируя по частям (учитываем, конечно, что $\delta u = 0$ на ∂G), получаем для первой вариации функционала выражение:

$$\delta F[u(\cdot), \delta u(\cdot)] = -\iint (\Delta u + \operatorname{div} p + \alpha)\, \delta u\, dx\, dy,$$

т.е. левая часть (15) есть производная Фреше для $F[u(\cdot)]$ в точке $u(\cdot)$. Уравнение (15) так или иначе решается, и для вычисления $\Phi[p(\cdot)]$ имеется эффективный алгоритм.

Перейдем к вычислению производной Ф. Проварьируем задачу, обозначив

$$u(\cdot, p(\cdot)) = \arg\min_{u(\cdot)} F[u(\cdot), p(\cdot)].$$

Тогда можно написать «явное» выражение:

$$\Phi[p(\cdot)] = F[u(\cdot, p(\cdot)), p(\cdot)].$$

Дифференцируя его, получаем

$$\frac{\partial \Phi[p(\cdot)]}{\partial p(\cdot)} = \frac{\partial F}{\partial u(\cdot)} \frac{\partial u(\cdot, p(\cdot))}{\partial p(\cdot)} + \frac{\partial F}{\partial p(\cdot)}.$$

Здесь нас выручает то обстоятельство, что

$$\frac{\partial F[u(\cdot), p(\cdot)]}{\partial u(\cdot)} = 0 \quad \text{в точке } u(\cdot, p(\cdot)).$$

Поэтому сложный и практически трудно вычисляемый объект — производная $u(\cdot, p(\cdot))$ по $p(\cdot)$ — нам не нужен.

Что же касается производной от $F[u(\cdot), p(\cdot)]$ по $p(\cdot)$ в точке $u(\cdot, p(\cdot))$, то здесь вычисления очень просты. В самом деле, варьируя аргумент p в определении (13) функционала $F[u(\cdot), p(\cdot)]$, получаем очевидное выражение

$$F[u(\cdot), p(\cdot) + \delta p(\cdot)] = F[u, p] + \iint_G (\delta p_1\, u_x + \delta p_2\, u_y)\, dx\, dy,$$

означающее, что функциональные производные F по $p(\cdot)$ суть

$$\frac{\partial F[u(\cdot), p(\cdot)]}{\partial p_1(x, y)} = u_x(x, y), \qquad \frac{\partial F}{\partial p_2} = u_y(x, y).$$

Теперь мы имеем все, чтобы описать алгоритм решения задачи.

0. Пусть имеется некоторое приближение $p(\cdot)$.

1. При фиксированном $p(x, y)$ решается задача (15), находится $u(x, y)$, вычисляются производные $u_x(x, y)$, $u_y(x, y)$.

2. Находится значение $\Phi[p(\cdot)] = F[u(\cdot), p(\cdot)]$.

3. Делается шаг по двойственным переменным:

$$p_1(x, y) = p_1(x, y) + S\,u_x(x, y), \qquad p_2 = p_2 + S\,u_y(x, y).$$

Вектор $p(x, y)$ проецируется на единичную сферу. Процесс повторяется до стабилизации двойственных переменных. Здесь S — шаг подъема по градиенту. Его выбор существенно влияет на успех всей процедуры.

Вышеприведенный алгоритм естественно трактовать как решение задачи $\max \Phi[p(\cdot)]$ по p. При этом необходимо контролировать ход вычислений, следя за эволюцией значения Φ при изменении p. В работах Ж. Лионса и сотрудников разработаны некоторые рекомендации по назначению шага S. Возможно и автоматическое регулирование S, опирающееся на сопоставление фактического приращения функционала $\Delta\Phi = \Phi[p + \delta p] - \Phi[p]$ с его первой вариацией $\delta\Phi = \Phi_p\,\delta p$. Опыт показал, что этот алгоритм быстро вырабатывает достаточно эффективное значение шага S.

§ 30. Псевдодифференциальные уравнения

Познакомимся с методами приближенного решения специфических задач, возникающих в линейной теории трещин. Начнем с постановки характерной математической задачи. В заданной области G ищется функция $u(x, y)$, удовлетворяющая интегральному уравнению

$$-\frac{1}{2\pi}\Delta \iint\limits_G \frac{u(x', y')\,dx'\,dy'}{\sqrt{(x'-x)^2+(y'-y)^2}} = f(x, y), \quad (x, y) \in G. \qquad (1)$$

Здесь Δ — оператор Лапласа по переменным x, y; функция f — заданная сила. Поясним механический смысл задачи: G — плоская область разрыва в сплошной трехмерной среде; $u(x, y)$ — «нормальный отрыв», т.е. смещение верхней границы трещины в направлении, ортогональном ее плоскости (нижняя смещается на $-u(x, y)$).

Заметим, что за пределами задачи остались такие важные вопросы, как определение самой плоской области G, по которой происходит разрыв вещества, определение тангенциальных к поверхности трещины смещений ее границ. Эти задачи в линейной теории решаются независимо от определения нормального отрыва. В частности, определение тангенциального смещения приводит к уравнению типа (1). При его решении возникают те же проблемы и применяются те же методы, но в более сложной форме, так как u становится двумерной вектор-функцией, а Δ заменяется на дифференциальный матричный оператор. Ограничимся этими разъяснениями и перейдем к чисто математическим вопросам.

Псевдодифференциальный оператор. Прежде всего отметим, что в (1) оператор Δ не внесен под знак интеграла не случайно: этому препятствует появление в ядре слишком сильной особенности (типа $1/r^3$, где $r=\sqrt{x^2+y^2}$). По форме уравнение (1) — интегральное, по математическим свойствам — дифференциальное первого порядка.

Подействуем интегральным оператором (1) на пробную функцию $u(x, y)=e^{i(kx+my)}$ (считая G полной двумерной плоскостью). Результатом будет функция $\sqrt{k^2+m^2}\,e^{i(kx+my)}$. Именно асимптотика «символа» оператора (1) $\sqrt{k^2+m^2}$ и определяет его порядок (символ оператора $\partial/\partial x$ есть ik, символ $\partial^2/\partial x^2$ есть $-k^2$, и т.д.). В то же время оператор $\Delta\iint r^{-1}$ не локальный: значение левой части (1) в точке (x, y) определяется всей функцией $u(\cdot)$ в G, а не ее значениями в сколь угодно малой окрестности (x, y), как в обычных дифференциальных операторах. Ядро интегрального преобразования имеет сильную особенность в точке $(x=x', y=y')$ и быстро убывает при удалении от точки (x, y). Это характерная для «псевдодифференциальных» операторов картина.

Как в прикладных задачах появляются уравнения с псевдодифференциальными операторами? Типичным источником таких задач является следующий прием. Предположим, что решается простое дифференциальное (как правило, эллиптического типа) уравнение в очень простой области, например уравнение Лапласа $\Delta u=0$ в круге. Простота уравнения и области понимаются в том смысле, что для некоторой простой краевой задачи известно эффективное выражение функции Грина Γ. Обычно такой краевой задачей является задача Дирихле. Если задано значение u на границе (обозначим его $u^*(s)$, где s — параметр на границе круга), то решение простой краевой задачи выписывается явно:

$$u(x, y)=\oint \Gamma(x, y; s)\,u^*(s)\,ds. \tag{2}$$

Но это не та задача, которая нас интересует. Требуется решить задачу с гораздо более сложными краевыми условиями, например с условиями

$$\alpha(s)\,u(x(s), y(s))+\beta(s)\,\frac{\partial u(x(s), y(s))}{\partial n}=\gamma(s), \quad \{x(s), y(s)\}\in\partial G. \tag{3}$$

Для этой задачи явного выражения функции Грина нет. Используем такой прием. Введем $u^*(s)$ в качестве неизвестной функции. Выразим решение в «явном виде» через искомую функцию u^* по формуле (2). Подставляя это выражение в краевое условие (3), получаем сингулярное интегральное уравнение относительно u^*. Примерно таким способом было получено уравнение (1).

Обобщенное решение. Первый вопрос, который, естественно, возникает (и ответ на него существен при построении численного метода решения (1)): что следует считать решением уравнения (1)? Здесь используется стандартная процедура, введенная Б. Г. Галеркиным с целью приближенного решения некоторых уравнений и превратившаяся в современную теорию обобщенных решений. Умножим (1) на некоторую достаточно гладкую финитную функцию $v(x, y)$ и проинтегрируем полученное выражение по G. Используя гладкость v, «перебросим» на нее часть дифференциального оператора $\Delta = \operatorname{div}\operatorname{grad}$, $\operatorname{div}^* = -\operatorname{grad}$. В результате получаем некоторое соотношение, которое должно (если u — решение) выполняться для всех пробных функций v.

Конечно, следует еще определить пространство, из которого может выбираться $u(x, y)$. Исследования показали, что с математической точки зрения можно ограничиться классом непрерывных $u(x, y)$, имеющих кусочно-непрерывные в G первые производные. Такой класс приемлем и с механической точки зрения. Обозначая $r = \sqrt{(x-x')^2 + (y-y')^2}$, имеем

$$-\frac{1}{2\pi}\iint\limits_G dx\,dy\,v(x, y)\operatorname{div}\operatorname{grad}\iint\limits_G r^{-1}u(x', y')\,dx'\,dy' =$$
$$= \frac{1}{2\pi}\iint\limits_G (\operatorname{grad} v)(x, y)\,dx\,dy \operatorname{grad}\iint\limits_G r^{-1}\,u(x', y')\,dx'\,dy'. \quad (4)$$

Внося оператор grad под внутренний интеграл, получаем векторное ядро

$$\{\Gamma_1(x-x', y-y'), \Gamma_2(x-x', y-y')\} = \{(r^{-1})_x, (r^{-1})_y\}$$

с допустимой (при оговоренных свойствах u) особенностью. Разумеется, мы использовали финитность в G функции v, опустив «краевые члены». В результате преобразований выражение (4) принимает вид

$$\frac{1}{2\pi}\iint\limits_G\iint\limits_G \{v_x(x, y)\,\Gamma_1(x-x', y-y') + v_y(x, y)\,\Gamma_2(x-x', y-y')\} \times$$
$$\times u(x', y')\,dx\,dy\,dx'\,dy'.$$

Мы получили некоторую билинейную форму от u, v (обозначим ее $l(u, v)$). Вместе с тем та же операция интегрирования, примененная к правой части (1), даст скалярное произведение функций f и v.

Итак, вместо «псевдодифференциального» уравнения (1) мы имеем обычное в современной теории соотношение, определяющее обобщенное решение:

$$l(u, v) = (f, v), \quad \forall v. \quad (5)$$

Вышеприведенные выкладки были проделаны потому, что именно соотношение (5) используется при «дискретизации» задачи.

Метод конечных элементов. Эффективный метод приближенного решения уравнений типа (1) разработан механиками на основе метода конечных элементов. Его применение в данных задачах требует некоторых предосторожностей. Введем в плоскости (x, y) квадратную сетку с шагом h. Пометим узлы этой сетки парой индексов (k, m), их геометрические координаты $x_k = kh$, $y_m = mh$. Определим в узлах искомую сеточную функцию $u_{k,m}$. С каждым узлом свяжем элементарную область $\omega_{k,m}$, состоящую из четырех примыкающих к узлу ячеек $h \times h$, и определенную на $\omega_{k,m}$ базисную функцию $\varphi_{k,m}(x, y)$. Эта функция равна единице в центре $\omega_{k,m}$, нулю — на ее границе. Внутрь $\omega_{k,m}$ функция $\varphi_{k,m}$ продолжается билинейной интерполяцией, т.е., например, обозначая $\xi = (x - x_k)/h$, $\eta = (y - y_m)/h$, в правой верхней ячейке $h \times h$, имеем выражение

$$\varphi_{k,m}(\xi, \eta) = \begin{cases} 1 - \xi - \eta + \xi\eta & \text{при } 0 \leqslant \xi, \eta \leqslant 1, \\ 0 & \text{вне } \omega_{k,m}. \end{cases}$$

Определим счетную область G_h. Будем считать точку (k, m) счетной, если $(x_k, y_m) \in G$. В этих точках определена сеточная функция $u_{k,m}$. Тогда $G_h = \bigcup\limits_{k,m} \omega_{k,m}$. Здесь, как и в дальнейшем, без специальных указаний предполагается, что индексы (k, m) пробегают значения, соответствующие счетным точкам и только им. Приближенное решение ищем в виде функции

$$u(x, y) = \sum_{k,m} u_{k,m}\, \varphi_{k,m}(x, y). \tag{6}$$

Составим уравнение (5), заменив $\forall\, v$ на $\forall\, \varphi_{k,m}$:

$$l\left(\sum_{k,m} u_{k,m}\varphi_{k,m}, \varphi_{i,j}\right) = (f, \varphi_{i,j}), \quad \forall\, i, j.$$

Используем билинейность формы l и вынесем коэффициенты $u_{k,m}$ и суммирование за пределы операции l. Обозначая $f_{i,j} = (f, \varphi_{i,j})$, $A_{i,j}^{k,m} = l(\varphi_{k,m}, \varphi_{i,j})$, получаем систему линейных алгебраических уравнений, аппроксимирующую уравнение (1) (или (4)):

$$\sum_{k,m} A_{i,j}^{k,m}\, u_{k,m} = f_{i,j}, \quad \forall\, i, j. \tag{7}$$

Очевидно, число уравнений равно числу неизвестных. В такой форме можно считать $\varphi_{k,m}$ любым базисом, а не только тем, который был описан выше.

Мы получили систему уравнений (7) с четырехиндексной матрицей. Хранение ее в памяти ЭВМ, если число базисных функций не очень мало, — проблема, впрочем, для современных ЭВМ (серии ЕС, например) уже преодолимая. Но вот вычисление матрицы остается проблемой: ведь каждый элемент A — интеграл по G (хуже того, интеграл от функции с достаточно сильной особенностью). Именно это заставляет сузить общую галеркинскую конструкцию, используя специальные конечные элементы.

Заметим, что ядро интегрального уравнения (1) зависит не от четырех аргументов (x, y, x', y'), а только от двух $(x-x', y-y')$, что дает основание ожидать соответствующего характера матрицы: $A^{k,m}_{i,j} = A^{0,0}_{k-i,j-m}$. Это весьма полезное свойство действительно удается получить, но только за счет использования одинаковых (с точностью до сдвига аргументов) конечных элементов. Очевидно,

$$\varphi_{k,m}(x, y) = \varphi_{0,0}(x - x_k, y - y_m).$$

С учетом этого имеем $l(\varphi_{k,m}, \varphi_{i,j}) = l(\varphi_{k-i,m-j}, \varphi_{0,0})$.

Теперь уравнение (7) можно записать в форме

$$\sum_{k,m} A^h_{k-i,m-j}\, u_{k,m} = f_{i,j}, \quad \forall\, i, j. \tag{8}$$

Здесь A^h означает, что матрица A вычислена для сетки $h \times h$. Несложный анализ (хотя бы размерностей) показывает, что можно вычислить «универсальную» матрицу на сетке 1×1 один раз и после этого пользоваться формулой $A^h = h^{-1}A^1$.

