В. Ф. ДЬЯЧЕНКО

ОСНОВНЫЕ ПОНЯТИЯ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ

Допущено Министерством высшего и среднего специального образования СССР в качестве учебного пособия для студентов высших технических учебных заведений

ИЗДАТЕЛЬСТВО «НАУКА»
ГЛАВНАЯ РЕДАКЦИЯ
ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ
МОСКВА 1972

518
Д 93
УДК 518

Основные понятия вычислительной математики. В. Ф. Дьяченко, Главная редакция физико-математической литературы изд-ва «Наука», 1972, 120 стр.

В книге рассматриваются простейшие понятия и идеи, лежащие в основе современных численных методов решения задач механики и математической физики, вопросы построения и исследования соответствующих вычислительных алгоритмов.

Характер изложения материала не предполагает высокой математической подготовленности читателя. Книга рассчитана на студентов естественных факультетов и вузов, а также на специалистов широкого диапазона физико-технических профессий, и может быть использована для первоначального знакомства с предметом вычислительной математики.

22 илл.

2-2-4
24-72

ОГЛАВЛЕНИЕ

Предисловие 4
Введение 7

ГЛАВА I

§ 1. Вычисление корней уравнений 11
§ 2. Функции и таблицы 17
§ 3. Обыкновенные дифференциальные уравнения 25

ГЛАВА II

§ 4. Уравнения в частных производных 34
§ 5. Аппроксимация и устойчивость 41
§ 6. Спектральный признак устойчивости 48
§ 7. Построение расчетных формул 58
§ 8. Неявные разностные схемы 69
§ 9. Решение разностных уравнений 77

ГЛАВА III

§ 10. Расчет разрывных решений 89
§ 11. Многомерные задачи 96
§ 12. Стационарные задачи 106

ПРЕДИСЛОВИЕ

Возможность постановки вычислительного эксперимента на электронной машине существенно ускорила процесс математизации науки и техники. Расширяется круг профессий, для которых математическая грамотность становится необходимой. Потребность в соответствующем образовании приводит к появлению различной литературы — от справочников до монографий. К ней принадлежит и эта книга.

Она посвящена вопросам разработки численных методов решения задач механики и математической физики. Соответствующая теория находится в стадии интенсивного развития и еще далека от завершения. Однако некоторые, относительно простые идеи и понятия уже выкристаллизовались. Именно они и только они излагаются в книге. Конечно, овладение азбукой численных методов не сделает читателя квалифицированным вычислителем. Для этого необходимо изучение более глубокой литературы и особенно личный опыт решения конкретных задач. Но эффективность того и другого резко возрастает, если самые простые вопросы ясны.

Разнообразие профессий и уровня подготовки предполагаемого читателя заставляют автора отказаться от традиционного для математической литературы, строго формализованного стиля изложения. Поэтому, возможно, некоторые рафинированные математики-профессионалы назовут книгу банальной и даже вульгарной. Нам придется согласиться с ними. Но полная математическая вооруженность, учитывающая все логические возможности, необходима, если в нашем распоряжении нет ничего, кроме математической логики. Реальные объекты не таковы, что́ и позволяет писать книги, подобные этой. С другой стороны, основное назначение любой книги — быть про-

читанной. Мы старались лишить читателя повода захлопнуть книгу.

Изложение ведется на «физическом» уровне строгости. Автор ориентируется не столько на математическую подготовленность читателя, сколько на его здравый смысл и сообразительность. Как правило, рассмотрение того или иного вопроса проводится на простом типичном примере, а затем намечаются пути обобщения полученных результатов на более сложные случаи.

В конце каждого параграфа помещены задачи различной степени трудности. Для решения большинства из них требуется ясное понимание принципов, излагаемых в основном тексте, и способность самостоятельно развить их.

Первые три параграфа носят вводный характер и касаются классических вопросов вычислительной математики — итерационных способов решения уравнений, интерполяционных и квадратурных формул, численного интегрирования обыкновенных дифференциальных уравнений.

Остальные параграфы, составляющие основное содержание книги, посвящены численным методам решения уравнений в частных производных, построению и исследованию соответствующих вычислительных алгоритмов. Здесь рассматриваются конкретные приемы проверки аппроксимации и устойчивости разностных задач, свойства явных и неявных схем, способы построения расчетных формул, методы решения систем разностных уравнений, возможности реализации алгоритмов и т. д.

Итак: мини-теория современных численных методов решения задач механики, газовой динамики и вообще математической физики, требующих интегрирования дифференциальных уравнений.

Книга родилась как результат многолетней работы автора в области прикладной математики и чтения соответствующего курса лекций в Московском физико-техническом институте. Она может быть использована студентами естественных факультетов и вузов для первоначального знакомства с предметом и методами вычислительной математики.

По вопросам, затронутым в первых трех параграфах, имеется обширная литература, и подробное изложение всех их можно найти почти в любом курсе приближенных вычислений.

Для более глубокого изучения разностных методов решения уравнений в частных производных рекомендуем следующие книги:

Рихтмайер Р. Д., Разностные методы решения краевых задач, ИЛ, 1960.

Годунов С. К., Рябенький В. С., Введение в теорию разностных схем, Физматгиз, 1962.

Вазов В., Форсайт Дж., Разностные методы решения дифференциальных уравнений в частных производных, ИЛ, 1963.

Яненко Н. Н., Метод дробных шагов решения многомерных задач математической физики, изд-во «Наука», СО, 1967.

Самарский А. А., Введение в теорию разностных схем, изд-во «Наука», 1971.

Автор благодарен В. С. Рябенькому, роль которого в появлении этой книги чрезвычайно велика.

В. Дьяченко

ВВЕДЕНИЕ

Численный метод решения задачи — это определенная последовательность операций над числами, т. е. вычислительный алгоритм, язык которого — числа и арифметические действия. Такая примитивность языка позволяет реализовывать численные методы на вычислительных машинах, что делает эти методы мощным и универсальным инструментом исследования.

Однако задачи, подлежащие решению, формулируются, как правило, на обычном математическом языке (уравнений, функций, дифференциальных операторов и т. п.). Поэтому разработка численного метода необходимо предполагает замену, аппроксимацию исходной задачи другой, близкой к ней и сформулированной в терминах чисел и арифметических операций. Несмотря на все разнообразие способов такой замены, некоторые общие свойства присущи им всем.

Обратимся к простейшему примеру. Требуется найти решение уравнения

$$x^2 - a = 0, \quad a > 0, \tag{1}$$

т. е. извлечь квадратный корень из заданного числа a. Можно, конечно, написать $x = \sqrt{a}$, но символ $\sqrt{\ }$ не решает задачи — не дает способа вычисления величины x.

Поступим следующим образом. Зададимся каким-либо начальным приближением x_0 (например, $x_0 = 1$) и будем последовательно, с помощью формулы

$$x_n = \frac{1}{2}\left(x_{n-1} + \frac{a}{x_{n-1}}\right) \tag{2}$$

вычислять значения $x_1, x_2, \ldots$ Прервем этот процесс на некотором $n = N$, и полученное в результате x_N объявим

приближенным решением исходной задачи (1), т. е. положим

$$\sqrt{a} \sim x_N.$$

Правомерность этого допущения зависит, очевидно, от требований, предъявляемых к точности решения, от величины a и от параметра N. Если иметь в виду любые требования, то нужно доказать, что для всякого a соответствующим выбором N можно добиться любой близости x_N к точному значению $\sqrt{a}$.

Докажем, что наш алгоритм (2) удовлетворяет этому условию. Положим

$$\frac{x_n}{\sqrt{a}} = 1 + \varepsilon_n. \tag{3}$$

Разделим равенство (2) на $\sqrt{a}$ и подставим в него (3), получим

$$1 + \varepsilon_n = \frac{1}{2}\left(1 + \varepsilon_{n-1} + \frac{1}{1+\varepsilon_{n-1}}\right),$$

откуда

$$\varepsilon_n = \frac{1}{2}\left(\varepsilon_{n-1} - 1 + \frac{1}{\varepsilon_{n-1}+1}\right) = \frac{1}{2}\,\frac{\varepsilon_{n-1}^2}{\varepsilon_{n-1}+1}. \tag{4}$$

Так как $1 + \varepsilon_0 = 1/\sqrt{a} > 0$, то из последнего равенства следует, что все ε_n, начиная с первого, положительны. А значит,

$$\frac{\varepsilon_{n-1}}{\varepsilon_{n-1}+1} < 1.$$

Используя это, получаем из (4)

$$\varepsilon_n < \frac{1}{2}\,\varepsilon_{n-1}, \tag{5}$$

т. е. ε_n убывает с ростом n быстрее, чем геометрическая прогрессия со знаменателем $^1/_2$. Следовательно,

$$x_N \to \sqrt{a} \quad \text{при} \quad N \to \infty, \tag{6}$$

и наше утверждение доказано.

Рис. 1 иллюстрирует итерационный процесс (2). Здесь даны графики левой — $y_л(x)$ и правой — $y_п(x)$ частей (2). Поскольку, очевидно, $y_л(\sqrt{a}) = y_п(\sqrt{a})$, то эти графики

пересекаются в точке $x = \sqrt{a}$. Проведение итераций по формуле (2) эквивалентно движению по изображенной на рисунке ломаной линии, зажатой между $y_л(x)$ и $y_п(x)$. Это еще раз убеждает нас в сходимости итераций к $\sqrt{a}$ при $N \to \infty$.

При исследовании сходимости мы допустили некоторую идеализацию алгоритма, молчаливо предположив возможность точной реализации вычислений по формуле (2). Но ни человек, ни машина не могут оперировать с произвольными действительными числами. Вычисления всегда ведутся с ограниченным числом десятичных знаков, и точность результата не может превосходить точность расчета. Важно установить, в каком отношении эти точности находятся, не будут ли ошибки, допускаемые при округлении, накапливаясь, лишать результат расчета какой-либо ценности.

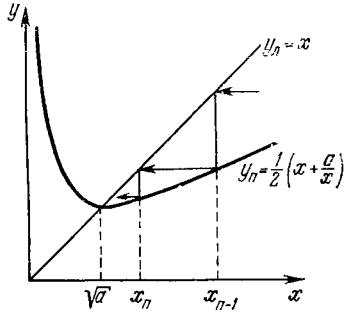

Рис. 1.

Хотя почти очевидно, что в нашем примере все обстоит благополучно, проведем формальную проверку влияния указанного фактора. Роль округлений сводится к тому, что фактически вместо формулы (2) мы пользуемся формулой

$$\tilde{x}_n = \frac{1}{2}\left(\tilde{x}_{n-1} + \frac{a}{\tilde{x}_{n-1}}\right)(1 + \delta_n), \qquad (7)$$

где множитель $1 + \delta_n$ эффективно учитывает ошибку, вводимую округлениями на данном n-м шаге расчета, а $\tilde{x}_n$ — фактически получаемая последовательность. Величина $\delta \ll 1$ характеризует точность вычислений. Заменяя $\tilde{x}_n$ на $\sqrt{a}\,(1 + \varepsilon_n)$, получим вместо (4)

$$\varepsilon_n = \frac{1}{2}\,\frac{\varepsilon_{n-1}^2}{\varepsilon_{n-1} + 1}(1 + \delta_n) + \delta_n.$$

Отсюда видно, что с ростом n ε_n убывает до величины порядка δ_n, т. е. точность результата соответствует точности вычислений.

Несмотря на свою элементарность, рассмотренный пример вполне отчетливо демонстрирует следующие принципы, общие для всех численных методов.

Во-первых, исходная задача (1) заменяется другой задачей — вычислительным алгоритмом (2).

Во-вторых, задача (2) содержит параметр N, которого нет в исходной задаче.

В-третьих, выбором этого параметра можно добиться, в принципе, любой близости решения второй задачи к решению первой, x_N к $\sqrt{a}$.

Наконец, в-четвертых, неточная реализация алгоритма, вызванная округлениями, не меняет существенно его свойств.

ГЛАВА I

§ 1. ВЫЧИСЛЕНИЕ КОРНЕЙ УРАВНЕНИЙ

Задача состоит в нахождении действительного корня уравнения

$$f(x) = 0. \tag{8}$$

Будем предполагать, что этот корень существует и располагается внутри некоторого известного интервала (может быть, большого).

Рассмотренная выше задача (1) является частным случаем данной. Примененный там метод итераций можно распространить и на общий случай — уравнение (8). Для этого нужно построить итерационный процесс вида

$$x_n = \varphi(x_{n-1}) \tag{9}$$

(ср. с (2)), сходящийся к корню уравнения (8), который обозначим X.

Рассмотрим условия, которым должна удовлетворять функция $\varphi(x)$, и некоторые способы ее построения.

Допустим, что итерационный процесс (9) сходится, т. е. $x_n \to X$ при $n \to \infty$. Тогда, переходя к пределу в левой и правой частях (9), получим $X = \varphi(X)$, т. е. $x = X$ должен быть общим корнем уравнений (8) и

$$x = \varphi(x). \tag{10}$$

Поэтому для получения итерационной формулы достаточно просто переписать уравнение (8) в виде (10) (ср. (1) и (2)). Разумеется, это можно сделать не единственным образом. Так, уравнение (1) можно представить, наряду с (2), в виде

$$x = \frac{a}{x}$$

или

$$x = 2x - \frac{a}{x}.$$

Нетрудно установить, что ни та, ни другая формула не годится для итераций. Первая дает $x_1 = a/x_0$, $x_2 = x_0$, $x_3 = a/x_0$, ..., а вторая порождает последовательность x_n, стремящуюся к бесконечности. Таким образом, далеко не всякая запись уравнения (8) в виде (10) приводит к цели.

Выясним, какие свойства $\varphi(x)$ влияют на сходимость процесса. Поскольку последняя означает, что $x_n - X \to 0$ при $n \to \infty$, то нужно, чтобы $|x_n - X|$ убывало с ростом n. Пусть для любого n справедливо неравенство

$$|x_n - X| \leqslant \Theta |x_{n-1} - X|, \qquad \Theta < 1. \tag{11}$$

Тогда, очевидно, $|x_n - X|$ убывает как геометрическая прогрессия со знаменателем Θ, т. е. сходимость имеет место. Подставим в левую часть (11) вместо x_n и X соответственно $\varphi(x_{n-1})$ и $\varphi(X)$. Получим

$$|\varphi(x_{n-1}) - \varphi(X)| \leqslant \Theta |x_{n-1} - X|, \qquad \Theta < 1. \tag{12}$$

Так как корень X неизвестен, то последнее условие непосредственно проверить нельзя и приходится несколько усилить его. Выше мы предполагали, что корень локализован внутри некоторого интервала. Если для любой пары точек x', x'' из этого интервала выполнено условие

$$|\varphi(x'') - \varphi(x')| \leqslant \Theta |x'' - x'|, \qquad \Theta < 1, \tag{13}$$

то заведомо выполнено и (12). Отображение $x = \varphi(x)$, удовлетворяющее (13), называют *сжатым* отображением (оно сжимает отрезки $x'' - x'$). Очевидно, условие (13) есть достаточное условие сходимости итерационного процесса (9).

Для конкретной оценки величины Θ, определяющей скорость сходимости, проще всего пользоваться очевидной формулой

$$\Theta = \max_x |\varphi'(x)|, \tag{14}$$

где max берется по интервалу локализации корня. Если отображение $x = \varphi(x)$ — сжатое, то этот интервал с ростом n будет уменьшаться, и оценку скорости сходимости можно уточнять.

Итак, любая запись уравнения (8) в виде (10), удовлетворяющая условию (13), дает итерационный процесс, сходящийся к корню.

Качество того или иного выбора $\varphi(x)$, очевидно, следует оценивать по скорости сходимости. Лучшим в этом смысле естественно считать тот, для которого величина Θ будет наименьшей.

Пусть $\varphi(x)$ таково, что уравнения $f(x)=0$ и $\varphi(x)-x=0$ имеют общий корень X. Тогда, если этот корень не кратный, $f'(X)\neq 0$, в некоторой окрестности его, где нет других нулей функции $f(x)$, отношение

$$\frac{\varphi(x)-x}{f(x)}=r(x)$$

есть ограниченная функция. Каждой функции $\varphi(x)$ соответствует $r(x)$ и, обратно, каждая ограниченная $r(x)$ порождает

$$\varphi(x)=x+r(x)f(x). \tag{15}$$

Нас интересуют $\varphi(x)$ с минимальным $|\varphi'(x)|$ (в силу (14)). Продифференцируем выражение (15),

$$\varphi'=1+r(x)f'(x)+r'(x)f(x).$$

В районе корня величина $f(x)$ мала, поэтому пренебрежем последним слагаемым и приравняем нулю оставшееся выражение. Это даст

$$r(x)=-\frac{1}{f'(x)},$$

т. е., в соответствии с (15), функцию

$$\varphi(x)=x-\frac{f(x)}{f'(x)}, \tag{16}$$

порождающую итерационный процесс, известный как *метод Ньютона*,

$$x_n=x_{n-1}-\frac{f(x_{n-1})}{f'(x_{n-1})}. \tag{17}$$

Скорость сходимости этого процесса очень высока, так как

$$\varphi'(x)=\frac{f''(x)}{(f'(x))^2}f(x)\to 0$$

по мере приближения к корню.

Формулу (17) можно получить и другим путем (рис. 2). А именно, имея некоторое приближение для корня x_{n-1}, при вычислении следующего приближения x_n будем

вместо уравнения $f(x) = 0$ рассматривать линейное уравнение

$$f(x_{n-1}) + f'(x_{n-1})(x - x_{n-1}) = 0,$$

учитывающее лишь два первых члена разложения функции $f(x)$ в ряд Тейлора около точки x_{n-1}. Решая последнее уравнение относительно x, получаем формулу (17).

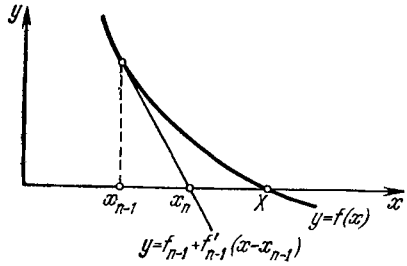

Рис. 2.

Преимущества метода Ньютона сказываются только в окрестности корня, где φ' мало. Поэтому в конкретных расчетах следует сначала локализовать корень каким-нибудь более простым и грубым способом, а затем для быстрого достижения высокой точности использовать метод Ньютона. Точность вычислений (количество удерживаемых десятичных знаков) рационально сделать переменной, увеличивая ее по мере приближения к корню. При этом, поскольку переход от x_{n-1} к x_n является одновременно проверкой уравнения $x = \varphi(x)$, эквивалентного $f(x) = 0$, то достигнутую точность можно оценивать по числу не меняющихся при этом переходе знаков.

Изложенный итерационный метод нахождения корня, использующий сжатые отображения, можно без принципиальных изменений распространить на многие более сложные задачи.

Пусть требуется решить систему M уравнений с M неизвестными

$$f^{(m)}(x^{(1)}, x^{(2)}, \ldots, x^{(M)}) = 0, \qquad m = 1, 2, \ldots, M. \quad (18)$$

Обозначим вектор с компонентами $f^{(1)}, f^{(2)}, \ldots, f^{(M)}$ через f, соответственно $x^{(1)}, x^{(2)}, \ldots, x^{(M)}$ через x и запишем систему (18) в виде векторного уравнения

$$f(x) = 0, \quad (19)$$

по форме совпадающего с (8). Все сказанное относительно уравнения (8) переносится на систему (19), нужно только, учитывая векторный характер величин, придать новый смысл употребляемым обозначениям.

Для оценки величины скаляра используется его модуль, для вектора — его *норма*. Мы определим последнюю как максимум модуля компонент

$$\|x\| = \max_m |x^{(m)}|. \tag{20}$$

Очевидно, равенство нормы вектора нулю означает равенство нулю всех его компонент.

Перепишем систему (19) в виде

$$x = \varphi(x), \tag{21}$$

где φ — вектор с компонентами $\varphi^{(1)}, \varphi^{(2)}, \ldots, \varphi^{(M)}$. Как и прежде, формула (21) порождает итерационный процесс, условием сходимости которого является сжатость отображения (21), т. е.

$$\|\varphi(x'') - \varphi(x')\| \leqslant \Theta \|x'' - x'\|, \qquad \Theta < 1. \tag{22}$$

Оценим величину Θ. Рассмотрим функцию $\varphi^{(m)}(x)$ на прямой, соединяющей точки x' и x''. Получим функцию скалярного аргумента s

$$\psi(s) = \varphi^{(m)}(x' + s(x'' - x')). \tag{23}$$

Точки x' и x'' соответствуют значениям $s = 0$ и $s = 1$. Очевидно,

$$\varphi^{(m)}(x'') - \varphi^{(m)}(x') = \psi(1) - \psi(0) = \psi'(\tilde{s}), \tag{24}$$

причем $0 \leqslant \tilde{s} \leqslant 1$. Дифференцируя (23) по s, получим

$$\psi'(s) = \sum_{k=1}^{M} \frac{\partial \varphi^{(m)}}{\partial x^{(k)}} ((x^{(k)})'' - (x^{(k)})').$$

Вместе с (24) это дает

$$|\varphi^{(m)}(x'') - \varphi^{(m)}(x')| \leqslant$$
$$\leqslant \max_x \sum_{k=1}^{M} \left| \frac{\partial \varphi^{(m)}}{\partial x^{(k)}} \right| \max_k |(x^{(k)})'' - (x^{(k)})'|.$$

Используя определение нормы (20), получаем отсюда

$$\|\varphi(x'') - \varphi(x')\| \leqslant \max_m \max_x \sum_{k=1}^{M} \left| \frac{\partial \varphi^{(m)}}{\partial x^{(k)}} \right| \|x'' - x'\|. \tag{25}$$

Роль φ' теперь, естественно, играет матрица производных $\frac{\partial \varphi^{(m)}}{\partial x^{(k)}}$ — производная вектор-функции по векторному аргументу. Если мы определим норму этой матрицы формулой

$$\|\varphi'\| = \max_m \sum_k \left| \frac{\partial \varphi^{(m)}}{\partial x^{(k)}} \right|, \tag{26}$$

то для оценки величины Θ в условии (22), как следует из (25), можно воспользоваться равенством

$$\Theta = \max_x \|\varphi'(x)\|, \tag{27}$$

обобщающим (14) на векторный случай.

Далее, построение $\varphi(x)$ можно по-прежнему производить с помощью (15)

$$\varphi(x) = x + r(x) f(x), \tag{28}$$

где $r(x)$ — произвольная матрица функций. При

$$r(x) = -(f'(x))^{-1}, \tag{29}$$

где $(f')^{-1}$ — матрица, обратная матрице производных $f' = \partial f^{(m)}/\partial x^{(k)}$, получим метод Ньютона. Он и здесь сохраняет свои преимущества, так как соответствующая ему матрица φ' в точке корня будет нулевой. В этом легко убедиться, дифференцируя (28) и используя (29). При применении метода Ньютона теперь, вместо деления на функцию f', требуется обращение матрицы f', т. е. решение на каждой итерации системы линейных уравнений

$$f(x_{n-1}) + f'(x_{n-1})(x - x_{n-1}) = 0.$$

При значительном порядке системы M это может оказаться довольно трудоемким. Поэтому обычно данный метод используется лишь при наличии хорошего приближения для корней, когда одна-две итерации дают резкое повышение точности.

Относительно решения систем линейных уравнений заметим следующее. Если нужно решить систему общего вида, то ничего лучше известного метода исключения не существует. Однако, используя специфику конкретных систем, часто удается построить достаточно эффективные итерационные методы.

Задачи

1. Построить итерационный процесс для извлечения корня степени p.

2. Определить скорость сходимости метода Ньютона в окрестности k-кратного корня.

3. Построить итерационный процесс решения уравнения $ax + b = 0$, где $0{,}1 < a < 1$, не использующий операции деления.

4. Дана система линейных уравнений $Ax + b = 0$ с матрицей A, все собственные значения которой действительны, различны, положительны и заведомо заключены внутри некоторого известного интервала (α, β). Подобрать число r так, чтобы итерационный процесс

$$x_n = x_{n-1} + r(Ax_{n-1} + b)$$

сходился возможно быстрее.

§ 2. ФУНКЦИИ И ТАБЛИЦЫ

Функция является фундаментальным математическим понятием и, естественно, присутствует в формулировках большинства задач. На языке численных методов функцию могут представлять, очевидно, только числовые последовательности. Возможны различные способы такого представления, например, с помощью коэффициентов ряда по каким-либо функциям или значений параметров в формуле заданного вида. Наиболее употребительным и универсальным является представление функции в виде таблицы ее значений. Так все «элементарные» функции $\sin x$, $\ln x$, и т. д.— это таблицы.

Рассмотрим функцию $F(x)$ и соответствующую ей таблицу — пару последовательностей x_k, f_k $(k = 0, 1, 2, \ldots)$. Качество этого соответствия будем оценивать по точности, с которой таблица x_k, f_k позволяет восстановить значение $F(x)$ в любой точке x.

Очевидно, эта точность зависит от близости f_k к $F(x_k)$ и плотности расположения узлов таблицы x_k. Роль первого фактора ясна — отклонение f_k от $F(x_k)$ ставит предел точности воспроизведения, который, не исправляя таблицы, превзойти нельзя. Идеализируем задачу и будем считать, что

$$f_k = F(x_k), \quad k = 0, 1, 2, \ldots \tag{30}$$

В этом случае точность восстановления $F(x)$ будет определяться тем, насколько таблица подробна, насколько

хорошо она описывает детали поведения функции $F(x)$ и, конечно, способом восстановления.

Пусть, например, способ — самый грубый и состоит в том, что мы каждое значение f_k распространяем на весь прилегающий интервал $x_k \leqslant x < x_{k+1}$ (рис. 3), т. е. для вычисления $F(x)$ используем кусочно-постоянную функцию

$$P_0(x) = f_k, \; x_k \leqslant x < x_{k+1}, \; k = 0, 1, 2, \ldots \qquad (31)$$

Чтобы оценить величину ошибки, $P_0(x) - F(x)$, нужно иметь какую-либо дополнительную информацию о функции $F(x)$, кроме (30). Будем считать, что $F(x)$ — гладкая

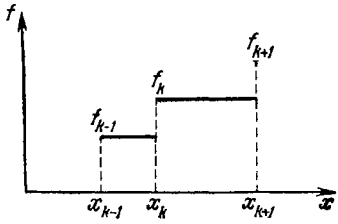

Рис. 3.

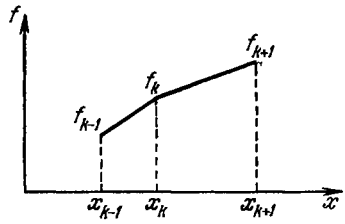

Рис. 4.

функция, имеющая непрерывную производную. Тогда на интервале (x_k, x_{k+1}) ее можно представить в виде

$$F(x) = F(x_k) + F'(\xi(x))(x - x_k).$$

Отсюда сразу следует, что на этом интервале

$$|P_0(x) - F(x)| \leqslant \max|F'|(x_{k+1} - x_k), \qquad (32)$$

т. е. ошибка *порядка* величины шага таблицы.

Более точным представляется способ вычисления $F(x)$ с помощью *линейной интерполяции*, т. е. путем построения (вместо (31)) кусочно-линейной функции, использующей и левое f_k, и правое f_{k+1} значения F на каждом интервале (рис. 4), вида

$$P_1(x) = f_k + \frac{f_{k+1} - f_k}{x_{k+1} - x_k}(x - x_k), \qquad x_k \leqslant x \leqslant x_{k+1}. \qquad (33)$$

Можно проводить интерполяцию с помощью квадратичных, кубичных и т. д. функций, используя для их построения тройки, четверки и т. д. точек таблицы. Рассмотрим общий случай.

Имеется $n+1$ точка $x_0, x_1, \ldots, x_n$ — *узлы интерполяции*, которым соответствуют значения $f_0, f_1, \ldots, f_n$. Построим полином n-го порядка

$$P_n(x) = \sum_{m=0}^{n} a_m x^m, \tag{34}$$

принимающий в точках x_i $(i = 0, 1, \ldots, n)$ значения f_i. Полагая в (34) $x = x_i$, получим систему $n+1$ уравнений для определения $n+1$ неизвестных — коэффициентов полинома

$$\sum_{m=0}^{n} a_m x_i^m = f_i, \qquad i = 0, 1, \ldots, n. \tag{35}$$

Определитель этой линейной системы $|x_i^m|$ (определитель Ван-дер-Монда) равен $\prod_{i,j}(x_i - x_j)$, т. е. отличен от нуля, так как все x_i различны. Следовательно, система (35) имеет единственное решение — набор a_m. Мы доказали, что существует единственный полином n-го порядка (не выше), принимающий в $n+1$ точке заданные значения. Он называется *интерполяционным полиномом*.

Конкретная форма записи его может быть различна. Вид (34) малоупотребителен из-за громоздкости выражения коэффициентов a_m через x_i, f_i. Запишем интерполяционный полином в *форме Лагранжа*

$$P_n(x) = \sum_{i=0}^{n} f_i \frac{q_i(x)}{q_i(x_i)}, \tag{36}$$

где

$$q_i(x) = \\ = (x - x_0)(x - x_1) \ldots (x - x_{i-1})(x - x_{i+1}) \ldots (x - x_n).$$

Так как $q_i(x_k) = 0$ при $i \neq k$, то, очевидно, $P_n(x_k) = f_k$. С другой стороны, каждое $q_i(x)$ — полином степени n. Следовательно, их линейная комбинация (36) есть интерполяционный полином.

Для случая $n = 1$ формула (36) превращается в

$$P_1(x) = f_0 \frac{x - x_1}{x_0 - x_1} + f_1 \frac{x - x_0}{x_1 - x_0} \tag{37}$$

— формулу линейной интерполяции (ср. с (33)).

