ОСНОВЫ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

И. Б. Петров
А. И. Лобанов

ЛЕКЦИИ ПО ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКЕ

Основы информационных технологий

И. Б. Петров, А. И. Лобанов

ЛЕКЦИИ ПО ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКЕ

Учебное пособие

Интернет-Университет
Информационных Технологий
www.intuit.ru

БИНОМ.
Лаборатория знаний
www.lbz.ru

Москва
2006

УДК 519.6(075.8)
ББК 22.19я73
П30

П30 **Петров Игорь Борисович**
Лекции по вычислительной математике: Учебное пособие / И.Б. Петров, А. И. Лобанов. — М.: Интернет-Университет Информационных Технологий; БИНОМ. Лаборатория знаний, 2006. — 523 с.: ил., табл.— (Серия «Основы информационных технологий»)

ISBN 5-94774-542-9 (БИНОМ. ЛЗ)
ISBN 5-9556-0065-5 (ИНТУИТ.РУ)

В курсе лекций рассматриваются основные понятия и методы вычислительной математики. Курс содержит как лекции, посвященные классическим численным методам анализа и линейной алгебры, так и решению дифференциальных уравнений. Особое внимание уделяется решению систем уравнений в частных производных гиперболического типа. Большинство лекций снабжено задачами для рассмотрения на семинарских занятиях и для самостоятельного решения.

УДК 519.6(075.8)
ББК 22.19я73

Издание осуществлено при финансовой и технической поддержке издательства «Открытые Системы», «РМ Телеком» и Kraftway Computers.

По вопросам приобретения обращаться:
«БИНОМ. Лаборатория знаний»
Телефон (495) 157-1902, 157-5272,
e-mail: Lbz@aha.ru, http://www.Lbz.ru

ISBN 5-94774-542-9 (БИНОМ. ЛЗ)
ISBN 5-9556-0049-3 (ИНТУИТ.РУ)

О проекте

Интернет-Университет Информационных Технологий — это первое в России высшее учебное заведение, которое предоставляет возможность получить дополнительное образование во Всемирной сети. Web-сайт университета находится по адресу www.intuit.ru.

Мы рады, что вы решили расширить свои знания в области компьютерных технологий. Современный мир — это мир компьютеров и информации. Компьютерная индустрия — самый быстрорастущий сектор экономики, и ее рост будет продолжаться еще долгое время. Во времена жесткой конкуренции от уровня развития информационных технологий, достижений научной мысли и перспективных инженерных решений зависит успех не только отдельных людей и компаний, но и целых стран. Вы выбрали самое подходящее время для изучения компьютерных дисциплин. Профессионалы в области информационных технологий сейчас востребованы везде: в науке, экономике, образовании, медицине и других областях, в государственных и частных компаниях, в России и за рубежом. Анализ данных, прогнозы, организация связи, создание программного обеспечения, построение моделей процессов — вот далеко не полный список областей применения знаний для компьютерных специалистов.

Обучение в университете ведется по собственным учебным планам, разработанным ведущими российскими специалистами на основе международных образовательных стандартов Computer Curricula 2001 Computer Science. Изучать учебные курсы можно самостоятельно по учебникам или на сайте Интернет-Университета, задания выполняются только на сайте. Для обучения необходимо зарегистрироваться на сайте университета. Удостоверение об окончании учебного курса или специальности выдается при условии выполнения всех заданий к лекциям и успешной сдачи итогового экзамена.

Книга, которую вы держите в руках, — очередная в многотомной серии «Основы информационных технологий», выпускаемой Интернет-Университетом Информационных Технологий. В этой серии будут выпущены учебники по всем базовым областям знаний, связанным с компьютерными дисциплинами.

Добро пожаловать в
Интернет-Университет Информационных Технологий!

Анатолий Шкред
anatoli@shkred.ru

Об авторах

Петров Игорь Борисович, доктор физико-математических наук, профессор, заведующий кафедрой информатики МФТИ. Область научных интересов: математическое моделирование в динамике деформируемого твердого тела, математическое моделирование в биологии и медицине, численные методы, современные методы преподавания информатики и вычислительной математики. Автор более 200 научных и учебно-методических работ.

Лобанов Алексей Иванович, доктор физико-математических наук, профессор. Область научных интересов: математическое моделирование в динамике высокотемпературной плазмы, математическое моделирование в биологии и медицине, численные методы. Автор более 150 научных и учебно-методических работ.