В универсальной матрице A^1 наиболее сложно вычисляются элементы $A_{i,j}$ с малыми значениями $|i| + |j| \leqslant 3$, так как именно в этих случаях сказывается сингулярность подынтегральной функции. Для вычисления таких элементов были разработаны специальные программы аккуратного интегрирования, с помощью которых вычислены элементы $A_{i,j}$ для малых i, j (их можно найти в соответствующих работах). Остальные элементы $A_{i,j}$ легко вычисляются по простым асимптотическим формулам. При этом используется очевидное свойство симметрии: $A_{i,j} = A_{-i,j} = A_{i,-j}$, позволяющее хранить в памяти только «четверть» матрицы $(i, j = 0, 1, \ldots, K$, где K — целое число, такое, что Kh превосходит линейный размер $G)$.

Итак, составление системы уравнений (8) проблем не содержит, если известны значения $A_{i,j}$ при малых $|i| + |j|$. Следующий вопрос — решение этой системы. Число уравнений не так уж мало. В расчетах, которые обсуждаются ниже, число узлов было от 500 до

1000, причем матрица системы полностью заполнена. Применение стандартных программ решения линейных уравнений потребовало бы порядка 10^9 операций, 10^6 памяти, что уже представляет серьезную проблему для современных машин ЕС (типа 1040, 1045, 1050), не говоря уже о БЭСМ-6, для которой это просто непосильная задача (а именно эта ЭВМ применялась в расчетах).

Достаточно эффективным средством решения системы (8) является простейший итерационный метод, известный под названием «релаксационный». Он состоит в поочередном пересчете $u_{k,m}$ по формуле

$$u_{k,m} = A_{0,0}^{-1}\left(f_{k,m} - \sum_{i,j}{}' A_{i-k,j-m}\, u_{i,j}\right). \tag{9}$$

Здесь в сумме по i, j пропускается слагаемое $i = k$, $j = m$. Используется также ускорение сходимости методом сверхрелаксации. По формуле (9) вычисляется «предварительное» значение $\tilde{u}_{k,m}$, окончательное значение находится по формуле $u_{k,m} = u_{k,m} + \omega(\tilde{u}_{k,m} - u_{k,m})$. Параметр ω определялся экспериментально.

В табл. 22 приведена эволюция нормы невязки уравнения (8) в зависимости от номера итерации ν при разных значениях ω. Видно, что оптимальное значение $\omega \approx 1.4$, при этом скорость сходимости достаточно высока. Начиная с тривиального приближения $u_{i,j} \equiv 0$, за 15 итераций можно получить невязку, существенно меньшую f. Такое приближенное решение системы (8) можно трактовать как точное, соответствующее незначительно измененной правой части.

Таблица 22

ν \ ω	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7
1	4.6e1	5.0e1	5.4e1	5.8e1	6.4e1	7.0e1	7.8e1	8.8e1
4	1.5e1	1.8e1	1.5e1	1.3e1	1.0e1	7.3e0	5.6e0	6.5e0
7	8.8e0	6.8e0	4.8e0	3.0e0	1.4e0	4.7e−1	8.3e−1	2.0e0
10	4.1e0	2.6e0	1.4e0	6.1e−1	1.4e−1	5.3e−2	1.9e−1	7.6e−1
13	1.9e0	9.9e−1	4.2e−1	1.2e−1	1.2e−2	8.4e−3	5.4e−2	3.1e−1
16	8.5e−1	3.7e−1	1.2e−1	2.3e−2	1.2e−3	1.5e−3	1.5e−2	1.4e−1
19	3.9e−1	1.4e−1	3.5e−2	4.5e−3	1.1e−4	4.8e−4	4.4e−3	5.7e−2

Возникает вопрос о точности расчета при приемлемом числе узлов сетки. В самом деле, 1000 узлов на двумерную область — это не так уж много (примерно по 30 узлов на линейный размер области). Уменьшить шаг h для повышения точности здесь не так просто: вы-

численение суммы в правой части (9) для каждого узла (k, m) «стоит» $O(h^{-2})$ операций. Число узлов $O(h^{-2})$, да и число итераций можно оценивать как $O(h^{-1})$ или, в лучшем случае, $O(h^{-1/2})$.

Вышеприведенные оценки сделаны по аналогии с хорошо изученным уравнением Лапласа. Там мы имеем дело с оператором второго порядка, для которого обусловленность матрицы конечномерной аппроксимации есть $O(h^{-2})$, сходимость простых итераций имеет скорость $1 - O(h^2)$ (сверхрелаксация при оптимальном ω доводит коэффициент сходимости до $1 - O(h)$). Мы же имеем дело с оператором первого порядка, с числом обусловленности матрицы A порядка $O(h^{-1})$. Эксперимент подтверждает это предположение.

Итак, «стоимость» приближенного решения есть в лучшем случае $O(h^{-4.5})$ операций и она резко возрастает при незначительном уменьшении шага (например, раз в 25 при переходе от h к $h/2$). На БЭСМ-6 расчет трещины с числом узлов порядка 800 стоит около 5 мин машинного времени. Достаточно ли доступного нам числа узлов для получения решения с нужной для приложений точностью? Следует иметь в виду, что требования к точности не очень высоки: погрешность $1 \div 5$ % допустима для многих технических приложений. Конечно, ответ зависит от дифференциальных свойств решения, последние, в свою очередь, — от гладкости правой части и свойств оператора в (1).

В прикладных расчетах f, как правило, — не очень сложная функция. Это понятно. Ведь трещина есть объект малых размеров, находящийся «в глубине» некоторой конструкции. Функция f создается системой внешних нагрузок. «Создать» функцию f, сильно меняющуюся на малом расстоянии, не так-то просто, тем более что никто к этому специально не стремится. Псевдодифференциальный оператор в (1) по своей природе близок к «эллиптическому дифференциальному оператору первого порядка», т.е. при его обращении гладкость решения повышается на один порядок по сравнению с f. Поэтому есть основание ожидать, что искомое решение $u(x, y)$ — достаточно просто устроенная функция.

Действительно, расчеты простых задач (G — эллипс, $f = \text{const}$) с известным точным решением показывают, что $u(x, y)$ есть колоколообразная функция простого вида, и при числе узлов $20 \div 30$ на линейный размер области точность приближенного решения очень высока почти всюду в области G. Исключение составляет узкая (шириной $(2 \div 3)h$) полоса узлов, примыкающих к контуру ∂G. В этой полосе погрешность составляет $10 \div 30$ %, вне ее — погрешность около 1 %; и чем дальше от границы, тем она меньше. Это обстоятельство не случайное, и его можно было бы предвидеть, используя достаточно развитую математическую теорию уравнения (1). Основной результат, который необходимо учесть при построении численного метода, состоит в следующем. При гладкой правой

части f решение u есть непрерывная функция, обращающаяся в нуль на ∂G, гладкая всюду, кроме окрестности контура ∂G.

Известен и характер $u(x, y)$ в этой окрестности. Вводя в точках контура локальную систему координат (ξ, η), где ξ — длина дуги на контуре, η — расстояние от контура по внутренней нормали, имеем для u асимптотику

$$u(\xi, \eta) = N(\xi)\,\eta^{1/2} + O(\eta^{3/2}). \qquad (10)$$

Иными словами, u имеет на контуре корневую особенность. Теперь понятны источники погрешностей численного решения вблизи контура: негладкость самого решения и грубая аппроксимация контура «ступенчатой» линией. Именно такой является граница счетной области G_h, на границе которой решение в форме (6) обращается в нуль.

Казалось бы, можно пренебречь полученной погрешностью: ведь в подавляющем числе узлов точность приближенного решения высока. Однако она высока для величин, особенного интереса для приложений не представляющих. Дело в том, что наиболее важным для приложений является так называемый коэффициент концентрации напряжений на контуре трещины ∂G — это функция $N(\xi)$, входящая в асимптотику (10). Именно ради определения $N(\xi)$ затеваются расчеты, так как этой функцией определяется дальнейшая судьба трещины. Будет ли она расти и с какой скоростью (т.е. представляет ли трещина опасность для конструкции), через какое время ее рост приведет к разрушению тела с трещиной, и т.п.?

Итак, получив численное решение, нужно проанализировать ход изменения $u_{k,m}$ вблизи контура и извлечь из него оценку $N(\xi)$. Ясно, что при этом анализе нужно отступить от контура внутрь области на $(2 \div 3)h$, однако там (если шаг h не очень мал) уже начинают сказываться величины порядка $O(\eta^{3/2})$. В трещинах простой формы такой анализ давал неплохие результаты. Но при переходе к трещинам «произвольной» формы ситуация осложняется и требуется уточнение описанной выше расчетной схемы.

Кстати, поясним смысл краевого условия «$u = 0$ на ∂G». Если подействовать оператором (1) на функцию u с асимптотикой (10), получим функцию $f(x, y)$, непрерывную в G вплоть до границы. Если же $u(x, y)$ непрерывна в G и не равна нулю на ∂G, $f(x, y)$ обращается в бесконечность на ∂G.

Метод граничных сеток. Перейдем к описанию метода уточнения расчетов около ∂G. Сначала, однако, обсудим возможности стандартных способов такого уточнения. Простое уменьшение шага, как уже отмечалось, делает расчеты слишком дорогими. В технике метода конечных элементов проблемы адаптации к сложному, нерегулярному виду границы области решают, используя конечные элементы с носителями неправильной формы. Это позволяет аппроксимировать

границу не ступенчатой кривой, а, например, набором малых хорд. Сложный характер решения вблизи границы (10) учитывается тем, что вместо стандартных базисных функций в ячейках сетки, примыкающих к границе, вводятся функции, уже содержащие особенность требуемого типа (корневую, в данном случае).

Однако, как нетрудно понять, применение таких методов разрушает «теплицеву» форму матрицы A, и приходится работать с общей четырехиндексной матрицей $A^{k,m}_{i,j}$, причем эта матрица — индивидуальная для каждой области. К таким же последствиям приводит и прием «регуляризации». Можно искать решение в форме $u(x, y) = \sqrt{\Phi(x, y)}\, v(x, y)$, где $\Phi(x, y)$ — известная функция, положительная в G и без касания обращающаяся в нуль на ∂G (построение такой функции в общем случае не так-то просто!), $v(x, y)$ — подлежащая расчету функция, уже не содержащая корневой особенности.

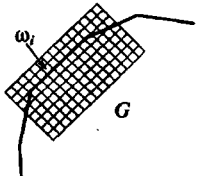


Рис. 56

Изложенный ниже метод позволяет заметно уточнить расчет u около границ, достаточно надежно определить $N(\xi)$ и сделать первые шаги в расчете роста трещин без существенного увеличения объема вычислений. Начнем с описания области G. Она задается своим контуром ∂G, последний — набором вершин $\{X_i, Y_i\}$ $(i = 1, 2, \ldots, I)$. Каждые две соседние вершины контура (i-я и $(i+1)$-я) соединяются отрезком прямой (который называется i-м ребром). Таким образом (так как I-я вершина совпадает с первой) мы получаем замкнутую кривую («полигон»). При достаточно большом I (около 100 в расчетах) эта кривая хорошо аппроксимирует те, в конце концов не такие уж сложные контуры, которые встречаются в прикладных расчетах. Первый этап расчета проводится так, как это было описано выше, на сетке с некоторым относительно грубым шагом h. Он дает нам «грубое решение» $U_{k,m}$, которое по интерполяционной формуле (6) восполняется до непрерывной в G_h функции $U(x, y)$.

Следующий этап расчета — уточнение решения около контура. Он состоит из I отдельных задач. Около каждого i-го ребра строится своя локальная малая область ω_i, покрытая сеткой с шагом h_i. Сетка строится так, что i-е ребро проходит по линии сетки (рис. 56), а область ω_i покрывает некоторую окрестность i-го ребра. После этого решается задача (1) в предположении, что в $G \setminus \omega_i$ решение уже известно (это «грубое» решение U). Такая локальная задача имеет форму

$$-\frac{1}{2\pi}\Delta \iint\limits_{\omega_i} \frac{1}{r}\, u(x', y')\, dx'\, dy' = f_i(x, y), \quad (x, y) \in \omega_i, \qquad (11)$$

$$f_i(x, y) = f(x, y) + \frac{1}{2\pi}\Delta \iint\limits_{G \setminus \omega_i} \frac{1}{r}\, U(x', y')\, dx'\, dy'. \qquad (12)$$

Задача (11) решается стандартным образом. Хотя число таких уточняющих задач велико (около 100), число узлов в ω_i относительно мало (около 100 при $h_i \approx h/3$). Объем вычислений так сильно зависит от числа узлов, что решение всех вспомогательных задач требует примерно того же машинного времени, что и получение грубого решения. Следует подчеркнуть, что наиболее трудоемкий элемент вышеизложенной методики — не решение системы (11), а ее формирование, т.е. вычисление f_i по формуле (12). Именно этому элементу надо уделить особое внимание.

Функции f_i вычислялись так. Сетка с шагом h_i продолжалась на несколько шагов за пределы ω_i. Вдоль границы ω_i выделялся некоторый «пояс», покрытый сеткой с малым шагом h_i. Интеграл по поясу от грубого решения вычислялся достаточно аккуратно с помощью матрицы A^{h_i}. В этом поясе значение U интерполировалось в узлы сетки по формуле (6). Такая аккуратная процедура связана с тем, что при вычислении этой части интеграла нужно учитывать сингулярность ядра интегрального преобразования (1). В остальной части $G\backslash\omega_i$ ситуация проще. Ядро уже гладкое, оно быстро убывает (как $1/r^3$). Функция $U(x, y)$ тоже достаточно гладкая. Поэтому соответствующая часть интеграла в (12) вычисляется на более грубой сетке с шагом $(2 \div 3)\,h$. Разумеется, точность подобной методики нуждается в тщательном контроле.

Проверка методики. Описанный выше метод расчета трещин проходит стадию становления; опыт его применения пока ограничен. В такой ситуации естественно встает вопрос о доверии к полученным результатам, о контроле самой методики. В вычислительной физике эта проблема всегда возникает при освоении нового класса задач. И решается она специфическими, не очень строгими методами. Математически строгие оценки либо отсутствуют, либо настолько грубы, что реального представления о точности расчетов не дают. Ниже мы опишем и обсудим те средства контроля, которые использовались в этой задаче. По своему характеру они типичны в вопросах подобного рода.

Основное средство контроля — сопоставление расчетов с некоторыми известными точными решениями (например, известны решения (1) в эллипсе при линейной зависимости f от x, y). Такие сравнения проводились, их результат был признан положительным. Что это означает, мы обсуждать не будем, отсылая читателя к специальным подробным публикациям. С методической точки зрения здесь все ясно. То же самое относится и к сравнению с некоторыми приближенными аналитическими решениями (например, для трещин в форме вытянутого прямоугольника известны асимптотики решения вблизи середины длинных сторон границы).

Перейдем к так называемым «внутренним» средствам контроля. Это очень важный элемент контроля, постоянно используемый не только при создании новой методики, но и при проведении серийных производственных расчетов.