Оценим точность, с которой интерполяционный полином воспроизводит функцию $F(x)$. Так как разность $P_n(x) - F(x)$ в узлах интерполяции $x_0, x_1, \ldots, x_n$ обращается в нуль, то частное от деления этой разности на функцию

$$q(x) = (x - x_0)(x - x_1) \ldots (x - x_n) \tag{38}$$

есть ограниченная функция, и мы можем написать

$$F(x) = P_n(x) + R(x)\, q(x). \tag{39}$$

Оценим $R(x)$. Для этого рассмотрим вспомогательную функцию

$$v(\xi) = F(\xi) - P_n(\xi) - R(x) q(\xi) = (R(\xi) - R(x))\, q(\xi). \tag{40}$$

Очевидно, функция $v(\xi)$ обращается в нуль по крайней мере в $n + 2$ точках $x_0, x_1, \ldots, x_n, x$. Следовательно, найдется хотя бы одна точка $\xi(x)$, где обращается в нуль $(n + 1)$-я производная от $v(\xi)$, разумеется, если эта производная существует и непрерывна. Дифференцируя (40) $n + 1$ раз и подставляя $\xi = \xi(x)$, получим

$$0 = F^{(n+1)}(\xi(x)) - R(x)(n + 1)!,$$

поскольку $(n + 1)$-я производная от полинома n-й степени $P_n(\xi)$ есть нуль, а $q(\xi)$ — полином вида $\xi^{n+1} + \ldots$ Итак,

$$R(x) = \frac{F^{(n+1)}(\xi(x))}{(n + 1)!},$$

и мы получаем выражение для *ошибки интерполяции*

$$F(x) - P_n(x) = \frac{1}{(n + 1)!} F^{(n+1)}(\xi(x))\, q(x). \tag{41}$$

Зависимость $\xi(x)$ остается, конечно, неопределенной.

Если шаг таблицы не превосходит некоторого h, т. е.

$$x_{k+1} - x_k \leqslant h, \quad k = 0, 1, \ldots, n - 1,$$

то $q(x) \sim h^{n+1}$ и из (41) следует

$$|F(x) - P_n(x)| \leqslant c_n \max_x |F^{(n+1)}(x)|\, h^{n+1}, \tag{42}$$

где c_n — некоторая константа. Отсюда заключаем, что при $h \to 0$ ошибка убывает как h^{n+1}.

Зависимость величины ошибки от степени интерполяционного полинома n более сложная. На практике интерполяционные формулы со сколько-нибудь большим n употребляются крайне редко. Причины этого следующие. Во-первых, как видно из (42), увеличение степени полинома может привести к уменьшению ошибки интерполяции лишь для очень *гладких* функций, имеющих достаточно большое число производных. Но такой информацией о свойствах $F(x)$ мы, как правило, не обладаем. Во-вторых, зачения f_k являются всегда приближенными значениями для $F(x_k)$, хотя бы из-за округления. Поэтому полиномы, построенные по f_k и $F(x_k)$, в лучшем случае будут отличаться друг от друга на величину порядка $f_k - F(x_k)$. К тому же ошибки, содержащиеся в f, носят всегда случайный характер, а это можно интерпретировать как сильную негладкость представляемой ими функции.

Описанный способ восстановления функции $F(x)$ по таблице x_k, f_k путем построения интерполяционного полинома не является единственно возможным. Мы строили $P_n(x)$ (34) с помощью системы степенных функций $x^m (m = 0, 1, \ldots)$. Но для этой цели годятся и многие другие системы $\varphi_m(x)$. В этом случае вместо (34) следует рассмотреть

$$\Phi_n(x) = \sum_{m=0}^{n} a_m \varphi_m(x) \tag{43}$$

и произвести соответствующее исследование возможности и качества аппроксимации. Можно идти еще дальше — конструировать аппроксимирующую функцию в виде какой-либо нелинейной комбинации опорных функций $\varphi_m(x)$. Но, разумеется, это имеет смысл делать только при достаточно обоснованной необходимости, так как выигрыш на этом пути может достигаться лишь за счет сужения области применимости метода и использования существенной дополнительной информации о $F(x)$, кроме таблицы ее значений.

К вопросу о восстановлении функции по набору ее значений можно подойти и по-другому. При построении интерполяционной функции мы требовали точного совпадения ее значений в узлах x_k с f_k. Но часто можно ограничиться требованием минимальности отклонения этих значений от табличных. Например, если f заведомо содержат значительные ошибки, или простой вид аппроксими-

рующей функции для нас важнее точности аппроксимации, то такой подход закономерен.

Опишем один из способов такого рода — *способ наименьших квадратов*. Имеем таблицу x_k, f_k $(k = 0, 1, \ldots, n)$. Требуется построить функцию

$$\Phi_M(x) = \sum_{m=0}^{M} a_m \varphi_m(x), \tag{44}$$

где $\varphi_m(x)$ — заданная система функций (например, $\varphi_m(x) = x^m$), так, чтобы величина

$$\sum_{k=0}^{n} (\Phi_M(x_k) - f_k)^2 = \delta \tag{45}$$

была минимальной. Если $M \geqslant n$, то задача решается построением интерполяционной функции (44), для которой $\delta = 0$. Нас интересует случай $M < n$.

Весь произвол заключается в выборе коэффициентов $a_0, a_1, \ldots, a_M$, поэтому δ есть функция от них. Для нахождения минимума функции $\delta(a_0, a_1, \ldots, a_M)$ приравняем нулю производные этой функции по $a_0, a_1, \ldots, a_M$ — получим систему уравнений

$$\sum_{l=0}^{M} a_l \sum_{k=0}^{n} \varphi_m(x_k)\varphi_l(x_k) = \sum_{k=0}^{n} \varphi_m(x_k) f_k, \qquad m = 0, 1, \ldots, M.$$

Решив эту систему линейных уравнений, найдем значения $a_0, a_1, \ldots, a_M$, полностью определяющие функцию $\Phi_M(x)$ (44), которая наилучшим образом среди функций этого вида аппроксимирует таблицу x_k, f_k, если за меру отклонения принять (45).

Остановимся на некоторых применениях полученных результатов. Тот или иной способ соответствия между таблицей и функцией дает возможность совершать над таблицей различные функциональные операции — интегрирование, дифференцирование.

Так, если требуется вычислить интеграл от функции, заданной таблицей x_k, f_k $(k = 0, 1, \ldots, K)$, то используя на каждом интервале (x_k, x_{k+1}) формулу линейной интерполяции (33), будем иметь

$$\int_{x_k}^{x_{k+1}} P_1(x)\,dx = \frac{f_k + f_{k+1}}{2}(x_{k+1} - x_k). \tag{46}$$

Суммируя эти выражения по всем интервалам, получаем способ вычисления интеграла. Для случая, когда шаг таблицы постоянен, $x_{k+1} - x_k = h$, имеем

$$\int_{x_0}^{x_K} P_1(x)\,dx = \left(\frac{1}{2} f_0 + f_1 + f_2 + \ldots + f_{K-1} + \frac{1}{2} f_K\right) h \tag{47}$$

— известную *квадратурную формулу трапеций*.

Можно оценить точность, с которой формула (47) дает величину интеграла от функции $F(x)$. Ошибка интерполяции в силу (41) при $n = 1$ есть

$$\varepsilon(x) = \frac{1}{2} F''(\xi(x))(x - x_k)(x - x_{k+1}), \qquad x_k \leqslant x \leqslant x_{k+1}. \tag{48}$$

Отсюда

$$\left| \int_{x_k}^{x_{k+1}} \varepsilon(x)\,dx \right| \leqslant \text{const} \max_{x_k \leqslant x \leqslant x_{k+1}} |F''| (x_{k+1} - x_k)^3.$$

Суммируя это неравенство по всем интервалам, учитывая, что $Kh = x_K - x_0$, получим

$$\left| \int_{x_0}^{x_K} \varepsilon(x)\,dx \right| \leqslant \text{const} \max_{x_0 \leqslant x \leqslant x_K} |F''| h^2, \tag{49}$$

т. е. формула трапеций (47) имеет точность *порядка* h^2. Используя другие интерполяционные функции $P_0(x)$, $P_2(x)$ и т. д., получим квадратурные формулы прямоугольников, Симпсона и т. д.

Столь же просто решается вопрос о вычислении производных от табличной функции. Используя, например, линейную интерполяцию (33), получим

$$\frac{dF}{dx} \sim \frac{dP_1}{dx} = \frac{f_{k+1} - f_k}{x_{k+1} - x_k}, \qquad x_k \leqslant x \leqslant x_{k+1}. \tag{50}$$

Дифференцируя выражение для ошибки интерполяции (48), имеем

$$\frac{d\varepsilon}{dx} = \frac{1}{2} F'''(\xi(x))\,\xi'(x)(x - x_k)(x - x_{k+1}) + \\ + \frac{1}{2} F''(\xi(x))(2x - x_k - x_{k+1}).$$

Второе слагаемое в правой части порядка $x_{k+1} - x_k$, т. е. h. Такова точность вычисления производной по формуле (50). Исключением является центральная точка интервала $x = (x_k + x_{k+1})/2$. В ней второе слагаемое обращается в нуль и, следовательно, формула (50) дает значение производной в этой точке с точностью h^2.

Попытка определения с помощью линейной интерполяции второй производной приводит к

$$\frac{d^2F}{dx^2} \sim \frac{d^2P_1}{dx^2} = 0,$$

что, очевидно, непригодно. Это согласуется с тем, что

$$\frac{d^2\varepsilon}{dx^2} = F''(\xi(x)) + \ldots,$$

т. е. ошибка конечна, не убывает с уменьшением h. Для вычисления старших производных необходимо использовать интерполяцию более высокой степени.

Мы рассмотрели только случай функции одного переменного. При переходе к многомерным задачам принципиальная сторона изложенных методов сохраняется, но появляется масса новых проблем.

Задачи

1 Оценить зависимость c_n от n в формуле (42).

2. Дана таблица

x	0	1	2
f	0	1	1

Аппроксимировать ее линейной функцией по методу наименьших квадратов. Затем аппроксимировать ее линейной функцией, используя для оценки отклонения не (45), а

$$\max_k |\Phi(x_k) - f_k| = \delta.$$

Сравнить полученные функции с интерполяционным полиномом второй степени, построенным по тем же точкам.

3. Получить квадратурные формулы, используя интерполяционные полиномы нулевой, второй степени. Оценить, возможно точнее, порядок ошибки этих формул.

4. Построить формулу вычисления второй производной табличной функции. Оценить ошибку.

5. Функция двух переменных $F(x, y)$ задана таблицей

x	0	h	0	$-h$	0
y	0	0	h	0	$-h$
F	$f_{0,\,0}$	$f_{1,\,0}$	$f_{0,\,1}$	$f_{-1,\,0}$	$f_{0,\,-1}$

Построить квадратурную формулу для вычисления интеграла

$$\iint\limits_{x^2+y^2\leqslant h^2} F(x, y)\,dx\,dy,$$

используя в каждом октанте линейную интерполяцию по двум переменным.

§ 3. ОБЫКНОВЕННЫЕ ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ

Рассмотрим не самую сложную, но типичную для таких уравнений задачу. Требуется найти функцию $U(t)$, удовлетворяющую при $t > 0$ уравнению

$$\frac{dU}{dt} = F(t, U) \tag{51}$$

и принимающую при $t = 0$ заданное значение

$$U(0) = U_0. \tag{52}$$

Как известно из теории обыкновенных дифференциальных уравнений, если правая часть (51), $F(t, U)$, удовлетворяет как функция своих аргументов определенным условиям гладкости, то решение задачи (51), (52), $U(t)$, существует, единственно и является гладкой функцией. Мы будем предполагать, что эти условия выполнены.

Случаи, когда $U(t)$ может быть выражена через элементарные функции или интегралы от них, являются исключительными. Как правило, единственным средством решения задачи (51), (52) оказываются численные методы. Они, конечно, дают ограниченную и приближенную информацию о решении, но зато являются универсальными. Опишем простейший из них — *метод Эйлера*.

Прежде всего, область непрерывного изменения аргумента $t \geqslant 0$ заменяем дискретным множеством точек

$$t_n = n\tau, \quad n = 0, 1, 2, \ldots, \tag{53}$$

где τ — некоторое фиксированное малое число — *параметр численного метода*. Вместо функции $U(t)$ будем рассматривать таблицу

$$t_n, u_n, n = 0, 1, 2, \ldots \tag{54}$$

Далее, так как, по определению производной, dU/dt есть предел отношения $(U(t+\tau) - U(t))/\tau$ при $\tau \to 0$, то заменяя производную этим *конечным отношением*, получим вместо дифференциального уравнения (51) *разностное уравнение*

$$\frac{u_{n+1} - u_n}{\tau} = F(t_n, u_n), \qquad n = 0, 1, 2, \ldots \tag{55}$$

или

$$u_{n+1} = u_n + \tau F(t_n, u_n), \quad n = 0, 1, 2, \ldots \tag{55'}$$

Таким образом, полагая, в силу (52),

$$u_0 = U_0, \tag{56}$$

мы с помощью (55′) можем найти последовательно все u_n. Построение вычислительного алгоритма закончено.

Его ценность, естественно, будет определяться тем, насколько хорошо таблица t_n, u_n воспроизводит точное решение исходной задачи (51), (52) — функцию $U(t)$. Для выяснения этого вопроса положим

$$u_n = U(t_n) + \delta u_n \tag{57}$$

и попытаемся оценить величину ошибки δu_n.

Подставляя (57) в разностное уравнение (55), получим

$$\frac{\delta u_{n+1} - \delta u_n}{\tau} = F(t_n, U(t_n) + \delta u_n) - \frac{U(t_{n+1}) - U(t_n)}{\tau}. \tag{58}$$

Оценим правую часть. Поскольку $U(t)$ — гладкая функция и удовлетворяет (51), то

$$U(t_{n+1}) = U(t_n) + \tau \left(\frac{dU}{dt}\right)_{t_n} + O(\tau^2) =$$

$$= U(t_n) + \tau F(t_n, U(t_n)) + O(\tau^2),$$

откуда

$$\frac{U(t_{n+1})-U(t_n)}{\tau}=F(t_n, U(t_n))+O(\tau). \qquad (59)$$

Сравнивая это равенство с (55), отметим, что точное решение исходной дифференциальной задачи удовлетворяет разностному уравнению (55) с точностью $O(\tau)$. Подставляя теперь (59) в правую часть (58), можем написать

$$\frac{\delta u_{n+1}-\delta u_n}{\tau}=F(t_n, U(t_n)+\delta u_n)-F(t_n, U(t_n))+O(\tau). \qquad (60)$$

Из условий гладкости, наложенных на $F(t, U)$, следует справедливость неравенства

$$|F(t_n, U(t_n)+\delta u_n)-F(t_n, U(t_n))|\leqslant M|\delta u_n|, \qquad (61)$$

где M — некоторая константа. Используя (61), находим из (60)

$$|\delta u_{n+1}|\leqslant(1+\tau M)|\delta u_n|+O(\tau^2). \qquad (62)$$

Формула (62) дает оценку роста ошибки за один шаг. Поэтому для ошибки, накопленной за N шагов, имеем

$$\begin{aligned}\delta u_N|&\leqslant O(\tau^2)+(1+\tau M)|\delta u_{N-1}|\leqslant\\ &\leqslant O(\tau^2)+(1+\tau M)(O(\tau^2)+(1+\tau M)|\delta u_{N-2}|)\leqslant\\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ &\leqslant O(\tau^2)+(1+\tau M)O(\tau^2)+(1+\tau M)^2O(\tau^2)+\\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ &\qquad+(1+\tau M)^{N-1}O(\tau^2)+(1+\tau M)^N|\delta u_0|,\end{aligned}$$

откуда, суммируя геометрическую прогрессию со знаменателем $1+\tau M$, имеем

$$|\delta u_N|\leqslant\frac{(1+\tau M)^N-1}{(1+\tau M)-1}O(\tau^2)+(1+\tau M)^N|\delta u_0|. \qquad (63)$$

Нас интересует величина ошибки при малых τ для любого фиксированного t. Полагая $t=N\tau$ и учитывая, что

$$(1+\tau M)^N=(1+\tau M)^{t/\tau}\sim e^{Mt} \qquad \text{при} \quad \tau\sim 0,$$

получаем из (63)

$$|\delta u(t)|\leqslant e^{Mt}O(\tau)+e^{Mt}|\delta u_0|. \qquad (64)$$

Поскольку $\delta u_0=u_0-U_0=0$, то отсюда следует, что

при $\tau \to 0$ ошибка δu на любом конечном интервале t стремится к нулю.

Итак, мы доказали, что при достаточно малом τ таблица t_n, u_n, получаемая методом Эйлера, как угодно точно аппроксимирует решение исходной задачи (51), (52).

Мы не получили критерия — какое τ следует считать «достаточно малым» для достижения заданной точности. Мы установили лишь факт *сходимости* приближенного решения к точному при $\tau \to 0$.

Можно было бы, используя вместо обозначений $O(\tau)$, $O(\tau^2)$ более содержательные выражения, дать количественную оценку величине ошибки — для этого пришлось бы преодолеть лишь чисто технические трудности. Но дело в том, что такая оценка была бы сильно завышенной и все равно малопригодна для вычисления фактической ошибки.

Поэтому, в случае необходимости, для определения достигнутой точности поступают следующим образом. Так как $u \to U$ при $\tau \to 0$, то начиная с некоторого τ первые значащие цифры u перестают меняться — они отвечают точному решению. Произведя несколько расчетов с различными τ и сравнив результаты, получают представление о достигнутой точности.

С этой точки зрения более удобны методы с р е з к о й зависимостью величины ошибки от параметра τ — методы *высокого порядка точности*. Порядок точности метода Эйлера — минимальный, $u - U = O(\tau)$. Очевидно, это связано с довольно грубым способом аппроксимации дифференциального уравнения разностным.

Качество *аппроксимации* оценивают по точности, с которой решение исходной задачи удовлетворяет разностному уравнению. Сравнение (55) и (59) показывает, что в данном случае аппроксимация имеет порядок $O(\tau)$. Такой способ оценки порядка аппроксимации содержит некоторый произвол. Например, если вместо (55) использовать (55′), то порядок аппроксимации будет $O(\tau^2)$ — за счет другой нормировки разностного уравнения (умножения его на τ). Мы будем использовать, так сказать, естественную нормировку разностного уравнения, при которой оно в пределе переходит в исходное дифференциальное уравнение. Так, (55) (вернее (59)) при $\tau \to 0$ переходит в (51), в то время как (55′) — в равенство $U = U$, никак не отражающее индивидуальность исходной задачи.

Рассмотрим разностное уравнение

$$\frac{u_{n+1}-u_n}{\tau}=\frac{1}{2}\left(F(t_n, u_n)+F(t_{n+1}, u_{n+1})\right) \qquad (65)$$

и оценим порядок, с которым оно аппроксимирует дифференциальное уравнение (51). Подставляя в (65) вместо u_n, u_{n+1} значения $U(t_n)$, $U(t_{n+1})$ и используя разложения

$$U(t_{n+1})=U(t_n)+\tau\left(\frac{dU}{dt}\right)_{t_n}+\frac{\tau^2}{2}\left(\frac{d^2U}{dt^2}\right)_{t_n}+O(\tau^3),$$

$$F(t_{n+1}, U(t_{n+1}))=F(t_n, U(t_n))+\tau\left(\frac{dF}{dt}\right)_{t_n}+O(\tau^2),$$

придем к равенству

$$\left(\frac{dU}{dt}+\frac{\tau}{2}\frac{d^2U}{dt^2}\right)_{t_n}=\left(F+\frac{\tau}{2}\frac{dF}{dt}\right)_{t_n}+O(\tau^2).$$

Так как $U(t)$ удовлетворяет дифференциальному уравнению (51) и следствию из него

$$\frac{d^2U}{dt^2}=\frac{dF}{dt},$$

то отсюда следует, что (65) при $u=U$ удовлетворяется с точностью $O(\tau^2)$.

Итак, разностное уравнение (65) аппроксимирует исходное (51) с точностью $O(\tau^2)$. Сохранение этого же порядка точности и для решения, т. е. утверждение $u-U=O(\tau^2)$, не является очевидным и требует доказательства. Но мы на этом останавливаться не будем.

В отличие от (55), разностная формула (65) не позволяет, вообще говоря, явно выразить u_{n+1} через u_n. Она является уравнением относительно u_{n+1}. Для его решения можно применить тот или иной итерационный процесс, тем более, что всегда имеется хорошее начальное приближение u_n. Однако труд, затраченный на получение точного значения u_{n+1}, будет напрасной потерей времени, так как точность самого уравнения порядка $O(\tau^2)$. Поэтому можно ограничиться двумя итерациями следующего вида. Сначала вычисляем первое приближение $\tilde{u}_{n+1}$ по формуле метода Эйлера

$$\tilde{u}_{n+1}=u_n+\tau F(t_n, u_n), \qquad (66)$$

а затем, подставляя его в правую часть (65), находим

окончательное значение u_{n+1},

$$u_{n+1} = u_n + \frac{\tau}{2}(F(t_n, u_n) + F(t_{n+1}, \tilde{u}_{n+1})). \qquad (67)$$

Фактически это означает, что вместо (65) мы пользуемся разностным уравнением

$$\frac{u_{n+1} - u_n}{\tau} = \frac{1}{2}(F(t_n, u_n) + F(t_{n+1}, \ u_n + \tau F(t_n, u_n))). \qquad (68)$$

Нетрудно показать, что это уравнение, как и (65), аппроксимирует исходное (51) с точностью $O(\tau^2)$ и в то же время дает явное выражение u_{n+1} через u_n.

Заметим, что при проведении расчета с помощью (66), (67) имеется возможность осуществлять определенный контроль точности получаемого решения путем сравнения значений $\tilde{u}_{n+1}$ и u_{n+1} на каждом шаге вычислений. Слишком большая (малая) величина «пересчета» $u_{n+1} - \tilde{u}_{n+1}$ будет сигнализировать о необходимости уменьшения (увеличения) шага τ. Меняя в соответствии с этим шаг, можно поддерживать точность на определенном уровне.

Обобщая изложенный способ построения разностных уравнений, получают методы еще более высоких порядков точности. Если ввести в рассмотрение несколько различных промежуточных значений $\tilde{u}$, вычисляемых последовательно, то придем к методам типа Рунге —Кутта (см. задачу 2). Другой путь обобщения основывается на использовании для получения u_{n+1} не только u_n, но и уже известных значений $u_{n-1}, u_{n-2}, \ldots$ Общий принцип построения таких методов (типа Адамса) выглядит следующим образом. Имея последовательность $F(t_n, u_n), F(t_{n-1}, \ u_{n-1}), \ldots, F(t_{n-k}, u_{n-k})$, строим по ним интерполяционный полином $P_k(t)$. Используя его вместо F на интервале $t_n, \ t_{n+1}$ (экстраполяция), можем написать, интегрируя (51),

$$u_{n+1} - u_n = \int_{t_n}^{t_{n+1}} P_k(t)\, dt.$$

Поскольку $P_k(t)$ линейно выражается через значения F в упомянутых точках, то выполнив интегрирование, получим

$$u_{n+1} = u_n + \tau \sum_{i=0}^{k} c_i F(t_{n-i}, u_{n-i}), \qquad (69)$$

т. е. разностную формулу. При $k = 0$ (69) превращается в формулу метода Эйлера (55′). Имея u_{n+1}, можно, как и в методе Эйлера, сделать «пересчет», используя интерполяционный полином, учитывающий только что полученное значение u_{n+1}. Этим способом нетрудно построить численные методы любого порядка точности, причем довольно экономные, так как повышение порядка точности достигается за счет использования уже имеющихся значений u.

Однако эти популярные в эпоху ручных расчетов методы сейчас почти не применяются. Причина очень характерна для современной вычислительной математики. Дело в том, что использовать формулу (69) можно только начиная с $n = k$. Следовательно, значения $u_1, u_2, \ldots, u_k$ должны быть получены по каким-либо другим, «нестандартным» формулам. Аналогичные затруднения возникают и при изменении шага интегрирования. Необходимость расширения алгоритма для учета нескольких исключительных случаев и приводит к непопулярности этих методов.

Все изложенное выше естественным образом обобщается на системы дифференциальных уравнений (а, следовательно, и на уравнения высших порядков). В частности, если под U, F, u понимать не скалярные величины, а векторные, то (51) превращается в систему дифференциальных уравнений, а разностные формулы (55) и др. описывают методы интегрирования этой системы. Сохраняют силу и результаты исследования сходимости, аппроксимации, хотя само исследование этих вопросов несколько усложняется.

При построении численных методов мы систематически и существенно использовали предположение о гладкости функции $F(t, U)$. Если отказаться от него и допустить наличие каких-либо особенностей у функции $F(t, U)$, то это приведет, очевидно, к необходимости разработки специального способа решения лишь в окрестности этих особенностей. Например, если функция $F(t, U)$ имеет особую точку, типа неопределенности 0/0, то следует, выделив главные члены $F(t, U)$, найти аналитическое приближенное решение в окрестности этой точки и использовать его для прохождения через особую точку путем «склейки» с таблицей t_n, u_n, получаемой численным методом.

Наконец, остановимся на роли ошибок округления. Как мы отмечали, всякое разностное уравнение содержит

ошибку аппроксимации. Если неточность, возникающая на отдельном шаге из-за округлений, не превосходит неточность аппроксимации, то такое соотношение сохранится и для накопленных ошибок, порожденных тем и другим источником. Следовательно, если вместе с уменьшением τ соответственно увеличивать точность, с которой ведутся вычисления, то сходимость не нарушится. Рассмотрим, например, метод Эйлера. Фактический вычислительный процесс даст нам не последовательность u_n, удовлетворяющую (55′), а некоторую, близкую к ней, последовательность $\tilde{u}_n$. Способ получения последней можно представить формулой

$$\tilde{u}_{n+1} = \tilde{u}_n + \tau F(t_n, \tilde{u}_n) + \delta_n,$$

где δ_n — погрешность, которую мы допускаем при вычислении правой части. Она складывается из неточности вычисления $F(t_n, \tilde{u}_n)$, неточности умножения на τ. Очевидно, δ_n характеризует точность расчета. Если $\delta_n = O(\tau^2)$, то $\tilde{u}_n$ удовлетворяет тому же уравнению, что и $U(t_n)$,— (59). Следовательно, $\tilde{u}_n - u_n = O(\tau)$, и реальный вычислительный процесс, если вести его с небольшим запасом точности $\delta = O(\tau^2)$, позволяет получить решение с точностью $O(\tau)$.

Задачи

1. Доказать, что решение, получаемое методом Эйлера с пересчетом ((66), (67)), сходится к точному с порядком $O(\tau^2)$.

2. Метод Рунге — Кутта интегрирования уравнения (51) описывается следующими формулами:

$$u_{n+1} = u_n + \frac{\tau}{6}(f_1 + 2f_2 + 2f_3 + f_4),$$

где

$$f_1 = F(t_n, u_n),$$

$$f_2 = F\left(t_n + \frac{\tau}{2}, u_n + \frac{\tau}{2} f_1\right),$$

$$f_3 = F\left(t_n + \frac{\tau}{2}, u_n + \frac{\tau}{2} f_2\right),$$

$$f_4 = F(t_n + \tau, u_n + \tau f_3).$$

Определить порядок аппроксимации этого метода.

3. Доказать, что метод Эйлера для линейной системы

$$\frac{dU_i}{dt} = \sum_{j=1}^{I} a_{ij} U_j, \quad i = 1, 2, \ldots, I,$$

дает решение с точностью $O(\tau)$.

4. Описать вычислительный алгоритм для нахождения решения системы

$$\frac{dx}{dt} = \frac{x}{t} + yf(t),$$

$$\frac{dy}{dt} = x + t,$$

удовлетворяющего начальным данным $t = 0$, $x = 0$, $y = 0$ и принимающего, при некотором $t = T$, заданные значения $x = X$, $y = Y$. Функция $f(t)$ — гладкая и задана таблицей, допускающей линейную интерполяцию.

5. Для решения задачи

$$\frac{dU}{dt} + MU = 0, \quad U(0) = 1$$

рассмотреть две разностные схемы

$$\frac{u_{n+1} - u_n}{\tau} + Mu_n = 0$$

и

$$\frac{u_{n+1} - u_n}{\tau} + Mu_{n+1} = 0.$$

Оценить порядок аппроксимации обеих схем. Сравнить приближенные решения, получаемые при конечных τ, с точным. Которой из этих схем отдать предпочтение при больших M, если требуется лишь общее качественное отражение характера точного решения?

ГЛАВА II

§ 4. УРАВНЕНИЯ В ЧАСТНЫХ ПРОИЗВОДНЫХ

Все дальнейшее будет посвящено численным методам решения дифференциальных уравнений в частных производных. При переходе от одной к двум и более независимым переменным разнообразие и сложность задач резко возрастают. Не имея возможности сколько-нибудь полно охватить все аспекты и методы этой интенсивно развивающейся теории, мы ограничимся рассмотрением лишь некоторых принципиальных вопросов.

Начнем с самой простой задачи. В области

$$-\infty < x < \infty, \quad t \geqslant 0 \tag{70}$$

требуется найти функцию $U(x, t)$, удовлетворяющую при $t > 0$ дифференциальному уравнению

$$\frac{\partial U}{\partial t} + a\frac{\partial U}{\partial x} = F(x, t) \tag{71}$$

и принимающую при $t = 0$ заданные значения

$$U(x, 0) = \Phi(x). \tag{72}$$

Точное решение этой задачи дается формулой

$$U(x, t) = \Phi(x - at) + \int_0^t F(x - at + at', t')\,dt', \tag{73}$$

в чем легко убедиться. Но сейчас эта задача интересует нас лишь как пример, на котором, несмотря на его элементарность, можно продемонстрировать многие существенные свойства численных методов интегрирования дифференциальных уравнений в частных производных. Ниже мы обратимся к нему еще не раз.