Оглавление

Предисловие

В последнее время в России издается довольно много книг по вопросам вычислительной математики. При написании учебников их создателям приходится решать противоречия между строгостью изложения и компактностью, сжатостью информации, между классическими темами и новым материалом, между необходимостью описывать те разделы, которые далеки от научных интересов авторов, и желанием уделить больше внимания любимой теме. Все эти противоречия встали и перед авторами данного курса. Насколько успешно их удалось преодолеть, судить читателям.

За основу данной книги были взяты курсы вычислительной математики, которые в течение ряда лет читались студентам факультетов общей и прикладной физики, молекулярной и биологической физики, проблем физики и энергетики МФТИ. Слушатели этих курсов — не профессионалы-вычислители, а исследователи, специалисты в предметной области. Но логика современного развития науки привела к тому, что успех исследования во многом определяется эффективностью применения вычислительной техники и численных методов. В этой связи авторы посчитали необходимым дать студентам представление и о сравнительно современных численных методах. В силу быстрых изменений, происходящих сейчас в вычислительной математике, издать курс, на 100 процентов удовлетворяющий потребностям сегодняшнего дня, просто невозможно.

В книгу включены идеи, методы, разделы, которые не являются обязательными и при первом прочтении могут быть опущены. Такие лекции помечены звездочками. Отметим, что термин «лекция» несколько условен. В книгу входит годовой курс, число лекций в году составляет 30–33. В книге 19 лекций, в силу того что под лекцией здесь понимается тематический раздел, который может включать в себя материал нескольких реально читаемых лекций.

Большинство лекций книги (все основные и некоторые необязательные) снабжены задачами для разборов на семинарских занятиях и для самостоятельного решения на компьютере с использованием либо пакетов программ, либо оригинальных программ, составленных самими обучающимися. По мнению авторов, без самостоятельной реализации основных алгоритмов и простых вычислительных процедур невозможно глубокое понимание предмета. В конце каждой лекции приведен список литературы — это источники, которыми пользовались авторы при написании курса, и специальная литература, более подробно освещающая те

или иные разделы. Списки литературы не претендуют на полноту, а лишь отражают вкусы и научные пристрастия авторов.

Все замечания можно направлять по электронной почте по адресам alexey@crec.mipt.ru или petrov@mipt.ru.

Авторы выражают свою искреннюю благодарность всем коллегам по кафедре вычислительной математики МФТИ за внимание и помощь в работе, первому заведующему кафедрой академику О. М. Белоцерковскому за доброжелательную поддержку. Особая благодарность В. С. Рябенькому, Р. П. Федоренко, А. С. Холодову, учителям авторов, которые передали им свои знания и любовь к предмету. Доценты Е. Н. Аристова, О. А. Пыркова, Т. К. Старожилова и старший преподаватель В. Д. Иванов прочли части книги в рукописи и высказали ряд конструктивных замечаний и предложений. Много сил и энергии потратили Е. А. Евсюкова, Д. В. Кибардина и Е. Р. Павлюкова при технической подготовки рукописи. Авторы также благодарны Ж. И. Утюшевой и М. С. Гриневой за предоставленные конспекты лекций.

Лекция 1. Предмет вычислительной математики. Обусловленность задачи, устойчивость алгоритма, погрешности вычислений. Задача численного дифференцирования

Первая лекция носит вводный характер. На простейших примерах иллюстрируются понятия численного алгоритма, устойчивость и обусловленность задачи. На примере задачи численного дифференцирования вводится метод неопределенных коэффициентов для получения приближенных формул. Рассматривается некорректность задачи численного дифференцирования.

Ключевые слова: алгоритм, обусловленность задачи, устойчивость алгоритма, погрешности вычислений, задача численного дифференцирования, метод неопределенных коэффициентов.