1. Один из таких способов контроля — расчеты на разных сетках. Хотя возможности менять сетку, как было уже объяснено, в данном случае весьма ограничены, небольшое число контрольных расчетов с уменьшенным вдвое шагом было все же проведено.

2. При решении вспомогательных задач на каждом ребре размещается до 10 узлов h_i-сетки, т.е. можно говорить о функции $N(\xi)$, представленной на сетке, содержащей более 500 узлов. На каждом ребре $N(\xi)$ вычисляется по своей сетке (эти сетки между собой не согласованы), и при наличии грубых погрешностей в вычислении N функция $N(\xi)$ на полном контуре может «распасться» на несогласованные друг с другом куски (соответствующие разным ребрам). Таким образом, одно из средств контроля — просто визуальный анализ функции $N(\xi)$ на контуре.

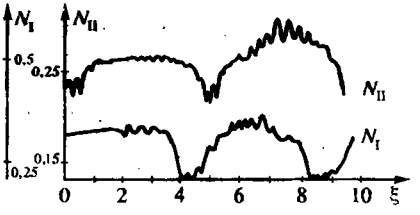


Рис. 57

На рис. 57 показаны графики функций $N_I(\xi)$ и $N_{II}(\xi)$, полученные при решении одной и той же задачи, в которой область G имела форму «банана». Два графика отвечают разному числу ребер полигона, аппроксимирующего контур (в расчете I их больше, чем в расчете II). Видно, что участки $N(\xi)$, соответствующие разным ребрам, хорошо согласуются между собой, хотя заметны и небольшие разрывы. Следует учесть, что в этих расчетах начало отсчета параметра ξ оказалось в разных точках контура.

Обратим внимание на то, что участки, соответствующие разным ребрам контура, имеют форму выпуклых вверх или вниз дуг с достаточно большим перепадом, который, однако, существенно уменьшается при расчете с более короткими ребрами. Это не случайный эффект. Дело в том, что функция $N(\xi)$ очень чувствительна к кривизне контура. Угловые точки контура с внутренним углом, бо́льшим π, являются точками локального максимума $N(\xi)$. Если этот угол меньше π (угол «выступает» из области), такая точка является точкой локального минимума. Грубо говоря, значение $N(\xi)$ тем больше, чем больше около граничной точки ξ области трещины (и наоборот). Таким образом, «волнистый» вид $N(\xi)$ есть счетный эффект, связанный с аппроксимацией контура кусочно-линейной кривой. Величина этих паразитических колебаний уменьшается при более точной аппроксимации гладкого контура «полигоном».

3. Еще один способ контроля — проверка некоторых асимптотик. Известно, что при постоянной силе f в области, имеющей форму бесконечного угла с раствором β, решение имеет вид (в цилиндрических координатах r, φ)

$$u(r, \varphi) = v(\varphi)\, r^{\gamma(\beta)}.$$

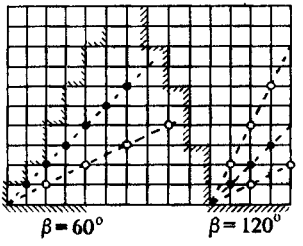


Рис. 58

Функция $\gamma(\beta)$ известна. Приведенное выражение является асимптотикой решения около угловой точки контура. Для проверки точности метода проводились расчеты трещин, контуры которых содержали угловые точки с внутренним углом 60°, 90°, 120°. Расчетные данные подвергались анализу с целью проверки асимптотики. Проверка производилась с использованием данных, полученных на уточняющей сетке, соответствующей одному из ребер контура, примыкающих к угловой точке.

Рисунок 58 показывает характер такой сетки для внутренних углов 60° и 120°. Одно ребро, примыкающее к угловой точке, проходит по координатной линии вспомогательной сетки, второе ребро аппроксимируется ступенчатой линией. Выбиралось несколько лучей, исходящих из угловых точек (два для угла 60°, три для углов 90°, 120°). Эти лучи проходят по узлам сетки, и в них вычислялись значения u_n/n^γ (n — номер точки на луче). Согласно асимптотике значения должны быть почти постоянными на луче, что видно из табл. 23. Для угла 60° приведены два варианта расчета (они построены по двум вспомогательным сеткам, примыкающим к углу, но значения в них соответствуют геометрически одинаковым точкам). Из табл. 23 хорошо видно, с какой точностью выполняется постоянство $u(r, \varphi)/n^\gamma$. Исключение составляют два-три

Таблица 23

n	$\beta = 60°$ $\gamma = 0.915$ u_n/n^γ	$\beta = 60°$ $\gamma = 0.915$ u_n/n^γ	$\beta = 90°$ $\gamma = 0.815$ u_n/n^γ	$\beta = 120°$ $\gamma = 0.71$ u_n/n^γ
1	430 440	440 440	244 205 243	316 266 333
2	302 408	297 408	229 175 228	318 252 364
3	256 413	256 413	229 176 228	326 268 376
4	236 411	233 411	227 175 225	328 272 380
5	238 401	238 404	225 174 222	329 275 382
6	243 404	241 398	222 173 219	328 276 382
7	239 396	238 405	220 172 216	327 277 382
8	237 394	233 404	218 171 213	326 277 382
9	236 392	228 415	216 170 214	324 277 382
10	240 392	223 431	214 168	320 276
11	242 398	218 461	213 167	320 275
12	245	216	165	319 274
13	250	216	165	273
14	256			

ближайших к угловой точке значения. Причина и здесь кроется в грубой аппроксимации угла ступенчатой ломаной. Дополнительный контроль — в расчете при $\beta = 90°$ два крайних столбца должны совпадать, так как соответствующие им лучи симметричны относительно биссектрисы угла.

Динамика трещины. Выше было указано, что основным результатом, извлекаемым из расчета трещины, является коэффициент интенсивности напряжений на контуре трещины $N(\xi)$, так как именно эта функция определяет рост трещины и, в конечном счете, время разрушения конструкции, содержащей трещину. При решении задач динамики трещин использовалась теория, согласно которой считается известной функция $v(N)$, имеющая смысл скорости продвижения границы трещины (по направлению нормали к ней) в зависимости от значения N в данной точке контура. Итак, если функция $N(\xi)$ известна, то контур трещины движется со скоростью $v(N(\xi))$.

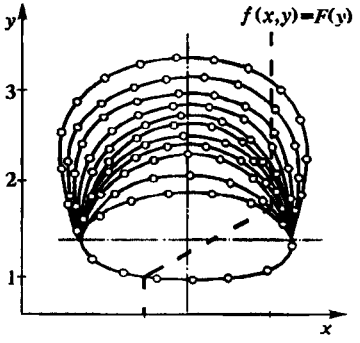


Рис. 59

Расчет динамики контура $\partial G(t)$ проводится по схеме, напоминающей простейшую схему интегрирования Эйлера. Пусть известен контур на момент времени t. При расчете трещины с данным контуром вычисляются $N(\xi)$, $v[\xi] = v(N(\xi))$. Точнее, вычисляется последовательность v_i (эти величины имеют смысл смещения середины i-го ребра в ортогональном к нему направлении). Обозначая точкой (X_i, Y_i) середину i-го ребра контура $\partial G(t)$, при некотором шаге численного интегрирования τ получаем точки ребра $\partial G(t+\tau)$ по очевидным формулам

$$X_i = \tilde{X}_i + \tau v_i n_x^i, \qquad Y_i = \tilde{Y}_i + \tau v_i n_y^i,$$

где $\{n_x^i, n_y^i\}$ — вектор единичной нормали к i-му ребру $\partial G(t)$.

Ограничимся этим общим описанием, в котором опущены многие технические подробности процедуры. Она не так проста, как это может показаться на основании того, что выше изложено. Сложности (и весьма значительные) связаны с характером зависимости $v(N)$. Например, для пластмасс $v(N) \approx N^{10} \div N^{20}$. Такая резкая зависимость v от N приводит к тому, что погрешности вычисления N, неизбежные и, быть может, не очень существенные для величины N, при вычислении v резко возрастают и описанная выше процедура интегрирования подвержена сильной неустойчивости. Если какая-то точка на одном шаге «вырвалась вперед», она, как было указано вы-

ше, становится точкой локального минимума $N(\xi)$; на следующем шаге интегрирования она практически стоит на месте, и т.д. Рисунок 59 дает представление о том, как происходит расчет динамики трещины под воздействием неравномерной нагрузки F. Показана эпюра нагрузки $f(x, y) = F(y)$ и представлена форма трещины для моментов времени 0, 5.14, 6.60, 7.152, 7.215, 7.259, 7.298, 7.330, 7.355, 7.373, 7.391, 7.407 (в последовательности снизу-вверх соответственно).

Одним из средств визуального контроля является проверка симметричности. Решение должно быть симметричным относительно прямой, проходящей через центр трещины (начальный контур — симметричный эллипс). Но с расчетной точки зрения правая и левая части области, конечно, несимметричны. При столь сильной зависимости $v(N)$ и не такой уж высокой точности расчета $N(\xi)$ можно было бы опасаться сильного проявления этой расчетной несимметричности. Следы ее видны на рис. 58, но они едва заметны.

§ 31. Метод конечных суперэлементов

Метод конечных элементов в настоящее время прочно вошел в арсенал фундаментальных вычислительных средств. Основная идея метода, его теория и практика применения описаны в многочисленных монографиях. Некоторые сведения о методе приведены в § 3, обсуждается он также в § 30 и в настоящем параграфе, посвященном специальной конструкции, которую можно трактовать как некоторое развитие идей метода конечных элементов. Она ориентирована на очень специфический класс задач, однако в различных приложениях все чаще возникают задачи с подобного рода особенностями. Представление об их характере дает следующая задача.

Задача о трещине гидроразрыва. Эта задача связана с проблемой использования геотермического тепла. Имеется «трещина» — разрыв сплошной среды; считается, что разрыв произошел по некоторой плоской области G. К трещине подводятся две (или больше) скважины, через которые протекает вода. Под действием давления воды трещина раскрывается, приобретает некоторый объем, заполненный водой. Давление на скважинах поддерживается различным, благодаря чему одна из них является нагнетающей, другая — отбирающей. Возникает течение воды в трещине: вода входит через нагнетающую скважину, некоторое время (в течение которого происходит нагрев воды) течет внутри трещины, затем выходит из трещины через отбирающую скважину.

Перейдем к математической формулировке задачи. Задана двумерная область G, в которой ищутся две функции: $u(x, y)$, имеющая смысл раскрытия трещины (см. § 30) и $p(x, y)$, имеющая смысл давления воды. Рассматривается установившееся течение: функции

не зависят от времени. Скорость однозначно связана с градиентом давления теорией течения вязкой жидкости в узком канале (формула Буссинеска): $v(x, y) = ((2u)^2/12\,\mu)\,\mathrm{grad}\,p$, где μ — коэффициент вязкости. Итак, давление (и скорость течения) воды зависят от раскрытия трещины, последнее же зависит от давления. В результате получается следующая связанная система уравнений для u и p (см. § 30):

$$-\frac{1}{2\pi}\Delta \iint\limits_G \frac{1}{r}\,u(x', y')\,dx'dy' = p(x, y) - p_0, \tag{1}$$

где p_0 — заданная величина. Уравнение для p имеет вид

$$\mathrm{div}\,[u^3\,\mathrm{grad}\,p] = Q(x, y), \quad (x, y) \in G\backslash Ug_i. \tag{2}$$

Здесь $Q(x, y)$ — сток, связанный с фильтрацией воды через стенки трещины. Будем считать Q заданной величиной, хотя в действительности Q определяется решением специального уравнения, в которое входят p и u.

Перейдем к постановке краевых условий. Для u краевые условия обычные (см. § 30): $u|_{\partial G} = 0$. Сложнее обстоит дело с p. Уравнение (2) определено не в G, а в $G\backslash Ug_i$, где g_i — площади, занимаемые скважинами. Это круги малого радиуса r, однако на их границах поставлены краевые условия

$$p = P_i \text{ на } \partial g_i \tag{3}$$

(P_i — заданное давление на i-й скважине). Сложность задачи состоит в том, что скважины имеют размер, малый не только относительно размера трещины, но даже относительно приемлемого шага сетки H.

В расчетах, которые обсуждаются ниже, трещина имела форму круга радиусом $R = 250$ м, шаг сетки $H = 17$ м, а радиус скважины $r = 0{,}1$ м. Первая проблема в том, как учесть влияние скважины (а оно определяет всю картину рассчитываемого явления) в расчетной схеме со столь большим шагом. Собственно, ради таких ситуаций и разрабатывается расчетная схема метода конечных суперэлементов (МКСЭ).

В задаче есть еще одна нестандартная деталь — отсутствие явно заданных краевых условий для p на внешней границе ∂G. Это связано с тем, что в (2) $u^3(x, y)$ играет роль коэффициента диффузии. Но $u(x, y)$ уменьшается при приближении (x, y) к ∂G, как корень квадратный от расстояния до ∂G (см. § 30). Следовательно, u^3 стремится к нулю, как расстояние до ∂G в степени 3/2. Таким образом, уравнение (2) вырождается на границе. Теорию таких уравнений разрабатывал М. В. Келдыш. Основной его результат состоит в том, что при определенной скорости обращения в нуль коэффициента диффузии в окрест-

ности границы области классические краевые условия для уравнения (2) ставиться не могут, их заменяет условие ограниченности решения. Мы имеем дело именно с таким случаем, и численная реализация условия ограниченности потребовала определенных изобретений. Эта особенность вычислительной схемы, однако, к методу конечных суперэлементов отношения не имеет. Итак, сейчас основной вопрос: как учесть влияние скважины размером $r \ll H$ на сетке $H \times H$?

Метод конечных суперэлементов (одномерный вариант). Начнем с разбора более простой ситуации. Пусть требуется решить одномерное уравнение диффузии

$$\frac{d}{dx}\left(D(x)\frac{du}{dx}\right) - A(x)\,u = 0, \quad x \in [0, X], \tag{4}$$

с краевыми условиями, для простоты, первого рода. Интервал $[0, X]$ состоит из M интервалов длиной H: $X = MH$, где $M \approx 20 \div 30$, например. На интервале H функции $D(x)$, $A(x)$ имеют достаточно сложное строение. Например, интервал H разбит на какое-то число частей, в каждой из которых D и A имеют постоянные значения. Пусть, наконец, имеется небольшое число типов таких отрезков длиной H, а вся система длиной X каким-то образом скомпонована из стандартных кусков длиной H. Эта ситуация моделирует (конечно, упрощенно) некоторые характерные трудности, с которыми сталкиваются при расчете такого важного объекта, как современный энергетический атомный реактор.

Можно ли в таких условиях, когда описание физической структуры области явно требует сетки с шагом $h \ll H$, тем не менее построить расчетную схему стандартного типа с шагом H? Оказывается, можно, хотя это связано с некоторыми затратами.