Для построения численного метода решения задачи (70), (71), (72) прежде всего заменим область непрерыв-

ного изменения аргументов (70) *расчетной сеткой* (рис. 5) — дискретным множеством точек с координатами

$$\left.\begin{aligned} x_k &= kh, \qquad k = 0, \pm 1, \pm 2, \ldots \\ t^n &= n\tau, \qquad n = 0, 1, 2, \ldots \end{aligned}\right\} \tag{74}$$

Тем самым в рассмотрение вводятся два параметра τ и h — *шаги сетки*. Вместо функций $U(x, t)$, $F(x, t)$, $\Phi(x)$

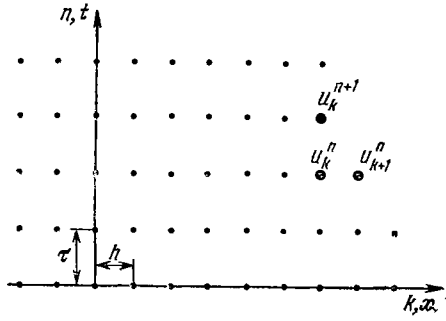

Рис. 5.

будем рассматривать *сеточные функции* — числовые последовательности u_k^n, f_k^n, φ_k, соответствующие точкам сетки x_k, t^n (74). Далее, заменяя частные производные, входящие в (71), *разностными отношениями*, напишем

$$\frac{u_k^{n+1} - u_k^n}{\tau} + a\,\frac{u_{k+1}^n - u_k^n}{h} = f_k^n \tag{75}$$

для каждой пары индексов k, n, т. е. для каждой точки расчетной сетки (74). Наконец, вместо (72) напишем

$$u_k^0 = \varphi_k. \tag{76}$$

Итак, задачу (70), (71), (72) мы заменили *разностной, численной задачей* (74), (75), (76). Разумеется, использованный способ замены не является единственным. Возможностей здесь много, больше, чем в случае обыкновенных дифференциальных уравнений. Но пока мы остановимся на этом.

Вычислительный алгоритм получения значений u_k^n весьма прост. Разрешим (75) относительно u_k^{n+1},

$$u_k^{n+1} = \left(1 + a\frac{\tau}{h}\right) u_k^n - a\frac{\tau}{h} u_{k+1}^n + \tau f_k^n. \tag{77}$$

Если величины u_k^n для некоторого n-го слоя точек известны, то формула (77) позволяет вычислить их для следующего $(n+1)$-го слоя точек. Поскольку u_k^0 известны, то находим по ним u_k^1, затем u_k^2 и т. д.

Перейдем теперь к выяснению основного вопроса — будет ли полученное численным методом решение u_k^n близко к точному решению исходной задачи $U(x, t)$. Очевидно, надеяться на это можно лишь при малых τ и h.

Положим

$$u_k^n = U(x_k, t^n) + \delta u_k^n \tag{78}$$

и подставим это выражение в разностную формулу (75). Получим

$$\frac{\delta u_k^{n+1} - \delta u_k^n}{\tau} + a\frac{\delta u_{k+1}^n - \delta u_k^n}{h} = \\ = f_k^n - \left(\frac{U_k^{n+1} - U_k^n}{\tau} + a\frac{U_{k+1}^n - U_k^n}{h}\right). \tag{79}$$

Здесь и далее U_k^n означает $U(x_k, t^n)$. Оценим правую часть (79). Считая $U(x, t)$ гладкой функцией, при малых τ и h можем написать

$$\left.\begin{aligned} U_k^{n+1} &= U_k^n + \tau\left(\frac{\partial U}{\partial t}\right)_k^n + O(\tau^2), \\ U_{k+1}^n &= U_k^n + h\left(\frac{\partial U}{\partial x}\right)_k^n + O(h^2), \end{aligned}\right\} \tag{80}$$

откуда

$$\frac{U_k^{n+1} - U_k^n}{\tau} + a\frac{U_{k+1}^n - U_k^n}{h} = \left(\frac{\partial U}{\partial t} + a\frac{\partial U}{\partial x}\right)_k^n + O(\tau, h).$$

Так как $U(x, t)$ удовлетворяет (71), а $F(x_k, t^n) = f_k^n$, то последнее равенство означает, что

$$\frac{U_k^{n+1} - U_k^n}{\tau} + a\frac{U_{k+1}^n - U_k^n}{h} = f_k^n + O(\tau, h). \tag{81}$$

Сравнивая (81) с (75), заключаем, что решение исходной задачи $U(x, t)$ удовлетворяет разностному уравнению (75) с точностью $O(\tau, h)$, т. е. *аппроксимация* имеет место.

Подставляя (81) в правую часть (79), получим уравнение для определения δu

$$\frac{\delta u_k^{n+1} - \delta u_k^n}{\tau} + a\,\frac{\delta u_{k+1}^n - \delta u_k^n}{h} = O(\tau, h) \tag{82}$$

или

$$\delta u_k^{n+1} = \left(1 + a\,\frac{\tau}{h}\right)\delta u_k^n - a\,\frac{\tau}{h}\,\delta u_{k+1}^n + \tau O(\tau, h). \tag{83}$$

Сначала рассмотрим случай, когда a, τ, h удовлетворяют неравенствам

$$0 \leqslant -a\,\frac{\tau}{h} \leqslant 1. \tag{84}$$

В этом случае коэффициенты при δu_k^n и δu_{k+1}^n в правой части (83) положительны, и можно написать

$$\begin{aligned} |\delta u_k^{n+1}| &\leqslant \\ &\leqslant \left(1 + a\,\frac{\tau}{h}\right)|\delta u_k^n| + \left(-a\,\frac{\tau}{h}\right)|\delta u_{k+1}^n| + \tau O(\tau, h) \leqslant \\ &\leqslant \max(|\delta u_k^n|, |\delta u_{k+1}^n|) + \tau O(\tau, h). \end{aligned}$$

Введем обозначение

$$\|\delta u^n\| = \max_k |\delta u_k^n|, \tag{85}$$

тогда предыдущее неравенство дает

$$\|\delta u^{n+1}\| \leqslant \|\delta u^n\| + \tau O(\tau, h), \tag{86}$$

т. е. максимальное отклонение δu за один шаг τ увеличивается не более чем на $\tau O(\tau, h)$. Соответственно за N шагов это даст

$$\|\delta u^N\| \leqslant \|\delta u^0\| + N\tau O(\tau, h). \tag{87}$$

Зафиксируем любое конечное $t = N\tau$ и устремим τ, h к нулю, а N соответственно к бесконечности. Поскольку $\delta u_k^0 = \Phi_k - \varphi_k = 0$, то из (87) следует

$$\|\delta u(t)\| = O(\tau, h). \tag{88}$$

Итак, мы доказали, что если при стремлении τ, h к нулю условия (84) выполнены, то решение разностной задачи (74), (75), (76) *сходится* к решению исходной задачи (70), (71), (72).

Рассмотрим теперь противоположный случай, когда стремление τ, h к нулю происходит таким образом, что хотя бы одно из условий (84) нарушено. Оказывается, что в этом случае сходимости, вообще говоря, нет. Покажем это с помощью следующего простого рассуждения.

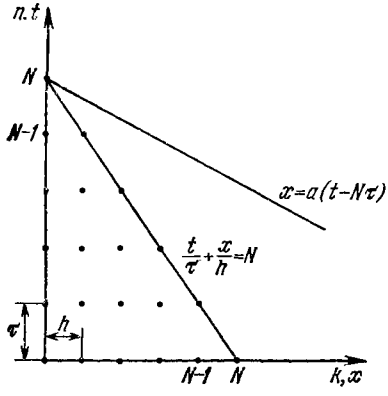

Рис. 6.

Определим область зависимости для u_0^N. Поскольку u_0^N выражается (77) через f_0^{N-1}, u_0^{N-1}, u_1^{N-1}, а последние, в свою очередь, через f_0^{N-2}, f_1^{N-2}, u_0^{N-2}, u_1^{N-2}, u_2^{N-2} и т. д. (рис. 6), то продолжая этот процесс, можно исключить все промежуточные значения u_k^n и выразить u_0^N непосредственно через f_k^n и φ_k. Точки k, n, значения f и φ в которых будут при этом использованы, очевидно, образуют треугольник, изображенный на рис. 6, который и является областью зависимости u_0^N. Его сторонами являются отрезки прямых

$$x = 0, \quad t = 0, \quad \frac{t}{\tau} + \frac{x}{h} = N, \tag{89}$$

т. е. он определяется двумя параметрами $N\tau$ и τ/h. Если они остаются постоянными при τ, $h \to 0$, то область зависимости не меняется.

Определим теперь область зависимости точного решения в этой же точке, т. е. $U(0, N\tau)$. Как следует из (73), $U(0, N\tau)$ полностью определяется значениями $F(x, t)$ на прямой

$$x = a(t - N\tau), \qquad 0 \leqslant t \leqslant N\tau, \tag{90}$$

и значением Φ в точке пересечения этой прямой с осью x.

Допустим, что прямая (90) располагается вне треугольника (89). В этом случае $U(0, N\tau)$ и u_0^N будут определяться различными, независимыми факторами, и рассчитывать на их близость мы оснований не имеем. Меняя, например, значения F и Φ только в окрестности прямой (90), мы будем получать различные $U(0, N\tau)$. В то же время u_0^N на эти изменения реагировать не будет.

Найдем условия принадлежности прямой (90) треугольнику (89). При данном $t < N\tau$ внутренности треугольника (89) отвечает

$$0 \leqslant x \leqslant h\left(N - \frac{t}{\tau}\right).$$

Подставляя вместо x его выражение (90), получаем

$$0 \leqslant a(t - N\tau) \leqslant \frac{h}{\tau}(N\tau - t)$$

или, сокращая на $N\tau - t$,

$$0 \leqslant -a \leqslant \frac{h}{\tau}.$$

Очевидно, последнее тождественно (84).

Итак, мы показали, что в случае нарушения условий (84) сходимости, вообще говоря, нет.

Подчеркнем следующий существенный факт. Разностная задача (74), (75), (76) *аппроксимирует* исходную дифференциальную задачу (70), (71), (72) вне зависимости от того, выполнены или нет условия (84). Однако оказывается, что одной лишь аппроксимации недостаточно для *сходимости* u_k^n к точному решению $U(x, t)$. Дополнительными условиями, обеспечивающими сходимость, являются в данной задаче условия (84). Невыполнение их может привести к любому отклонению u_k^n от $U(x, t)$.

Тем не менее, интересно выяснить, какой характер имеет решение u_k^n, получаемое при данных конечных τ, h, и что с ним происходит при уменьшении τ, h. Чтобы понять это, не проводя конкретных расчетов, поступим следующим образом. Обратимся к формуле (83), которая описывает процесс эволюции ошибки от слоя к слою. Наличие в правой части члена $\tau O(\tau, h)$ указывает на то, что порядок ошибки во всяком случае не меньше этой величины. Сама ошибка δu_k^n есть некоторая сложная

функция от индекса k. Допустим, что ее можно представить в виде суммы, одно из слагаемых которой имеет вид $\varepsilon(-1)^k$, где, разумеется, ε имеет порядок $\tau O(\tau, h)$. Проследим за развитием только этой компоненты ошибки, т. е. положим

$$\delta u_k^n = \varepsilon(-1)^k, \tag{91}$$

и найдем δu_k^{n+1}, δu_k^{n+2}, ... При этом, несколько идеализируя задачу, не будем учитывать вклад, даваемый членом $\tau O(\tau, h)$ на следующих шагах. Получим

$$\delta u_k^{n+1} = \left(1 + a\frac{\tau}{h}\right)\varepsilon(-1)^k - a\frac{\tau}{h}\varepsilon(-1)^{k+1} = \\ = \left(1 + 2a\frac{\tau}{h}\right)\varepsilon(-1)^k,$$

т. е. функция вида (91) переходит на следующем слое в себя, приобретая множитель $1 + 2a\tau/h$. Очевидно, через N слоев она приобретет множитель $(1 + 2a\tau/h)^N$,

$$\delta u_k^{n+N} = \left(1 + 2a\frac{\tau}{h}\right)^N \varepsilon(-1)^k, \tag{92}$$

и, следовательно, развитие рассматриваемой компоненты ошибки определяется величиной $1 + 2a\tau/h$. Если

$$\left|1 + 2a\frac{\tau}{h}\right| < 1, \tag{93}$$

то ошибка затухает, в противном случае нарастает экспоненциально.

Последнее приводит к тому, что решение u_k^n, содержащее эту ошибку, довольно быстро теряет какой-либо смысл, становясь хаотической последовательностью очень больших чисел. Малость ε, очевидно, не спасает положения и может лишь несколько оттянуть наступление катастрофы. Описанный эффект получил название *неустойчивости*.

Нетрудно заметить, что (93) есть опять все то же условие (84). А это значит, что отсутствие сходимости и неустойчивость имеют одну и ту же причину. Используя (92) для конечного отрезка $t = N\tau$, получаем, что в случае нарушения условия (93), т. е. (84),

$$|\delta u_k^{n+1}| = \left|1 + 2a\frac{\tau}{h}\right|^{t/\tau} \tau O(\tau, h) \to \infty \quad \text{при } \tau, h \to 0.$$

Расчет по неустойчивой разностной схеме не только не дает решения близкого к точному, но вообще невозможен. При этом уменьшение τ, h только ухудшает положение.

Задачи

1. Если $a > 0$, то условия (84) не выполняются ни при каких τ, h. Как изменить разностную формулу (75) для этого случая?

2. Построить и исследовать разностный метод решения задачи (70), (71), (72) в общем случае переменного a, $a = a(x, t)$.

3. Для решения задачи

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2}, \quad U(x, 0) = \Phi(x)$$

рассмотреть разностную схему

$$\frac{u_k^{n+1} - u_k^n}{\tau} = \frac{u_{k+1}^n - 2u_k^n + u_{k-1}^n}{h^2}, \quad u_k^0 = \varphi_k.$$

Исследовать ее аппроксимацию, сходимость (при $2\tau/h^2 \leqslant 1$) и устойчивость (на функции $\varepsilon(-1)^k$).

§ 5. АППРОКСИМАЦИЯ И УСТОЙЧИВОСТЬ

Принципиальная схема проведенного выше исследования сходимости является типичной для широкого класса задач и может быть описана следующим образом. Пусть некоторая исходная дифференциальная задача заменена разностной задачей вида

$$lu = f. \tag{94}$$

Здесь u, f — сеточные функции u_k^n, f_k^n, а l — линейный разностный оператор, зависящий от параметров τ, h. В частности, рассмотренная задача (75), (76) записывается в виде (94), если положить

$$lu = \begin{cases} \dfrac{u_k^{n+1} - u_k^n}{\tau} + a\,\dfrac{u_{k+1}^n - u_k^n}{h}, \\ u_k^0 \end{cases} \tag{95}$$

и

$$f = \begin{cases} f_k^n, \\ \varphi_k. \end{cases} \tag{96}$$

Для выяснения вопроса о сходимости решения задачи (94) к решению исходной задачи — к функции U полагаем

$$u = U + \delta u$$

и, подставляя это выражение в (94), получаем уравнение для определения δu

$$l\,\delta u = f - lU. \tag{97}$$

Очевидно, *сходимость*, т. е. стремление δu к 0, будет иметь место, если, во-первых, выбором параметров τ, h правая часть (97) может быть сделана как угодно малой, и, во-вторых, эта малость сохранится для решения уравнения (97) — сеточной функции δu.

Первое из этих условий

$$lU - f \to 0 \text{ при } \tau,\ h \to 0 \tag{98}$$

есть уже знакомое нам условие *аппроксимации* (ср. с (81)). Оно характеризует связь между разностной и дифференциальной задачами и устанавливает факт близости этих задач.

Выполнение второго условия зависит только от свойств разностной задачи. А именно, разностный оператор l должен быть таков, чтобы при любых τ, h решение задачи (97) имело тот же порядок, что и правая часть, т. е.

$$\delta u \sim f - lU. \tag{99}$$

Этим свойством обладает далеко не каждый разностный оператор. Как было показано выше, для оператора l (95) условие (99) справедливо, если a, τ, h удовлетворяют соотношениям (84). В противном случае условие (99) не выполняется. Поскольку при этом наблюдается явление неустойчивости, то вкладывая в последний термин несколько большее содержание, условие (99) называют условием *устойчивости*.

Заметим, что уравнения (97) и (94) отличаются только обозначениями: δu, $f - lU$ вместо u, f. Поэтому условие устойчивости (99) можно переписать в виде

$$u \sim f, \tag{100}$$

которому должно удовлетворять решение задачи (94). Подчеркнем, что оператор l зависит от параметров τ, h, и устойчивость означает, что соотношение (100) выполняется при любых сколь угодно малых τ, h.

Чтобы придать соотношениям (98), (99), (100) точный смысл, необходимо указать способ оценки входящих в них величин u, f, lU и т. д., т. е. ввести *норму* этих сеточных функций: $\|u\|$ и т. д. В рассмотренном выше примере мы в качестве нормы использовали максимум модуля значений функции, т. е.

$$\delta u = O(\tau, h)$$

означало

$$\|\delta u\| = \max_{k,\, n} |\delta u_k^n| = O(\tau, h).$$

Возможны и другие определения нормы, важно лишь, чтобы она удовлетворяла нас как способ измерения действительной величины сеточной функции. Так, норма $\|\delta u\| = \max\limits_{k,\, n} h\,|\delta u_k^n|$ не годится, поскольку в этом случае $\|\delta u\| \to 0$ при $h \to 0$, даже если δu_k^n остаются конечными. Всякая норма есть обобщение понятия абсолютной величины числа и должна удовлетворять соотношениям

$$\|u + v\| \leqslant \|u\| + \|v\|, \quad \|\alpha u\| = |\alpha|\,\|u\|$$

(где α — число).

Мы нигде не предполагали, что сеточные функции u, f, фигурирующие в (94), (98), (100), скалярны. Если исходная задача состоит в интегрировании системы дифференциальных уравнений, решение которой является системой функций, т. е. вектор-функцией, то, очевидно, под u, f, lu следует понимать сеточные вектор-функции. При определении нормы это должно быть учтено.

Итак, для любой линейной разностной задачи вида (94) сходимость

$$\|u - U\| \to 0 \qquad \text{при} \quad \tau,\, h \to 0 \tag{101}$$

вытекает из аппроксимации

$$\|lU - f\| \to 0 \qquad \text{при } \tau,\, h \to 0 \tag{102}$$

и устойчивости

$$\|u\| \sim \|f\|.$$

Последнее означает, что при любом f для соответствующе-

го ему решения u справедлива оценка

$$\|u\| \leqslant \text{const}\ \|f\|, \tag{103}$$

где const не зависит от τ и h.

Очевидно, если задача устойчива, то стремление к нулю в (101) и (102) одинаково, т. е. порядок точности решения совпадает с порядком аппроксимации задач.

Проверка аппроксимации даже для сложных задач не вызывает затруднений. Используя гладкость точного решения и дифференциальные уравнения, которым оно удовлетворяет, тем же способом, что и в рассмотренном выше простейшем примере, определяем величину $lU - f$.

При исследовании устойчивости большую роль играет специфика задачи. Прямая оценка решения u через правую часть f, как это было сделано в примере, удается редко. Для широкого класса задач более перспективным оказывается способ проверки устойчивости с помощью частных решений, обобщающий прием, которым в том же примере была выявлена неустойчивость. Ниже мы вернемся к этому вопросу.

Перейдем к нелинейным задачам. Чтобы представить характер возникающих здесь трудностей, начнем с примера — задачи

$$\left.\begin{aligned} \frac{\partial U}{\partial t} + \left(\frac{\partial U}{\partial x}\right)^2 &= F(x,\ t), \\ U(x,\ 0) &= \Phi(x). \end{aligned}\right\} \tag{104}$$

Для ее решения естественно использовать разностную схему

$$\left.\begin{aligned} \frac{u_k^{n+1} - u_k^n}{\tau} + \left(\frac{u_{k+1}^n - u_k^n}{h}\right)^2 &= f_k^n, \\ u_k^0 &= \varphi_k. \end{aligned}\right\} \tag{105}$$

Эта «естественность» основывается, конечно, на факте аппроксимации задачи (104) задачей (105), на том, что решение первой U почти удовлетворяет второй. Действительно, используя гладкость $U(x,\ t)$ и уравнение (104), без труда получаем

$$\left.\begin{aligned} \frac{U_k^{n+1} - U_k^n}{\tau} + \left(\frac{U_{k+1}^n - U_k^n}{h}\right)^2 &= f_k^n + O(\tau,\ h), \\ U_k^0 &= \varphi_k. \end{aligned}\right\} \tag{106}$$

Для оценки отклонения u от U подставляем в (105) $u = U + \delta u$ и, учитывая (106), получаем

$$\left.\begin{aligned} \frac{\delta u_k^{n+1} - \delta u_k^n}{\tau} + 2\,\frac{U_{k+1}^n - U_k^n}{h}\cdot\frac{\delta u_{k+1}^n - \delta u_k^n}{h} + \\ + \left(\frac{\delta u_{k+1}^n - \delta u_k^n}{h}\right)^2 = O(\tau,\ h), \\ \delta u_k^0 = 0. \end{aligned}\right\} \qquad (107)$$

Наличие квадратичного члена не позволяет эффективно оценить поведение δu. Допустим, однако, что сходимость есть, и δu мало, например, $\delta u \ll U$. В этом случае поведение δu будет определяться главными линейными членами (107), т. е. уравнением

$$\frac{\delta u_k^{n+1} - \delta u_k^n}{\tau} + 2\,\frac{U_{k+1}^n - U_k^n}{h}\,\frac{\delta u_{k+1}^n - \delta u_k^n}{h} = O(\tau,\ h), \qquad (108)$$

которое с точностью до коэффициента совпадает с рассмотренным ранее (82). Чтобы сделать это совпадение полным и тем самым иметь возможность использовать уже полученные результаты, ограничимся рассмотрением (108) лишь в некоторой окрестности данной точки x_k, t^n. В этой окрестности

$$2\,\frac{U_{k+1}^n - U_k^n}{h} \sim 2\left(\frac{\partial U}{\partial x}\right)_k^n = a,$$

и (108) можно заменить на (82):

$$\frac{\delta u_k^{n+1} - \delta u_k^n}{\tau} + a\,\frac{\delta u_{k+1}^n - \delta u_k^n}{h} = O(\tau,\ h).$$

Как было установлено, решение этого уравнения остается малой величиной лишь при выполнении условия (84), которое сейчас означает

$$-1 \leqslant 2\,\frac{U_{k+1}^n - U_k^n}{h}\,\frac{\tau}{h} \leqslant 0. \qquad (109)$$

Итак, предположение о сходимости не приводит к противоречию только в случае выполнения условий (109).

Мы не можем утверждать, что вопрос о сходимости для задачи (105) полностью решен, но и полученное, необходимое условие (109) дает довольно много. В частности,

если точное решение имеет в некоторой области положительную производную, $\frac{\partial U}{\partial x} > 0$, то разностные уравнения (105) использовать нельзя. Для проверки условия (109) вместо значений U, которые неизвестны, очевидно, следует брать u_k^n, получаемые в расчете.

Повторим проведенные рассуждения применительно к общей нелинейной разностной задаче, которую запишем в виде

$$\mathscr{L}(u) = 0. \tag{110}$$

Поскольку оператор $\mathscr{L}$ нелинеен, то нет смысла выделять правую часть f.

Прежде всего проверяем аппроксимацию, т. е. выполнение условия

$$\mathscr{L}(U) \to 0 \qquad \text{при } \tau,\ h \to 0. \tag{111}$$

Если оно не выполнено, то надеяться на сходимость, очевидно, оснований нет.

Можно было бы обобщить понятие устойчивости и на нелинейные задачи, определив его соотношением

$$u - U \sim \mathscr{L}(u) - \mathscr{L}(U). \tag{112}$$

Но ввиду практической непроверяемости последнего, это вряд ли имеет смысл, хотя для сходимости недостает именно соотношения (112), поскольку малость $\mathscr{L}(u) - \mathscr{L}(U)$ обеспечивается аппроксимацией.

Оператор $\mathscr{L}$ нелинеен, и его свойства на различных функциях могут быть различны. Естественно, в первую очередь нас интересуют функции, близкие к точному решению. Если уже здесь свойства $\mathscr{L}$ окажутся неудовлетворительными, то сходимости ожидать нельзя. Но для малых $u - U = \delta u$ информацию о $\mathscr{L}(u)$ можно получить, рассматривая *главную линейную часть* его, т. е.

$$\mathscr{L}(U) + \mathscr{L}'(U)\,\delta u, \tag{113}$$

где $\mathscr{L}'(U)$ — линейный оператор (вариация $\mathscr{L}$), действующий на δu и зависящий от U как от параметра. Заметим, что используемая при этом малость δu формально не имеет отношения к малости τ, h.

Итак, подставляя в (110) $u = U + \delta u$, заменяя $\mathscr{L}(U + \delta u)$ на (113), учитывая (111), получаем

$$\mathscr{L}'(U)\,\delta u \to 0 \qquad \text{при } \tau,\ h \to 0. \tag{114}$$

Если эта задача, порождаемая линейным оператором $\mathcal{L}'(U)$, устойчива, то можно ожидать, что сходимость имеет место. Если неустойчива, то сходимости нет.

Рассматривая оператор $\mathcal{L}'(U)$ *локально*, в небольшой области плоскости x, t, где $U(x, t)$ меняется мало, можно еще более упростить задачу, сведя ее к исследованию устойчивости линейной задачи с постоянными коэффициентами. Упрощая, моделируя задачу, важно не потерять какой-либо существенной черты ее. Это требует определенного искусства, максимального учета специфики задачи.

Подведем итог. Первый необходимый этап исследования сходимости любой разностной задачи состоит в проверке условия аппроксимации. Если результат положительный, то *линеаризацией* задачи, с последующим «*замораживанием*» коэффициентов, вопрос сводится к анализу устойчивости линейной разностной схемы с постоянными коэффициентами. Коротко, этот прием упрощения, моделирования задачи можно изобразить формулой

$$\mathcal{L}(u) \to \mathcal{L}'(U)\,\delta u \to l\delta u. \tag{115}$$

Задачи

1. Доказать непосредственной оценкой, что задача (108) устойчива при выполнении условий (109).

2. Найти условия устойчивости (сходимости) для разностных схем, аппроксимирующих задачи

$$\frac{\partial U}{\partial t} = U\frac{\partial^2 U}{\partial x^2}, \quad U(0, x) = U_0(x)$$

и

$$\frac{dU}{dt} = F(t, U), \qquad U(0) = U_0.$$

3. Построить разностные схемы, аппроксимирующие задачи

$$\frac{\partial U}{\partial t} = \frac{\partial}{\partial x}\mu(U)\frac{\partial U}{\partial x}, \qquad U(0, x) = U_0(x)$$

и

$$\frac{\partial U}{\partial t} + \frac{\partial p(V)}{\partial x} = 0, \qquad U(0, x) = U_0(x),$$

$$\frac{\partial V}{\partial t} + \frac{\partial U}{\partial x} = 0, \qquad V(0, x) = V_0(x).$$

Произвести линеаризацию полученных разностных уравнений.

§ 6. СПЕКТРАЛЬНЫЙ ПРИЗНАК УСТОЙЧИВОСТИ

Для линейных разностных задач вида $lu = f$ устойчивость означает, что $u \sim f$, т. е. порядок решения и порядок правой части при τ, $h \to 0$ совпадают. Ограничимся рассмотрением операторов l слоистой структуры вида

$$lu = \begin{cases} \dfrac{u^{n+1} - Ru^n}{\tau}, & n = 0, 1, \ldots, \\ u^0. \end{cases} \tag{116}$$

Здесь u^n обозначает *сеточную функцию на n-м слое*, т. е. набор u_k^n для фиксированного n, R — некоторый линейный оператор, переводящий функцию на слое в функцию на слое и зависящий от параметров τ, h. Для операторов l вида (116) задача $lu = f$ может быть записана в форме

$$\left.\begin{aligned} u^{n+1} &= Ru^n + \tau f^n, \quad n = 0, 1, \ldots, \\ u^0 &= v, \end{aligned}\right\} \tag{117}$$

где f^n, v — заданные сеточные функции на слоях. В этом и заключается расслоение оператора l — сеточные функции u^1, u^2, ... могут быть получены последовательно, одна за другой, с помощью одного и того же оператора R, *оператора перехода от слоя к слою*.

Наша задача состоит в выяснении соотношения между величинами u и f, v. Для любого N формулы (117) позволяют выразить u^N через f^n ($n = N - 1$, $N - 2$, ...) и v. Действительно,

$$\begin{aligned} u^N &= \tau f^{N-1} + Ru^{N-1} = \\ &= \tau f^{N-1} + R(\tau f^{N-2} + Ru^{N-2}) = \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ &= \tau(f^{N-1} + Rf^{N-2} + R^2 f^{N-3} + \ldots + R^{N-1} f^0) + R^N v = \\ &= \tau \sum_{m=0}^{N-1} R^m f^{N-m-1} + R^N v, \end{aligned} \tag{118}$$

где R^m означает *m-ю степень оператора R* и также является линейным оператором.

Для оценки величины сеточных функций на слое v, u^n, f^n, Rv, ... введем какую-либо норму (например $\|u^n\| = \max\limits_k |u_k^n|$). Поскольку u^N выражается через f^n, v с помощью

R^m, то, очевидно, соотношение между величинами u^N и f^n, v будет зависеть от метрических свойств операторов R^m, от того, насколько применение R^m меняет норму сеточных функций. Пусть операторы R^m таковы, что для любой сеточной функции на слое v справедливо неравенство

$$\|R^m v\| \leqslant \rho_m \|v\|, \tag{119}$$

где ρ_m — числа, уменьшить которые, без нарушения (119) хотя бы для одной функции v, нельзя (ρ_m называется *нормой оператора* R^m).