Первое применение вычислительных методов принадлежит древним египтянам, которые умели вычислять диагональ квадрата за конечное количество действий. Они также могли находить квадратный корень из 2, скорее всего, с помощью алгоритма, в дальнейшем получившего название формулы Герона, а еще позднее — метода Ньютона:

$$u_{k+1} = \frac{1}{2}\left(u_k + \frac{2}{u_k}\right), \quad u_0 = a.$$

С именем среднеазиатского врача, философа и математика Аль-Хоремзи связано понятие алгоритма. Разработкой вычислительных методов занимались Л. Эйлер, которому принадлежит, по-видимому, первый численный метод для решения обыкновенных дифференциальных уравнений, И. Ньютон, О. Л. Коши, Ж. Л. Лагранж, А. М. Лежандр, П. С. Лаплас, А. Пуанкаре, П. Л. Чебышёв и многие другие известные математики. Решающую роль в развитии вычислительной математики как самостоятельной науки сыграли немецкий математик Карл Рунге и русский математик, механик и кораблестроитель А. Н. Крылов.

В наше время вычислительная математика получила значительный импульс в 1950-е годы, что было связано с развитием ядерной физики, механики полета, аэродинамики спускаемых космических аппаратов. В дальнейшем решались задачи, связанные не только с расчетами действия ядерного взрыва и обтеканием боеголовок стратегических ракет. Численные методы нашли свое применение в

таких областях как динамика атмосферы, термогидрография, физика плазмы, механика горных пород и ледников, синергетика, биомеханика, теория оптимизации, математическая экономика и др. Наиболее наукоемки и требуют максимальных вычислительных ресурсов задачи физики, механики и электродинамики сплошных сред. К ним относятся системы уравнений в частных производных Эйлера, Лагранжа, Максвелла и др., кинетической теории газов (уравнения Власова, Берда), а также задачи многомерной оптимизации. Развитие многих вычислительных методов — заслуга ученых, неразрывно связанных с МФТИ. Среди них следует упомянуть академиков А. А. Дородницына, О. М. Белоцерковского, А. А. Самарского, членов-корреспондентов РАН А. С. Холодова, Б. Н. Четверушкина, Ю. П. Попова, С. П. Курдюмова, профессоров В. С. Рябенького, Э. Э. Шноля, Р. П. Федоренко, Л. А. Чудова, В. Ф. Дьяченко, И. М. Гельфанда, Г. А. Тирского, А. П. Фаворского.

Вычислительная математика отличается от других математических дисциплин и обладает специфическими особенностями.

1. Вычислительная математика имеет дело не только с непрерывными, но и с дискретными объектами. Так, вместо отрезка прямой часто рассматривается система точек $\{t_k\}_{k=0}^{K}$, вместо непрерывной функции $f(x)$ — табличная функция $\{f_k\}_{k=0}^{K}$, вместо первой производной — ее разностная аппроксимация, например,

$$\frac{f_{k+1} - f_k}{x_{k+1} - x_k}, \quad k = 0 \div K, \quad x_{k+1} > x_k.$$

Такие замены, естественно, порождают погрешности метода.

2. В машинных вычислениях присутствуют числа с ограниченным количеством знаков после запятой из-за конечности длины мантиссы при представлении действительного числа в памяти ЭВМ. Другими словами, в вычислениях присутствует машинная погрешность (округления) δ_M. Это приводит к вычислительным эффектам, неизвестным, например, в классической теории обыкновенных дифференциальных уравнений, уравнений математической физики или в математическом анализе.

3. В вычислительной практике большое значение имеет *обусловленность задачи*, т. е. чувствительность ее решения к малым изменениям входных данных.

4. В отличие от «классической» математики выбор вычислительного алгоритма влияет на результаты вычислений.

5. Существенная черта численного метода — *экономичность* вычислительного алгоритма, т. е. минимизация числа элементарных операций при выполнении его на ЭВМ.

6. Погрешности при численном решении задач делятся на две категории — неустранимые и устранимые. К первым относят погрешности, связанные с построением математической модели объекта и приближенным заданием входных данных, ко вторым — погрешности метода решения задачи и ошибки округления, которые являются источниками малых возмущений, вносимых в решение задачи.

Специфику машинных вычислений можно пояснить на нескольких элементарных примерах.