Рассмотрим стандартный интервал длиной H, который входит в компоновку всей задачи. Оснастим его двумя базисными функциями, обозначив их $\varphi_1(x)$, $\varphi_2(x)$. Функцию φ_1 определим как решение уравнения (4) с краевыми условиями $\varphi_1(0) = 1$, $\varphi_1(H) = 0$. Функцию φ_2 определим точно так же, но с краевыми условиями $\varphi_2(0) = 0$, $\varphi_2(H) = 1$. Если быть аккуратным, то эти функции следует отметить еще одним индексом — номером типа той стандартной ячейки длиной H, для которой они рассчитаны: $\varphi_i^t(x)$ $(i = 1, 2)$. Говоря о решении (4), мы имеем в виду решение по стандартной схеме с шагом $h \ll 1$, обеспечивающим нужную точность в обычном смысле слова. Такие базисы должны быть рассчитаны для всех типов ячеек, которые могут встретиться в компоновке исходной задачи. Эта конструкция отличается от стандартной конструкции метода конечных элементов только тем, что в методе конечных элементов базис строится из общих, не связанных с решаемой задачей соображений и функции φ_i выписываются явно: например, $\varphi_1(x) = 1 - x/H$, $\varphi_2(x) = x/H$.

Область полного расчета $[0, X]$ покрывается стандартными конечными элементами (в данном случае, отрезками длины H, оснащенными своими базисами). Получается обычная сетка точек $x_m = mH$, в которых определена сеточная функция u_m. Теперь мы имеем процедуру восполнения сеточной функции $\{u_m\}_{m=0}^{M}$ до непрерывной $u(x)$. Это в сущности есть специфическая интерполяция:

$$u(x) = u_m \varphi_1^t(x) + u_{m+1} \varphi_2^t(x), \quad x \in [x_m, x_{m+1}]. \tag{5}$$

Здесь t — тип элемента, помещаемого на интервале $[x_m, x_{m+1}]$.

Заметим, что в силу специального выбора базиса функция $u(x)$ непрерывна и почти всюду, за исключением точек x_m, удовлетворяет решаемому уравнению (мы отвлекаемся от погрешностей расчета базиса). Для того чтобы эта функция была решением исходной задачи, нужно обеспечить непрерывность потока в каждом внутреннем узле x_m. Соответствующие «балансные соотношения» образуют систему уравнений, имеющую форму обычной трехточечной разностной схемы, решением которой должны быть u_m (для того чтобы (5) было просто точным решением).

Для составления таких соотношений нам нужны не базисные функции φ, а некоторые функционалы от них. Определим эти функционалы (потоки):

$$\Pi_{i,1}^t = -D^t \frac{\partial \varphi_i^t}{\partial x}\Big|_{x=0}, \qquad \Pi_{i,2}^t = D^t \frac{\partial \varphi_i^t}{\partial x}\Big|_{x=H}, \qquad i = 1, 2. \tag{6}$$

Они имеют смысл потоков внутрь ячейки через ее левую и правую границы. Очевидно, при интерполяции (5) поток в точку x_m определяется значениями u_{m-1}, u_m, u_{m+1}, и точное решение характеризуется тем, что он равен нулю (ввиду отсутствия δ-образных источников). Это, впрочем, равносильно непрерывности потока: $Du_x|_{x_m-0} = Du_x|_{x_m+0}$.

В терминах функционалов (6) условие непрерывности потока в узле x_m записывается следующим образом:

$$u_{m-1} \Pi_{1,2}^1 + u_m \Pi_{2,2}^1 + u_m \Pi_{1,1}^2 + u_{m+1} \Pi_{2,1}^2 = 0, \tag{7}$$

или в стандартной трехточечной форме:

$$a_m u_{m-1} + 2b_m u_m + c_m u_{m+1} = 0, \tag{8}$$

где $a_m = \Pi_{1,2}^1$, $b_m = 0.5(\Pi_{2,2}^1 + \Pi_{1,1}^2)$, $c_m = \Pi_{2,1}^2$. Для простоты предполагаем, что на $[x_{m-1}, x_m]$ помещен элемент первого типа, на $[x_m, x_{m+1}]$ — элемент второго типа.

Таким образом, (8) есть так называемая «точная разностная схема» (точная в той мере, в которой точно нахождение базисных функций φ). Если бы исходная задача (4) была неоднородной и содержала в правой части не нуль, а $f(x)$, причем f — гладкая на H-сетке функция (т.е. на интервалах $[x_m, x_{m+1}]$ можно пренебречь отличием f от среднего значения $f_{m+1/2}$), то следовало бы оснастить каждый элемент еще одной базисной функцией $\varphi_0^t(x)$, определяемой краевыми значениями $\varphi(0) = \varphi(H) = 0$ но с правой частью, тождественно равной единице. Тогда в схеме (8) добавится слагаемое $f_{m-1/2}\Pi^1_{0,2} + f_{m+1/2}\Pi^2_{0,1}$.

Подчеркнем еще раз отличие метода конечных суперэлементов от метода конечных элементов. В методе конечных суперэлементов базис состоит из решений исходной задачи, и это определяет его преимущество. В частности, метод конечных суперэлементов претендует на расчет с большим шагом H. Однако с этим же связана и его определенная ограниченность. Он требует предварительного расчета базисов и сохранения используемой в дальнейшем информации — потоков Π, что реально можно сделать только для ограниченного числа стандартных «элементов», из которых и должна компоноваться рассчитываемая система. Метод конечных элементов, конечно, гораздо универсальнее, но он требует достаточно малых размеров конечных элементов, геометрическая форма которых может быть достаточно разнообразной.

Гомогенизация. Как подчеркивалось выше, существуют важные приложения, в которых посильными для расчетов на современных ЭВМ являются сетки с шагом H, однако требующие расчета реальные системы имеют пространственную структуру с много меньшими H размерами. Конкретно это можно представить себе, например, следующим образом. Решается задача (4) в слоистой среде, т.е. интервал $[0, X]$ заполнен слоями разных веществ, толщины которых много меньше H. Для этих веществ известны коэффициенты D и A. Используется такой подход. Для каждой ячейки размером H пытаются найти свои эффективные коэффициенты $D_{m+1/2}$, $A_{m+1/2}$, с которыми составляют стандартную разностную схему. Такая операция весьма распространена, например, в математической теории атомных реакторов. Она получила название «*гомогенизация*», а эффективные коэффициенты уравнения (4) называют «гомогенизированными».

Часто гомогенизация осуществляется простым осреднением коэффициентов «на решении». Поясним суть дела на примере задачи (4). Прежде всего следует конкретизировать «типичное» решение, по которому производится осреднение. Так как шаг H в некотором смысле в дальнейшем считается не очень большим (в том смысле, что можно ожидать малого изменения искомого решения

полной задачи при переходе от точки x_m к x_{m+1}), то «типичное» решение в ячейке длиной H определяется решением задачи (4) с краевыми условиями $\tilde{u}(0) = \tilde{u}(H) = 1$.

Определив решение $\tilde{u}(x)$, осредняют на нем коэффициенты (4), придерживаясь физических соображений. Так, в уравнении (4) и ему подобных важную роль играет число поглощенных частиц в ячейке

$$\int_0^H A(x)\, u(x)\, dx$$

(здесь $u(x)$ означает решение исходной задачи (4)). Исходя из стремления правильно передать физическую сущность процесса (а число поглощенных частиц есть одна из основных характеристик при расчетах атомных ректоров, их защит, прохождения излучения через оптически толстые среды и т.п.), осредненный (гомогенизированный) коэффициент A определяют, например, формулой

$$A = \int_0^H \tilde{u}(x)\, A(x)\, dx \Big/ \int_0^H \tilde{u}(x)\, dx.$$

Близкие соображения используют и при осреднении коэффициента диффузии.

Точная разностная схема. Если отвлечься от погрешностей расчета базисных функций, то при любых значениях u_m функция (5) является «кусочно-точным» решением уравнения (4): оно удовлетворяется во всех точках, кроме узлов x_m. Если к тому же значения u_m удовлетворяют разностному уравнению (7), т.е. в узлах x_m выполняется условие непрерывности потока, то функция (5) является просто точным решением (4).

Описанная выше процедура построения «точной» схемы является наиболее обоснованной процедурой «гомогенизации». Она дает нетривиальные результаты даже при постоянных коэффициентах D и A только за счет «большого» шага. Поучительно привести характерный пример. Пусть A и D в (4) постоянны, но шаг H большой. Тогда точная схема (7) получается стандартным образом:

$$(\tilde{D}/H^2)(u_{m-1} - 2u_m + u_{m+1}) - \tilde{A}u_m = 0,$$

если используются коэффициенты

$$\tilde{D} = D\alpha/\operatorname{sh}\alpha, \qquad \tilde{A} = 2(D/H^2)\alpha \operatorname{th}(\alpha/2), \qquad \alpha = \sqrt{AH^2/D} > 1.$$

Простое усреднение не учитывает большого шага и дает в данном случае тривиальный результат. Кстати (предоставим убедиться в этом читателю), результат гомогенизации на базе точной схемы

зависит и от вида схемы (например, можно записать член Au в виде $(\tilde{A}/6)(u_{m-1}+4u_m+u_{m+1})$; тогда будут другие значения $\tilde{D}$, $\tilde{A}$. К сожалению, такую схему гомогенизации можно реализовать только в не очень-то интересном, одномерном, случае. Однако некоторые высказанные выше соображения можно в какой-то мере использовать и в более реальных двумерных и трехмерных задачах.

Двумерный метод конечных суперэлементов. Имея в виду расчет задачи о трещине гидроразрыва и другие аналогичные задачи, рассмотрим в качестве суперэлемента квадрат $H\times H$, из которого исключен расположенный в центре круг радиусом $r\ll H$ («скважина»). Занумеруем вершины квадрата так, как показано на рис. 60. Границу квадрата разделим на четыре части σ_j, каждую из которых ассоциируем с соответствующей вершиной. Оснастим такой элемент базисом из пяти функций $\varphi_i(x, y)$ $(i=0, 1, \ldots, 4)$. Функцию $\varphi_1(x, y)$ определим как решение уравнения Лапласа $\Delta\varphi=0$ при следующих краевых условиях. Положим в первой вершине элемента $\varphi=1$, в остальных — $\varphi=0$. С вершин на грани элемента значения φ проинтерполируем линейно. На внутренней границе $x^2+y^2=r^2$ (на скважине) положим $\varphi=0$.

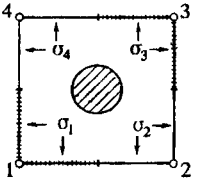


Рис. 60

Аналогичным образом определим базисные функции φ_2, φ_3, φ_4. Что касается φ_0, то она определяется решением того же уравнения Лапласа при нулевых значениях на внешней границе квадрата и при значении $\varphi=1$ на границе скважины. Таким базисом оснащается элемент первого типа. Элемент второго типа — обычная однородная ячейка $H\times H$, в которой подобная процедура построения базиса приводит к элементарным билинейным функциям типа

$$\varphi(x, y)=1-x/H-y/H+xy/H^2.$$

Именно таким базисом оснащается элемент (ячейка счетной сетки) в стандартном методе конечных элементов.

Расчетная область покрывается сеткой с шагом H, в каждую ее ячейку помещается элемент первого или второго типа в зависимости от того, содержится в ней скважина или нет. В узлах сетки $H\times H$ определяется сеточная функция $p_{k,m}$. Внутрь ячейки, содержащей скважину с заданным давлением P, функция p интерполируется с помощью базиса элемента данного типа:

$$p(x, y)=P\,\varphi_0(x, y)+\sum_{i=1}^{4} p_i\,\varphi_i(x, y), \qquad (9)$$

где p_i — значения $p_{k,m}$, но иначе занумерованные ($p_1 = p_{k,m}$, $p_2 = p_{k+1,m}$ и т.д.). В (9) ради простоты опущен индекс у базисных функций — номер типа суперэлемента, помещенного в данную ячейку. Конечно, эта формула действует лишь внутри ячейки $(k + 1/2, m + 1/2)$, т.е. при $x \in (x_k, x_{k+1})$, $y \in (y_m, y_{m+1})$.

Функция (9) всюду, кроме координатных линий сетки, удовлетворяет уравнению $\Delta p = 0$, на границах скважин она принимает заданные значения. Здесь мы, конечно, отвлекаемся от того, что вычисление базисных функций осуществляется на специальной неравномерной сетке, шаг которой вблизи скважины существенно меньше r. Кроме того, если требуется решать неоднородное уравнение (2), это делается аналогично одномерному случаю. Для того чтобы функция всюду была гармонической, нужно потребовать условий сшивки на координатных линиях сетки. Такие условия, как известно, состоят из требований непрерывности самой функции (это требование автоматически выполняется конструкцией (9) при любых $p_{k,m}$) и непрерывности нормальных к линиям сетки производных. Последнее требование, разумеется, точно выполнить невозможно, распоряжаясь лишь значениями $p_{k,m}$). Однако можно потребовать выполнения его в некотором «слабом» смысле.

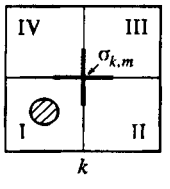


Рис. 61

С каждым внутренним узлом сетки (k, m) свяжем специальную область $\sigma_{k,m}$, имеющую нулевую площадь. Эта область имеет форму креста (рис. 61), но каждый его луч имеет две стороны, обращенные в соседние ячейки. Функцию $p(x, y)$ вида (9) будем считать приближенным решением, если для всех внутренних узлов (k, m) выполняется условие равенства нулю потока в области $\sigma_{k,m}$. Введя обозначения $\partial\sigma_{k,m}$ для двусторонней крестообразной линии (границы $\sigma_{k,m}$) и n для направления нормали к $\partial\sigma_{k,m}$, условие равенства потока в $\sigma_{k,m}$ запишем в виде

$$\oint\limits_{\partial\sigma_{k,m}} D \frac{\partial p}{\partial n}\, ds = 0. \tag{10}$$

Можно показать, что такое же условие (10) получается, если проинтегрировать уравнение $\operatorname{div} D \operatorname{grad} p = 0$ по квадрату $H \times H$ с центром в узле (k, m), исключив из него, разумеется, скважину, если она имеется в какой-то из примыкающих к узлу ячеек.

Уравнение (10) следует превратить в разностное уравнение, связывающее девять значений $p_{k+\alpha, m+\beta}$, где $\alpha, \beta \in [-1, 1]$. Для этого при расчете базисных функций вычислим и сохраним значения потоков через элементарные контуры σ_j, показанные на рис. 60:

$$\Pi^t_{i,j} = \int\limits_{\sigma_j} D^t \frac{\partial \varphi^t_i}{\partial n}\, ds, \quad i = 0, 1, \ldots, 4, \quad j = 1, 2, \ldots, 4.$$

Напомним смысл индексов: t — номер типа ячейки, i — номер базисной функции (i-я функция соответствует i-й вершине, в которой i-я функция равна единице), j — номер участка границы. Если коэффициент диффузии в ячейке постоянен (или очень мало меняется, т.е. в расчетной схеме полагается постоянным), то потоки вычисляются для решений уравнения $\Delta\varphi = 0$ и затем умножаются на D. Если D в ячейке имеет сложный характер, базисная функция ищется решением уравнения $\operatorname{div} D \operatorname{grad} \varphi = 0$.