Оценим величину u^N, используя (118), (119),

$$\|u^N\| \leqslant \tau \sum_{m=0}^{N-1} \|R^m f^{N-m-1}\| + \|R^N v\| \leqslant$$

$$\leqslant \tau \sum_{m=0}^{N-1} \rho_m \|f^{N-m-1}\| + \rho_N \|v\| \leqslant$$

$$\leqslant \tau N \max_m \rho_m \max_m \|f^m\| + \rho_N \|v\|. \tag{120}$$

Поскольку нас интересуют только конечные t, то

$$0 \leqslant m\tau < N\tau \leqslant t \text{ и } 0 \leqslant m < N \leqslant t/\tau.$$

Устойчивость будет иметь место, если коэффициенты при $\max \|f^m\|$ и $\|v\|$ в (120) будут оставаться ограниченными при τ, $h \to 0$. Поскольку $\tau N \leqslant t$, то условие устойчивости записывается в виде

$$\rho_m \leqslant \text{const при } \tau, h \to 0, \ m\tau \leqslant t. \tag{121}$$

Разумеется, const не должна зависеть от τ, h, хотя ρ_m, как и R, R^m, зависят от этих параметров. В этом существо дела, так как для каждых фиксированных τ, h, m величина ρ_m, естественно, конечна.

Итак, для операторов l слоистой структуры (116) вопрос об устойчивости разностной задачи сводится к оценке норм степеней оператора R — к проверке условия (121).

Рассмотренная в § 4 задача дает, очевидно, пример оператора l слоистой структуры. Там мы, с одной стороны, установили выполнение (121) при условии $-1 \leqslant a\tau/h \leqslant 0$ непосредственной оценкой, а с другой стороны, выявили

неустойчивость с помощью частного решения $u_k \sim (-1)^k$. В более сложных случаях первое удается далеко не всегда, в то время как второй способ допускает обобщение на широкий класс задач.

Поведение $R^m v$ в зависимости от m проще всего исследовать на *собственных функциях* оператора R, т. е. сеточных функциях на слое v, применение к которым R тождественно умножению на числа (*собственные значения* λ)

$$Rv = \lambda v. \tag{122}$$

Для собственных функций $R^m v = \lambda^m v$ и, если их достаточно много, то по величине λ^m можно судить о норме ρ_m.

Рассмотрим сеточные функции на слое $v = \{v_k\}$, определенные и ограниченные для всех значений дискретного аргумента k, $-\infty < k < \infty$. Пусть линейный оператор R задается на этих функциях формулой

$$(Rv)_k = \sum_p \alpha_p v_{k+p}, \qquad k = 0, \pm 1, \pm 2, \ldots, \tag{123}$$

где α_p — заданные коэффициенты, зависящие от параметров τ, h, а p пробегает некоторое множество значений. В частности, рассмотренный в § 4 пример получается при $\alpha_0 = 1 + a\tau/h$, $\alpha_1 = -a\tau/h$ и $\alpha_p = 0$ для остальных p.

Собственные функции оператора R вида (123) должны удовлетворять соотношению (122), т. е.

$$\sum_p \alpha_p v_{k+p} = \lambda v_k, \qquad k = 0, \pm 1, \pm 2, \ldots \tag{124}$$

Решение этого линейного разностного уравнения будем искать в виде

$$v_k = v_0 q^k, \tag{125}$$

где q — некоторое число, а v_0 — нормировочный множитель. Подставляя (125) в (124), получаем, после сокращения на $v_0 q^k$,

$$\sum_p \alpha_p q^p = \lambda, \tag{126}$$

т. е. при любом q функция v (125) удовлетворяет (122), с $\lambda = \lambda(q)$ (126). Из этого обилия функций мы отберем только ограниченные (по k) сеточные функции (125). Если $|q| \neq 1$, то $|v_k| \to \infty$ либо при $k \to \infty$, либо при $k \to -\infty$. Следовательно, $|q| = 1$.

Саму разностную задачу мы рассматриваем только в действительной области. Однако всякую действительную функцию можно представить в виде комбинации комплексных. Поэтому привлечение последних может дать эффективную информацию о метрических свойствах оператора R в действительной области, в частности, о его норме. В данном случае мы имеем всего две действительные собственные функции, при $q = 1$ и $q = -1$. Комплексных же собственных функций существенно больше, и характер оператора проявляется именно на них. Полагая $q = e^{i\varphi}$, перепишем (125), (126) в виде

$$v_k = v_0 e^{ik\varphi}, \tag{127}$$

$$\lambda = \sum_p \alpha_p e^{ip\varphi}. \tag{128}$$

Таким образом, каждому φ из интервала $(0, 2\pi)$ соответствует собственная функция v_k (127) с собственным значением λ (128). Очевидно, величина v_0 не играет роли, и можно положить $v_0 = 1$.

Поскольку для собственных функций $R^m v = \lambda^m v$, то на них

$$\|R^m v\| = \|\lambda^m v\| \leqslant \max_\varphi |\lambda|^m \|v\|, \tag{129}$$

и величина $\max |\lambda|^m$ является, так сказать, нормой оператора R^m на системе собственных функций. Эта система является частью всего множества сеточных функций, поэтому норма оператора R^m может быть только больше, чем $\max |\lambda|^m$,

$$\max_\varphi |\lambda|^m \leqslant \rho_m. \tag{130}$$

Сравнивая это неравенство с условием устойчивости (121), приходим к выводу, что для выполнения последнего, во всяком случае, необходимо, чтобы

$$\max_\varphi |\lambda|^m \leqslant \text{const} \qquad \text{при } \tau,\ h \to 0,\ m\tau \leqslant t. \tag{131}$$

Собственные значения λ, как и α_p, через которые они выражаются с помощью (128), зависят от параметров τ, h. Поскольку $m \sim 1/\tau \to \infty$, то условие (131) эквивалентно требованию

$$\max_\varphi |\lambda| \leqslant 1 + O(\tau) \qquad \text{при } \tau,\ h \to 0. \tag{132}$$

В противном случае, $|\lambda|^m \to \infty$. Если же $|\lambda| = 1 + c\tau$,

$$|\lambda|^n \sim (1 + c\tau)^{t/\tau} \sim e^{ct}.$$

Последняя величина, хоть и конечна, но может быть довольно большой. Поэтому иногда вместо (132) рассматривают более сильное условие

$$\max_{\varphi} |\lambda| \leqslant 1 \qquad \text{при } \tau,\ h \to 0. \tag{133}$$

Итак, мы получили необходимое условие устойчивости задачи (117) с оператором R вида (123). Множество собственных значений λ называют спектром оператора, а величину $\max |\lambda|$ — спектральным радиусом. В связи с этим условие (132) или (133) называют *спектральным признаком устойчивости*.

Для примера § 4 равенство (128) дает

$$\lambda = 1 + a\frac{\tau}{h} - a\frac{\tau}{h}e^{i\varphi},$$

т. е. спектр есть окружность (в комплексной плоскости λ) радиуса $|a\tau/h|$ с центром в точке $1 + a\tau/h$. Нетрудно проверить, что $|\lambda| \leqslant 1$ лишь в случае выполнения все тех же неравенств

$$-1 \leqslant a\frac{\tau}{h} \leqslant 0.$$

Достоинством спектрального признака устойчивости является то, что он легко распространяется на многие более сложные задачи, в частности, на системы уравнений. Вернемся к задаче (116), (117) и будем считать, что сеточные функции u^n, f^n, v являются сеточными вектор-функциями, т. е. определяются в каждой расчетной точке несколькими величинами. Все проведенные рассуждения сохраняют силу, и лишь заключительная часть подлежит корректировке. А именно, поскольку (123) есть теперь векторное равенство, то α_p следует считать квадратными матрицами того же порядка, что и векторы v_k. Подстановка

$$v_k = v_0 q^k = v_0 e^{ik\varphi}, \tag{134}$$

где v_0 — вектор, в (124) приводит к

$$\sum_p \alpha_p e^{ip\varphi} v_0 = \lambda v_0$$

— системе линейных однородных уравнений относительно компонент вектора v_0. Условием существования решения ее является равенство нулю детерминанта

$$\left|\sum_p \alpha_p e^{ip\varphi} - \lambda I\right| = 0, \tag{135}$$

где I — единичная матрица. Все отличие от скалярного случая состоит в том, что теперь спектр состоит из нескольких ветвей $\lambda_1, \lambda_2, \ldots$, определяемых как корни уравнения (135) вместо (128).

Итак, спектральный признак позволяет свести исследование устойчивости к задаче об оценке величины корней алгебраического уравнения.

Он дает, вообще говоря, лишь необходимое условие устойчивости. Однако, если для некоторого класса функций, в некоторой норме, система собственных функций является *полной*, т. е. любую функцию из этого класса можно аппроксимировать комбинацией собственных, то спектральный признак будет и достаточным условием устойчивости. Рассмотрим, например, скалярный случай оператора R вида (123) и ограничимся сеточными функциями $u = \{u_k\}$, для которых ряд

$$\sum_k u_k e^{-ik\varphi} = w(\varphi) \tag{136}$$

является сходящимся. Каждая сеточная функция u_k порождает, таким образом, периодическую функцию $w(\varphi)$, для которой ряд (136) есть ряд Фурье, а u_k — коэффициенты этого ряда. Последние, как известно, удовлетворяют соотношениям

$$u_k = \frac{1}{2\pi}\int_0^{2\pi} w(\varphi)\, e^{ik\varphi}\, d\varphi \tag{137}$$

и

$$\sum_k u_k^2 = \frac{1}{2\pi}\int_0^{2\pi} |w(\varphi)|^2\, d\varphi. \tag{138}$$

В справедливости (137), (138) можно убедиться и непосредственно. Для этого нужно умножить (136) на $e^{im\varphi}$ в первом случае, на $\overline{w}(\varphi) = \sum_m u_m e^{im\varphi}$ — во втором и выполнить интегрирование.

Применим к (137) оператор R (123), получим

$$(Ru)_k = \sum_p \alpha_p \frac{1}{2\pi} \int_0^{2\pi} w(\varphi) e^{i(k+p)\varphi} d\varphi =$$

$$= \frac{1}{2\pi} \int_0^{2\pi} w(\varphi) e^{ik\varphi} \sum_p \alpha_p e^{ip\varphi} d\varphi =$$

$$= \frac{1}{2\pi} \int_0^{2\pi} w(\varphi) \lambda(\varphi) e^{ik\varphi} d\varphi,$$

последнее в силу (128). Сравнивая это равенство с (137), замечаем, что $(Ru)_k$ являются коэффициентами ряда Фурье функции $w(\varphi)\lambda(\varphi)$. Следовательно, для них справедливо соотношение вида (138), т. е.

$$\sum_k (Ru)_k^2 = \frac{1}{2\pi} \int_0^{2\pi} |w(\varphi)\lambda(\varphi)|^2 d\varphi.$$

Вынося $\max\limits_\varphi |\lambda(\varphi)|^2$ за знак интеграла и заменяя оставшийся интеграл на $\sum\limits_k u_k^2$, получаем

$$\sum_k (Ru)_k^2 \leqslant \max_\varphi |\lambda(\varphi)|^2 \sum_k u_k^2. \tag{139}$$

Определим норму сеточной функции равенством

$$\|u\| = \left(\sum_k u_k^2 h\right)^{1/2}, \tag{140}$$

где множитель h введен для того, чтобы в пределе, при $h \to 0$, норма сохраняла смысл. Используя (140), перепишем (139) в виде

$$\|Ru\| \leqslant \max_\varphi |\lambda| \, \|u\|. \tag{141}$$

Последнее означает, что на указанном классе функций, в норме (140), величина $\max|\lambda|$ не меньше нормы оператора R и, следовательно, спектральный признак является достаточным признаком устойчивости.

Остановимся теперь на одном вопросе, существенном для практического применения спектрального признака

устойчивости. Во всякой вычислительно реальной задаче количество расчетных точек конечно, и индекс k пробегает конечное множество значений. В граничных точках оператор R не может сохранять свой стандартный вид (123), хотя бы из-за отсутствия полного набора величин u_{k+p}. В этих точках оператор R выражается с помощью особых формул, реализующих то или иное граничное условие. Чтобы представить, насколько такое искажение может менять свойства стандартного, «безграничного» оператора R (123), обратимся опять к примеру § 4.

Пусть оператор R задан на сеточных функциях u_k ($k = 0, 1, 2, \ldots, K$) формулами

$$\left.\begin{aligned} &(Ru)_k = \left(1 + a\frac{\tau}{h}\right) u_k - a\frac{\tau}{h} u_{k+1}, \quad k=0, 1, \ldots, K-1, \\ &(Ru)_K = 0. \end{aligned}\right\} \tag{142}$$

Очевидно, это соответствует дифференциальной задаче на конечном интервале $0 \leqslant x \leqslant X = Kh$ с нулевым граничным условием на правом конце, решение которой, при $a < 0$, существует и единственно.

Найдем собственные функции оператора R (142). Полагая $Rv = \lambda v$, приходим к системе линейных однородных уравнений относительно $v_0, v_1, \ldots, v_k$:

$$\left(\lambda - 1 - a\frac{\tau}{h}\right) v_k + a\frac{\tau}{h} v_{k+1} = 0, \qquad k = 0, 1, \ldots, K-1,$$
$$\lambda v_K = 0.$$

Нетрудно убедиться, что она имеет лишь два нетривиальных решения

$$v_k = \left(\frac{1 + a\frac{\tau}{h}}{a\frac{\tau}{h}}\right)^k \qquad \text{при } \lambda = 0$$

и

$$v_0 = 1, \; v_1 = v_2 = \ldots = v_K = 0 \quad \text{при } \lambda = 1 + a\frac{\tau}{h}.$$

Хотя второе и порождает условие $\left|1 + a\frac{\tau}{h}\right| \leqslant 1$, но ясно, что столь бедный запас собственных функций не дает возможности судить по ним о свойствах оператора.

В то же время представляется маловероятным, чтобы искажение оператора лишь в одной крайней точке могло

кардинально изменить его свойства, так отчетливо проявившиеся на функциях $e^{ik\varphi}$. Последние не удовлетворяют условию $u_K = 0$, поэтому «исправим» их, т. е. положим

$$\left.\begin{aligned} v_k &= e^{ik\varphi}, \qquad k = 0, 1, \ldots, K-1, \\ v_K &= 0 \end{aligned}\right\} \tag{143}$$

и применим к ним оператор R (142). Очевидно, для всех $k < K - 1$ функция (143) перейдет в себя, приобретя множитель

$$\lambda = 1 + a\,\frac{\tau}{h} - a\,\frac{\tau}{h}\,e^{i\varphi}, \tag{144}$$

и лишь в приграничной точке этот закон нарушится. К полученной функции опять применим оператор R (142) — влияние исправления распространится и на точку $K - 2$. Продолжая этот процесс, получим, что для функции v_k (143)

$$(R^m v)_k = \lambda^m v_k \qquad \text{при } k = 0, 1, \ldots, K - m. \tag{145}$$

Нас интересует R^m при τ, $h \to 0$, т. е. при $m \sim 1/\tau \to \infty$, $K \sim 1/h \to \infty$. Если при этом h/τ остается ограниченным, т. е. m стремится к бесконечности не быстрее K, то всегда найдется интервал значений k, где (145) справедливо. Следовательно, нижняя оценка для нормы $\|R^m\|$ через $|\lambda|^m$, где λ дается (144), т. е. является собственным значением невозмущенного оператора, сохраняет смысл.

Можно рассуждать и иначе, рассматривая вместо (143) функцию

$$v_k = \left(1 - \frac{k}{K}\right) e^{ik\varphi}. \tag{146}$$

Используя (142), (144), найдем, что в этом случае

$$(Rv - \lambda v)_k = \frac{a\tau}{hK}\,e^{i(k+1)\varphi}, \quad k = 0, 1, \ldots, K-1,$$

$$(Rv - \lambda v)_K = 0.$$

Поскольку $hK = X = \text{const}$, то последние равенства означают, что

$$Rv - \lambda v = O(\tau), \tag{147}$$

хотя сама функция v (146) конечна ($v_0 = 1$). Можно сказать, что v (146) является «почти собственной» функцией, а λ (144) — «почти собственным» значением оператора R (142). И мы опять приходим к выводу, что наличие гранич-

ных условий следует рассматривать как возмущение оператора, при котором сохраняются многие существенные свойства его. Разумеется, задав «дикие» граничные условия, можно испортить стандартный оператор, но исправить его недостатки с помощью граничных условий нельзя.

Изучение этих вопросов в общих случаях не столь просто. Однако проведенные рассуждения убеждают нас в том, что спектральный признак, примененный к стандартному, безграничному оператору, остается необходимым условием устойчивости и для задач на ограниченном интервале.

Задачи

1. Из определения нормы оператора (119) следует

$$\| R^m u \| = \| R R^{m-1} u \| \leqslant \rho_1 \| R^{m-1} u \| \leqslant \ldots \leqslant \rho_1^m \| u \|,$$

т. е. $\rho_m \leqslant \rho_1^m$. Следовательно, для выполнения (121) достаточно, чтобы $\rho_1^m \leqslant \text{const}$. Это эквивалентно условию на норму оператора R

$$\rho_1 \leqslant 1 + O(\tau), \tag{148}$$

которое является, тем самым, достаточным условием устойчивости.

Показать, что для операторов R вида (123), в норме $\| u \| = \max_k | u_k |$, условие (148) равносильно условию

$$\sum_p | \alpha_p | \leqslant 1 + O(\tau). \tag{149}$$

Для оператора

$$(Ru)_k = u_k - \frac{a\tau}{2h} (u_{k+1} - u_{k-1})$$

исследовать устойчивость с помощью (149) и спектрального признака. Сравнить результаты. Определить дифференциальную задачу, которой соответствует этот оператор.

2. Для задачи

$$\frac{\partial U}{\partial t} + \frac{\partial V}{\partial x} = 0, \qquad U(0, x) = U_0(x),$$

$$\frac{\partial V}{\partial t} + \frac{\partial U}{\partial x} = 0, \qquad V(0, x) = V_0(x)$$

рассмотреть различные разностные схемы и исследовать их устойчивость.

3. Построить разностную схему для решения задачи

$$\frac{\partial U}{\partial t} + \frac{\partial f(U)}{\partial x} = \frac{\partial}{\partial x} \mu(U) \frac{\partial U}{\partial x}, \qquad U(0, x) = U_0(x),$$

где $f(U)$, $\mu(U)$ — заданные функции. Исследовать аппроксимацию и устойчивость (на линейной модели).

§ 7. ПОСТРОЕНИЕ РАСЧЕТНЫХ ФОРМУЛ

Теперь, когда у нас есть определенная ясность — каким требованиям должна удовлетворять разностная задача, перейдем к вопросу о способах построения ее. Во всех конкретных примерах мы просто заменяли каждую производную, входящую в исходные дифференциальные уравнения, соответствующим разностным отношением. Однако это не единственный и не самый короткий путь.

Составление разностной задачи начинается с выбора расчетной сетки — дискретного множества точек, заменяющего непрерывную область изменения независимых переменных.

В принципе это множество может быть произвольным. Можно увеличивать концентрацию точек на наиболее важных участках, с целью получения здесь большей точности. Можно задавать закон построения сетки, зависящий от решения, получаемого в процессе расчета. Но если нет специальных показаний, лучше брать сетку регулярной, определяемой минимальным числом параметров. Это существенно облегчает исследование разностной задачи.

Пусть расчетная сетка или закон ее образования заданы. Если неравномерность сетки выражена слабо, т. е. ее параметры мало меняются от точки к точке, то на небольших участках она может быть смоделирована равномерной. В дальнейших рассмотрениях мы будем иметь в виду регулярную (не обязательно прямоугольную) сетку, определяемую для задач с двумя независимыми переменными x, t всего двумя параметрами h, τ — шагами сетки. На этой сетке определяются функции (или вектор-функции) u_k^n, f_k^n, φ_k, ... — сеточные функции дискретных аргументов k, n (номеров точек).

Следующий этап — построение разностных уравнений, т. е. арифметических соотношений между величинами τ, h, u_k^n, f_k^n, φ_k, ... Поскольку мы исходим из принципа сходимости решения дискретной задачи к решению дифференциальной, то разностные уравнения должны удовлетворять определенным требованиям. Выше мы сформулировали их в виде условий аппроксимации и устойчивости.

Используя формулы численного дифференцирования для замены производных конечными разностями, легко написать те или иные соотношения, которые в пределе,

для любой гладкой функции, будут переходить в исходные дифференциальные уравнения. Однако успех здесь приходит не сразу, так как большинство логически возможных разностных схем оказываются неустойчивыми.

Поиск удовлетворительной разностной схемы можно сделать более эффективным с помощью приема, который мы продемонстрируем сначала все на той же простой задаче

$$\left.\begin{aligned} \frac{\partial U}{\partial t} + a\frac{\partial U}{\partial x} = 0, \\ U(x, 0) = U_0(x). \end{aligned}\right\} \qquad (150)$$

Зададимся формой *расчетной ячейки*, т. е. укажем точки, значения сеточной функции в которых мы хотим связать разностными соотношениями. Для данной задачи, на сетке $x_k = kh$, $t^n = n\tau$, в качестве такой ячейки (рис. 7) возьмем четыре точки с номерами $(k, n+1)$, $(k-1, n)$, (k, n), $(k+1, n)$.

Рис. 7.

Исходная задача (150) линейна. Естественно разностные уравнения также строить в виде линейных соотношений. Общий вид такого соотношения между значениями сеточной функции в указанных точках есть

$$u_k^{n+1} = \alpha_{-1}u_{k-1}^n + \alpha_0 u_k^n + \alpha_1 u_{k+1}^n = \sum_{p=-1}^{1} \alpha_p u_{k+p}^n. \qquad (151)$$

Полагая $u_k^0 = U_0(x_k)$, получаем, при любом наборе α_p, некоторую разностную задачу. Подберем коэффициенты α_p так, чтобы задача (151) аппроксимировала исходную и была устойчивой.

Если для проверки устойчивости мы намерены использовать спектральный признак, то задачу (151) следует представить в стандартной форме $lu = f$, положив

$$lu = \begin{cases} \dfrac{u_k^{n+1} - \sum\limits_p \alpha_p u_{k+p}^n}{\tau}, \\ u_k^0, \end{cases} \qquad f = \begin{cases} 0, \\ U_0(x_k). \end{cases} \qquad (152)$$

Как мы знаем, необходимое условие устойчивости записывается в этом случае в виде неравенства

$$\left|\sum_p \alpha_p e^{ip\varphi}\right| \leqslant 1. \tag{153}$$

Таким образом, одно условие на коэффициенты α_p у нас уже есть.

Перейдем к аппроксимации, которая означает, что $lU - f \to 0$ при $\tau, h \to 0$. Для определения порядка $lU - f$ воспользуемся, как всегда, разложением решения (150) — функции $U(x, t)$ в ряд по степеням τ, h в окрестности центральной точки расчетной ячейки x_k, t^n. Имеем

$$\left.\begin{aligned} U_k^{n+1} &= U + \tau U_t + \frac{\tau^2}{2} U_{tt} + \frac{\tau^3}{6} U_{ttt} + \ldots, \\ U_{k+p}^n &= U + phU_x + \frac{p^2h^2}{2} U_{xx} + \frac{p^3h^3}{6} U_{xxx} + \ldots, \end{aligned}\right\} \tag{154}$$

где U, U_t, ... берутся в центральной точке. Поскольку $U(x, t)$ — решение уравнения (150), то для него справедливы равенства

$$\left.\begin{aligned} U_t &= -aU_x, \\ U_{tt} &= a^2U_{xx}, \\ U_{ttt} &= -a^3U_{xxx}, \\ &\ldots\ldots\ldots\ldots \end{aligned}\right\} \tag{155}$$

получающиеся дифференцированием (150).

Используя (154), (155), составим необходимую нам комбинацию значений U:

$$\begin{aligned} U_k^{n+1} - \sum_p \alpha_p U_{k+p}^n = \left(1 - \sum \alpha_p\right) U - \left(\tau a + h \sum p\alpha_p\right) U_x + \\ + \frac{1}{2}\left(\tau^2a^2 - h^2 \sum p^2\alpha_p\right) U_{xx} - \\ - \frac{1}{6}\left(\tau^3a^3 + h^3 \sum p^3\alpha_p\right) U_{xxx} + \ldots, \end{aligned} \tag{156}$$

отличающуюся от $lU - f$ лишь множителем $1/\tau$. Порядок аппроксимации будет зависеть от порядка первых ненулевых членов этого разложения. Величины U, U_x, U_{xx},... следует считать произвольными и независимыми. Приравнивая коэффициенты при них нулю, получим цепочку

равенств

$$\left.\begin{aligned} 1-\sum\alpha_p&=0,\\ \tau a+h\sum p\alpha_p&=0,\\ \tau^2a^2-h^2\sum p^2\alpha_p&=0,\\ \tau^3a^3+h^3\sum p^3\alpha_p&=0,\\ \ldots\ldots\ldots\ldots\ldots\ldots \end{aligned}\right\} \qquad (157)$$

— уравнений для определения α_p, возрастающего порядка по τ, h. Чем большему числу этих уравнений мы сможем удовлетворить, тем выше будет порядок аппроксимации.

В нашем распоряжении всего лишь три неопределенных коэффициента α_p. Поэтому самое большее, чего мы можем добиться,— это удовлетворить трем из уравнений (157). Обозначив $a\tau/h$ через r, выпишем эти уравнения:

$$\alpha_{-1}+\alpha_0+\alpha_1=1, \qquad \alpha_{-1}-\alpha_1=r, \qquad \alpha_{-1}+\alpha_1=r^2.$$

Решая их, получим

$$\alpha_{-1}=\frac{r^2+r}{2}, \qquad \alpha_0=1-r^2, \qquad \alpha_1=\frac{r^2-r}{2}. \qquad (158)$$

При этих значениях α_p первый, отличный от нуля, член в разложении (156) имеет порядок τ^3, rh^3. Следовательно, $lU-f=O(\tau^2, h^2)$, т. е. аппроксимация второго порядка по τ, h.

Остается удовлетворить условию устойчивости (153). Подставим (158) в (153):

$$\begin{aligned}\sum\alpha_p e^{ip\varphi}&=\frac{r^2+r}{2}e^{-i\varphi}+1-r^2+\frac{r^2-r}{2}e^{i\varphi}=\\ &=1-r^2+r^2\cos\varphi-ir\sin\varphi=\\ &=1-2r^2\sin^2\frac{\varphi}{2}-i\,2r\sin\frac{\varphi}{2}\cos\frac{\varphi}{2}.\end{aligned}$$

Следовательно,

$$\begin{aligned}\left|\sum\alpha_p e^{ip\varphi}\right|^2&=\left(1-2r^2\sin^2\frac{\varphi}{2}\right)^2+\left(2r\sin\frac{\varphi}{2}\cos\frac{\varphi}{2}\right)^2=\\ &=1-4r^2\sin^2\frac{\varphi}{2}+4r^4\sin^4\frac{\varphi}{2}+4r^2\sin^2\frac{\varphi}{2}\cos^2\frac{\varphi}{2}=\\ &=1-4r^2\sin^4\frac{\varphi}{2}+4r^4\sin^4\frac{\varphi}{2}=\\ &=1-4r^2(1-r^2)\sin^4\frac{\varphi}{2}.\end{aligned}$$

Очевидно, для выполнения неравенства (153) нужно, чтобы $r^2 \leqslant 1$, т. е.

$$|a|\frac{\tau}{h} \leqslant 1. \tag{159}$$

Это и есть условие устойчивости разностной задачи (151), (158).

Если для нахождения коэффициентов α_p ограничиться первыми двумя уравнениями (157), то полученная схема будет иметь, очевидно, первый порядок аппроксимации $O(\tau, h)$. Поскольку эти два уравнения содержат три неизвестных α_p, то таких схем много. Одна из них была рассмотрена в § 4. Использование лишь первого из уравнений (157) не обеспечивает аппроксимации, так как ошибка в этом случае будет конечной.

Понятно, что описанный способ построения расчетных формул может быть применен и для любой другой задачи Прежде всего, руководствуясь какими-либо априорными соображениями, выбираем форму расчетной ячейки и вид разностных соотношений, содержащих *неопределенные коэффициенты*. Требуя затем выполнения условий аппроксимации и устойчивости, сводим задачу к решению алгебраической системы уравнений и неравенств

При применении способа неопределенных коэффициентов к сложным задачам приходится проводить значительную аналитическую работу.

Для упрощения выкладок следует, не стремясь к наибольшей общности, ограничиваться частными видами разностных соотношений. Например, в рассмотренной задаче можно было вместо (151) искать разностную схему в виде

$$\frac{u_k^{n+1} - u_k^n}{\tau} = c_1 \frac{u_{k+1}^n - u_k^n}{h} + c_2 \frac{u_k^n - u_{k-1}^n}{h} \tag{160}$$

с неопределенными коэффициентами c_1, c_2.

Очевидно, что если существуют разностные уравнения, удовлетворяющие поставленным условиям (форма расчетной ячейки, порядок аппроксимации, вид разностного соотношения), то все они могут быть получены описанным способом.

Коснемся еще одного вопроса. При исследовании разностной схемы (151) мы записали ее в виде $lu = f$, определив l формулой (152), «зачем-то» разделив (151) на τ.

Более естественным представляется определение l равенством

$$lu = u_k^{n+1} - \sum_{p=-1}^{1} \alpha_p u_{k+p}^n. \tag{161}$$

Заметим, что в этом случае (при тех же α_p) мы получим $lU - f = \tau O(\tau^2, h^2)$, т. е. аппроксимацию более высокого порядка. Однако свойства задачи, очевидно, не зависят от обозначений, формы записи и способа исследования ее. Выигрыш в аппроксимации оказывается фиктивным и погашается «более слабой» устойчивостью. А именно, для задачи $lu = f$ с l, определяемым (161), мы, как нетрудно видеть, получим оценку $u \sim f/\tau$, что означает неустойчивость, при нашем определении ее. Тем не менее сходимости это не мешает, поскольку имеется лишний порядок в аппроксимации.

Приведенное замечание указывает на формальную неоднозначность разделения вопроса о сходимости на аппроксимацию и устойчивость.