1.1. Обусловленность задачи

Пример 1.1. Вычислить все корни уравнения

$$x^4 - 4x^3 + 8x^2 - 16x + 15.\underbrace{99999999}_{8} = (x-2)^4 - 10^{-8} = 0.$$

Точное решение задачи легко найти:

$$(x-2)^2 = \pm 10^{-4},$$
$$x_1 = 2{,}01; \quad x_2 = 1{,}99; \quad x_{3,4} = 2 \pm 0{,}01i.$$

Если компьютер работает при $\delta_M > 10^{-8}$, то свободный член в исходном уравнении будет округлен до 16,0 и, с точки зрения представления чисел с плавающей точкой, будет решаться уравнение $(x-2)^4 = 0$, т. е. $x_{1,2,3,4} = 2$, что, очевидно, неверно. В данном случае малые погрешности в задании свободного члена $\approx 10^{-8}$ привели, независимо от метода решения, к погрешности в решении $\approx 10^{-2}$.

Пример 1.2. Решается задача Коши для обыкновенного дифференциального уравнения 2-го порядка:

$$u''(t) = u(t), u(0) = 1, u'(0) = -1.$$

Общее решение имеет вид

$$u(t) = 0{,}5[u(0) + u'(0)]e^t + 0{,}5[u(0) - u'(0)]e^{-t}.$$

При заданных начальных данных точное решение задачи: $u(x) = e^{-t}$, однако малая погрешность δ в их задании приведет к появлению члена δe^t, который при больших значениях аргумента может существенно исказить решение.

Пример 1.3. Пусть необходимо найти решение обыкновенного дифференциального уравнения

$$\dot{u} = 10u, \quad u = u(t),$$
$$u(t_0) = u_0, \quad t \in [0, 1].$$

Его решение: $u(t) = u_0 e^{10(t-t_0)}$, однако значение $u(t_0)$ известно лишь приближенно: $u(t_0) \approx u_0^*$, и на самом деле $u^*(t) = u_0^* e^{10(t-t_0)}$.

Соответственно, разность $u^* - u$ будет

$$u^* - u = (u_0^* - u_0)e^{10(t-t_0)}.$$

Предположим, что необходимо гарантировать некоторую заданную точность вычислений $\varepsilon > 0$ всюду на отрезке $t \in [0, 1]$. Тогда должно выполняться условие

$$|u^*(t) - u(t)| \leqslant \varepsilon.$$

Очевидно, что $\max\limits_{t\in[0,1]} |u^*(t) - u(t)| = |u*(1) - u(1)| = |u_0^* - u_0|\, e^{10(1-t_0)}$.

Отсюда можно получить требования к точности задания начальных данных δ: $|u_0^* - u_0| < \delta, \delta \leqslant \varepsilon\, e^{-10}$ при $t_0 = 0$.

Таким образом, требование к заданию точности начальных данных оказываются в e^{10} раз выше необходимой точности результата решения задачи. Это требование, скорее всего, окажется нереальным.

Решение оказывается очень чувствительным к заданию начальных данных. Такого рода задачи называются *плохо обусловленными.*

Пример 1.4. Решением системы линейных алгебраических уравнений (СЛАУ)

$$\begin{cases} u + 10v = 11 \\ 100u + 1001v = 1101 \end{cases}$$

является пара чисел $\{1, 1\}$.

Изменив правую часть системы на 0,01, получим возмущенную систему

$$\begin{cases} u + 10v = 11.01 \\ 100u + 1001v = 1101 \end{cases}$$

с решением $\{11.01; 0.00\}$, сильно отличающимся от решения невозмущенной системы.Эта система также плохо обусловлена.

Пример 1.5. Рассмотрим полином

$$(x-1)(x-2)\ldots(x-20) = x^{20} - 210x^{19} + \ldots,$$

корни которого $x_1 = 1, x_2 = 2, \ldots, x_{20} = 20$.

Положим, что коэффициент (-210) при x_{19} увеличен на $\approx 10^{-7}$. В результате вычислений с 11-ю значащими цифрами получим совершенно иные корни: $x_1 = 1,00; x_2 = 2,00; x_3 = 3,00; x_4 = 4,00; x_5 = = 5,00; x_6 = 6,00; x_7 = 7,00; x_8 = 8,01; x_9 = 8,92; x_{10,11} = 10,1 \pm 0,644i;$

$x_{12,13} = 11,8 \pm 1,65i; x_{14,15} = 14,0 \pm 2,52i; x_{16,17} = 16,7 \pm 2,81i; x_{18,19} = 19,5 \pm 19,4i; x_{20} = 20,8$.

Причина значительного расхождения также заключается в плохой обусловленности задачи вычисления корней рассматриваемого выражения.