Теперь для представления явного вида разностной схемы нужно рассмотреть все точки $(k+\alpha, m+\beta)$, каждая из которых дает свой вклад в интеграл (10). Выпишем соответствующую формулу, полагая, что читатель без труда поймет принцип ее образования:

$$p_{k-1,m-1}\Pi^1_{1,3} + p_{k,m-1}(\Pi^1_{2,3} + \Pi^2_{1,4}) + p_{k+1,m-1}\Pi^2_{2,4} + \\ + p_{k-1,m}(\Pi^1_{4,3} + \Pi^4_{1,2}) + p_{k+1,m}(\Pi^2_{3,4} + \Pi^3_{2,1}) + \\ + p_{k,m}(\Pi^1_{3,3} + \Pi^2_{4,4} + \Pi^3_{1,1} + \Pi^4_{2,2}) + p_{k-1,m+1}\Pi^4_{4,2} + \\ + p_{k,m+1}(\Pi^4_{3,2} + \Pi^3_{4,1}) + p_{k+1,m+1}\Pi^3_{3,1} + P^1\Pi^1_{0,3} = 0, \quad (11)$$

где верхние индексы у Π есть номера типов суперэлементов, помещенных в ячейки, примыкающие к узлу (k, m). В (11) мы ограничились случаем, когда только в ячейке 1 имеется скважина с заданным на ней давлением P^1.

Если попытаться сформулировать класс задач, в которых методы расчета с «большим шагом» окажутся достаточно эффективными, то, видимо, следует принять важное предположение: на координатных линиях сетки с шагом H искомая функция хорошо аппроксимируется с помощью линейной интерполяции. Успех вероятен в том случае, когда дифференциальные свойства искомой функции $p(x, y)$ и ее ограничения на координатные линии H-сетки существенно различны. Для функции, рассматриваемой во всей двумерной области, шаг H очень велик, ограничение ее на H-сетку пропускает существенные детали (в данном случае, сильные деформации решения в малой окрестности скважины). Но функция, рассматриваемая лишь на координатных линиях, может быть достаточно гладкой, и значения ее в узлах (k, m) дают весьма полное представление о поведении функции на линиях сетки, но не во всей плоскости.

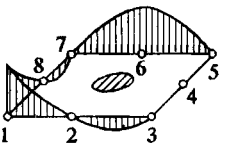


Рис. 62

Метод конечных суперэлементов второго порядка. Описанную выше конструкцию естественно называть схемой «первого порядка», учитывая линейность базиса на линиях сетки. Можно уточнить ее, построив в том же духе схему «второго порядка», с квадратичным базисом. В этом случае вводятся еще четыре счетные точки — в серединах сторон ячейки (рис. 62). Теперь уже будет во-

семь базисных функций, которые определяются специальным выбором значений φ на внешней границе ячейки, а внутрь продолжаются, так же как и раньше, решением уравнения Лапласа с учетом условий на скважине (если она есть). Краевые условия, например для φ_1 и φ_2, формулируются так ($\xi = x/H$, $\eta = y/H$):

$$\varphi_1(x, 0) = 1 - 3\xi + 2\xi^2, \quad \varphi_1(0, y) = 1 - 3\eta + 2\eta^2,$$

$$\varphi_1 \equiv 0 \text{ на остальных гранях,}$$

$$\varphi_2(x, 0) = 4\xi(1 - \xi), \quad \varphi_2 = 0 \text{ на остальных гранях.}$$

На рис. 62 показаны граничные значения φ_1 и φ_6. С каждой счетной точкой на границе ячейки нужно связать свою часть ее контура σ_i. Теперь, кроме «целых» точек (k, m), появляются «полуцелые» точки $(k, m + 1/2)$, $(k + 1/2, m)$ и соответствующие двусторонние контуры $\sigma_{k,m}$, $\sigma_{k+1/2,m}$, $\sigma_{k,m+1/2}$ (рис. 63). Опыт применения схемы второго порядка показал заметное повышение точности расчета. Но это связано и с заметным возрастанием сложности схемы: в целой точке она 21-точечная, в полуцелой — 13-точечная. При выборе элементарных контуров σ и двусторонних контуров $\sigma_{k,m}, \ldots$, как показал опыт, следует придерживаться естественного правила: контуры $\sigma_{k,m}$, $\sigma_{k+1/2,m}$, $\sigma_{k,m+1/2}$ не должны перекрываться и должны покрывать все координатные линии сетки (не допуская «пустот»).

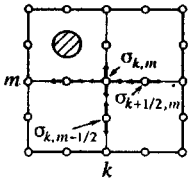


Рис. 63

Заметим, наконец, что, кроме функционалов, используемых при составлении разностной схемы, часто возникает необходимость при подготовке исходной информации вычислить и сохранить некоторые дополнительные функционалы, необходимые для содержательной интерпретации полученного сеточного решения. В задаче о трещине гидроразрыва таким функционалом является поток в скважину:

$$\Pi_{i,0} = \oint_{\partial\gamma} \frac{\partial \varphi_i}{\partial n}\, ds, \quad i = 0, 1, \ldots, 4.$$

Имея такие функционалы, поток в скважину на полученном решении можно вычислить по очевидной формуле

$$D_{k+1/2,\, m+1/2} \{p_{k,m} \Pi_{1,0} + p_{k+1,m} \Pi_{2,0} + p_{k+1,m+1} \Pi_{3,0} + p_{k,m+1} \Pi_{4,0} + P \Pi_{0,0}\},$$

где P — давление на скважине в ячейке $(k + 1/2, m + 1/2)$ $D_{k+1/2,\, m+1/2}$ — коэффициент диффузии в ней.

Решение вырожденного уравнения диффузии. Вернемся к уравнению (2) для давления, причем $u(x, y)$ считается известной функцией. Реально эта функция известна в узлах сетки, т.е. в виде сеточной функции $u_{k,m}$, полученной как «грубое решение» (см. § 30) при расчете раскрытия трещины. Кроме того, известно, что около внешней границы u представляется в виде

$$u(\xi, \eta) = N(\xi)\sqrt{\eta} + O(\eta^{3/2}), \tag{12}$$

где ξ — длина дуги на ∂G, η — расстояние от ∂G по нормали. При столь сильном вырождении уравнения (2) «краевое условие» на ∂G ставится в очень неопределенной форме — в виде требования ограниченности решения. Этим «условием» надо замкнуть систему разностных уравнений. Ниже в общих чертах мы опишем численную реализацию такого замыкания.

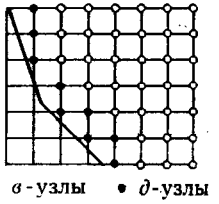


Рис. 64

Начнем с некоторых терминов. Пусть контур ∂G задан, как и в § 30, набором точек $\{X_i, Y_i\}$, соединяемых отрезками прямых. Плоскость (x, y) покрыта сеткой узлов (k, m). Точки, попавшие внутрь G, называются счетными. В них определена сеточная функция $p_{k,m}$ (и $u_{k,m}$, используемая для вычисления коэффициента диффузии, который полагался постоянным в пределах ячейки; эта постоянная величина вычислялась усреднением u по четырем вершинам ячейки). Разделим счетные узлы на два типа (рис. 64). Узлы $в$-типа («внутренние») — это узлы (k, m), у которых все восемь соседних узлов являются счетными. В этих узлах используется стандартная разностная аппроксимация на 9-точечном шаблоне. Остальные счетные узлы суть узлы $д$-типа («дефектные»). В этих узлах неприменимо стандартное разностное уравнение и нет привычного краевого условия, с помощью которого в невырожденных задачах записываются разностные уравнения в узлах $д$-типа.

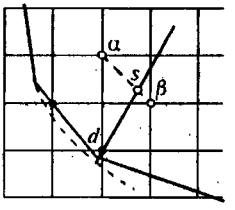


Рис. 65

Для того чтобы разобраться в том, как следует поступать в таком случае, рассмотрим решение в окрестности границы, вводя местную систему координат (рис. 65). В этих переменных уравнение можно записать в форме

$$\frac{\partial}{\partial\eta}\left(N^3\,\eta^{3/2}\frac{\partial p}{\partial\eta}\right) + \frac{\partial}{\partial\xi}\left(N^3\,\eta^{3/2}\frac{\partial p}{\partial\xi}\right) = Q + \ldots \tag{13}$$

Здесь мы используем только главные члены асимптотики (12), при этом Q можно считать постоянной величиной. Преобразуем

выражение (13) к виду:

$$N^3 \eta^{3/2} \frac{\partial^2 p}{\partial \eta^2} + \frac{3}{2} N^3 \eta^{1/2} \frac{\partial p}{\partial \eta} + N^3 \eta^{3/2} \frac{\partial^2 p}{\partial \xi^2} = Q + \ldots$$

Считая величину $\eta^{3/2} \partial^2 p / \partial \xi^2$ пренебрежимо малой по сравнению с остальными, рассмотрим асимптотическое уравнение

$$\eta^{3/2} \frac{\partial^2 p}{\partial \eta^2} + \frac{3}{2} \eta^{1/2} \frac{\partial p}{\partial \eta} = q = \frac{Q}{N^3}.$$

Оно легко интегрируется. Его общее решение есть

$$p(\eta) = C_1 + 2C_0 \eta^{-1/2} + 2q\eta^{1/2},$$

где C_0, C_1 — произвольные постоянные.

Очевидно, ограниченное решение можно получить только при $C_0 = 0$. (Именно это сокращение числа произвольных постоянных в общем решении уравнения второго порядка и является причиной того, что нельзя ставить классическое краевое условие при $\eta = 0$, т.е. на границе области G.) Итак, мы получили асимптотику решения около границы:

$$p(\xi, \eta) = C(\xi) + 2q(\xi)\eta^{1/2} + o(\eta^{1/2}). \qquad (14)$$

Подставляя в (13) такое решение (предположим, что $C(\xi)$, $q(\xi)$ — гладкие функции), можно оценить отброшенные выше члены и оправдать эти действия. В формуле (14) функция $C(\xi)$, конечно, неизвестна, она определяется некоторой процедурой «сшивки» с решением внутри области.

Алгоритмическая реализация асимптотики состоит в следующем. Рассмотрим некоторую точку ∂-типа (точка d на рис. 65). Пусть (X_i, Y_i) — ближайшая к ней вершина границы. Через вершины с номерами $i-1$, i, $i+1$ проведем окружность, которую будем считать участком границы ∂G. Через точку ∂-типа проведем нормаль к ∂G, пересекающую вертикальные и горизонтальные границы ячеек или их диагонали. Среди этих отрезков выберем ближайший к точке ∂-типа, такой, что оба его конца суть точки a-типа (на рис. 65 они обозначены α и β). Пересечение нормали с отрезком обозначим буквой s.

Значение p в точке s можно проинтерполировать по значениям в точках α и β:

$$p_s = wp_\alpha + (1-w)\, p_\beta, \quad w \in [0, 1].$$

Используя асимптотику (14) запишем соотношения для p_d, p_s:

$$p_d = C + 2q\sqrt{\eta_d}, \qquad p_s = C + 2q\sqrt{\eta_s}.$$

Исключая из них C, получаем связь

$$p_d = wp_\alpha + (1-w)\, p_\beta + \frac{2Q}{N^3}(\sqrt{\eta_s} - \sqrt{\eta_d}). \qquad (15)$$

Формулу (15) в итерационном процессе решения системы разностных уравнений можно использовать двумя способами. Первый способ состоит в следующем. Пусть получено некоторое приближение p во внутренних узлах $в$-типа. Используя (15), доопределим его в точках $д$-типа. Таким образом, значения $p_{k,m}$ известны во всех счетных точках. Теперь по стандартной формуле простейшего метода релаксации новые значения $p_{k,m}$ можно получить во всех внутренних точках. Далее процесс повторяется.

Уточним некоторые детали. Представим разностную схему в форме

$$\sum_{i=-1}^{1}\sum_{j=-1}^{1} C_{k,m}^{i,j}\, p_{k+i,m+j} + c_{k,m} = 0. \tag{16}$$

Метод релаксации с ускорением состоит в следующем. В каждой внутренней точке (k, m) поочередно пересчитывается значение

$$\tilde{p}_{k,m} = -\Big(c_{k,m} + \sum_{i,j=-1}^{1}{}^{*}\, C_{k,m}^{i,j}\, p_{k+i,m+j}\Big)\Big/C_{k,m}^{0,0} \tag{17}$$

(звездочка отмечает пропуск слагаемого $i = j = 0$). Это предварительное значение.

Окончательное значение есть

$$p_{k,m} := p_{k,m} + \omega\,(\tilde{p}_{k,m} - p_{k,m}), \tag{18}$$

где параметр ω, ускоряющий сходимость, подбирается экспериментально. Теория уравнения Лапласа с постоянными коэффициентами указывает диапазон оптимального значения: $\omega \approx 1.7 \div 1.8$. Найденное по (18) значение $p_{k,m}$ сразу записывается в массив p, так что формула (18) является «полунеявной»: часть входящих в сумму значений $p_{k+i,m+j}$ относится к ν-й итерации, часть — к $(\nu+1)$-й. Вычислительный эксперимент показал слабую расходимость этого итерационного процесса.

Второй способ состоит в том, что связь (15) заранее вносится в разностные уравнения. Если точка $д$-типа d входит в шаблон в точке (k, m), причем соответствующие точки α и β тоже входят в этот шаблон, то точка $д$-типа исключается из схемы. Выражение (15) подставляется в (16), пересчитываются коэффициенты схемы, соответствующие точкам α и β, коэффициент, соответствующий точке d, делается нулевым. Далее итерационный процесс выполняется точно так же, как было описано выше, но значения в точках $д$-типа в расчете фактически не участвуют. После выполнения итерации формула (15) используется для расчета значений p в точках $д$-типа. Редко, но все же встречаются ситуации, когда хотя бы одна из точек α и β не входит в шаблон в точке (k, m). Тогда операция исключения дефектной точки из схемы не производится и работает первый способ учета асимптотики (15).

Итак, алгоритм построен. Он основан не на точной теории, а на соображениях, привлекаемых из близких ситуаций (в которых эти соображения являются результатом достаточно аккуратной теории). Такой способ действия характерен для вычислительной физики. В ней редко встречаются чистые, укладывающиеся в уже готовую теорию задачи. Специалист по вычислительной физике обычно начинает построение алгоритма «по аналогии» с тем, что он уже знает, и начинает именно с практического использования своего алгоритма, а не с развития соответствующей ему теории. Это понятно: ведь алгоритм может оказаться неудачным и стоит ли тогда строить теорию? Если же он оказался удачным, наступает время решать прикладные задачи, а теория может и подождать.

В данном случае, конечно, в первую очередь хочется понять, удачен ли алгоритм или надо в нем что-то менять. Естественным средством проверки алгоритма является решение задачи, имеющей точное решение и содержащей характерные для данного случая трудности. Такая задача может быть построена. В круге единичного радиуса в центре помещена скважина малого радиуса. Коэффициент диффузии берется в виде $(1-r)^{3/2}$. Поскольку задача цилиндрически-симметрична (ее решение зависит только от r), уравнение диффузии становится обыкновенным уравнением:

$$\frac{1}{r}\frac{d}{dr}\left[r(1-r)^{3/2}\frac{dp}{dr}\right] = f = \text{const}, \quad r \in [0, 1].$$

Оно элементарно интегрируется:

$$p(r) = f\sqrt{1-r} + \frac{f+2C_1}{\sqrt{1-r}} + C_1 \ln\frac{1-\sqrt{1-r}}{1+\sqrt{1+r}} + C_2,$$

где C_1, C_2 — произвольные постоянные.