Опишем еще один способ построения расчетных формул. Он имеет более узкую область применимости, но часто достаточно эффективен.

Наш подход к аппроксимации дифференциальных задач разностными можно охарактеризовать как л о к а л ь н ы й. При построении разностных уравнений все наше внимание сосредоточено на отдельной расчетной ячейке, на окрестности расчетной точки с размерами порядка τ, h. Но, как всякую гладкую функцию можно локально считать линейной, так и всякую задачу, если ее рассматривать в малой области, где решение меняется мало, можно приблизить линейной задачей с постоянными коэффициентами.

Например, вместо уравнения

$$\frac{\partial U}{\partial t} + a(x, t, U)\frac{\partial U}{\partial x} = 0 \tag{162}$$

можно в окрестности точки x_k, t^n рассматривать уравнение

$$\frac{\partial U}{\partial t} + a_k^n \frac{\partial U}{\partial x} = 0, \tag{163}$$

где $a_k^n = a(x_k, t^n, U_k^n)$. Действительно, если решение (гладкое) первого уравнения подставить во второе, то

оно удовлетворится с точностью до $O(\tau, h)$, так как

$$(a(x, t, U(x, t)) - a_k^n) \frac{\partial U}{\partial x} = O(\tau, h).$$

Добавим к сказанному, что исследование устойчивости разностных задач (§ 5) мы проводим на их линейных моделях. Вся теория устойчивости есть, по существу, теория устойчивости линейных разностных задач.

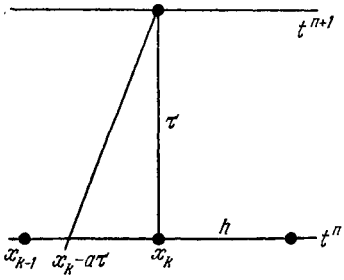

Рис. 8.

Приведенные соображения показывают, что имеет смысл рассматривать вопрос о способах построения расчетных формул, ограничиваясь линейными задачами с постоянными коэффициентами.

Выше, демонстрируя различные положения на примере задачи (150), мы систематически уклонялись от использования того факта, что ее точное решение нам известно и дается явной формулой

$$U(x, t) = U_0(x - at). \tag{164}$$

Используя его, мы рисковали получить утверждения, справедливые только для задач, решение которых известно, и лишить, тем самым, наши рассмотрения смысла. Мы обращались к задаче (150) как представителю некоторого класса задач и использовали те свойства ее, которые типичны для этого класса. Наличие же формулы, дающей явное выражение решения, типично для линейных задач с постоянными коэффициентами.

Поскольку в настоящий момент нас интересуют именно такие задачи, то мы имеем право этот факт использовать. Как это можно сделать, выясним сначала на примере уравнения (163).

Нам нужно получить соотношение, выражающее u_k^{n+1} через величины n-го слоя u_k^n, $u_{k\pm1}^n$, ... Формула точного решения (164) в применении к данному случаю дает

$$U(x_k, t^{n+1}) = U(x_k - a_k^n\tau, t^n). \tag{165}$$

Отсюда вытекает, что для получения u_k^{n+1} нам требуется

значение решения в точке $x = x_k - a_k^n\tau$ n-го слоя (рис. 8). Но на этом слое мы имеем дискретный набор значений u_k^n, $u_{k\pm1}^n$, ...— сеточную функцию. Для вычисления нужного нам значения естественно применить интерполяцию.

Допустим, что

$$x_{k-1} \leqslant x_k - a_k^n\tau \leqslant x_k. \tag{166}$$

Тогда линейная интерполяция дает

$$u^n(x_k - a_k^n\tau) = a_k^n \frac{\tau}{h} u_{k-1}^n + \left(1 - a_k^n \frac{\tau}{h}\right) u_k^n, \tag{167}$$

и, в соответствии с формулой (165), получаем

$$u_k^{n+1} = u_k^n - a_k^n \frac{\tau}{h} (u_k^n - u_{k-1}^n)$$

— уже знакомое нам разностное уравнение, устойчивое при

$$0 \leqslant a_k^n \frac{\tau}{h} \leqslant 1.$$

Мы видим, что условие устойчивости в данном случае совпадает с условием (166), которое обеспечивает интерполяционность формулы (167). Это не удивительно, так как для устойчивости необходим правильный учет области зависимости решения. Именно последнее и является существенной стороной предлагаемого метода.

Общие черты метода теперь ясны, и мы, не стремясь к максимальной общности, изложим их для случая, когда исходная задача состоит в интегрировании системы дифференциальных уравнений в частных производных вида

$$\frac{\partial U}{\partial t} = \mathcal{D}(U) \tag{168}$$

при заданных начальных данных. Здесь U — искомая вектор-функция от x, t, а $\mathcal{D}$ — дифференциальный (по x) оператор, для которого поставленная задача корректна, т. е. ее решение существует, единственно и непрерывно зависит от начальных данных. Такие задачи называют *эволюционными* — они описывают эволюцию во времени некоторого начального состояния.

Первый этап — построение линейной модельной системы, аппроксимирующей локально систему (168).

Оператор $\mathscr{D}$ есть некоторая заданная вектор-функция от векторных аргументов U, U_x, U_{xx}, ...,

$$\mathscr{D}(U) = f(U, U_x, U_{xx}, \ldots).$$

Поэтому в окрестности значений U_k^n, $(U_x)_k^n$, $(U_{xx})_k^n$,... оператор $\mathscr{D}$ можно аппроксимировать линейным выражением

$$\mathscr{D}(U) \sim f_k^n + (f_U)_k^n (U - U_k^n) + (f_{U_x})_k^n (U_x - (U_x)_k^n) + \ldots \tag{169}$$

Здесь f_U, f_{U_x}, ... обозначают матрицы производных компонент вектора f по компонентам векторов U, U_x, ..., а f_k^n, U_k^n, $(U_x)_k^n$,... — значения соответствующих величин в точке x_k, t^n.

Считая решение $U(x, t)$ гладким, т. е. разности $U - U_k^n$, $U_x - (U_x)_k^n$, ... малыми, рассмотрим вместо (168) систему линейных дифференциальных уравнений с постоянными коэффициентами

$$\frac{\partial U}{\partial t} = DU + C, \tag{170}$$

где D — линейный дифференциальный оператор, выражающийся однородной частью (169), а C — константа, объединяющая остальные члены (169):

$$DU = (f_U)_k^n U + (f_{U_x})_k^n U_x + \ldots,$$

$$C = f_k^n - (f_U)_k^n U_k^n - (f_{U_x})_k^n (U_x)_k^n + \ldots$$

Из теории дифференциальных уравнений в частных производных известно, что решение системы линейных уравнений с постоянными коэффициентами (170) можно выразить формулой

$$U(x, t) = QU(x, 0) + Ct, \tag{171}$$

где Q — линейный интегральный оператор

$$QU(x, 0) = \int_{-\infty}^{\infty} q(\xi, t)\, U(x + \xi, 0)\, d\xi, \tag{172}$$

а q — соответствующая данной системе (170) матрица функций. На методах ее получения мы здесь останавли-

ваться не будем. Укажем, что, например, для *уравнения теплопроводности*

$$\frac{\partial U}{\partial t} = a\frac{\partial^2 U}{\partial x^2}, \qquad a > 0, \tag{173}$$

функция q имеет вид

$$q(x, t) = \frac{1}{2\sqrt{\pi a t}} e^{-\frac{x^2}{4at}}. \tag{174}$$

Формулу (164) для решения задачи (150) также можно представить в виде (171), (172), если положить

$$q(x, t) = \delta(x + at),$$

где $\delta(x)$ — дельта-функция.

С помощью (171) мы можем решение в момент $t^{n+1} = t^n + \tau$ выразить через функцию $U(x, t^n)$. Последняя представлена у нас сеточной функцией u_k^n. Поэтому для того, чтобы использовать формулу (171), построим предварительно интерполяционную функцию $Pu^n(x)$. Поскольку все рассмотрения мы ведем в окрестности точки x_k, t^n, то запишем эту функцию в виде

$$Pu^n(x) = \sum_m u_{k+m}^n P_m(x), \tag{175}$$

где $P_m(x)$ — соответствующие полиномы.

Интерполяционная функция (175) может быть использована и для определения значений $(U_x)_k^n$, ..., входящих в выражение (169), а следовательно, в коэффициенты D и правую часть C (170).

Подставляя $Pu^n(x)$ в (172), получаем

$$QPu^n(x) = \sum_m \alpha_m u_{k+m}^n,$$

где

$$\alpha_m = \int_{-\infty}^{\infty} q(\xi, \tau) P_m(x_k + \xi)\, d\xi,$$

и в соответствии с (171) можем написать

$$u_k^{n+1} = \sum_m \alpha_m u_{k+m}^n + C\tau,$$

т. е. нужную нам расчетную формулу. Очевидно, последняя является просто некоторой квадратурной формулой для (171), (172).

Сделаем некоторые замечания. Система (170) аппроксимирует исходную (168) с точностью до второго порядка по τ, h, так как ошибка линейной аппроксимации (169) — второго порядка по $U - U_k^n$, $U_x - (U_x)_k^n$, ... С целью упрощения линейного оператора D можно, представляя $\mathscr{D}(U)$ в виде (169), учитывать его зависимость лишь от старших производных (как в (162), (163)), поскольку устойчивость в основном зависит от вида тех членов разностных уравнений, которые соответствуют старшим производным. Правда, это понижает порядок аппроксимации до $O(\tau, h)$.

По тем же соображениям при вычислении величин U_k^n, $(U_x)_k^n$, ..., входящих в коэффициенты и правую часть (170), не обязательно пользоваться интерполяционной функцией (175). Для этого годятся любые разностные выражения, конечно, аппроксимирующие эти величины.

Все получаемые изложенным методом разностные схемы отличаются друг от друга лишь способом локальной аппроксимации (169) и видом интерполяции (175). Если примененная интерполяция достаточно точна и эффективно учитывает область зависимости решения, то разностная схема оказывается удовлетворительной. Исследование ее проводится обычными способами.

Задачи

1. Какую неточность, в смысле порядка по τ, h, можно допустить при решении системы (157)?

2. Найти все устойчивые схемы вида (151) и (160) для задачи (150) первого порядка точности по τ, h.

3. Способом неопределенных коэффициентов построить разностную схему вида (151) для решения уравнения теплопроводности (173). При каком соотношении между τ и h эта схема будет иметь максимальный порядок точности?

4. Используя формулу точного решения (171), (172), (174) уравнения теплопроводности (173), построить разностную схему, применяя при $t = t^n$:

а) линейную интерполяцию по двум точкам x_{k-1}, x_{k+1},

б) кусочно-линейную интерполяцию по x_{k-1}, x_k, при $x < x_k$, и по x_k, x_{k+1}, при $x > x_k$,

в) квадратичную интерполяцию по трем точкам x_{k-1}, x_k, x_{k+1}.

Исследовать аппроксимацию и устойчивость каждой из полученных схем.

5. Решение линейной системы уравнений

$$\frac{\partial U}{\partial t}+\frac{\partial V}{\partial x}=0,$$

$$\frac{\partial V}{\partial t}+\frac{\partial U}{\partial x}=0$$

дается формулами

$$U(x, t)=\frac{1}{2}(U(x+t, 0)+U(x-t, 0)-V(x+t, 0)+V(x-t, 0)),$$

$$V(x, t)=\frac{1}{2}(V(x+t, 0)+V(x-t, 0)-U(x+t, 0)+U(x-t, 0)).$$

Использовать их для построения разностной схемы, соответствующей системе

$$\frac{\partial U}{\partial t}+\frac{\partial p(V)}{\partial x}=0,$$

$$\frac{\partial V}{\partial t}+\frac{\partial U}{\partial x}=0,$$

где $p(V)$ — заданная функция.

§ 8. НЕЯВНЫЕ РАЗНОСТНЫЕ СХЕМЫ

При применении численного метода для решения конкретной задачи всегда возникает вопрос о выборе значений параметров метода — шагов сетки τ, h, т. е. количества и расположения расчетных точек. Решение этого вопроса зависит от многих разнородных факторов: свойств решения задачи, требований к точности расчета, наличных средств реализации метода, т. е. мощности вычислительной машины и т. д. Поэтому нельзя дать общих рецептов, пригодных для сколько-нибудь широкого класса задач.

Теоретическое исследование сходимости метода проводится лишь для выяснения этого вопроса в принципе. Стремление τ и h к нулю для реальных задач неосуществимо. Поэтому, выбирая расчетную сетку из соображений «достаточной точности», обычно руководствуются некоторой информацией о характерных свойствах решения и требуют, чтобы густота сетки позволяла удовлетворительно отразить эти свойства, выявить интересующие нас закономерности в поведении решения. А для этого, как правило, много расчетных точек не нужно.

Однако не для всякого численного метода такое рассуждение правомерно, поскольку оно не учитывает индивидуальных, внутренних свойств метода. Не касаясь всех аспектов этого вопроса, остановимся здесь лишь на одном, самом простом.

Все конкретные разностные задачи, которые были рассмотрены выше, оказывались доброкачественными (устойчивыми) лишь при выполнении определенных ограничений на шаги сетки вида

$$\frac{\tau}{h} < \text{const}, \qquad \frac{\tau}{h^2} < \text{const}. \tag{176}$$

Очевидно, при выборе параметров расчетной сетки они должны быть учтены. При этом типичной оказывается следующая ситуация. Для значений τ, h, продиктованных соображениями точности, условия (176) не выполнены. Поэтому приходится уменьшать τ (и существенно), что значительно увеличивает количество вычислительной работы.

Поскольку условия (176) формулируются в терминах параметров численного метода, а не исходной задачи, и специфичны для него, то можно попытаться строить методы с более слабыми ограничениями на шаги сетки или вообще без них.

Исследованные выше разностные схемы давали явное выражение решения в каждой точке данного временного слоя, u_k^{n+1}, через значения решения в нескольких ближайших точках предыдущего слоя u_k^n, $u_{k\pm1}^n$. По этой причине такие схемы называют *явными*. Для них условия устойчивости вида (176) отражают необходимость правильного учета области зависимости решения. Эта связь не всегда проявляется столь отчетливо, как в примере § 4.

Так, рассмотрим задачу

$$\left.\begin{aligned} &\frac{\partial U}{\partial t} = a\frac{\partial^2 U}{\partial x^2}, \qquad U(0, x) = U_0(x), \\ &-\infty < x < \infty, \qquad 0 \leqslant t \leqslant T \end{aligned}\right\} \tag{177}$$

и аппроксимирующую ее разностную схему

$$\left.\begin{aligned} &\frac{u_k^{n+1} - u_k^n}{\tau} = a\,\frac{u_{k+1}^n - 2u_k^n + u_{k-1}^n}{h^2}, \qquad u_k^0 = U_0(x_k), \\ &k = 0, \ \pm1, \ \pm2, \ldots, \qquad n = 0, 1, \ldots, N. \end{aligned}\right\} \tag{178}$$

Нетрудно убедиться с помощью спектрального признака, что условие устойчивости задачи (178) записывается в виде

$$\frac{\tau}{h^2} \leqslant \frac{1}{2a}. \qquad (179)$$

Сравним области зависимости решения задач (177) и (178). Как известно (см. (172) — (174)), для уравнения теплопроводности (177) этой областью будет вся прямая $-\infty < x < \infty$. В то же время в случае разностной задачи

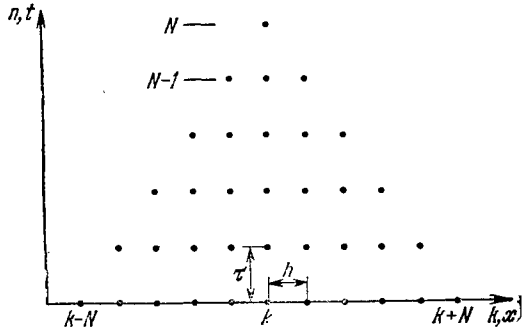

Рис. 9.

(178) областью зависимости для точек N-го слоя будут точки нулевого слоя, заполняющие интервал конечной ширины $2Nh$ (рис. 9). Таким образом, несмотря на неполный учет области зависимости точного решения, задача (178) при выполнении (179) устойчива. Противоречие разрешается тем, что при τ, $h \to 0$ с соблюдением (179) ширина указанного интервала неограниченно возрастает. Действительно, так как $\tau = T/N$, то из (179) следует $T/Nh^2 \leqslant 1/2a$ и

$$Nh \geqslant \frac{2aT}{h} \to \infty \qquad \text{при } \tau, h \to 0.$$

Следовательно, в пределе область зависимости решения учитывается правильно, и это опять связано с ограничением на шаги сетки — неравенством (179).

Итак, разностные схемы, не имеющие существенных ограничений на шаги сетки, должны быть довольно громоздкими. Чтобы эффективно учесть область зависимости, необходимо при вычислении величин в точках данного слоя использовать большое число точек предыдущего слоя.

Но подойдем к вопросу о причине неустойчивости и возникновении условия (179) с другой, может быть формальной, стороны.

Проверяя устойчивость разностной схемы (178) с помощью спектрального признака, мы, фактически, исследуем поведение частных решений вида

$$u_k^n = \lambda^n e^{ik\varphi}, \tag{180}$$

т. е. используем метод Фурье. Подстановка (180) в разностное уравнение (178) приводит к

$$\frac{\lambda^{n+1} - \lambda^n}{\tau} = -M\lambda^n, \tag{181}$$

где

$$M = 2a\,\frac{1 - \cos\varphi}{h^2}, \tag{182}$$

откуда следует

$$\lambda^{n+1} = (1 - \tau M)\,\lambda^n = (1 - \tau M)^{n+1}, \tag{183}$$

и поскольку нас устраивает только случай $|\lambda| \leqslant 1$, то мы получаем условие $|1 - \tau M| \leqslant 1$, т. е. (179).

Показатель степени n в (181) можно интерпретировать как индекс (номер шага по t), а само соотношение (181) как разностное уравнение, соответствующее обыкновенному дифференциальному уравнению

$$\frac{d\lambda}{dt} = -M\lambda, \qquad \lambda(0) = 1, \tag{184}$$

где M — положительная константа, сколь угодно большая, из-за малости h.

Точное решение уравнения (184) есть

$$\lambda = e^{-Mt}, \tag{185}$$

и приближенное решение, даваемое (181), т. е. (183), будет к нему сходиться при $\tau \to 0$, так как $(1 - \tau M)^{t/\tau} \to \to e^{-Mt}$. С этой точки зрения метод (181) (а это есть просто метод Эйлера) интегрирования уравнения (184) вполне удовлетворителен. Плох он другим. В то время как для точного решения (185) оценка $|\lambda| \leqslant 1$ справедлива при любом $M > 0$, для приближенного решения (183) она справедлива лишь при небольших M, удовлетворяющих неравенству $|1 - \tau M| \leqslant 1$. А в данном случае нас интересует именно это свойство решения.

Причина указанного дефекта понятна — точное решение (185) есть резко убывающая (при больших M) функция, а метод (181) использует для вычисления производной моментально устаревающее значение ее ($-M\lambda^n$ в правой части (181)) (рис. 10). Положение можно исправить, если взять упрежденное значение производной $-M\lambda^{n+1}$, т. е. вместо (181) рассмотреть разностное уравнение

$$\frac{\lambda^{n+1}-\lambda^n}{\tau} = -M\lambda^{n+1}. \qquad (186)$$

В этом случае вместо (183) получаем

$$\lambda^{n+1} = \frac{\lambda^n}{1+\tau M} = \frac{1}{(1+\tau M)^{n+1}}. \qquad (187)$$

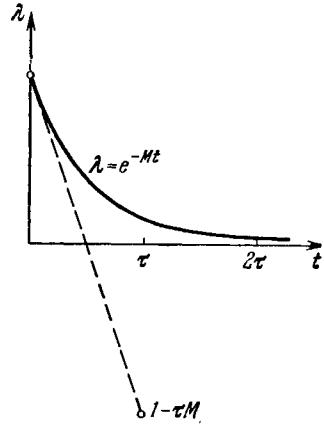

Рис. 10.

Это решение, как и (183), сходится, очевидно, к точному, при $\tau \to 0$, но в отличие от (183) удовлетворяет условию $|\lambda^{n+1}| \leqslant 1$ для любого сколь угодно большого $M > 0$, т. е. хорошо отражает основное свойство точного решения — резкое убывание его.

Возвращаясь к задаче (177), можно сказать, что если для ее решения мы построим разностную схему, соответствующую (186), а не (181), то условие устойчивости (179) снимается. Во всяком случае, проверка такой схемы на функциях вида (180) не обнаруживает неустойчивости. Этой схемой, очевидно, является следующая:

$$\left.\begin{aligned} &\frac{u_k^{n+1}-u_k^n}{\tau} = a\,\frac{u_{k+1}^{n+1}-2u_k^{n+1}+u_{k-1}^{n+1}}{h^2}, \\ &u_k^0 = U_0(x_k), \\ &k = 0, \pm 1, \pm 2, \ldots, \qquad n = 0, 1, \ldots, N. \end{aligned}\right\} \qquad (188)$$

Она отличается от (178) лишь тем, что величины в правой части берутся не с n-го, а с $(n+1)$-го слоя — соответствующая расчетная ячейка изображена на рис. 11. Это приводит к тому, что каждое из уравнений (188), связывая значения u_k^{n+1} в трех точках, не дает явного выражения для u_k^{n+1}, и для нахождения последних необходимо решать

систему если не бесконечного, то очень большого числа линейных уравнений (188). Такие схемы называют *неявными*. На способах решения системы (188) мы остановимся далее, а сейчас продолжим ее исследование.

Очевидно, задача (188), как и (178), аппроксимирует исходную дифференциальную задачу (177). Проверка устойчивости с помощью функций (180) основывается на спектральном признаке. Он был сформулирован нами лишь для явных схем, и применение его к неявным схемам требует обоснования. Но для нашей задачи (188) устойчивость можно доказать непосредственно.

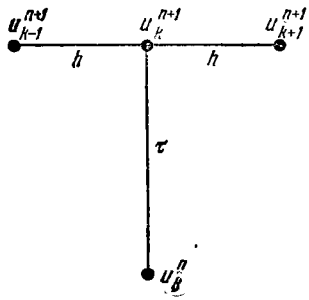

Рис. 11.

Несколько видоизменим задачу, сделав ее вычислительно реальной. А именно, будем рассматривать разностные уравнения (188) лишь для конечного (пусть очень большого) множества значений k, дополнив их в крайних точках граничными условиями, задав, например, значения функции в этих точках. Получим систему

$$\left.\begin{aligned} &\frac{u_k^{n+1}-u_k^n}{\tau} - a\,\frac{u_{k+1}^{n+1}-2u_k^{n+1}+u_{k-1}^{n+1}}{h^2} = 0, \\ &u_k^0 = U_0(x_k), \qquad k = 1, 2, \ldots, K-1, \\ &u_0^{n+1} = \alpha^{n+1}, \qquad u_K^{n+1} = \beta^{n+1}, \end{aligned}\right\} \tag{189}$$

где α^{n+1}, β^{n+1} заданы. Разумеется, задача (189) аппроксимирует дифференциальную задачу вида (177) на конечном интервале x с соответствующими условиями на его краях.

Для доказательства устойчивости разностной задачи (189) нужно записать ее в виде $lu = f$ и убедиться, что решение u имеет тот же порядок, что и f, при произвольном f. Хотя разностные уравнения (189) однородны, мы должны в правые части дописать f_k^n и показать, что решение имеет тот же порядок, что и f, α, β, U_0 (неоднородность появится при исследовании сходимости, из-за ошибки аппроксимации).

Оценим $|u_k^{n+1}|$. Допустим, что максимум этой величины на слое достигается в точке k. Тогда, если $u_k^{n+1} > 0$, то

$$u_{k+1}^{n+1} - 2u_k^{n+1} + u_{k-1}^{n+1} \leqslant 0,$$

поскольку левое и правое значения мажорируются средним. Если $u_k^{n+1} < 0$, то последнее выражение $\geqslant 0$. В силу (189), с дописанными f_k^n, знак этого выражения сохраняется и для величины

$$\frac{u_k^{n+1} - u_k^n}{\tau} - f_k^n.$$

Следовательно, в обоих случаях

$$|u_k^{n+1}| \leqslant |u_k^n + \tau f_k^n| \leqslant |u_k^n| + \tau |f_k^n|.$$

По предположению, в левой части стоит максимальное из значений $|u_k^{n+1}|$ на $(n+1)$-м слое. Используя, как обычно, обозначение

$$\|u^n\| = \max_k |u_k^n|,$$

получим из последнего неравенства

$$\|u^{n+1}\| \leqslant |u_k^n| + \tau |f_k^n| \leqslant \|u^n\| + \tau \|f^n\|.$$

Если же $\max |u_k^{n+1}|$ достигается на границе, то он равен $|\alpha^{n+1}|$ или $|\beta^{n+1}|$. Итак,

$$\|u^{n+1}\| \leqslant \max(|\alpha^{n+1}|, |\beta^{n+1}|, \|u^n\| + \tau \|f^n\|).$$

Применив это неравенство для последовательной оценки u^n через u^{n-1} и т. д., получим

$$\|u^{n+1}\| \leqslant \max(\|\alpha\|, \|\beta\|, \|U_0\| + (n+1)\tau \|f\|), \qquad (190)$$

где использованы обозначения

$$\|f\| = \max_n \|f^n\|, \qquad \|\alpha\| = \max_n |\alpha^n|, \ldots$$

Поскольку $(n+1)\tau \leqslant T = \text{const}$, то оценка (190) означает устойчивость разностной задачи (189). Заметим, что результат, полученный ранее с помощью спектрального признака, полностью подтвердился. Разностная схема (189) устойчива при л ю б ы х соотношениях шагов τ, h.

С точки зрения учета области зависимости решения этот результат неудивителен. Чтобы вычислить каждое из u_k^{n+1}, нужно решить систему уравнений (189), при этом, формально, влияние каждого u_k^n будет отражено.

Неявные абсолютно устойчивые разностные схемы могут быть построены и для других задач, связанных с интегрированием эволюционных уравнений и систем уравнений. Возможность выбирать шаги τ, h, исходя лишь из требований точности, дает очень часто большую экономию вычислительной работы, даже если учесть увеличение количества арифметических операций на одну расчетную точку, вызванное необходимостью решать системы уравнений.

Как мы увидим ниже, специфика этих систем позволяет, в случае их линейности, применять сравнительно простые и в то же время эффективные методы решения. Поэтому, при аппроксимации нелинейных дифференциальных задач неявными разностными схемами следует стремиться к линейности последних относительно величин неизвестного $(n+1)$-го слоя.

Допустим, исходная задача *квазилинейна*, т. е. линейна относительно старших производных. В этом случае, применяя неявную аппроксимацию лишь для последних, получим разностную задачу, условия устойчивости которой будут определяться явной аппроксимацией лишь младших членов и коэффициентов. А эти условия обычно необременительны.

Так, если в рассмотренной задаче коэффициент теплопроводности a считать функцией от U, $a = a(U)$, то используя аппроксимацию (188) с $a = a(u_k^n)$, получим неявную разностную схему, по-прежнему линейную относительно неизвестных u_k^{n+1}. Сама разностная задача (188), разумеется, становится нелинейной, как и исходная дифференциальная. Применяя для исследования обычные способы (§§ 5, 6), можно показать, что ограничение на шаги сетки — условие (179) — и в этом случае оказывается снятым.

Не следует думать, что наличие неявных схем лишает смысла использование схем явных. Простота и компактность последних оказываются во многих случаях чрезвычайно ценными, особенно для сложных нелинейных многомерных задач.

Задачи

Построить и исследовать различные неявные разностные схемы для решения на интервале $0 \leqslant x \leqslant 1$ следующих задач:

1. $$\frac{\partial U}{\partial t} + a\frac{\partial U}{\partial x} = 0, \quad a > 0,$$

$$U(0, x) = U_0(x), \quad U(t, 0) = \alpha(t).$$

2. $$\frac{\partial U}{\partial t} + U\frac{\partial U}{\partial x} = \frac{\partial}{\partial x} U^2 \frac{\partial U}{\partial x},$$

$$U(0, x) = U_0(x), \quad U(t, 0) = \alpha(t), \quad U(t, 1) = \beta(t).$$

3. $$\frac{\partial U}{\partial t} + c^2\frac{\partial V}{\partial x} = 0,$$

$$\frac{\partial V}{\partial t} + \frac{\partial U}{\partial x} = 0,$$

$$U(0, x) = U_0(x), \quad V(0, x) = V_0(x),$$
$$U(t, 0) = \alpha(t), \quad V(t, 1) = \beta(t).$$

§ 9. РЕШЕНИЕ РАЗНОСТНЫХ УРАВНЕНИЙ

При использовании неявных разностных схем приходится на каждом слое решать систему уравнений. Только после того, как указан способ ее решения, можно считать, что описание вычислительного алгоритма закончено.

Дело в том, что число уравнений и неизвестных очень велико — порядка $1/h$. Если не учитывать специфику этих систем и решать их как системы общего вида, то это потребует огромного количества арифметических операций, намного больше, чем при использовании явных схем. Можно показать, что для линейных систем число операций будет порядка $1/h^3$.

Кроме того, нужно помнить, что никакой алгоритм нельзя реализовать точно, так как вычисления всегда ведутся с ограниченным числом десятичных знаков. При большом порядке системы накопление ошибок округления может иметь катастрофические последствия.

Обратимся к линейной системе (189) и перепишем ее в виде

$$\left.\begin{aligned} u_0 &= \alpha, \\ -ru_{k-1} + (1+2r)u_k - ru_{k+1} &= u_k^n, \quad k = 1, 2, \ldots, K-1, \\ u_K &= \beta. \end{aligned}\right\} \tag{191}$$

Мы опустили индекс $n+1$ у неизвестных и ввели обозначение

$$r = a\frac{\tau}{h^2}.$$

Как было установлено при доказательстве устойчивости задачи (189), решение системы (191) должно удовлетворять неравенству

$$|u_k| \leqslant \max(|\alpha|,\ |\beta|,\ \max_k |u_k^n|).$$

Отсюда вытекает, что соответствующая однородная система, получающаяся при $\alpha = \beta = u_k^n = 0$, имеет только тривиальное решение $u_k = 0$. Следовательно, решение системы уравнений (191) существует и единственно.