Ограниченное решение находим при $C_1 = -f/2$. Постоянная C_2 не существенна, положим $C_2 = 0$. Это решение имеет корневую особенность на внешней границе ($r = 1$) и логарифмическую — в центре. Преобразованием подобия получим решение уравнения (2) в области $\rho \leqslant r(x, y) \leqslant R$, где $\rho = 0.1$, $R = 250$. Постоянную f легко подобрать так, чтобы выполнялось внутреннее краевое условие: $p = 1.2$ на границе скважины. Задача решалась с помощью метода конечных суперэлементов на квадратной сетке с шагом $H = 17$. Число счетных узлов было около 650. Заметим, что граница исходной области (окружность $r = R$) аппроксимируется контуром $\{X_i, Y_i\}$ достаточно аккуратно, хотя область определения счетных величин $p_{k,m}$ (множество счетных узлов) аппроксимирует круг $r \leqslant R$ очень грубо.

Вычислительный эксперимент имел целью выяснить два обстоятельства: как сходятся итерации и какова точность разностного решения? Сходимость итерационного процесса (использовался второй способ) иллюстрирует табл. 24, в которой представлены: ν — число итераций, ε — невязка (максимум по (k, m) модуля левой части (16)),

ω — значение параметра релаксации в (18). Итерации начинались с $p_{k,m} = 0$, значения $c_{k,m} \approx 4 \div 5$. Из таблицы видно, что при $\omega = 1$ невязка быстро достигает малых значений, затем сходимость становится очень медленной: за первые 200 итераций невязка уменьшается почти в 10^4 раз, в дальнейшем за 200 итераций она уменьшается примерно вдвое.

Таблица 24

ω	ν	ε	p'	p''
1	200	$6.5 \cdot 10^{-4}$	0.568	−0.081
1	400	$3.1 \cdot 10^{-4}$	0.626	0.018
1	700	$1.0 \cdot 10^{-4}$	0.661	0.079
1.5	400	$1.8 \cdot 10^{-5}$	0.675	0.103
1.7	200	$2.8 \cdot 10^{-5}$	0.675	0.101
1.75	400	$4.6 \cdot 10^{-8}$	0.678	0.107
1.8	400	$1.6 \cdot 10^{-9}$	0.678	0.107

Как расценить этот результат? С одной стороны, погрешность в правой части (порядка 0.01 %), кажется, не требует существенного улучшения, с другой стороны, медленная сходимость внушает какую-то тревогу. Обычно она свидетельствует о том, что оператор (здесь $\operatorname{div} u^3 \operatorname{grad}$) имеет собственное значение, близкое к нулю. В § 14 специально отмечалось, что связь между невязкой и погрешностью определяется минимальным собственным значением: погрешность есть величина порядка невязки, деленной на $|\lambda_{\min}|$.

В рассматриваемой задаче есть основания подозревать наличие очень малого собственного значения. В самом деле, легко угадать функцию, которая является «почти собственной» с собственным числом $\lambda = 0$. Это есть функция $p(x, y) \equiv 1$. Она удовлетворяет «почти всем уравнениям» $\operatorname{div}(u^3 \operatorname{grad} p) = 0$. Не выполнено только однородное краевое условие на скважине: $p = 0$ на ∂g. Конечно, $\lambda = 0$ не является собственным значением, но приведенное выше рассуждение служит основанием ожидать близкого к нулю собственного значения, тем более близкого, чем меньше радиус скважины.

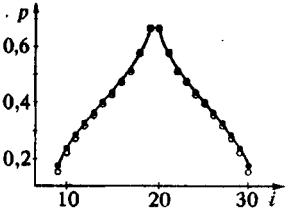


Рис. 66

Действительно, результаты, приведенные в табл. 24, показывают, что дальнейшее уменьшение невязки (казалось бы, излишнее) сопровождается заметным изменением важных величин p', p''. Это значения $p_{k,m}$, взятые на луче, выходящем из центра скважины по диагонали сетки: p' — первое значение p на луче (на расстоянии $H/\sqrt{2}$ от центра скважины), p'' — последнее, почти граничное значение p на луче. Видно, что уточнение в процессе итераций решения системы разностных уравнений до значений $\varepsilon \approx 10^{-6}$ имело смысл.

Таблица 24 дает представление и о точности самой разностной схемы (при шаге $H = 17$). Значения точного решения в соответст-

вующих точках суть 0.680 и 0.124. На рис. 66 представлено точное решение в сечении по линии $x = y$, проходящей через центр скважины. Кстати, значение p на скважине есть 1.2. Приближенное решение отмечено кружками.

Расчет трещины гидроразрыва. Метод Пикара. Расчет трещины гидроразрыва привел к системе двух уравнений, содержащих неизвестные u и p, причем уравнение (1) при известном p решается методом, описанным в § 30, уравнение (2) при известном u решается так, как было описано выше. Это типичная ситуация, в которой естественно начинать работу с самого простого итерационного метода — метода Пикара. В данном случае он оказался удовлетворительно сходящимся. В основных чертах стандартная итерация состоит из следующих операций.

0. Пусть имеется некоторое приближение u, p.

1. Фиксируя p, решаем уравнение (1), причем находится как грубое решение, так и уточненные на локальных сетках решения, позволяющие достаточно аккуратно оценивать коэффициент концентрации напряжений на контуре $N(\xi)$. Он используется в дальнейшем при аппроксимации уравнения (2) вблизи контура.

2. Рассчитываем коэффициенты схемы метода конечных суперэлементов (16): $C^{i,j}_{k,m}$, $c_{k,m}$, $\forall\, k, m$, $i, j = -1, 1$. Эти десять двумерных массивов хранятся в памяти ЭВМ: расчет коэффициентов достаточно сложен, чтобы его производить заново при каждом обращении к точке (k, m). (В методе конечных разностей коэффициенты схемы часто вычисляются так просто, что их не имеет смысла вычислять заранее и хранить.)

3. С учетом асимптотики (14) в соответствии с формулой (15) корректируем коэффициенты схемы в тех внутренних узлах сетки, в которых шаблон 9-точечной схемы включает дефектные счетные точки.

4. Решаем уравнение (2) для p при фиксированном u. Далее процесс повторяется до стабилизации результата.

Под решением уравнений (1), (2) выше понимается некоторое число итераций, причем в качестве начального приближения, естественно, берутся имеющиеся к этому моменту приближенные значения u, p. О достигнутой точности можно судить по невязкам в уравнениях (1), (2). Следует только подчеркнуть, что эти невязки вычисляются дважды: на входе в итерационный u- и p-процессы и на их выходе. Невязки на выходе позволяют контролировать сходимость внутренних итерационных процессов, но ничего не говорят о сходимости внешнего, полного, итерационного процесса. О ней информацию дают именно невязки, вычисленные на входах в каждый внутренний процесс. Если они достаточно малы, это свидетельствует о том, что полученное приближение хорошо удовлетворяет совместной системе уравнений (1), (2).

СПИСОК ЛИТЕРАТУРЫ

1. *Абрамов А. А.* Вариант метода прогонки //ЖВМиМФ. 1961. Т. 1, № 2

2. *Абрамов А. А.* О переносе граничных условий для систем линейных обыкновенных уравнений //ЖВМиМФ. 1961. Т. 1, № 3

3. *Азатян В. В., Коган А. М., Нейгауз М. Г.* Роль самовозгорания при горении водорода вблизи предела воспламенения //Кинетика и катализ. 1973. Т. XVI, Вып. 3.

4. *Алалыкин Г. Б., Годунов С. К., Кireева И. Л., Плинер Л. А.* Решение одномерных задач газовой динамики. — М.: Наука, 1970

5. *Альберг Дж., Нельсон Э., Уолш Дж.* Теория сплайнов и ее приложения. — М.: Мир, 1972

6. *Алексеев В. М., Тихомиров В. М.* Оптимальное управление. — М.: Наука, 1979

7. *Арнольд В. И.* Математические методы классической механики — М.: Наука, 1974

8. *Астраханцев Г. П.* Об одном итерационном методе //ЖВМиМФ. 1971. Т. 11, № 2

9. *Бабенко К. И.* Основы численного анализа. — М.: Наука, 1975

10. *Бабенко К. И., Воскресенский Г. П., Любимов А. Н., Русанов В. В.* Пространственное обтекание гладких тел идеальным газом. — М.: Наука, 1964

11. *Бабенко К. И.* (ред.) Теоретические основы и конструирование алгоритмов решения задач математической физики /Анучина Н. Н., Бабенко К. И., Годунов С. К. и др. — М.: Наука, 1979

12. *Бабушка И., Соболев С. Л.* Оптимизация численных методов //Aplicace Matem. Svazee. 1965. № 10. С. 96

13. *Багриновский К. А., Годунов С. К.* Разностные схемы для многомерных задач //ДАН СССР. 1957. Т. 115, № 3

14. *Байдин Г. В., Федоренко Р. П.* О приближенном решении некоторых негладких вариационных задач. — Препринт ИПМ им. М. В. Келдыша. 1985, № 76

15. *Байдин Г. В., Федоренко Р. П.* Опыт приближенного решения задач о стационарном течении вязкопластической среды. — Препринт ИПМ им. М. В. Келдыша. 1985, № 145

16. *Баничук Н. В.* Введение в оптимизацию конструкций. — М.: Наука, 1986

17. *Бахвалов Н. С.* О сходимости одного релаксационного метода //ЖВМиМФ. 1966. Т. 6, № 5

18. *Бахвалов Н. С.* Численные методы. — М.: Наука, 1975

19. *Бахвалов Н. С., Жидков Н. П., Кобельков Г. М.* Численные методы. — М.: Наука, 1987

20. *Беллман Р., Калаба Р.* Квазилинеаризация и нелинейные уравнения. — М.: Мир, 1968

21. *Белоцерковский О. М.* Численное моделирование в механике сплошных сред. — М.: Наука, 1984

22. *Белоцерковский О. М., Давыдов Ю. М.* Метод крупных частиц. — М.: Наука, 1982

23. *Блехер П. М., Турчанинов В. И.* Построение собственных функций двумерного оператора Шредингера с периодическим потенциалом. — Препринт ИПМ АН СССР. 1978, № 1

24. *Боголюбов Н. Н., Митропольский Ю. А.* Асимптотические методы в теории нелинейных колебаний. — М.: Физматгиз, 1955

25. *Болтянский В. Г.* Математические методы оптимального управления. — М.: Наука, 1969

26. *Бутковский А. Г.* Методы управления системами с распределенными параметрами. — М.: Наука, 1975

27. *Борис Дж. П., Бук Д. Л.* (*Boris J. P., Book D. L.*) Flux-corrected transport I: SHASTA — a fluid transport algorithm that works //J. Comp. Phys. 1973. 11. P. 38

28. *О'Брайен Г., Хайман М. А., Каплан С.* (*O'Brien G. G., Hyman M. A., Kaplan S.*) A study of numerical solution of partial differential equations //J. Math. Phys. 1951. 29B. P. 4

29. *Бригс Л. Л., Льюис Е. Е.* (*Briggs L. L., Lewis E. E.*) A two-dimensional constrained method for nonuniform latice problems //Nucl. Scl. Eng. 1980. V. 75. P. 76

30. *Вазов В., Форсайт Дж.* Разностные методы решения дифференциальных уравнений в частных производных. — М.: ИЛ, 1963

31. *Васильев Ф. П.* Методы решения экстремальных задач. — М.: Наука, 1981

32. *Васильева А. Б., Бутусов В. Ф.* Асимптотическое разложение решений сингулярно-возмущенных уравнений. — М.: Наука, 1977

33. *Витушкин А. Г.* Оценки сложности задачи табулирования. — М.: Физматгиз, 1959

34. Вычислительные методы в гидродинамике /Под ред. Б. Олдера, С. Фернбаха, М. Ротенберга. — М.: Мир, 1967

35. *Вашпресс Е. Л.* (*Wachspreess E. L.*) Iterative solution of elleptic systems and application to the neutron diffusion equation of reactor physic. — N. Y.: Prentice-Hall, Inc., 1966

36. *Вашпресс Е. Л.* (*Wachspress E. L.*) Optimal alternating-direction-implicite iteration parameters //SIAM J. 1962. V. 10, No. 2

37. *Гельфанд И. М., Зуева Н. М., Локуциевский О. В. и др.* К теории нелинейных колебаний электронной плазмы //ЖВМиМФ. 1967. Т. 7, № 2

38. *Герасимов Б. П., Семушин С. А.* Анализ некоторых численных методов газовой динамики на неподвижных эйлеровых сетках. — Препринт ИПМ им. Келдыша. 1983, № 38

39. *Герасимов Б. П., Семушин С. А.* Расчет на неподвижной эйлеровой сетке обтекания тел изменяющейся формы //Дифференциальные уравнения. 1981. Т. 17, № 7

40. *Годунов С. К., Забродин А. В., Иванов М. Я. и др.* Численное решение многомерных задач газовой динамики /Под ред. С. К. Годунова. — М.: Наука, 1976

41. *Годунов С. К., Прокопов Г. П.* О расчетах конформных отображений и построении разностных сеток //ЖВМиМФ. 1967. Т. 7, № 5

42. *Годунов С. К., Прокопов Г. П.* Об использовании подвижных сеток в газодинамических расчетах //ЖВМиМФ. 1972. Т. 12, № 2

43. *Годунов С. К., Рябенький В. С.* Введение в теорию разностных схем. — М.: Физматгиз, 1962

44. *Годунов С. К., Рябенький В. С.* Разностные схемы. — М.: Наука, 1973

45. *Гольдин В. Я., Калиткин Н. Н., Шишова Т. В.* Нелинейные разностные схемы для гиперболических уравнений //ЖВМиМФ. 1965. Т. 5, № 5

46. *Гольдштейн Р. В., Зазовский А. Ф., Спектор А. А., Федоренко Р. П.* Решение вариационными методами пространственных контактных задач качения //Успехи механики. 1982. Т. 5, № 2/3

47. *Гольдштейн Р. В., Отрощенко И. В., Федоренко Р. П.* Метод уточняющих граничных сеток в задачах о трещинах в упругих телах. — Препринт Ин-та проблем механики АН СССР. 1984, № 230

48. *Гловински Р., Лионс Ж.-Л., Тремольер Р.* Численное исследование вариационных неравенств. — М.: Мир, 1979

49. *Давиденко Д. Ф.* О приложении метода вариации параметра к теории нелинейных функциональных уравнений //Укр. мат. ж. 1953. Т. 7. С. 18

50. *Дегтярев Л. М., Фаворский А. П.* Потоковый вариант метода прогонки //ЖВМиМФ. 1968. Т. 8, № 3

51. *Деккер К., Вервер Я.* Устойчивость методов Рунге–Кутты для жестких нелинейных дифференциальных уравнений. — М.: Мир, 1988