Специфика системы (191) заключается в том, что каждое k-е уравнение содержит только три неизвестных u_{k-1}, u_k, u_{k+1}. Это дает возможность провести последовательное исключение u_0, u_1, u_2, ... следующим простым способом. Значение u_0 задано, поэтому уравнение, соответствующее $k = 1$, содержит фактически лишь два неизвестных u_1 и u_2, т. е. дает соотношение между ними. С помощью этого соотношения, используя следующее уравнение, исключаем u_1 и получаем соотношение между u_2 и u_3, и т. д. Пусть соотношение между u_{k-1} и u_k известно, а именно

$$u_{k-1} = L_k u_k + M_k. \tag{192}$$

Подставим это выражение в k-е уравнение и разрешим его относительно u_k. Получим

$$u_k = \frac{r u_{k+1} + u_k^n + r M_k}{1 + 2r - r L_k},$$

т. е. соотношение между следующей парой неизвестных u_k и u_{k+1}. Запишем его в виде (192), положив

$$L_{k+1} = \frac{r}{1 + 2r - r L_k}, \tag{193}$$

$$M_{k+1} = \frac{u_k^n + r M_k}{1 + 2r - r L_k}. \tag{194}$$

Эти две формулы позволяют перейти от L_k, M_k к L_{k+1}, M_{k+1} при любом k. Поскольку $u_0 = \alpha$, то в соответствии с (192) следует положить $L_1 = 0$, $M_1 = \alpha$ и по рекуррент-

ным формулам (193), (194) вычислить последовательно все L_k, M_k до L_K, M_K включительно. Затем, поскольку u_K нам известно, $u_K = \beta$, по формуле (192) находим последовательно все u_k.

Таким образом, процесс решения системы линейных уравнений (191) сводится к «прогонке» коэффициентов L_k, M_k (их последовательному вычислению по формулам (193), (194)) и обратной «прогонке» u_k (по формуле (192)). Отсюда и название этого процесса: *метод прогонки*. Главное достоинство его — высокая экономичность. Легко видеть, что при его использовании требуемое количество арифметических операций по порядку величины совпадает с числом неизвестных $\sim 1/h$, т. е. минимально.

Проверим теперь, насколько метод прогонки чувствителен к ошибкам округления — каково соотношение между точностью вычислений и точностью получаемого решения. Чтобы оценить развитие и накопление этих ошибок, будем считать, что приближенный (с округлениями) расчет по формулам (192), (193), (194) можно интерпретировать как точный расчет по некоторым другим, близким к ним формулам.

Начнем с формулы (193), дающей переход от L_k к L_{k+1}. Ошибка в фактически вычисленном значении L_{k+1} по сравнению с его точным значением возникает по двум причинам. Во-первых, используемое значение L_k содержит ошибку δL_k, порожденную предыдущими вычислениями. Ее вклад в ошибку δL_{k+1} определим, проварьировав (продифференцировав) выражение (193) по L_k. Это даст

$$\delta L_{k+1} = (L_{k+1})^2 \delta L_k.$$

Во-вторых, поскольку мы округляем результат каждой арифметической операции, то даже при использовании точного значения L_k все равно L_{k+1} будет содержать ошибку — ошибку, возникающую на данном цикле вычислений. Обозначим ее через δ_k.

Итак, во всяком случае при $\delta L \ll L$, имеем

$$\delta L_{k+1} = L_{k+1}^2 \delta L_k + \delta_k. \tag{195}$$

Величина δ_k характеризует точность вычислений.

Мы видим, что эволюция ошибок округления полностью определяется коэффициентами L_k. В частности, если $|L_k| > 1$, то имеет место экспоненциальное

нарастание δL_k и быстрая потеря точности. Оценим величину L_k.

Поскольку $L_1 = 0$, то (193) дает

$$L_2 = \frac{r}{1+2r} < 1.$$

Пусть $0 < L_k < 1$. Тогда, как нетрудно установить из формулы (193), справедлива оценка

$$0 < L_{k+1} < \frac{r}{1+r} < 1, \tag{196}$$

т. е. все L_k не превосходят $r/(1+r)$.

Пользуясь формулой (195) рекуррентно, найдем

$$\begin{aligned} |\delta L_N| &\leqslant |\delta_{N-1}| + \left(\frac{r}{1+r}\right)^2 |\delta L_{N-1}| \leqslant \\ &\leqslant |\delta_{N-1}| + \left(\frac{r}{1+r}\right)^2 |\delta_{N-2}| + \left(\frac{r}{1+r}\right)^4 |\delta L_{N-2}| \leqslant \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ &\leqslant \max_k |\delta_k| \frac{1}{1-\left(\frac{r}{1+r}\right)^2} + \left(\frac{r}{1+r}\right)^{2N} |\delta L_0|, \end{aligned}$$

откуда сразу следует, что ошибка δL порядка δ.

Формулы (194) и (192) исследуются аналогично. Вариация их дает

$$\delta M_{k+1} = L_{k+1}\delta M_k + \delta_k',$$
$$\delta u_{k-1} = L_k \delta u_k + \delta_k'',$$

где в δ', δ'' включены ошибки δu^n, $\delta\alpha$, $\delta\beta$. Опираясь на неравенство (196), получаем, что при вычислении M и u накопления ошибок округления также не происходит.

Мы доказали, что при решении системы линейных уравнений (191) методом прогонки точность результата с о в п а д а е т с точностью расчета и исходных данных.

Подобное утверждение справедливо не для всякого метода. Приведем пример. Переписав каждое из уравнений системы (191) в виде

$$u_{k+1} = -u_{k-1} + \frac{1+2r}{r} u_k - \frac{1}{r} u_k^n, \tag{197}$$

можем, имея пару u_{k-1}, u_k, вычислить u_{k+1}. Чтобы начать этот процесс, нам нужны значения u_0 и u_1. Первое из них

мы знаем, $u_0 = \alpha$. Зададим второе произвольно, например, $u_1 = 0$, и вычислим с помощью (197) остальные u_k. Обозначим полученное решение $u_k^{(1)}$. Очевидно, оно практически всегда не будет удовлетворять правому граничному условию $u_K = \beta$. Построим таким же способом второе решение $u_k^{(2)}$, положив $u_1^{(2)} = 1$. Рассмотрим линейную комбинацию этих решений

$$u_k = cu_k^{(1)} + (1 - c)u_k^{(2)}, \tag{198}$$

которая, очевидно, удовлетворяет левому граничному условию и уравнениям. Определим c так, чтобы удовлетворить правому граничному условию $u_K = \beta$. Получим

$$c = \frac{\beta - u_K^{(2)}}{u_K^{(1)} - u_K^{(2)}}. \tag{199}$$

При этом значении c формула (198) дает решение задачи.

На первый взгляд, описанный метод даже удобнее, чем метод прогонки. Однако он никуда не годится. Действительно, рассмотрим следующий частный случай, когда решение можно выписать явно. Положим $u_k^n = 0$ и будем искать решение в виде $u_k = \text{const} \cdot q^k$. Подставляя это выражение в (197), получим квадратное уравнение для определения q:

$$q^2 - \left(2 + \frac{1}{r}\right) q + 1 = 0.$$

Корни его при любом $r > 0$ действительны и, что существенно, один из них больше единицы, $q_1 < 1$, $q_2 > 1$. Любое решение задачи есть линейная комбинация q_1^k и q_2^k, и нетрудно показать, что в данном случае упоминаемые выше $u_k^{(1)}$ и $u_k^{(2)}$ имеют вид

$$u_k^{(1)} = \frac{\alpha q_2 q_1^k - \alpha q_1 q_2^k}{q_2 - q_1},$$

$$u_k^{(2)} = \frac{(\alpha q_2 - 1)\, q_1^k + (1 - \alpha q_1)\, q_2^k}{q_2 - q_1},$$

а решение (198), соответственно,

$$u_k = u_k^{(2)} - c\,\frac{q_2^k - q_1^k}{q_2 - q_1}. \tag{200}$$

Посмотрим, к чему могут привести ошибки округления при реализации процесса (197) — (199). Даже если идеализировать ситуацию и допустить, что мы округляем лишь величину c (199), а все остальные вычисления имеем возможность провести абсолютно точно, то все равно в решении u_k появится ошибка

$$\delta u_k \sim \delta c \frac{q_2^k - q_1^k}{q_2 - q_1},$$

как это следует из (200). Поскольку $q_2^k \gg 1$ при больших k, то малейшая неточность при вычислении c приведет к колоссальной ошибке в решении. Величина k порядка $1/h$. Следовательно, для достижения в решении заданной точности мы должны вести вычисления в $q_2^{1/h}$ раз точнее. Например, при $q_2 = 2$ и $k \sim 100$ мы получаем величину $q_2^k \sim 2^{100} \sim 10^{30}$. Очевидно, такой запас десятичных знаков обеспечить невозможно. И теоретически безупречный метод должен быть забракован.

Мы продемонстрировали метод прогонки на примере системы (191), соответствующей уравнению теплопроводности для конечного интервала $(0, X)$ с условиями $U(t,0) = \alpha(t)$, $U(t, X) = \beta(t)$ на краях. Чтобы представить себе возможные пути распространения этого метода на более сложные задачи, остановимся на следующих моментах, в которых выражается сущность его.

Любое краевое условие можно интерпретировать как эффективное выражение влияния процессов, происходящих во внешней, отброшенной части пространства. С этой точки зрения сами дифференциальные уравнения задачи описывают распространение этого влияния внутрь области. Разностная задача аппроксимирует дифференциальную. В частности, к ней применимы и соображения о роли краевых условий и уравнений. Метод прогонки явно выражает этот процесс переноса влияния краевых условий внутри области, правда, только в случае, если разностные уравнения линейны.

Общая схема метода прогонки для любой линейной задачи такая же, как и в рассмотренном случае уравнения теплопроводности. Разностная реализация граничных условий на одном из концов расчетного интервала (например, на левом) дает соотношения между значениями сеточной функции (или вектор-функции) в граничной и

приграничной точках. Используя разностные уравнения, производим последовательное исключение неизвестных, «прогоняем» соотношение между соседними значениями сеточной функции через всю область расчета, вплоть до ее правого конца. Последние соотношения вместе с граничными условиями, имеющимися здесь, позволяют найти значение функции в последней точке. Двигаясь теперь в обратном направлении, с помощью полученных ранее соотношений находим последовательно все значения искомой сеточной функции.

Конечно, это всего лишь схема метода, и конкретная форма его может существенно меняться от задачи к задаче, в соответствии со спецификой краевых условий и уравнений.

Для решения систем разностных уравнений можно применять и другие методы, в частности, и т е р а ц и о н н ы е. Обратимся к системе уравнений (191), перепишем ее в виде

$$\left.\begin{aligned} &u_0 = \alpha, \\ &u_k = \frac{r(u_{k-1} + u_{k+1}) + u_k^n}{1+2r}, \qquad k = 1, 2, \ldots, K-1, \\ &u_K = \beta, \end{aligned}\right\} \quad (201)$$

который и используем для итерационного процесса. Подставляя в правую часть ν-е приближение $u_k^{(\nu)}$, получим в левой части $u_k^{(\nu+1)}$. Исследуем сходимость итераций. Пусть

$$u_k^{(\nu)} = u_k + \delta_k^{(\nu)},$$

где u_k — точное решение системы (201). Нетрудно получить формулы

$$\begin{aligned} &\delta_0^{(\nu+1)} = 0, \\ &\delta_k^{(\nu+1)} = \frac{r}{1+2r}(\delta_{k-1}^{(\nu)} + \delta_{k+1}^{(\nu)}), \qquad k = 1, 2, \ldots, K-1, \\ &\delta_K^{(\nu+1)} = 0, \end{aligned}$$

определяющие поведение ошибки в зависимости от номера итерации. Непосредственная оценка $|\delta_k^{(\nu+1)}|$ дает

$$\max_k |\delta_k^{(\nu+1)}| \leqslant \frac{2r}{1+2r} \max_k |\delta_k^{(\nu)}|.$$

Следовательно, итерации сходятся, $\delta_k^{(\nu)} \to 0$, и скорость сходимости определяется величиной

$$\Theta = \frac{2r}{1+2r} < 1.$$

Остановимся на вопросе о числе итераций Допустим, что мы продолжаем процесс до совпадения двух последовательных итераций, $u_k^{(\nu+1)} = u_k^{(\nu)}$, т. е. получаем максимально точное решение системы (201). Однако в такой точности необходимости нет, поскольку сама система (201) является лишь приближением исходной дифференциальной задачи. При использовании метода прогонки эта точность получается «задаром», здесь же — за счет дополнительных итераций, дополнительной вычислительной работы, которая, таким образом, может оказаться лишней.

С другой стороны, сходимость численного метода (следующая из аппроксимации и устойчивости) доказана нами в предположении, что разностная задача решается точно. Нетрудно показать, что учет ошибок округления не меняет дела (устойчивость их гасит). Влияние же ошибок, возникающих из-за недоитерированности, требует особого рассмотрения.

Утрируя ситуацию, допустим, что мы ограничиваемся одной итерацией, беря в качестве начального приближения $u_k^{(0)} = u_k^n$. Это означает, что фактически для вычисления u_k^{n+1} мы пользуемся формулами [см. (201)]

$$u_k^{n+1} = \frac{r(u_{k-1}^n + u_{k+1}^n) + u_k^n}{1+2r}. \tag{202}$$

Оценивая $|u_k^{n+1}|$, получим

$$\max_k |u_k^{n+1}| \leqslant \max_k |u_k^n|,$$

т. е. разностная схема (202) устойчива (при любом r, хотя схема явная!).

Вспомним про аппроксимацию. Подставляя в (202) разложения

$$U_k^{n+1} = U + U_t\tau + O(\tau^2),$$

$$U_{k\pm1}^n = U \pm U_x h + U_{xx}\frac{h^2}{2} \pm U_{xxx}\frac{h^3}{6} + O(h^4),$$

получаем

$$U_t + O(\tau) = \frac{aU_{xx} + O(h^2)}{1 + 2r}.$$

Следовательно, разностное соотношение (202) аппроксимирует уравнение теплопроводности лишь при $r \to 0$. Если же r конечно, то используя (202), мы будем решать не то уравнение. Можно показать (см. задачу 4), что аналогичная ситуация будет и после ν-й итерации, т. е. ошибка аппроксимации является функцией от r и ν, стремящейся к нулю лишь при $r \to 0$ или $\nu \to \infty$. Итак, недоитерированность может приводить к отсутствию аппроксимации.

Возможен и другой случай. Пусть, например, для решения системы (191) вместо (201) применяется какой-либо другой итерационный процесс вида

$$u_k^{(\nu+1)} = \varphi(u_{k-1}^{(\nu)}, u_k^{(\nu)}, u_{k+1}^{(\nu)}, u_k^n),$$

при котором условие аппроксимации соблюдается. Очевидно, на этот раз опасность может заключаться в нарушении условий устойчивости. Чтобы оценить необходимое число итераций, будем рассуждать так. При первой итерации для определения $u_k^{(1)}$ мы учитываем, кроме u_k^n, только $u_{k-1}^n = u_{k-1}^{(0)}$ и $u_{k+1}^n = u_{k+1}^{(0)}$. При второй, используя $u_{k-1}^{(1)}$ и $u_{k+1}^{(1)}$, учитываем, тем самым, $u_{k-2}^n = u_{k-2}^{(0)}$ и $u_{k+2}^n = u_{k+2}^{(0)}$ и т. д. Каждая итерация расширяет область зависимости на одну расчетную точку в ту и другую стороны. Поэтому ν итераций эквивалентны некоторой явной схеме, использующей $2\nu + 1$ значений n-го слоя. Для устойчивости последней, как всегда, необходимо выполнение некоторого неравенства вида

$$\frac{\tau}{(\nu h)^2} \leqslant \text{const},$$

т. е. $r \leqslant \text{const} \cdot \nu^2$. Таким образом, необходимое число итераций

$$\nu \sim \sqrt{r}.$$

Мы видим, что при небольших r итерационные методы могут оказаться вполне удовлетворительными и даже конкурировать с методом прогонки, поскольку позволяют разделить вычислительный процесс на несколько парал-

лельных ветвей и не требуют запоминания больших массивов коэффициентов прогонки L, M. Иногда целесообразно комбинировать оба метода, применяя прогонку на отдельных участках и «склеивая» их итерациями.

Наконец, последнее замечание. Мы исследовали только случай линейных разностных уравнений. Если для итерационных методов это не является ограничением, то метод прогонки использует линейность уравнений по существу. Для решения нелинейных уравнений никаких методов, кроме итерационных, нет. Однако итерации можно делать по-разному. Выше мы рассмотрели явный алгоритм типа $u^{(\nu+1)} = \varphi(u^{(\nu)})$. При больших r он оказывается медленно сходящимся, из-за необходимости учета области зависимости решения. Последняя определяется свойствами задачи, которые в значительной степени проявляются уже в ее линейной модели. Обладая столь эффективным средством, как метод прогонки, мы можем при построении итерационного процесса использовать неявные, но линейные алгоритмы, сводя задачу к решению, на каждой итерации, линейной системы. Это должно дать быстро сходящийся процесс.

Пусть, например, в рассмотренной выше задаче коэффициент a зависит от искомой функции, $a = a(U)$. Конечно, можно, при аппроксимации задачи разностной, брать $a(u^n)$, но допустим, что это по каким-то причинам нас не устраивает и мы используем $a(u^{n+1})$. Тогда система разностных уравнений (191) оказывается нелинейной, $r = r(u_k)$. Естественно положить $r = r(u_k^{(\nu)})$ и, решая линейную систему (191) методом прогонки, получать $u_k^{(\nu+1)}$. Если в качестве начального приближения использовать $u_k^{(0)} = u_k^n$, то скорость сходимости итерационного процесса, $u_k^{(\nu)} \to u_k^{n+1}$, будет определяться величиной фактического изменения решения от слоя к слою.

Задачи

1. Оценить порядок количества арифметических операций, необходимого для решения системы N линейных уравнений общего вида методом исключения.

2. Если задан способ решения системы уравнений, возникающей при использовании неявной схемы, то тем самым определен оператор перехода от слоя к слою, и мы можем интерпретировать схему как явную, записав ее в виде

$$u^{n+1} = R(u^n + \tau f^n).$$

Используя устойчивость задачи, т. е. условие ограниченности степеней оператора R, $\| R^n \| \leqslant \text{const}$, исследовать поведение ошибок, возникающих в u^{n+1} из-за округления правой части, при возрастании n, $n\tau < t$.

3. Исследовать разностную задачу

$$u_0^{n+1} = \alpha^{n+1},$$

$$\left.\begin{aligned} \frac{u_k^{n+1} - u_k^n}{\tau} + \frac{v_{k+1/2}^{n+1} - v_{k-1/2}^{n+1}}{h} = 0, \\ \frac{v_{k-1/2}^{n+1} - v_{k-1/2}^n}{\tau} + \frac{u_k^{n+1} - u_{k-1}^{n+1}}{h} = 0, \end{aligned}\right\} \quad k = 1, 2, \ldots, K,$$

$$v_{K+1/2}^{n+1} = \beta^{n+1},$$

где $v_{k+1/2}$ означает, что это значение соответствует точке $x_{k+1/2} = (k + 1/2)h$. Применить метод прогонки, используя соотношение вида

$$u_{k-1} = L_k v_{k-1/2} + M_k,$$

или

$$u_{k-1} = L_k u_k + M_k.$$

Проверить вычислительную корректность алгоритма относительно ошибок округления.

Построить аналогичную схему для решения нелинейной задачи

$$\frac{\partial U}{\partial t} + \frac{\partial P(V)}{\partial x} = 0, \qquad U(0, x) = U_0(x),$$

$$\frac{\partial V}{\partial t} + \frac{\partial U}{\partial x} = 0, \qquad V(0, x) = V_0(x)$$

на интервале $0 \leqslant x \leqslant X$, на краях которого заданы граничные условия

$$U(t, 0) = \alpha(t), \qquad V(t; X) = \beta(t).$$

4. Пусть при проведении рассмотренного выше процесса (201)

$$u_k^{(\nu+1)} = \frac{r(u_{k-1}^{(\nu)} + u_{k+1}^{(\nu)}) + u_k^n}{1 + 2r}, \quad k = 0, \pm 1, \ldots,$$

мы ограничиваемся N итерациями, т. е. $u^{n+1} = u_k^{(N)}$, $u_k^{(0)} = u_k^n$ (для простоты отбросим граничные условия).

Проверить, что полученная разностная схема устойчива для любых r и N.

Доказать, что ошибка аппроксимации пропорциональна величине

$$\left(\frac{2r}{1 + 2r}\right)^N.$$

Для этого положить

$$u_k^{(\nu)} = a^{(\nu)} + hb^{(\nu)} k + \frac{h^2}{2} c^{(\nu)} k^2 + \ldots,$$

где, очевидно,

$$a^{(0)} = u_0^n, \quad b^{(0)} = (u_x)_0^n, \quad c^{(0)} = (u_{xx})_0^n,$$

и найти $a^{(N)}$ с помощью формулы итерационного процесса. Так как

$$u_0^{n+1} = a^{(N)} = u_0^n + \tau (u_t)_0^n + \ldots,$$

то, зная $a^{(N)}$, легко определить ошибку аппроксимации.

ГЛАВА III

§ 10. РАСЧЕТ РАЗРЫВНЫХ РЕШЕНИЙ

Во всех предыдущих рассмотрениях мы предполагали, что точное решение исходной задачи есть гладкая функция. При исследовании аппроксимации, при построении линейных моделей задач мы существенно использовали это допущение. Для дифференциальных задач оно естественно, так как, по самому смыслу решения, оно должно иметь хотя бы те производные, которые входят в уравнения. Наличие в решении каких-либо особенностей, нарушающих гладкость, требует дополнительного исследования и, при необходимости, соответствующего видоизменения метода. Ничего более определенного сказать нельзя, так как каждая особенность особенна по-своему.

Здесь мы ограничимся изучением этого вопроса только для одного практически важного случая, когда решение является кусочно-гладкой функцией, имеющей разрывы. Обратимся к примеру

$$\frac{\partial U}{\partial t} + U\frac{\partial U}{\partial x} = 0, \qquad U(0, x) = U_0(x). \tag{203}$$

Если решение $U(t, x)$ — гладкая функция, то это наша старая, хорошо изученная задача. Однако в некоторых случаях требования гладкости и существования решения могут оказаться исключающими друг друга. Так, решение задачи (203), очевидно, удовлетворяет системе

$$\frac{dx}{dt} = U,$$

$$\frac{dU}{dt} = 0,$$

т. е. постоянно вдоль прямых

$$x = x_0 + U_0(x_0)\, t.$$

Эти линии называют *характеристиками* — вдоль них уравнения задачи вырождаются в соотношения между дифференциалами функции (или функций). Наклон каждой характеристики в нашей задаче определяется значением функции $U_0(x)$ в точке x_0 — точке пересечения данной характеристики с линией начальных данных $t=0$ (рис. 12).

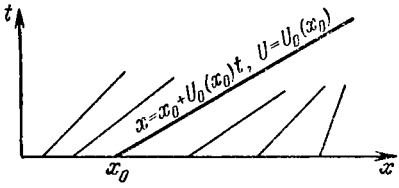

Рис. 12.

Нетрудно видеть, что в случае, если $U_0(x)$ хотя бы на небольшом участке оси x будет убывающей функцией, то характеристики, исходящие из точек этого участка, пересекутся. Поскольку каждая из них приносит свое, независимое, значение функции, то это приведет к неоднозначности решения в точке пересечения. Выход из положения подсказывается задачами, описывающими реальные физические процессы (гидродинамика). Он заключается в допущении разрывных решений.

В частности, указанная неоднозначность будет свидетельствовать о возникновении разрыва. Но если мы допускаем разрыв в решении, то, поскольку в точке разрыва производные не определены и дифференциальные уравнения теряют смысл, мы должны заменить последние конечными соотношениями, которые свяжут значения функции по обе стороны разрыва. Если продолжить аналогию с гидродинамическими задачами, то следует предположить, что эти соотношения должны выражать для разрывных решений те же физические законы, что и дифференциальные уравнения — для гладких. Формально-математическая процедура получения их выглядит следующим образом.

Пусть $U(x, t)$ — гладкая функция, удовлетворяющая в некоторой области плоскости x, t уравнению (203). Проинтегрируем (203) по любой части S этой области. Легко получить, что

$$\iint\limits_S \left(\frac{\partial U}{\partial t} + U\frac{\partial U}{\partial x}\right) dx\,dt = \iint\limits_S \left(\frac{\partial U}{\partial t} + \frac{\partial}{\partial x}\left(\frac{U^2}{2}\right)\right) dx\,dt =$$

$$= \iint\limits_S \frac{\partial U}{\partial t}\,dt\,dx + \iint\limits_S \frac{\partial}{\partial x}\left(\frac{U^2}{2}\right) dx\,dt = \oint\limits_\Gamma U\,dx - \frac{U^2}{2}\,dt,$$

где последний, криволинейный, интеграл берется по контуру Г — границе S. Следовательно, если $U(x, t)$ — решение (203), то

$$\oint_{\Gamma} U\,dx - \frac{U^2}{2}\,dt = 0 \tag{204}$$

для любого контура Г. При использовании (204) вместо (203) мы можем не требовать от функции $U(x, t)$ ни гладкости, ни непрерывности. Поэтому для получения соотношений на разрыве следует использовать именно его.

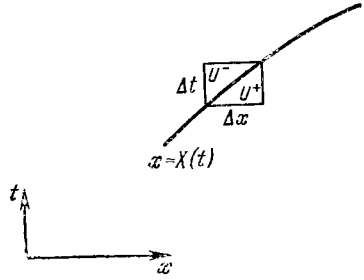

Рис. 13.

Будем рассуждать так. Разрыв в решении, возникнув, затем перемещается, описывая в плоскости x, t некоторую кривую $x = X(t)$. Рассмотрим маленький элемент этой кривой и построим на нем как на диагонали прямоугольный контур Г (рис. 13). Поскольку прямоугольник мал, то в каждой из его половин можно считать функцию U постоянной и равной U^- — слева, U^+ — справа. Вычисляя интеграл (204) для такого контура, получаем

$$U^+\Delta x - \left(\frac{U^2}{2}\right)^+ \Delta t - U^-\Delta x + \left(\frac{U^2}{2}\right)^- \Delta t = 0,$$

где Δx, Δt — стороны прямоугольника. При стягивании прямоугольника в точку отношение $\Delta x/\Delta t$ стремится к $X'(t)$ и последнее равенство в пределе дает

$$(U^+ - U^-)\,X' = \left(\frac{U^2}{2}\right)^+ - \left(\frac{U^2}{2}\right)^-. \tag{205}$$

Как мы видели выше, пересечение характеристик и возникновение разрыва происходит лишь при $U^- > U^+$. Можно показать, что только такие (для нашей задачи) разрывы могут существовать. Сокращая (205) на $U^+ - U^-$ и дополняя его указанным неравенством, получим

$$X' = \frac{U^- + U^+}{2}, \qquad U^- > U^+. \tag{206}$$

Это соотношение между U^-, U^+ и X' заменяет на линии разрыва $x = X(t)$ дифференциальное уравнение (203).

Итак, расширим постановку задачи (203) и допустим наличие разрывов на некоторых линиях $x = X(t)$, удовлетворяющих соотношениям (206).

Перейдем к конструированию численного метода решения задачи (203), (206). Самым естественным представляется просто использование рассмотренных ранее методов с соответствующей модификацией их лишь в непосредственной окрестности линий разрыва. Это не составляет проблемы, во всяком случае для нашей простой задачи, однако имеет ряд неудобств. Нестандартные формулы для расчета величин на линии разрыва и в ближайших к ней точках должны учитывать все возможные случаи расположения этой линии относительно точек основной сетки. Кроме того, нужно рассчитывать на возможность возникновения разрыва и, следовательно, проверять получаемое решение соответствующим образом (во всех расчетных точках). Ясно, что это намного увеличивает объем вычислительного алгоритма, он теряет свою простоту и компактность, по существу, из-за ничтожно малого количества расчетных точек. В связи с этим часто предпочитают другой путь построения расчетных формул, к описанию которого мы и перейдем.

С точки зрения разностной задачи понятие разрывности решения формально лишено смысла, поскольку сеточная функция определяется на дискретном множестве точек — расчетной сетке. С другой стороны, как мы видели, условие на разрыве (206) является следствием дифференциального уравнения (203) и, следовательно, «содержится» в нем. Из теории дифференциальных уравнений известно, что разрывное решение можно получить как предел гладкого решения возмущенного уравнения при стремлении параметра возмущения к нулю. Этот факт мы можем использовать, так как, применяя тот или иной численный метод, мы всегда заменяем исходную задачу другой — разностной, возмущенной задачей. Будем осуществлять аппроксимацию в два этапа. Сначала заменим исходную задачу, введя возмущение, на некоторую промежуточную, и уже затем перейдем к разностной задаче.

Вместо (203) рассмотрим уравнение

$$\frac{\partial U}{\partial t} + U\frac{\partial U}{\partial x} + \varepsilon^2 \frac{\partial}{\partial x}\left(\frac{\partial U}{\partial x}\right)^2 = 0, \qquad (207)$$

где ε — малый параметр. Очевидно, если ограничиваться только гладкими функциями, то, из-за малости ε, решения задач (207) и (203) при одинаковых или близких начальных данных будут близки. Чтобы представить себе их отличие в случае разрывных (для (203)) решений, рассмотрим следующий частный пример.

Пусть решение задачи (203), (206) есть ступенчатая функция

$$U = \begin{cases} U^- & \text{при} \quad x - \omega t < 0, \\ U^+ & \text{при} \quad x - \omega t > 0, \end{cases} \tag{208}$$

где

$$\omega = \frac{U^- + U^+}{2}, \qquad U^- > U^+, \tag{209}$$

а U^-, U^+ — константы. Очевидно, функция (208) удовлетворяет (203), (206).

Соответствующее решение уравнения (207) будем искать в виде

$$U_\varepsilon(x, t) = f(x - \omega t), \tag{210}$$

причем естественно принять, что

$$f(x) \to U^\pm \quad \text{при} \quad x \to \pm\infty, \tag{211}$$

поскольку вдали от разрыва решения U, U_ε — гладкие функции и, следовательно, близки. Подставляя (210) в (207), получим обыкновенное дифференциальное уравнение для f:

$$-\omega f' + ff' + \varepsilon^2 (f'^2)' = 0, \tag{212}$$

Нетрудно убедиться, что

$$f = \omega + \text{const} \cdot \sin \frac{x - \omega t}{\varepsilon \sqrt{2}}$$

есть решение (212). Поскольку

$$f = \text{const}$$

также удовлетворяет уравнению (212), то интересующее нас решение имеет вид (рис. 14):

$$U_\varepsilon = \begin{cases} U^- & \text{при} \quad \dfrac{x - \omega t}{\varepsilon \sqrt{2}} < -\dfrac{\pi}{2}, \\ \dfrac{U^+ + U^-}{2} + \dfrac{U^+ - U^-}{2} \sin \dfrac{x - \omega t}{\varepsilon \sqrt{2}} & \text{при} \quad \left| \dfrac{x - \omega t}{\varepsilon \sqrt{2}} \right| < \dfrac{\pi}{2}, \\ U^+ & \text{при} \quad \dfrac{x - \omega t}{\varepsilon \sqrt{2}} > \dfrac{\pi}{2}. \end{cases}$$

Это решение — гладкая функция, вместо разрыва U^-, U^+ оно имеет зону непрерывного перехода от U^- к U^+ ширины $\varepsilon\pi\sqrt{2}$. Если ε достаточно мало, то эта зона узка, и U_ε близко к разрывному решению (208) исходной задачи (203), (206).

Это дает основания для замены задачи (203), (206) задачей (207). Очень важно, что решение последней — гладкая функция. Поэтому при построении численного метода для нее мы можем использовать все рассмотренные ранее способы получения и исследования разностных задач. Усложнение, связанное с добавлением возмущающего члена (его называют *искусственной вязкостью*), полностью окупается возможностью проведения расчета по стандартным формулам, без какого-либо выделения особенностей и проб на возникновение разрывов.

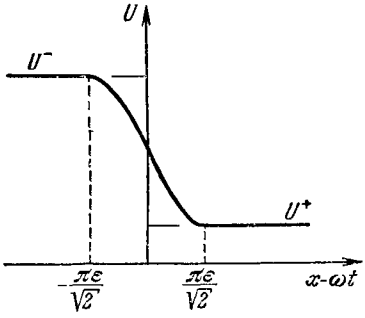

Рис. 14.

Рассмотренный способ введения искусственной вязкости (207) (не единственно возможный) удобен тем, что ширина зоны «размазывания» разрыва — порядка ε и не зависит от величины разрыва $U^- - U^+$. Тем самым все изменения сеточной функции на интервалах больше, чем $\sim\varepsilon$, имеют реальный смысл для любого решения. Очевидно, из этих соображений и следует выбирать величину ε.

Относительно выбора h (и τ для неявных схем) заметим, что введение вязкости дает нужный эффект только в том случае, если в зоне разрыва будет располагаться хотя бы несколько расчетных точек. В противном случае надеяться на удовлетворительную аппроксимацию нельзя, разностная задача может неправильно прореагировать на вязкость.

Таким образом, между параметрами ε, τ, h есть некоторая связь, т. е. $\varepsilon \sim \varepsilon(\tau, h)$. С другой стороны, всякая разностная задача содержит некоторую вязкость (ее называют *аппроксимационной*), поскольку разностная схема эквивалентна исходным уравнениям плюс ошибка аппроксимации, например, $O(\tau, h)$. Последняя на самом деле является некоторым дифференциальным выражением с ма-

лыми коэффициентами, т. е. может быть квалифицирована как искусственная вязкость. Поэтому, если угодно, вышеприведенный способ построения разностной задачи сводится к обычному, но имеющему ошибку аппроксимации специального вида.

Хотя мы рассмотрели лишь простейший пример (203), но все наиболее существенное по поводу численного решения такого рода задач при наличии разрывов, за одним исключением, сказано.

Таким исключением, как ни странно, являются линейные уравнения и, вообще, задачи, в которых разрыв может распространяться по характеристикам (т. е. линия разрыва $x = X(t)$ — характеристика). В этом случае разрывы не поддаются стабильному размазыванию с помощью искусственной вязкости. Поэтому, если есть необходимость, то для аккуратного расчета такого разрыва приходится использовать специальные формулы.

Для расчета любой особенности есть два пути. Либо детальное описание, либо ликвидация, «замазывание» ее. Пример второго и продемонстрирован в этом параграфе.

Задачи

1. Для задачи (203), (206) построить разностные формулы расчета величин на разрыве и в соседних с ним точках. Типичная расчетная ячейка изображена на рис. 15. Крестиками обозначены расчетные точки, расположенные на линии разрыва. В них вычисляются два значения решения, левое u^- и правое u^+.

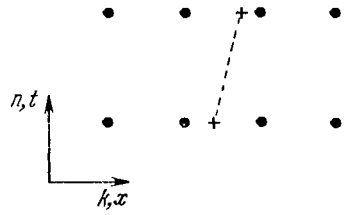

Рис. 15.

2. Для обеспечения расчета разрывных решений уравнения

$$\frac{\partial U}{\partial t} + \frac{\partial F(U)}{\partial x} = 0$$

с условиями на линии разрыва $x = X(t)$:

$$(U^+ - U^-)\, X' = F(U^+) - F(U^-),$$

$$F'_U(U^-) > F'_U(U^+),$$

где F — заданная функция U, применить различные способы введения искусственной вязкости. Наряду с

$$\frac{\partial U}{\partial t} + \frac{\partial F(U)}{\partial x} + \varepsilon^2 \frac{\partial}{\partial x}\left(\frac{\partial U}{\partial x}\right)^2 = 0$$

рассмотреть возмущенное уравнение вида

$$\frac{\partial U}{\partial t} + \frac{\partial F(U)}{\partial x} = \varepsilon \frac{\partial^2 U}{\partial x^2}.$$

В обоих случаях, на функциях $U_\varepsilon(x - \omega t)$, исследовать последствия введения вязкости для разрывного решения. Отдельно рассмотреть эти же вопросы при линейной $F(U)$, $F = aU$.

3. Составить разностную схему для уравнения (207). Провести исследование аппроксимации и устойчивости.

§ 11. МНОГОМЕРНЫЕ ЗАДАЧИ

Переход от обыкновенных дифференциальных уравнений к уравнениям в частных производных, т. е. переход от одной независимой переменной к двум, приводит, как мы видели, не только к количественному усложнению задач, но и к новым существенным проблемам. Основные из них были изучены в предыдущих параграфах. Переход к задачам с тремя и более независимыми переменными также не всегда тривиален. Однако почти все рассмотренные нами способы построения и исследования разностных задач допускают простое и естественное обобщение на этот случай. В этом смысле задачи с двумя независимыми переменными t, x являются хорошей моделью многомерных задач.

Остановимся, коротко, на основных вопросах, имея в виду задачи, в которых независимыми переменными будут t (время) и две пространственные координаты x, y. Такие задачи называют *двумерными*. Для демонстрации изберем двумерное уравнение теплопроводности, т. е. задачу

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2}; \qquad U(0, x, y) = U_0(x, y). \qquad (213)$$

В двумерном случае простейшая расчетная сетка будет состоять из точек с координатами

$$t^n = n\tau, \qquad x_k = kh_x, \qquad y_m = mh_y$$

и определяться, следовательно, тремя параметрами τ, h_x, h_y — шагами сетки. Соответствующие этим точкам значения сеточной функции обозначим $u^n_{k,m}$.

Все перечисленные в § 7 способы построения расчетных формул непосредственно обобщаются на многомерный случай.

В частности, для задачи (213) при использовании расчетной ячейки, изображенной на рис. 16, любой из этих

способов даст следующую разностную задачу:

$$\left.\begin{aligned} &\frac{u^{n+1}_{k,m}-u^{n}_{k,m}}{\tau}= \\ &=\frac{u^{n}_{k+1,m}-2u^{n}_{k,m}+u^{n}_{k-1,m}}{h_x^2}+\frac{u^{n}_{k,m+1}-2u^{n}_{k,m}+u^{n}_{k,m-1}}{h_y^2}, \\ &u^{0}_{k,m}=U_0(x_k,y_m), \end{aligned}\right\} \quad (214)$$

являющуюся прямым обобщением одномерной — (178).

Принципиальная схема исследования сходимости, сводящая этот вопрос к аппроксимации и устойчивости, сформулирована в § 5 в столь общей форме, что рассматриваемые сейчас двумерные задачи можно считать рядовым частным случаем. Следует лишь вместо двух параметров τ, h и двух аргументов сеточной функции n, k во всех формулировках иметь в виду три параметра τ, h_x, h_y и три аргумента n, k, m.

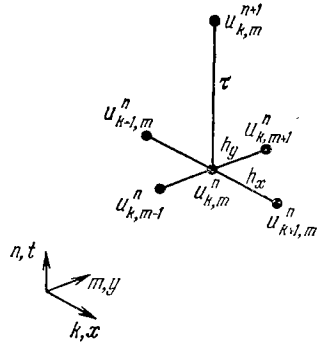

Рис. 16.

Для проверки аппроксимации разностной и дифференциальной задач используется тот же прием. Так, для задач (213), (214), предполагая гладкость точного решения $U(t, x, y)$, можем написать:

$$U^{n+1}_{k,m}=U^{n}_{k,m}+\tau\left(\frac{\partial U}{\partial t}\right)^{n}_{k,m}+O(\tau^2),$$

$$U^{n}_{k\pm1,m}=U^{n}_{k,m}\pm h_x\left(\frac{\partial U}{\partial x}\right)^{n}_{k,m}+\frac{1}{2}h_x^2\left(\frac{\partial^2 U}{\partial x^2}\right)^{n}_{k,m}\pm$$
$$\pm\frac{1}{6}h_x^3\left(\frac{\partial^3 U}{\partial x^3}\right)^{n}_{k,m}+O(h_x^4),$$

$$U^{n}_{k,m\pm1}=U^{n}_{k,m}\pm h_y\left(\frac{\partial U}{\partial y}\right)^{n}_{k,m}+\frac{1}{2}h_y^2\left(\frac{\partial^2 U}{\partial y^2}\right)^{n}_{k,m}\pm$$
$$\pm\frac{1}{6}h_y^3\left(\frac{\partial^3 U}{\partial y^3}\right)^{n}_{k,m}+O(h_y^4).$$

Подставляя эти выражения в (214), находим без труда,

что оно удовлетворяется с точностью $O(\tau, h_x^2, h_y^2)$, т. е. аппроксимация имеет место.

Что касается исследования устойчивости, то фактически нами был рассмотрен (в § 6) лишь один общий способ — спектральный признак устойчивости линейных разностных задач, имеющих слоистую структуру вида

$$u^{n+1} = Ru^n + \tau f^n.$$

Если под функцией на слое u^n понимать теперь $u_{k,m}^n$ — сеточную функцию двух индексов k, m, то все рассуждения, проведенные в начале § 6, сохраняют силу, и устойчивость по-прежнему сведется к ограниченности норм степеней оператора R. Для операторов R, действующих по формуле

$$(Ru)_{k,m} = \sum_{p,q} \alpha_{p,q} u_{k+p,m+q}, \tag{215}$$

обобщающей (123), можно оценить эти нормы с помощью спектрального радиуса операторов. А именно, легко убедиться, что сеточные функции

$$u_{k,m} = u_{0,0} e^{i(k\varphi + m\psi)} \tag{216}$$

при любых φ, ψ являются собственными функциями оператора R (215), а

$$\lambda = \sum_{p,q} \alpha_{p,q} e^{i(p\varphi + q\psi)}$$

— соответствующими собственными значениями. Требуя выполнения неравенства $|\lambda(\varphi, \psi)| \leqslant 1$, получаем необходимые условия устойчивости. Таким образом, и в двумерном случае спектральный признак устойчивости сохраняет свою эффективность.

В применении к задаче (214) эта процедура дает следующее условие устойчивости

$$\frac{\tau}{h_x^2} + \frac{\tau}{h_y^2} \leqslant \frac{1}{2}.$$

Далее, аппроксимация с помощью неявных разностных уравнений опять приводит к схемам, устойчивым при любых соотношениях между шагами сетки. Для задачи

(213) такой схемой будет, очевидно, следующая (рис. 17):

$$\frac{u_{k,m}^{n+1}-u_{k,m}^{n}}{\tau}=$$

$$=\frac{u_{k+1,m}^{n+1}-2u_{k,m}^{n+1}+u_{k-1,m}^{n+1}}{h_x^2}+\frac{u_{k,m+1}^{n+1}-2u_{k,m}^{n+1}+u_{k,m-1}^{n+1}}{h_y^2}. \quad (217)$$

Ее устойчивость можно доказать непосредственной оценкой u^{n+1}, как в § 8. Если же в качестве u^n взять функцию вида (216), то легко убеждаемся, что $u^{n+1}=\lambda u^n$, причем

$$\lambda=\frac{1}{1+\frac{4\tau}{h_x^2}\sin^2\frac{\varphi}{2}+\frac{4\tau}{h_y^2}\sin^2\frac{\psi}{2}},$$

и при любых τ, h_x, h_y имеем $|\lambda|\leqslant 1$.

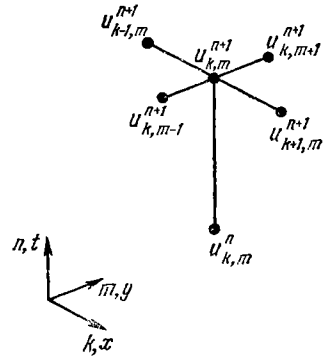

Рис. 17.

Таким образом, основные вопросы, связанные с построением и исследованием разностных задач, не претерпевают принципиальных изменений при увеличении размерности. В то же время увеличение объема задачи и усложнение вычислительных алгоритмов могут порождать и новые проблемы.

Остановимся, в связи с этим, на вопросе о способах решения систем разностных уравнений, возникающих при использовании неявных схем. Рассмотрим уравнения (217) в прямоугольнике (рис. 18)

$$1\leqslant k\leqslant K, \qquad 1\leqslant m\leqslant M, \quad (218)$$

задав значения u^{n+1} на краях его. Положив, для простоты, $h_x=h_y=h$, обозначив $r=\tau/h^2$ и отбросив индекс $n+1$ у неизвестных $u_{k,m}^{n+1}$, запишем эти уравнения в виде

$$-ru_{k,m+1}-$$
$$-ru_{k-1,m}+(1+4r)u_{k,m}-ru_{k+1,m}-$$
$$-ru_{k,m-1}=u_{k,m}^{n}. \quad (219)$$

Граничные значения $u_{k,m}$ заданы:

$$\left.\begin{array}{c} u_{k,M+1} = \delta_k, \\ u_{0,m} = \alpha_m, \qquad u_{K+1,m} = \beta_m, \\ u_{k,0} = \gamma_k. \end{array}\right\} \tag{220}$$

Индексы k, m пробегают множество значений (218). Таким образом, число уравнений и число неизвестных равны KM, каждой внутренней точке прямоугольника соответствует свое уравнение (219).

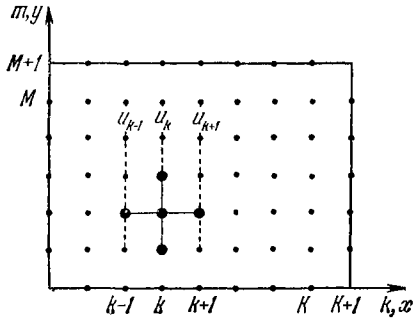

Рис. 18.

В одномерном случае для решения аналогичной системы уравнений мы смогли применить эффективный метод исключения — метод прогонки (§ 9). Это удалось благодаря тому, что там мы имели систему чрезвычайно простого вида — каждое уравнение связывало только три неизвестных с последовательными номерами, соответствующими их естественному упорядочиванию. Здесь этого нет. Тем не менее, можно построить метод решения системы уравнений (219), непосредственно обобщающий метод прогонки, так хорошо себя зарекомендовавший в одномерном случае.

Будем считать совокупность значений $u_{k,1}$, $u_{k,2}$, ..., $u_{k,M}$ при фиксированном k компонентами M-мерного вектора u_k (рис. 18). Отберем из системы (219) уравнения, соответствующие этому значению k. Очевидно, они будут связывать компоненты только трех векторов u_{k-1}, u_k, u_{k+1}. Запишем эти M уравнений в виде одного векторного уравнения

$$-Au_{k-1} + Bu_k - Cu_{k+1} = d_k, \tag{221}$$

где A, B, C — квадратные матрицы, а d_k — вектор порядка M.

Очевидно, $A = C = rI$ (I — единичная матрица), и

$$B = \begin{pmatrix} 1+4r & -r & 0 & & & 0 \\ -r & 1+4r & -r & & & \\ 0 & -r & 1+4r & \ddots & & \\ & \ddots & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1+4r & -r \\ 0 & & & & -r & 1+4r \end{pmatrix},$$

$$d_k = \begin{pmatrix} u^n_{k,1} + r\gamma_k \\ u^n_{k,2} \\ u^n_{k,3} \\ \vdots \\ u^n_{k,M-1} \\ u^n_{k,M} + r\delta_k \end{pmatrix}.$$

Поскольку система (219) свелась к K уравнениям (221), связывающим только тройки векторов u_{k-1}, u_k, u_{k+1}, то мы можем применить метод прогонки, учитывая, конечно, векторный характер уравнений (221).

Пусть между u_{k-1} и u_k имеется соотношение

$$u_{k-1} = L_k u_k + M_k, \tag{222}$$

где L_k — квадратная матрица, а M_k — вектор того же порядка, что и u_k. Подставив (222) в (221), исключим u_{k-1}, т. е. получим соотношение между следующей парой векторов

$$(B - AL_k)\, u_k - Cu_{k+1} = AM_k + d_k.$$

Разрешая последнее относительно u_k, т. е. умножая слева на матрицу, обратную к $B - AL_k$, получаем

$$u_k = (B - AL_k)^{-1}\ (Cu_{k+1} + AM_k + d_k).$$

Чтобы представить последнее соотношение в виде (222), положим

$$\left.\begin{aligned} L_{k+1} &= (B - AL_k)^{-1} C, \\ M_{k+1} &= (B - AL_k)^{-1} (AM_k + d_k). \end{aligned}\right\} \tag{223}$$

Метод решения теперь ясен. Граничное условие при $k = 0$ определяет $L_1 = 0$, $M_1 = \alpha$. По формулам (223) находим последовательно все L_k, M_k. Поскольку u_{K+1} известно, $u_{K+1} = \beta$, то, имея L_k, M_k, по формуле (222) получаем все u_k — решение нашей системы. Коэффициенты L_k — матрицы, поэтому изложенный метод называют *методом матричной прогонки*.

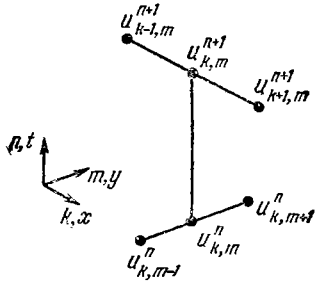

Рис. 19.

Внешне он не отличается от обычного, одномерного, метода прогонки. Однако, в противоположность последнему, употребляется крайне редко. Причиной этого является его колоссальная трудоемкость. На каждом цикле вычислительного процесса нужно обратить матрицу высокого $\sim 1/h$ порядка (и запомнить $\sim 1/h$ таких матриц). Поэтому часто применение итерационных способов решения систем (и даже использование явных схем) оказывается более эффективным.

Решение проблемы создания наиболее экономного алгоритма для рассматриваемого типа задач лежит на другом пути. Рассмотрим разностную схему (рис. 19):

$$\frac{u^{n+1}_{k,m} - u^{n}_{k,m}}{\tau} = \frac{u^{n+1}_{k+1,m} - 2u^{n+1}_{k,m} + u^{n+1}_{k-1,m}}{h_x^2} + \frac{u^{n}_{k,m+1} - 2u^{n}_{k,m} + u^{n}_{k,m-1}}{h_y^2}, \quad (224)$$

являющуюся «промежуточной» между схемами (214), (217) и также аппроксимирующую дифференциальное уравнение (213). Подстановка в (224) сеточных функций вида (216) дает $u^{n+1} = \lambda u^n$, где

$$\lambda = \frac{1 - \frac{4\tau}{h_y^2}\sin^2\frac{\psi}{2}}{1 + \frac{4\tau}{h_x^2}\sin^2\frac{\varphi}{2}},$$

и следовательно, разностная схема (224) может быть устой-

чива лишь при

$$\frac{\tau}{h_y^2} \leqslant \frac{1}{2}, \tag{225}$$

что не удивительно, так как эта схема по одной из переменных, y, — явная. Такие схемы применяют, если направления x и y неравноправны по существу задачи. Например, если решение слабо зависит от y, то h_y может быть выбран намного больше h_x, и условие (225) не будет обременительным.

Обратим внимание на следующие обстоятельства. С одной стороны, схема (224) не налагает никаких ограничений на соотношение между τ и h_x, а с другой — совокупность уравнений (224) при каждом фиксированном m образует систему, которую можно решить простым методом прогонки.

Таким образом, можно сказать, что вопрос о построении эффективного алгоритма «наполовину» решен. Остается лишь условие на τ и h_y (225). Число же арифметических операций, требуемое для получения решения, оказывается пропорциональным числу расчетных точек.

Поменяем направления x и y ролями, т. е. рассмотрим схему, явную по x и неявную по y. Получим схему, которая снимает вторую «половину» вопроса. Схемы исключают друг друга, но попробуем пользоваться ими поочередно — одной на четных, а другой на нечетных шагах по t. Поскольку элементарным циклом вычислительного процесса будет в этом случае пара шагов, то удобнее называть одним шагом по времени весь цикл, каждую схему использовать для продвижения на $\tau/2$ и решение, полученное на первой половине шага, понимать как некоторое промежуточное $\tilde{u}$.

Это дает следующую разностную схему:

$$\frac{\tilde{u}_{k,m} - u_{k,m}^n}{\tau/2} = \frac{\tilde{u}_{k+1,m} - 2\tilde{u}_{k,m} + \tilde{u}_{k-1,m}}{h_x^2} + \frac{u_{k,m+1}^n - 2u_{k,m}^n + u_{k,m-1}^n}{h_y^2}, \tag{226}$$

$$\frac{u_{k,m}^{n+1} - \tilde{u}_{k,m}}{\tau/2} = \frac{\tilde{u}_{k+1,m} - 2\tilde{u}_{k,m} + \tilde{u}_{k-1,m}}{h_x^2} + \frac{u_{k,m+1}^{n+1} - 2u_{k,m}^{n+1} + u_{k,m-1}^{n+1}}{h_y^2}. \tag{227}$$

Как и предыдущие, она, очевидно, аппроксимирует уравнение (213). Исследуем устойчивость этой схемы.

Используя в качестве u^n функцию вида (216), получим из (226) $\tilde{u} = \tilde{\lambda} u^n$, где

$$\tilde{\lambda} = \frac{1 - \frac{2\tau}{h_y^2} \sin^2 \frac{\psi}{2}}{1 + \frac{2\tau}{h_x^2} \sin^2 \frac{\varphi}{2}},$$

а из (227) $u^{n+1} = \tilde{\tilde{\lambda}} \tilde{u}$, где

$$\tilde{\tilde{\lambda}} = \frac{1 - \frac{2\tau}{h_x^2} \sin^2 \frac{\varphi}{2}}{1 + \frac{2\tau}{h_y^2} \sin^2 \frac{\psi}{2}}.$$

Нас интересует только произведение $\tilde{\lambda}\tilde{\tilde{\lambda}} = \lambda$, так как именно оно соответствует целому шагу по времени. Легко видеть, что λ есть произведение двух сомножителей

$$\lambda = \frac{1 - \frac{2\tau}{h_x^2} \sin^2 \frac{\varphi}{2}}{1 + \frac{2\tau}{h_x^2} \sin^2 \frac{\varphi}{2}} \cdot \frac{1 - \frac{2\tau}{h_y^2} \sin^2 \frac{\psi}{2}}{1 + \frac{2\tau}{h_y^2} \sin^2 \frac{\psi}{2}}, \tag{228}$$

каждый из которых не превосходит по модулю единицы при любых τ, h_x, h_y, φ, ψ.

Таким образом, разностная схема (226), (227), как и неявная схема (217), устойчива при любых соотношениях шагов сетки и, в отличие от (217), по количеству операций достаточно экономна. Действительно, процесс решения системы уравнений (226), (227) сводится, во-первых, к решению системы (226) при каждом фиксированном m, нахождению $\tilde{u}$, и, во-вторых, к решению системы (227) при каждом фиксированном k, что дает u^{n+1}. И то, и другое можно выполнить с помощью обычного метода прогонки. Разностные методы такого типа называют *методами переменных направлений* или *методами дробных шагов*.

Понятно, почему метод матричной прогонки менее эффективен по сравнению с изложенным — он слишком универсален. Действительно, как мы знаем, достоинства

обычного метода прогонки объясняются очень точным учетом взаимного влияния решения в различных точках. Взглянем на метод матричной прогонки с этой точки зрения. Запись соотношения между векторами u_{k-1}, u_k в виде (222) отражает формально существующие связи между всеми компонентами этих векторов. Очевидно, однако, что взаимное влияние различных компонент быстро ослабевает по мере удаления соответствующих им расчетных точек друг от друга, при увеличении разницы в номерах m. Метод матричной прогонки эту специфику системы уравнений не учитывает — он ориентирован на гораздо более широкий класс задач и потому в данном частном случае оказывается далеко не лучшим.

Мы рассмотрели лишь одну проблему, возникающую при переходе от одномерных задач к двумерным. Как и прежде, мы использовали для этого частный характерный пример, позволивший нам выявить существо дела. Увеличение размерности задач, конечно, ставит и другие проблемы, но мы на них останавливаться не будем. В основном они связаны с трудностями аппроксимации многомерных областей хорошими расчетными сетками и, разумеется, борьбой за простоту, экономичность вычислительного алгоритма.

Задачи

1. Построить и исследовать различные разностные схемы для решения уравнения

$$\frac{\partial U}{\partial t} + a\frac{\partial U}{\partial x} + b\frac{\partial U}{\partial y} = 0,$$

используя в качестве расчетной ячейку, изображенную на рис. 16.

Исследовать также разностную схему вида

$$\frac{u_{k,m}^{n+1} - u_{k,m}^{n-1}}{2\tau} + a\,\frac{u_{k+1,m}^{n} - u_{k-1,m}^{n}}{2h_x} + b\,\frac{u_{k,m+1}^{n} - u_{k,m-1}^{n}}{2h_y} = 0.$$

2. Рассмотреть итерационные способы решения системы уравнений (219), (220). Оценить количество итераций.

3. Оценить число арифметических операций, необходимых для решения задачи на конечном отрезке времени и в ограниченной области, при использовании:

а) явной схемы (214),

б) неявной схемы (217), решаемой методом матричной прогонки,

в) неявной схемы (217), применяя итерационные способы,

г) метода переменных направлений (226), (227).

В последних трех случаях принять $\tau \sim h$.

4. Исключить из формул (226), (227) промежуточную величину $\tilde{u}$. Сравнить полученное разностное уравнение с неявной схемой (217).

5. Показать, что для решения задачи (213) могут быть использованы разностные алгоритмы, определяемые формулами

$$\frac{\tilde{u}_{k,m} - u^n_{k,m}}{\tau} = \frac{\tilde{u}_{k+1,m} - 2\tilde{u}_{k,m} + \tilde{u}_{k-1,m}}{h_x^2} + \frac{u^n_{k,m+1} - 2u^n_{k,m} + u^n_{k,m-1}}{h_y^2},$$

$$\frac{u^{n+1}_{k,m} - \tilde{u}_{k,m}}{\tau} = \frac{u^{n+1}_{k,m+1} - 2u^{n+1}_{k,m} + u^{n+1}_{k,m-1}}{h_y^2} - \frac{u^n_{k,m+1} - 2u^n_{k,m} + u^n_{k,m-1}}{h_y^2}$$

и

$$\frac{\tilde{u}_{k,m} - u^n_{k,m}}{\tau} = \frac{\tilde{u}_{k+1,m} - 2\tilde{u}_{k,m} + \tilde{u}_{k-1,m}}{h_x^2},$$

$$\frac{u^{n+1}_{k,m} - \tilde{u}_{k,m}}{\tau} = \frac{u^{n+1}_{k,m+1} - 2u^{n+1}_{k,m} + u^{n+1}_{k,m-1}}{h_y^2}.$$

Провести сравнение их с неявной схемой (217) и методом переменных направлений (226), (227) (исключив $\tilde{u}$).

6. Рассмотреть возможности обобщения всех упоминаемых в этом параграфе методов на соответствующие трехмерные задачи.