52. *Демьянов В. Ф., Васильев Л. В.* Недифференцируемая оптимизация. — М.: Наука, 1981

53. *Демьянов В. Ф., Малоземов В. Н.* Введение в минимакс. — М.: Наука, 1972

54. *Джексон.* Теория аппроксимации. — М.: ИЛ., 1951

55. *Джордж А., Лю Дж.* Численное решение больших систем уравнений. — М.: Мир, 1984

56. *Дьяконов Е. Г.* Минимизация вычислительной работы. — М.: Наука, 1989

57. *Дьяконов Е. Г., Столяров Н. Н.* О реализации эффективных итерационных методов для разностных задач теории упругости и пластичности //Численные методы решения задач упругости и пластичности. Ч. II. — Новосибирск: ВЦ СО АН СССР, 1978

58. *Дьяченко В. Ф.* Об одном новом методе численного решения нестационарных пространственных задач газовой динамики с двумя пространственными переменными //ЖВМиМФ. 1965. Т. 5, № 5

59. *Дьяченко В. Ф.* Разностные методы на произвольном множестве точек. Препринт ИПМ АН СССР. 1969, № 37

60. *Дюво Г., Лионс Ж.-Л.* Неравенства в механике и физике. — М.: Наука, 1980

61. *Дефлхард П.* (*Deuflhard P.*) A modified Newton method for the solution of ill-conditioned systems of non-linear equations with application to multiple shooting //Num. Math. 1974. V. 22. P. 289

62. *Евтушенко Ю. Г.* Методы решения экстремальных задач и их применение в системах оптимизации. — М.: Наука, 1982

63. *Ермаков С. М.* Метод Монте–Карло и смежные вопросы. — М.: Наука, 1971

64. *Злотник А. А.* О скорости сходимости в W_2 вариационно-разностного метода для эллиптических уравнений //ДАН СССР. 1983. Т. 27, № 4

65. *Злотник А. А.* О скорости сходимости проекционно-разностной схемы с расщепляющимся оператором для гиперболических уравнений //ЖВМиМФ. 1980. Т. 20, № 2

66. *Злотник А. А.* Оценки скорости сходимости проекционно-сеточных методов для гиперболических уравнений второго порядка //Вычислительные процессы и системы. Вып. 8 /Под ред. Г. И. Марчука. — М.: Наука, 1991

67. *Зойтендейк Г.* Методы возможных направлений. — М.: ИЛ, 1963

68. *Зуева Н. М., Соловьев Л. С.* Нелинейная теория газодинамической неустойчивости. — Препринты ИАЭ. 1980, № 3290/1, 3300/1, 3289/1

69. *Иванов В. К.* Обратная задача потенциала для тела, близкого к данному //Изв. АН СССР, Сер. матем. 1956. Т. 20, № 6

70. *Калиткин Н. Н.* Численные методы. — М.: Наука, 1978

71. *Канторович Л. В.* Функциональный анализ и прикладная математика //УМН. 1948. Т. 3, № 6

72. *Канторович Л. В., Крылов В. И.* Приближенные методы высшего анализа. — М.; Л.: Физматгиз, 1962

73. *Колган И. П.* Применение принципа минимальности производных к построению конечно-разностных схем для расчета разрывных решений газовой динамики //Уч. зап. ЦАГИ. 1972. Т. 3, № 6

74. *Колмогоров А. Н.* О сохранении условно-периодических движений при малом возмущении гамильтониана //ДАН СССР. 1954. Т. 98, № 4

75. Крылов Н. М., Боголюбов Н. Н. Новые методы нелинейной механики и их применение к изучению работы электронных генераторов. — М.: ОНТИ, 1934

76. *Курант Р., Фридрихс К., Леви Г.* О разностных уравнениях математической физики //УМН. 1940. Т. 8

77. *Лаврентьев М. М., Романов В. Г., Шишатский С. П.* Некорректные задачи математической физики и анализа. — М.: Наука, 1980

78. *Лебедев В. И., Финогенов С. А.* О порядке выбора итерационных параметров в чебышевском циклическом методе //ЖВМиМФ. 1971. Т. 11, № 2

79. *Либовиц Г.* Разрушение. — М.: Мир, 1973. Т. 1; 1975. Т. 2; 1976. Т. 3; М.: Машиностроение, 1977. Т. 4, 5; 1978. Т. 6, 7

80. *Лидский В. Б., Нейгауз М. Г.* К методу прогонки в случае самосопряженной системы второго порядка //ЖФМиМФ. 1962. Т. 2, № 1

81. *Лионс Ж.-Л.* Некоторые методы решения нелинейных краевых задач. — М.: Мир, 1972

82. *Льюинс Дж.* Ценность. Сопряженная функция. — М.: Атомиздат, 1972

83. *Марчук Г. И.* Методы вычислительной математики. — М.: Наука, 1989

84. *Марчук Г. И.* Методы расщепления. — М.: Наука, 1988

85. *Марчук Г. И., Агошков В. И., Шутяев В. П.* Сопряженные уравнения и методы возмущений в нелинейных задачах математической физики. — М.: Наука, 1993

86. *Марчук Г. И., Лебедев В. И.* Численные методы в теории переноса нейтронов. — М.: Атомиздат, 1971

87. *Михайлов Г. А.* Некоторые вопросы теории методов Монте-Карло. — Новосибирск: Наука, 1974

88. *Мозер Ю.* Лекции о гамильтоновых системах. — М.: Мир, 1973

89. *Молчанов А. М.* Об устойчивости нелинейных систем. Дисс. д.ф.-м. наук. Матем. ин-т АН СССР, 1963

90. *Мосолов П. П., Мясников В. П.* Вариационные методы в теории течений вязкопластической среды //Прикл. мат. и мех. 1965. Т. 29, Вып. 3

91. *Минорский Н.* (*Minorsky N.*) Nonlinear oscilations. — N. Y.: Princeton, 1962

92. *Натансон И. П.* Конструктивная теория функций. — М.; Л.: Физматгиз, 1949

93. *Неймарк Ю. И.* Метод точечных отображений в теории нелинейных колебаний. — М.: Наука, 1972

94. *Никифоров А. Ф.* Численные методы решения некоторых задач квантовой механики. — М.: Изд-во Моск. ун-та, 1981

95. *Овсянников Л. В.* Лекции по основам газовой динамики. — М.: Наука, 1981

96. *Оран Э., Борис Дж.* Численное моделирование реагирующих потоков. — М.: Мир, 1990

97. *Ортега Дж., Рейнболдт В.* Итерационные методы решения нелинейных систем уравнений со многими неизвестными. — М.: Мир, 1975

98. *Отрощенко И. В., Федоренко Р. П.* Итерационное решение бигармонического уравнения //ЖФМиМФ. 1983. Т. 23, № 4

99. *Павлов Б. В., Повзнер А. Я.* Об одном методе численного интегрирования систем обыкновенных дифференциальных уравнений //ЖВМиМФ. 1973. Т. 13, № 5

100. *Пасконов В. М., Полежаев В. И., Чудов Л. А.* Численное моделирование процессов тепло- и массообмена. — М.: Наука, 1984

101. *Поляк Б. Т.* Введение в оптимизацию. — М.: Наука, 1983

102. *Понтрягин Л. С., Болтянский В. Г., Гамкрелидзе Р. В.,* Мищенко Е. Ф. Математическая теория оптимальных процессов. — М.: Физматгиз, 1969

103. *Попов В. С., Федоренко Р. П.* О стандартной программе решения задач оптимального управления. — Препринт ИПМ им. М. В. Келдыша. 1983, № 100

104. *Пропой А. И.* Элементы теории оптимальных дискретных процессов. — М.: Наука, 1973

105. *Пшеничный Б. Н.* Метод линеаризации. — М.: Наука, 1983

106. *Пшеничный Б. Н., Данилин Ю. М.* Численные методы в экстремальных задачах. — М.: Наука, 1975

107. *Писман Д., Рэчфорд Г.* (*Peaceman D. W., Rachford H. H.*) The numerical solution of parabolic and elliptic differential equations //SIAM J. 1955. V. 3, № 1.

108. *Ракитский Ю. В., Устинов С. М., Черноруцкий И. Г.* Численные методы решения жестких систем. — М.: Наука, 1979

109. *Растригин Л. А.* Статистические методы поиска. — М.: Наука, 1968

110. *Репях В. С.* Применение одного метода суперэлементов к решению плоской задачи теории упругости //ЖВМиМФ. 1986. Т. 26, № 11

111. *Рихтмайер Р.* Принципы современной математической физики. — М.: Мир, 1982

112. *Рихтмайер Р., Мортон К.* Разностные методы решения краевых задач. — М.: Мир, 1972

113. *Роуч П.* Вычислительная гидродинамика. — М.: Мир, 1980

114. *Рябенький В. С.* Метод разностных потенциалов для некоторых задач механики сплошной среды. — М.: Наука, 1987

115. *Рябенький В. С., Филиппов А. Ф.* Об устойчивости разностных уравнений. — М.: Гостехиздат, 1956

116. *Самарский А. А.* Введение в теорию разностных схем. — М.: Наука, 1971

117. *Самарский А. А.* Теория разностных схем. — М.: Наука, 1977

118. *Самарский А. А., Гулин А. В.* Численные методы. — М.: Наука, 1989

119. *Самарский А. А., Моисеенко Б. А.* Экономическая схема сквозного счета для многомерной задачи Стефана //ЖВМиМФ. 1965. Т. 5, № 5.

120. *Самарский А. А., Николаев Е. С.* Методы решения сеточных уравнений. — М.: Наука, 1978

121. *Самарский А. А., Попов Ю. П.* Разностные схемы газовой динамики. — М.: Наука, 1975

122. *Саульев В. В.* Интегрирование уравнений параболического типа методом сеток. — М.: Физматгиз, 1960

123. *Сигов Ю. С., Ходырев Ю. В.* К теории дискретных моделей плазмы //Численные методы механики сплошной среды. 1976. Т. 7, № 2

124. Соболев С. Л. Введение в теорию кубатурных формул. — М.: Наука, 1974

125. *Соболь И. М.* Численные методы Монте-Карло. — М.: Наука, 1973

126. *Стечкин С. Б., Субботин Ю. Н.* Сплайны в вычислительной математике. — М.: Наука, 1976

127. *Страховская Л. Г., Климов А. Д., Федоренко Р. П.* Метод расчета кинетики импульсного реактора. — Препринт ИПМ АН СССР. 1975, № 42

128. *Страховская Л. Г., Федоренко Р. П.* О методах расчета некоторых квазистационарных режимов работы ядерного реактора //ЖВМиМФ. 1979. Т. 19, № 5

129. *Страховская Л. Г., Федоренко Р. П.* Об одной специальной разностной схеме //Численные методы механики сплошной среды. 1974. Т. 5, № 1

130. *Страховская Л. Г., Федоренко Р. П.* Об одном варианте метода конечных элементов //ЖВМиМФ. 1979. Т. 19, № 4

131. *Стренг Г., Фикс Дж.* Теория метода конечных элементов. — М.: Мир, 1977

132. *Саусвелл Р. В.* (*Southwell R. V.*) Relaxation methods in theoretical physics. — Oxford, 1946. V. 1; 1956. V. 2

133. *Тихонов А. Н.* Об устойчивости обратных задач //ДАН СССР. 1943. Т. 39, № 5

134. *Тихонов А. Н., Арсенин В. Я.* Методы решения некорректных задач. — М.: Наука, 1979

135. *Тихонов А. Н., Арсенин В. Я., Тимонов А. А.* Математические задачи компьютерной томографии. — М.: Наука, 1987

136. *Тихонов А. Н., Гончарский А. В., Степанов В. В., Ягола А. Г.* Регуляризирующие алгоритмы и априорная информация. — М.: Наука, 1983

137. *Тихонов А. Н., Самарский А. А.* Об однородных разностных схемах //ЖВМиМФ. 1961. Т. 1, № 3

138. *Федоренко Р. П.* Вывод и обоснование уравнений в медленном времени //ЖВМиМФ. 1974. Т. 14, № 5

139. *Федоренко Р. П.* Жесткие системы обыкновенных дифференциальных уравнений //Вычислительные процессы и системы. Вып. 8 /Под ред. Г. И. Марчука. — М.: Наука, 1991

140. *Федоренко Р. П.* Итерационное решение разностных эллиптических уравнений //УМН. 1973. Т. 28, Вып. 2

141. *Федоренко Р. П.* Некоторые задачи и приближенные методы вычислительной механики //ЖВМиМФ. 1994. Т. 34, № 2

142. *Федоренко Р. П.* О минимизации негладких функций //ЖВМиМФ. 1981. Т. 21, № 3

143. *Федоренко Р. П.* О регулярных жестких системах обыкновенных дифференциальных уравнений //ДАН СССР. 1983. Т. 273, № 6

144. *Федоренко Р. П.* О скорости сходимости одного итерационного процесса //ЖВМиМФ. 1964. Т. 4, № 3

145. *Федоренко Р. П.* Приближенное решение задач оптимального управления. — М.: Наука, 1978

146. *Федоренко Р. П.* Применение разностных схем высокого порядка точности для гиперболических уравнений //ЖВМиМФ. 1962. Т. 2, № 6

147. *Федоренко Р. П.* Разностная схема для задачи Стефана //ЖВМиМФ. 1975. Т. 15, № 5

148. *Федоренко Р. П.* Разностный метод расчета течений газа в канале произвольной формы //Численные методы механики сплошной среды. 1974. Т. 5, № 1

149. *Федоренко Р. П.* Релаксационный метод решения разностных эллиптических уравнений //ЖВМиМФ. 1961. Т. 1, № 5

150. *Федоренко Р. П.* (*Fedorenko R. P.*) Stiff systems of ordinary differential equation //Numerical methods and applications / Ed. G. Marchuk. — N.Y.: CRC Press, Inc., 1994

151. *Филипп С.* (*Filipi S.*) Altrers und Neues zur Numerishen differentiation //Electronishe Datenverarbeitung. 1966. Bd 2

152. *Хайрер Э., Нерсетт С., Ваннер Г.* Решение обыкновенных дифференциальных уравнений. — М.: Мир, 1990

153. *Харлоу Ф. Х.* Численный метод частиц в ячейках для задач газовой динамики //Вычислительные методы в гидродинамике. — М.: Мир, 1967

154. *Хокни Р., Иствуд Дж.* Численное моделирование методом частиц. — М.: Мир, 1987

155. *Холл Дж., Уатт Дж.* Современные численные методы решения обыкновенных дифференциальных уравнений. — М.: Мир, 1979

156. *Холодов А. С.* О построении разностных схем с положительной аппроксимацией //ЖВМиМФ. 1984. Т. 24, № 9

157. *Хакбуш В.* (*Hackbusch V.*) Multigrid method and applications. — Berlin etc.: Springer-Verlag, 1985

158. *Хартен А., Ошер О.* (*Harten A., Osher S.*) Uniformly high-order accurate nonoscilating schemes //J. Numer. Analys. 1987. V. 24, № 2

159. *Ченцов Н. Н.* Статистические решающие правила. — М.: Наука, 1972

160. *Черноусько Ф. Л., Баничук В. П.* Вариационные задачи механики и управления. — М.: Наука, 1973

161. *Шайдуров В.* Многосеточные методы конечных элементов. — М.: Наука, 1989

162. *Шокин Ю. И., Яненко Н. Н.* Метод дифференциального приближения: Применение к газовой динамике. — Новосибирск: Наука, 1985

163. *Шор Н. З.* Методы минимизации недифференцируемых функций. — Киев: Наукова думка, 1979

164. *Янг Д., Хейгман Л.* Прикладные итерационные методы. — М., Мир, 1986

165. *Яненко Н. Н.* Метод дробных шагов решения многомерных задач математической физики. — Новосибирск: Наука, 1967

БИБЛИОГРАФИЧЕСКИЙ КОММЕНТАРИЙ

К § 1. Решение систем нелинейных уравнений

Конструкции итерационных методов и их теорию см. в [97]; там же имеется полная библиография и сведения по истории. Решение функциональных уравнений методом продолжения по параметру предложено Д. Ф. Давиденко [49], в частных задачах оно использовалось и раньше (см. [97], с. 230). Термин «инвариантное погружение» введен в [20]. Нормировка задачи введена в [61, 145]. Метод Ньютона в функциональных пространствах рассмотрен в [71, 72].

К § 2. Численное дифференцирование

Численное дифференцирование описано в [18, 19, 70, 118] и др. О современных алгоритмах надежного дифференцирования см. обзор [151]. Некорректная задача вычислений производной как функции, определенной на интервале, изучена в [134].

К § 3. Интерполяция функций

Теорию интерполяции см. в [18, 19, 70, 118] и в монографиях [9, 54, 92]. Вопросы наилучшей аппроксимации функций см. в [9, 33]. Теорию сплайн-интерполяции см. в [5, 126]. Об оригинальной конструкции локальных сплайнов, разработанной В. С. Рябеньким (1952 г.), см. в [115]. Метод конечных элементов описан в монографиях [131, 161]. Конструктивную теорию функций (включая теорию полиномов Чебышева) см. в [54, 92].

К § 4. Вычисление определенных интегралов

Теорию численного интегрирования см. в учебниках [9, 18, 19, 70, 118] и др. Современные исследования оценок минимального объема вычислений, необходимых для интегрирования с заданной точностью, см. в [124]. Теория и приложения метода Монте-Карло описаны в [83, 87, 125, 159]. Экстраполяция Ричардсона в сложных задачах математической физики рассмотрена в [83].

К § 5. Численное интегрирование задачи Коши для систем обыкновенных дифференциальных уравнений

Методы Адамса (1883 г.), Рунге (1895 г.), Кутты (1901 г.), составляющие основу современных алгоритмов, описаны в руководствах [9, 18, 19, 70, 83, 118] и др. О современных исследованиях повышения надежности, автоматизации выбора шага интегрирования, обеспечивающего заданную точность при возможно меньшем объеме вычислений, см. в [19, 44, 152, 155]. О развитии методов приближенного интегрирования уравнений с большим параметром (жестких систем) см. в [51, 108] и в § 17, 18.

К § 6. Абстрактная форма приближенного метода

Изучение приближенных методов с позиций функционального анализа проведено в [2, 72, 115, 117]. Исследование точности некоторых конкретных схем при возможно более слабых предположениях о гладкости решения см., например, в [64–66].

К § 7. Исследование сходимости методов Рунге–Кутты

См. список литературы к § 5.

К § 8. Приближенное решение краевых задач для систем обыкновенных дифференциальных уравнений

Методы решения краевых задач (включая вычисление спектра) см. в [9, 18, 19, 44, 70]. Реализация метода Ньютона в функциональном пространстве и пример решения взяты из [103]. Метод вычисления точек комплексного спектра применен при решении задачи, связанной с исследованием устойчивости атмосферы Венеры, в дипломной работе А. В. Лемехи (МФТИ).

К § 9. Метод дифференциальной прогонки

Прогонка, как устойчивый метод решения краевых задач с большим параметром, была введена и исследована И. М. Гельфандом и О. В. Локуциевским в 1952 г. (опубликована в приложении к [43]). Важнейшие обобщения принадлежат А. А. Абрамову [2] и С. К. Годунову [43]. См. также список литературы к § 10, 18.

К § 10. Прогонка в разностной задаче Штурма–Лиувилля

Прогонка является алгоритмом Гаусса с предписанным порядком исключения неизвестных, который обычно неустойчив (устойчив метод Гаусса с выбором максимального элемента матрицы). Прогонка была «открыта» И. М. Гельфандом и О. В. Локуциевским в 1952 г. именно как применение алгоритма, изложенного в школьном учебнике алгебры. Их заслугой является установление устойчивости и использование алгоритма при решении сложных задач. Примерно в то же время в связи с аналогичным работами прогонка была предложена другими авторами. В настоящее время она является одним из самых массовых алгоритмов. Этот алгоритм (и его обобщения) описаны практически в любом руководстве по численному анализу [9, 18, 19, 118, 120]. См. также список литературы к § 9, 15, 18, 22.

К § 11. Численное интегрирование задачи Коши для уравнений с частными производными

Методы построения и анализа разностных аппроксимаций уравнений математической физики на простейших сетках подробно описаны в [9, 19, 30, 43, 44, 70, 83, 112, 118]; там же рассмотрены вопросы реализации схем. Впервые на важность соотношения между шагами сетки было указано в [76]. Фундаментальный характер «условий Куранта» в полной мере был оценен позже, когда на ЭВМ стали решаться задачи, требующие проведения миллионов операций без контроля математика. В настоящее время практика вынуждает использовать сетки с нерегулярным расположением узлов. Построение аппроксимаций на таких сетках осуществляют не явными формулами, а алгоритмами вычисления коэффициентов схемы. Видимо, первым такие схемы использовал В. Ф. Дьяченко [58, 59] (см. § 23). См. также [130, 156].

К § 12. Спектральный признак устойчивости

Автором метода спектрального анализа устойчивости считают фон-Неймана, хотя он и не является автором первой публикации [28]. Это — пример работы, оказавшей огромное влияние на численный анализ, несмотря на крайнюю простоту используемого математического аппарата. Изложение теории устойчивости разностных схем и практики ее применения см. в [30, 43, 44]. Общую теорию устойчивости (необходимые и достаточные условия в терминах матричных неравенств) см. в [116–118]. Метод исследования разностных краевых условий был доложен К. И. Бабенко, И. М. Гельфандом и О. В. Локуциевским на конференции по функциональному анализу (Москва, 1956) и опубликован в [10]. Подробное изложение и дальнейшее развитие этого метода см. в [44].

К § 13. Метод переменных направлений

Метод переменных направлений, принадлежащий к небольшому числу алгоритмических изобретений, оказавших существенное влияние на развитие вычислительной математики, был предложен в 1955 г. Д. Писманом и Г. Рэчфордом [107]. Обобщение этой конструкции привело к созданию методов расщепления. В настоящее время эта конструкция широко используется для решения двумерных и трехмерных задач. Методы решения систем линейных уравнений с разреженными матрицами описаны в [55]. Теория схем со слабой аппроксимацией изложена в [117, 165].

К § 14. Решение эллиптических задач методом сеток

Теория приближенного решения эллиптических краевых задач разработана очень полно. Многие теоретические результаты (особенно расчет оптимальных итерационных параметров) используются в практической работе. До появления ЭВМ основным был

релаксационный метод Р. В. Саусвелла [132], используемый и сейчас. Современное его состояние описано в [164]. Метод переменных направлений, предложенный впервые в [107] и оптимизированный Е. Л. Вашпрессом [36], стал ярким событием в развитии численного анализа. Обобщения метода, расширяющие область его приложений, и развитие соответствующей теории выполнены Е. Г. Дьяконовым [56]. Подробно описание итерационных методов см. в [83, 120, 122, 140]. Теория устойчивого метода чебышевского ускорения предложена в [78, 116]. Устойчивый трехслойный вариант алгоритма, основанный на рекуррентном соотношении для полиномов Чебышева, изложен в [164]. Многосеточный метод предложен Р. П. Федоренко [149]; его теоретическое обоснование в простейшем случае (уравнение Пуассона в квадрате) дано в [144]. См. также [140, 161]. Независимость эффективности итераций от шага сетки в весьма общей ситуации доказана в [8, 17, 161]. Широкое распространение метод, названный Multigrid, получил после работ Хакбуша [157]. Применение метода к уравнениям упругости (бигармоническому, системе уравнений Ламе) см. в [98, 127].

К § 15. Спектральная задача Штурма–Лиувилля

Метод тригонометрической прогонки был предложен в [80, 94]. См. также список литературы к § 9, 11, 12, 18. Алгоритмы К. И. Бабенко см. в [9].

К § 16. Главная спектральная задача для краевых задач математической физики

Изложение теории и практики решения спектральной задачи в расчетах реакторов см. в [35, 86]. Метод решения уравнения Шредингера разработан П. М. Блехером и В. И. Турчаниновым [23]. Исследование равновесных конфигураций плазмы проводилось Н. М. Зуевой [68] и др. Метод расчета нестационарного процесса в реакторе разработан и реализован Л. Г. Страховской и Р. П. Федоренко [127, 128].

К § 17. Жесткие системы обыкновенных дифференциальных уравнений

Вопросы приближенного решения жестких систем см. в [51, 108, 152]. Теорию сингулярно-возмущенных систем см. в [32]. Системный метод изложен в [3, 99]. Асимптотическая теория жестких систем предложена Р. П. Федоренко [139, 143, 150]. В теория численного интегрирования подробно изложена в [51].

К § 18. Жесткие линейные краевые задачи

Параграф написан на основе [2, 49]. Теория корректных краевых задач разработана С. К. Годуновым и В. С. Рябеньким [44]. Периодическая прогонка предложена в [1]. См. также [120].

К § 19. Осреднение быстрых вращений

Теория метода построена Н. М. Крыловым и Н. Н. Боголюбовым [24, 75]. Стробоскопический метод введен Н. Минорским [91]; близкий «метод точечных отображений» см. в [93]. Изложение основано на обзоре [138] и диссертации А. М. Молчанова [89]. В [138] см. изложение идей А. Н. Колмогорова [74] (осреднение для гамильтоновых систем), развитых в работах В. И. Арнольда [7] и Ю. Мозера [88].

К § 20. Одномерные уравнения газовой динамики и их численное интегрирование

Методы приближенного решения одномерных уравнений см. в [4, 100, 112, 121]. Теорию уравнений и автомодельных решений см. в [95]. Метод С. К. Годунова описан в [40], конструкции характеристических схем — в [21]. Теорию дифференциальных приближений см. в [162]. Расчет разрывных решений рассмотрен в [44]. Гибридные схемы впервые предложены в [45, 73, 146]. Их широкое применение началось после [27] (см., например, [96]). Следует выделить TVD-схему А. Хартена [158] и схемы А. С. Холодова [156].

К § 21. Нелинейное уравнение теплопроводности

Параграф основан на опыте работы группы И. М. Гельфанда (ИПМ им. М. В. Келдыша, 50-е годы). Аппроксимация потока (21.6) была получена К. В. Брушлинским. Потоковая прогонка предложена в [51]. См. также [120, 121].

К § 22. Реализация разностной схемы для уравнений газовой динамики с теплопроводностью

Основу изложения составляет схема, разработанная И. М. Гельфандом, В. Ф. Дьяченко, О. В. Локуциевским. Полностью консервативные схемы введены в [121]. Проблема «монотонизации» схем впервые рассмотрена в [146], откуда взяты численные результаты.

К § 23. Приближенное решение двумерных задач газовой динамики

PIC-метод предложен Ф. Х. Харлоу (1955 г.) [153]. Метод крупных частиц и его применение см. в [21, 22]. Метод свободных точек описан в [11]; там же описаны и другие методы. Особое место занимает метод С. К. Годунова, А. В. Забродина и Г. П. Прокопова с выделением поверхностей разрыва [43, 48, 11]. Аппроксимация около границы предложена в [148]. В дальнейшем другие аппроксимации были построены в [38, 39].

К § 24. Приближенное интегрирование уравнения Власова

Одной из первых попыток интегрирования уравнения Власова была работа [37]. Развитие «метода заряженных облаков» Ю. С. Сиговым см. в [123]. Подробное изложение методов моделирования плазмы см. в [154].

К § 25. Некорректные задачи и их приближенное решение

Теория этих задач началась с работ [69, 133]; подробное изложение см. [134]. Реализация алгоритмов их решения описана в [136]. Решение обратной задачи теплопроводности изложено по [145]. Обратные задачи геофизики рассмотрены в [77], некорректные обратные задачи компьютерной томографии — в [135].

К § 26. Поиск минимума

Методы поиска минимума описаны в [31, 67, 101, 105, 106] и др. Оптимизации недифференцируемых функций посвящены работы [52, 53, 163]. Метод поиска минимакса предложен в [148].

К § 27. Дифференцирование функционалов

Техника дифференцирования функционалов описана в [82, 85, 145]. В [26] представлены вариационные задачи для уравнений с частными производными. Применение функциональных производных в задачах экологии рассмотрено в [83].

К § 28. Задачи оптимального управления

Современное вариационное исчисление изложено в [6, 25, 102, 104, 145] и др. Приближенные методы описаны в [62, 145, 160]. Решение задачи о развороте взято из [62, 103].

К § 29. Вариационные задачи механики с недифференцируемыми функционалами

Теорию задачи Бингема см. в [90]. Другие задачи в терминах вариационных неравенств см. в [48, 60]. Приближенные методы описаны в [14, 15, 48]. Теорию, метод приближенного решения задачи качения и обзор численных результатов см. в [46, 141].

К § 30. Псевдодифференциальные уравнения

Опыт решения задач теории трещин описан в [47, 141]. Задача о трещине гидровзрыва решалась под руководством Р. П. Федоренко в диссертации А. В. Лемехи (ИПМ им. М. В. Келдыша).

К § 31. Метод конечных суперэлементов

Разработка метода только начинается. Этот параграф написан по материалам оригинальных работ Л. Г. Страховской и Р. П. Федоренко [127–130]. См. также обзор [141].

Учебное издание

ФЕДОРЕНКО Радий Петрович

ВВЕДЕНИЕ В ВЫЧИСЛИТЕЛЬНУЮ ФИЗИКУ

Набор и верстка выполнены в издательстве.
Операторы *Л. Г. Быканова, А. К. Розанов, В. Н. Федотов*

Редактор ***Л. И. Гладнева.*** **Корректор** ***О. И. Холодкевич***
Художник ***М. В. Ивановский***

ЛР № 040060 от 21.08.91

ИБ № 2

Подписано в печать 30.09.94.
Формат 60×88/16. Бумага офсетная книжно-журнальная.
Гарнитура тип. «таймс». Печать офсетная. Тираж 5000 экз.
Заказ 1833 . С-006.

Издательство Московского физико-технического института.
141700, г. Долгопрудный Московской области, Институтский пер., д. 9.

Вторая типография ВО «Наука»
121099, Москва, Г-99, Шубинский пер., 6