§ 12. СТАЦИОНАРНЫЕ ЗАДАЧИ

Этот термин применяется для задач, описывающих стационарные, не меняющиеся во времени, состояния различных систем. Типичным представителем такого рода задач является следующая. Требуется найти функцию $U(x, y)$, удовлетворяющую, в некоторой ограниченной области G плоскости x, y, уравнению

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = f(x, y) \tag{229}$$

и принимающую на границе Г этой области заданные значения

$$U|_{\Gamma} = g. \tag{230}$$

Построение соответствующей разностной задачи не представляет труда. Покрываем область G расчетной сеткой, для простоты с равными шагами по x и y (рис. 20). Уравнение (229) заменяем на этой сетке разностным соотношением

$$\frac{u_{k+1,m}-2u_{k,m}+u_{k-1,n}}{h^2}+\frac{u_{k,m+1}-2u_{k,m}+u_{k,m-1}}{h^2}=f_{k,m}, \quad (231)$$

которое имеет смысл для каждой внутренней расчетной точки. Под внутренней будем понимать любую точку k, m, для которой все четыре соседние точки $k \pm 1$,

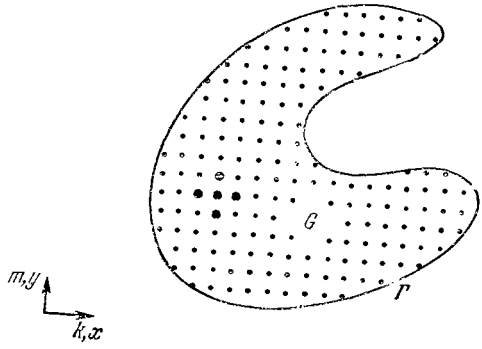

Рис. 20.

$m \pm 1$, используемые в (231), расположены внутри области G. Остальные расчетные точки k, m, принадлежащие G, объявим граничными, их совокупность обозначим через γ. Значения $u_{k,m}$ на γ получим просто переносом значения g из ближайшей точки границы Г

$$u\,|_{\gamma}=g\,|_{\Gamma}. \quad (232)$$

Разностная задача (231), (232) принадлежит к рассмотренному в § 5 классу. Ее исследование сводится к проверке аппроксимации и устойчивости. Первое очевидно, так как (231) в пределе при $h \to 0$ переходит в уравнение (229), а (232) в (230) (расстояние между γ и Г порядка h).

Докажем устойчивость построенной разностной задачи, т. е. совпадение порядков решения и правых частей (231), (232) при любом h. Для этого воспользуемся следующим приемом.

Пусть решение задачи u существует. Рассмотрим две вспомогательные функции v_+ и v_-, положив

$$v_\pm = \pm u + \alpha (x^2 + y^2) + \beta, \tag{233}$$

где α, β — пока произвольные постоянные. Обозначим левую часть (231) через Du. Подставим (233) в (231). Получим

$$Dv_\pm = \pm f + 4\alpha,$$

так как $Du = f$, $D(x^2 + y^2) = 4$, $D\beta = 0$.

Выберем α таким, чтобы всюду в области G выполнялось неравенство

$$Dv_\pm \geqslant 0. \tag{234}$$

Очевидно, для этого достаточно положить

$$\alpha = \frac{1}{4} \max_{k,m} |f_{k,m}|. \tag{235}$$

Определим теперь β так, чтобы

$$v_\pm |_\gamma \leqslant 0. \tag{236}$$

В соответствии с (232), (233) это будет при

$$\beta = -\max_\Gamma |g| - \alpha \max_G (x^2 + y^2). \tag{237}$$

Допустим, что $\max v_\pm$ достигается в некоторой внутренней точке k, m. Тогда хотя бы в одной из соседних с ней точек значение $v_\pm$ меньше максимального, и, как нетрудно увидеть из (231), $(Dv_\pm)_{k,m} < 0$. Это противоречит (234) и, следовательно, $\max v_\pm$ может достигаться только на границе γ. Но здесь, в силу (236), $v_\pm$ отрицательна. Значит, она отрицательна всюду. Таким образом,

$$\pm u_{k,m} + \alpha (x^2 + y^2)_{k,m} + \beta \leqslant 0,$$

т. е. в соответствии с (235), (237),

$$\max_{k,m} |u_{k,m}| \leqslant \max_\Gamma |g| + \frac{1}{4} \max_{k,m} |f_{k,m}| \max_G (x^2 + y^2). \tag{238}$$

Полученное неравенство означает устойчивость задачи (231), (232).

Заодно мы доказали существование и единственность решения этой задачи. Действительно, поскольку всякое решение должно удовлетворять (238), то при $g = f = 0$

возможно только тривиальное решение $u = 0$. Следовательно, решение неоднородной системы линейных уравнений (231), (232) существует и единственно.

Перейдем к вопросу о способах решения системы (231), (232). Исторически первыми были способы итерационные. Простейший из них следующий.

Разрешим каждое уравнение (231) относительно значения $u_{k,m}$ в центральной точке расчетной ячейки:

$$u_{k,m} = \frac{1}{4}(u_{k-1,m} + u_{k+1,m} + u_{k,m-1} + u_{k,m+1} - h^2 f_{k,m}) \quad (239)$$

и используем эту формулу для проведения итераций. Вычислительный процесс весьма прост — на каждой ν-й итерации вычисляем среднее арифметическое значений $u^{(\nu)}_{k\pm1,m\pm1}$ в точках, окружающих данную центральную, и получаем следующее приближение $u^{(\nu+1)}_{k,m}$.

Исследуем сходимость этого процесса. Положим

$$u^{(\nu)}_{k,m} = u_{k,m} + \delta^{(\nu)}_{k,m},$$

где $u_{k,m}$ — точное решение системы (231), (232). Тогда, очевидно, ошибка $\delta_{k,m}$ будет определяться следующим итерационным процессом:

$$\delta^{(\nu+1)}_{k,m} = \frac{1}{4}(\delta^{(\nu)}_{k+1,m} + \delta^{(\nu)}_{k-1,m} + \delta^{(\nu)}_{k,m+1} + \delta^{(\nu)}_{k,m-1}), \ \delta^{(\nu+1)}|_{\gamma} = 0. \quad (240)$$

Обозначим максимум модуля $\delta^{(0)}_{k,m}$ через $\bar{\delta}$ и будем рассуждать так. Поскольку $\delta^{(1)}_{k,m}$ есть среднее арифметическое четырех значений $\delta^{(0)}_{k\pm1,m\pm1}$, то $|\delta^{(1)}_{k,m}|$ также не превосходит $\bar{\delta}$, это справедливо для всех точек, на всех итерациях. Но для точек, соседних с граничными, можно сделать более точную оценку. А именно, если хотя бы одна из соседних для данной точки k, m — граничная, где $\delta = 0$, то в этой точке k, m

$$|\delta^{(1)}_{k,m}| \leqslant \frac{0 + 3\bar{\delta}}{4} = \frac{3}{4}\bar{\delta}.$$

Последнее неравенство справедливо для всего приграничного слоя точек. Перейдем ко второй итерации — влияние границы распространится еще на один слой точек,

в которых

$$|\delta^{(2)}_{k,m}| \leqslant \frac{\frac{3}{4}\bar{\delta} + 3\bar{\delta}}{4} = \frac{15}{16}\bar{\delta}.$$

Эта оценка заведомо справедлива и для первого приграничного слоя точек. Продолжая рассуждение, будем продвигаться с каждой итерацией в глубь области G, получая для пройденных слоев точек оценку

$$|\delta^{(\nu)}_{k,m}| \leqslant \left(1 - \frac{1}{4^{\nu}}\right)\bar{\delta}.$$

Наконец, на какой-то n-й итерации, $n \sim 1/h$, мы исчерпаем все расчетные точки. Это значит, что за n итераций ошибка δ уменьшится, по крайней мере, в $(1-4^{-n})$ раз. За следующие n итераций — еще во столько же раз, и т. д. Мы доказали, что при $\nu \to \infty$ ошибка $\delta^{(\nu)} \to 0$, т. е. итерационный процесс сходится.

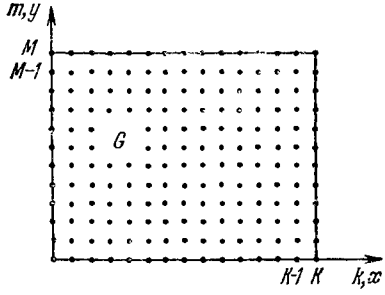

Рис. 21.

За n итераций ошибка убывает в $(1-4^{-n})$ раз, следовательно за одну итерацию, в среднем, в $(1-4^{-n})^{1/n}$ или (так как $n \sim 1/h$) в $(1-4^{-1/h})^h \sim 1-h4^{-1/h}$ раз. Принято *скорость сходимости* характеризовать величиной относительного убывания ошибки за одну итерацию, т. е. отношением $(\delta^{(\nu)} - \delta^{(\nu+1)})/\delta^{(\nu)} = \varkappa$. В данном случае для этой величины мы получили оценку

$$\varkappa \sim h4^{-1/h}. \tag{241}$$

Фактически для многих случаев эта оценка оказывается слишком завышенной, итерационный процесс сходится быстрее. Так, если область G является прямоугольником (рис. 21), то нетрудно получить более точную оценку. Для этого заметим, что формула (240), описывающая эволюцию ошибки $\delta^{(\nu)}$ от итерации к итерации, может быть интерпретирована как формула разностной задачи

$$\delta^{(\nu+1)} = R\delta^{(\nu}$$

уже знакомого нам класса (§ 6). Все отличие состоит лишь в том, что ν есть теперь номер итерации, а не номер временнóго слоя. Поэтому для исследования эволюции ошибки $\delta^{(\nu)}$ мы можем применить спектральный признак. Не будем пока принимать в расчет граничных условий $\delta|_\gamma = 0$ и положим

$$\delta^{(\nu)}_{k,m} = \delta^{(\nu)}_{0,0} e^{i(k\varphi + m\psi)}. \tag{242}$$

Как всегда, $\delta^{(\nu+1)}$ оказывается равным $\lambda\delta^{(\nu)}$, причем в данном случае легко получаем, что

$$\lambda = \frac{\cos\varphi + \cos\psi}{2}, \tag{243}$$

т. е. $|\lambda| \leqslant 1$. Такая оценка обеспечивала нам устойчивость эволюционных задач, но сейчас она недостаточна. Нас устроит только строгое неравенство $|\lambda| < 1$, только это гарантирует сходимость: $\delta^{(\nu)} \to 0$ при $\nu \to \infty$. Легко видеть, что $|\lambda| = 1$ соответствует функциям (242), которые не удовлетворяют граничным условиям $\delta|_\gamma = 0$ (они получаются при $\varphi = \psi = 0$ или $\varphi = \psi = \pi$), и потому оценка $|\lambda|$ может быть уточнена.

Оставим только те комбинации функций (242), которые обращаются в нуль на границах нашей прямоугольной области G (рис. 21), т. е. удовлетворяют условиям

$$\begin{aligned} \delta_{0,m} = \delta_{K,m} = 0, \qquad & m = 0, 1, \ldots, M, \\ \delta_{k,0} = \delta_{k,M} = 0, \qquad & k = 0, 1, \ldots, K. \end{aligned}$$

Поскольку экспоненты, фигурирующие в (242), выражаются через $\sin k_\varphi$, $\cos k\varphi$, $\sin m\psi$, $\cos m\psi$, то интересующие нас функции являются комбинациями последних. Чтобы удовлетворить левому граничному условию $\delta_{0,m} = 0$, эти функции должны содержать множитель $\sin k\varphi$. Требуя выполнения условия на правой границе при $k = K$, приходим к равенству $\sin K\varphi = 0$. Это возможно только при $K\varphi = p\pi$, где p — целое. Таким образом, мы должны рассматривать только дискретный набор φ, и даже конечный

$$\varphi_p = p\frac{\pi}{K}, \qquad p = 1, 2, \ldots, K-1, \tag{244}$$

так как при остальных p мы получим те же сеточные функции $\sin k\varphi_p$, а при $p = 0$ или $p = K$ — нулевую функцию на сетке.

По аналогичным соображениям, наши функции $\delta_{k,m}$ должны содержать множитель $\sin m\psi_q$, причем

$$\psi_q = q\frac{\pi}{M}, \qquad q = 1, 2, \ldots, M-1. \tag{245}$$

Итак, сеточные функции

$$\delta_{k,m} = \sin k\varphi_p \sin m\psi_q \tag{246}$$

при любых φ_p, ψ_q, определяемых равенствами (244), (245), удовлетворяют граничным условиям.

Подставим (246) в (240) вместо $\delta_{k,m}^{(\nu)}$. Получим после несложных выкладок

$$\delta_{k,m}^{(\nu+1)} = \frac{\cos\varphi_p + \cos\psi_q}{2} \sin k\varphi_p \sin m\psi_q,$$

т. е. $\delta_{k,m}$ (246) являются собственными функциями итерационного оператора, а соответствующие им собственные значения выражаются формулой

$$\lambda_{p,q} = \frac{\cos\varphi_p + \cos\psi_q}{2}, \tag{247}$$

совпадающей с (243). Запас собственных функций (246) большой (можно показать, что достаточно большой), и по величине $\lambda_{p,q}$ (247) можно судить о действительной скорости сходимости итерационного процесса.

Наибольшие значения $|\lambda_{p,q}|$ достигаются при крайних значениях φ_p, ψ_q, т. е. при $|\cos\varphi_p| = \cos(\pi/K)$ и $|\cos\psi_q| = \cos(\pi/M)$. Так как Kh и Mh определяют размеры области G, т. е. порядка единицы, то мы можем написать

$$\max_{p,q}|\lambda_{p,q}| = \frac{\cos(\pi/K) + \cos(\pi/M)}{2} \sim \cos h \sim 1 - \frac{h^2}{2}.$$

Следовательно, введенная выше характеристика скорости сходимости итераций

$$\varkappa \sim h^2, \tag{248}$$

что, конечно, лучше, чем (241), полученное грубой оценкой.

Мы уже отметили аналогию между итерационным процессом и эволюционной разностной задачей. Фактически

она простирается гораздо дальше. Всякую стационарную задачу можно рассматривать как частный случай эволюционной, нестационарной, где нас интересует лишь конечное, установившееся состояние, а не сам процесс установления. Это отражается и на методах решения стационарных задач. Все они итерационные, начиная от простейших — вычисления корней уравнений, все используют эволюцию, пусть фиктивную. Исключением являются линейные задачи, но если учесть, что даже деление есть операция итерационная, то ясно, что не это исключение подтверждает общее правило. Мы начали с того, что уравнение $f(x) = 0$ записали в виде $x = \varphi(x)$ (§ 1). Кончаем тем, что проделываем эту процедуру с задачей данного параграфа.

Итак, для решения стационарных задач мы можем применять весь аппарат, созданный для задач эволюционных. Рассмотренная формула итерационного процесса (239) может быть переписана в виде

$$\frac{u_{k,m}^{(\nu+1)} - u_{k,m}^{(\nu)}}{h^2/4} =$$

$$= \frac{u_{k-1,m}^{(\nu)} - 2u_{k,m}^{(\nu)} + u_{k+1,m}^{(\nu)}}{h^2} + \frac{u_{k,m-1}^{(\nu)} - 2u_{k,m}^{(\nu)} + u_{k,m+1}^{(\nu)}}{h^2} - f_{k,m}.$$

Если ν считать номером слоя по «времени», а $h^2/4$ — шагом τ, то это — известная нам явная разностная схема (214) для решения двумерного уравнения теплопроводности. Заметим, что условие устойчивости ее выполнено, так как $\tau/h^2 = 1/4$.

Как мы помним, наилучшим методом решения указанной задачи оказался метод переменных направлений. Он не налагает никаких ограничений на шаг τ, и следует ожидать, что применение его позволит нам быстрее достигнуть предельного стационарного состояния — решения нашей задачи (231), (232).

Поэтому обратимся опять к формулам (226), (227), добавив правую часть $f_{k,m}$, и используем их вместе с соответствующими граничными условиями для решения рассматриваемой стационарной задачи. Чтобы упростить исследование метода, ограничимся случаем, когда область G — квадрат $0 \leqslant x, y \leqslant X$, а шаги сетки одинаковы: $h_x = h_y = h$. Переход от ν-й итерации к $(\nu + 1)$-й

состоит в решении системы уравнений

$$\frac{\tilde{u}_{k,m} - u^{(\nu)}_{k,m}}{\tau/2} = \frac{\tilde{u}_{k+1,m} - 2\tilde{u}_{k,m} + \tilde{u}_{k-1,m}}{h^2} + \frac{u^{(\nu)}_{k,m+1} - 2u^{(\nu)}_{k,m} + u^{(\nu)}_{k,m-1}}{h^2} - f_{k,m} \tag{249}$$

относительно $\tilde{u}$, а затем системы уравнений

$$\frac{u^{(\nu+1)}_{k,m} - \tilde{u}_{k,m}}{\tau/2} = \frac{\tilde{u}_{k+1,m} - 2\tilde{u}_{k,m} + \tilde{u}_{k-1,m}}{h^2} + \frac{u^{(\nu+1)}_{k,m+1} - 2u^{(\nu+1)}_{k,m} + u^{(\nu+1)}_{k,m-1}}{h^2} - f_{k,m} \tag{250}$$

относительно $u^{(\nu+1)}$. Индексы k, m пробегают значения от 0 до $K = X/h$, причем величины $\tilde{u}$, $u^{(\nu+1)}$, как и $u^{(\nu)}$, на границе, т. е. при k, $m = 0$, K заданы, а каждой внутренней точке соответствует пара уравнений (249), (250).

Относительно вычислительного алгоритма решения систем (249) и (250) все уже было сказано, сейчас нас интересует только скорость сходимости итерационного процесса. Соответствующее исследование аналогично проведенному выше для простой итерации. Полагая $u^{(\nu)} = u + \delta^{(\nu)}$, получаем для ошибки $\delta^{(\nu)}$ ту же разностную задачу (249), (250), с $f = 0$ и нулевыми граничными условиями. Сеточные функции $\delta_{k,m}$ (246), с

$$\varphi_p = p\pi/K, \ \psi_q = q\pi/K; \qquad p, q = 1, 2, \ldots, K-1, \tag{251}$$

опять будут собственными функциями итерационного оператора. Собственные значения λ, которые и определяют скорость сходимости, выражаются формулой (228) с $\varphi = \varphi_p$, $\psi = \psi_q$, т. е. являются произведениями однотипных сомножителей, $\lambda = \lambda_p\lambda_q$. Выпишем один из них —

$$\lambda_p = \frac{1 - \frac{2\tau}{h^2}\sin^2\frac{\varphi_p}{2}}{1 + \frac{2\tau}{h^2}\sin^2\frac{\varphi_p}{2}}. \tag{252}$$

Второй, λ_q, получается отсюда заменой φ_p на ψ_q, и следовательно,

$$\max|\lambda| = \max|\lambda_p|\max|\lambda_q| = \max|\lambda_p|^2. \tag{253}$$

То, что $\max|\lambda|<1$ при любом положительном значении параметра τ, т. е. процесс итераций всегда сходится,— очевидно. Но для определения скорости сходимости нам нужна возможно более точная оценка величины $|\lambda|$.

Так как $Kh = X$, то φ_p меняется в пределах

$$h\pi/X \leqslant \varphi_p \leqslant \pi - h\pi/X, \tag{254}$$

и для λ_p при малых h справедлива оценка (рис. 22)

$$\frac{1-\tau\frac{\pi^2}{2X^2}[1+O(h^2)]}{1+\tau\frac{\pi^2}{2X^2}[1+O(h^2)]} \geqslant \lambda_p \geqslant \frac{1-\tau\frac{2}{h^2}[1-O(h^2)]}{1+\tau\frac{2}{h^2}[1-O(h^2)]}. \tag{255}$$

Мы заинтересованы в минимальности $\max|\lambda_p|$. В нашем распоряжении параметр τ. Он представляет фиктивное время, и, казалось бы, чем τ больше, тем сходимость должна быть лучше, так как мы быстрее будем приближаться к предельному состоянию. Действительно, с ростом τ левая часть (255) убывает. Однако при этом правая часть (255) будет приближаться к -1, что замедлит сходимость. Объяснить этот эффект можно тем, что хотя при больших τ мы продвигаемся по «времени» быстрее, но делаем это слишком грубо — неточность получаемого решения превосходит изменение его.

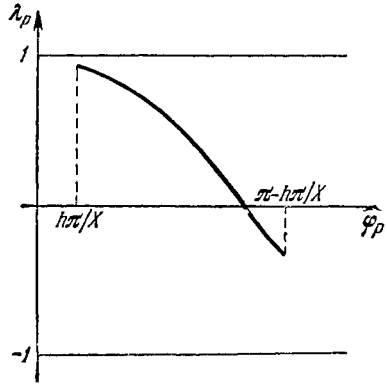

Рис. 22.

Очевидно, оптимальным значением τ будет то, при котором левая и правая части (255) равны по модулю. Отсюда, отбрасывая младшие члены $O(h^2)$, получаем уравнение для τ:

$$\frac{1-\tau\frac{\pi^2}{2X^2}}{1+\tau\frac{\pi^2}{2X^2}} = -\frac{1-\tau\frac{2}{h^2}}{1+\tau\frac{2}{h^2}},$$

решив которое, найдем

$$\tau = hX/\pi. \tag{256}$$

Подставляя это значение τ в неравенства (255), получим оценку

$$\max|\lambda_p| = 1 - h\pi/X + O(h^2),$$

т. е., в силу (253),

$$\max|\lambda| = 1 - 2h\pi/X + O(h^2),$$

и величина $\varkappa$, характеризующая скорость сходимости, оказывается порядка h,

$$\varkappa \sim h\frac{2\pi}{X}. \tag{257}$$

Таким образом, итерационный процесс, использующий метод переменных направлений с $\tau \sim h$, имеет скорость сходимости на порядок лучше, чем простая итерация, где $\varkappa \sim h^2$ (248).

Если представить себе ошибку δ разложенной на отдельные компоненты, гармоники — функции (246), (251), то $\lambda_p\lambda_q$ будет коэффициентом погашения соответствующей компоненты за одну итерацию. Из формулы (252) и рис. 22 видно, что различные компоненты гасятся неодинаково. Наиболее сильно подавляются гармоники с частотами φ_p, для которых

$$\frac{2\tau}{h^2}\sin^2\frac{\varphi_p}{2} \sim 1.$$

Очевидно, выбором τ можно этот диапазон частот регулировать. Так, при вышеуказанном значении τ (256) это — средние частоты. Однако их относительно сильное подавление ценности не представляет — скорость сходимости определяется коэффициентами погашения крайних частот, φ_p, $\psi_q \sim h$ и φ_p, $\psi_q \sim \pi$. Это наводит на мысль об использовании на различных итерациях различных значений τ, с целью равномерного гашения всех частот. Таким приемом, употребляя специальным образом построенные последовательности $\tau = \tau^{(\nu)}$, удается получить методы, имеющие еще большую, чем (257), скорость сходимости, например (см. задачу 2)

$$\varkappa \sim \frac{1}{\ln(1/h)}. \tag{258}$$

На описании этих методов мы останавливаться не будем.

Все итерационные методы решения системы разностных уравнений (231) имеют малую скорость сходимости, которая более или менее резко падает с уменьшением h: $\varkappa \to 0$ при $h \to 0$. Тем не менее они все равно оказываются выгоднее, чем прямые, неитерационные методы. Оценим количество арифметических операций N, требуемое на решение задачи.

Начнем с общего метода исключения. При использовании его число необходимых операций порядка куба числа неизвестных. Последнее порядка $1/h^2$, следовательно,

$$N \sim 1/h^6. \tag{259}$$

В предыдущем параграфе был описан метод матричной прогонки, очевидно, применимый и в данном случае. Он требует

$$N \sim 1/h^4 \tag{260}$$

арифметических операций, поскольку на обращение матрицы порядка $1/h$ затрачивается $\sim 1/h^3$ операций, и таких матриц $\sim 1/h$ штук.

Итерационные методы дают приближенное решение системы. Поэтому их трудоемкость оценивают по количеству операций, необходимому для уменьшения ошибки в заданное число раз. За одну итерацию ошибка убывает в $1 - \varkappa$ раз. Если ошибку начального приближения принять за единицу, то после n итераций она будет равна

$$(1 - \varkappa)^n = \varepsilon.$$

Следовательно, для достижения точности ε требуется

$$n = \frac{\ln \varepsilon}{\ln (1 - \varkappa)} \sim \frac{\ln (1/\varepsilon)}{\varkappa}$$

итераций. В рассмотренных методах количество арифметических операций на одной итерации порядка числа точек сетки, т. е. $\sim 1/h^2$. Следовательно, общее количество операций

$$N \sim \frac{\ln (1/\varepsilon)}{\varkappa h^2}.$$

Подставляя сюда различные значения $\varkappa$ — (248), (257), (258), получим: для простой итерации

$$N \sim \ln (1/\varepsilon)/h^4, \tag{261}$$

для метода переменных направлений

$$N \sim \ln(1/\varepsilon)/h^3, \tag{262}$$

а в случае специального выбора $\tau^{(\nu)}$

$$N \sim \ln(1/\varepsilon)\ln(1/h)/h^2. \tag{263}$$

Сравнение (259), (260) с (261) — (263) говорит не в пользу первых. Разумеется, экономичность итерационных способов объясняется тем, что при их использовании удается максимально учесть специфику системы уравнений.

Задачи

1. Определить оптимальное значение параметра τ при проведении итерационного процесса по формулам типа (249), (250) в случае, когда область G является не квадратом, а прямоугольником со сторонами X, Y и шаги сетки h_x, h_y не равны друг другу.

2. Доказать возможность достижения скорости сходимости итераций, указанной в формуле (258). Для этого рассмотреть итерационный процесс (249), (250) с переменным τ:

$$\tau^{(\nu)} = \tau^{(1)} z^{\nu-1}, \qquad \nu = 1, 2, \ldots, n,$$

$$\tau^{(\nu)} = \tau^{(\nu-n)}, \qquad \nu = n+1, n+2, \ldots$$

Подобрать значения $\tau^{(1)}$ и n так, чтобы величина

$$\xi_p^{(\nu)} = \frac{2\tau^{(\nu)}}{h^2}\sin^2\frac{\varphi_p}{2},$$

фигурирующая в выражении λ_p (252), для каждого допустимого p при некотором $\nu = \nu(p)$ попадала в интервал между $\sqrt{z}$ и $1/\sqrt{z}$. Тогда соответствующее этому ν значение $\lambda_p^{(\nu)}$ будет удовлетворять неравенству

$$|\lambda_p^{(\nu)}| < \left|\frac{1-\sqrt{z}}{1+\sqrt{z}}\right|,$$

используя которое, можно оценить среднее за цикл из n итераций значение λ_p и, следовательно, скорость сходимости.

3. Исследовать возможность применения методов, указанных в задаче 5 § 11, для решения стационарных задач.

4. Предложить вычислительные алгоритмы решения уравнения (229) в областях со сложной геометрией (фигура, составленная из прямоугольников, круг и др.), при различных граничных условиях.

5. Построить и исследовать разностные схемы для решения системы уравнений

$$\frac{\partial U}{\partial x}+\frac{\partial V}{\partial y}=f(x, y),$$
$$\frac{\partial V}{\partial x}-\frac{\partial U}{\partial y}=g(x, y)$$

в прямоугольнике $0 \leqslant x \leqslant X$, $0 \leqslant y \leqslant Y$ с граничными условиями на его сторонах

$$U(0, y)=\alpha(y), \qquad U(x, Y)=\beta(x),$$
$$V(x, 0)=\gamma(x), \qquad V(X, y)=\delta(y).$$

В частности, рассмотреть итерационный алгоритм, использующий разностные уравнения вида

$$\tilde{u}_{k, m-1/2}+\tau \frac{\tilde{u}_{k, m-1/2}-\tilde{u}_{k-1, m-1/2}}{h_x}=$$
$$=u^n_{k, m-1/2}-\tau \frac{v^n_{k-1/2, m}-v^n_{k-1/2, m-1}}{h_y}+\tau f_{k-1/2, m-1/2},$$
$$\tilde{v}_{k-1/2, m}-\tau \frac{\tilde{v}_{k+1/2, m}-\tilde{v}_{k-1/2, m}}{h_x}=$$
$$=v^n_{k-1/2, m}-\tau \frac{u^n_{k, m+1/2}-u^n_{k, m-1/2}}{h_y}-\tau g_{k, m},$$
$$u^{n+1}_{k, m-1/2}+\tau \frac{v^{n+1}_{k-1/2, m}-v^{n+1}_{k-1/2, m-1}}{h_y}=$$
$$=\tilde{u}_{k, m-1/2}-\tau \frac{\tilde{u}_{k, m-1/2}-\tilde{u}_{k-1, m-1/2}}{h_x}+\tau f_{k-1/2, m-1/2},$$
$$v^{n+1}_{k-1/2, m}+\tau \frac{u^{n+1}_{k, m+1/2}-u^{n+1}_{k, m-1/2}}{h_y}=$$
$$=\tilde{v}_{k-1/2, m}+\tau \frac{\tilde{v}_{k+1/2, m}-\tilde{v}_{k-1/2, m}}{h_x}-\tau g_{k, m},$$
$$k=1, 2, \ldots, K; \qquad m=1, 2, \ldots, M;$$
$$h_x=\frac{X}{K+1/2}, \qquad h_y=\frac{Y}{M+1/2},$$

где n — номер итерации, $\sim$ обозначает промежуточные значения неизвестных, τ — некоторый параметр.

Владимир Федотович Дьяченко

ОСНОВНЫЕ ПОНЯТИЯ
ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ

М., 1972, 120 стр. с илл.

Редактор *Г. Я. Пирогова*
Техн. редактор *Е. Н. Земская*
Корректоры *О. А. Бутусова*, *Н. Д. Дорохова*

Сдано в набор 31/I 1972 г.
Подписано к печати 15/IV 1972 г. Бумага $84 \times 108^1/_{32}$.
Физ. печ. л. 3,75. Условн. печ. л. 6,3. Уч.-изд. л. 5,93.
Тираж 45000 экз. Т-07218. Цена книги 22 к. Заказ 116.

Издательство «Наука»
Главная редакция
физико-математической литературы
117071, Москва В-71, Ленинский проспект, 15

2-я типография издательства «Наука».
Москва, Шубинский пер., 10