1.2. Влияние выбора вычислительного алгоритма на результаты вычислений

Пример 1.6. Пусть необходимо вычислить значение выражения $\left(\frac{\sqrt{2}-1}{\sqrt{2}+1}\right)^3$.

Избавившись от знаменателя, получаем $(\sqrt{2}-1)^6 = (3-2\sqrt{2})^3 = 99-70\sqrt{2}$.

Полагая а) $\sqrt{2} \approx \frac{7}{5} = 1,4$, в) $\sqrt{2} \approx \frac{17}{12} = 1,41(6)$ и рассматривая эти приближения как разные методы вычисления, получим следующие результаты:

$\sqrt{2}$	$(\sqrt{2}-1)^6$	$(3-2\sqrt{2})^3$	$99-70\sqrt{3}$
7/5	0,004096	0,008000	1
17/12	0,005233	0,004630	-0,1(6)

Очевидно, что столь значительное различие в результатах вызвано влиянием ошибки округления в задании $\sqrt{2}$.

Пример 1.7. Вычисление функции $\sin x$ с помощью ряда Тейлора.

Из курса математического анализа известно, что функция синус представляется своим рядом Тейлора

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots,$$

причем радиус сходимости ряда равен бесконечности — ряд сходится при любых значениях x.

Вычислим значения синуса при двух значениях аргумента. Пусть сначала $x = 0,5236(30°)$. Будем учитывать лишь члены ряда, большие, чем 10^{-4}. Выполнив вычисления с четырьмя значащими цифрами, получим $\sin(0.5236) = 0.5000$, что соответствует принятой точности.

Пусть теперь $x = 25,66(1470°)$. Если вычисления по данной формуле проводить с восемью значащими цифрами, то получим абсурдный результат: $\sin(25,66) \approx 24$ (учитывались члены ряда, большие, чем 10^{-8}).

Разумеется, выходом из создавшейся ситуации может быть использование формул приведения.

Пример 1.8. Вычисление функции e^x с помощью ряда Тейлора.

Из курса математического анализа известно, что экспонента представляется своим рядом Тейлора

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots,$$

радиус сходимости этого ряда также равен бесконечности.

Приведем некоторые результаты расчетов (e^x_M — значения экспоненты, вычисленные на компьютере).

x	e^x	e^x_M
1	2,718282	2,718282
20	$4,8516520 \cdot 10^8$	$4,8516531 \cdot 10^8$
-1	0,3678795	0,3678794
-10	$4,5399930 \cdot 10^{-5}$	$-1,6408609 \cdot 10^{-4}$
-20	$2,0611537 \cdot 10^{-9}$	1,202966

Выходом из этой ситуации может быть использование для отрицательных аргументов экспоненты формулы

$$e^{-x} = \frac{1}{e^x} = \frac{1}{1 + x + \frac{x^2}{2!} + \ldots}.$$

Естественно ожидать рост ошибок округления при вычислении рассматриваемой функции при больших значениях аргумента x. В этом случае можно использовать формулу $e^x = e^{n+a} = e^n e^a$, где $n = [x]$.

Пример 1.9. Рассмотрим следующий метод вычисления интеграла

$$I_n = \int_0^1 x^n e^{1-x} dx, n = 1, 2 \ldots$$

Интегрирование по частям дает

$$I_n = \int_0^1 x^n d(-e^{1-x}) - x^{n-1} \cdot e^{1-x} + \int_0^1 e^{1-x} \cdot d(x^n) = -1 + \int_0^1 n x^{n-1} e^{1-x} dx,$$

откуда следует $I_0 = \int_0^1 e^{1-x} dx = e - 1 \approx 1,71828$. $I_n = nI_{n-1} - 1, n \geqslant 1$. Тогда

$$I_1 = 1 \cdot I_0 - 1 \approx 0.71828, I_2 = 2I_1 - 1 \approx 0.43656, I_3 = 3I_2 - 1 \approx 0.30968,$$
$$I_4 = 4I_3 - 1 \approx 0.23872, I_5 = 5I_4 - 1 \approx 0.1936, I_6 = 6 \cdot I_5 - 1 \approx 0.16160,$$
$$I_7 = 7I_6 - 1 \approx 0.13120, I_8 = 8I_7 - 1 \approx 0.00496, I_9 = 9I_8 - 1 \approx -0.55360,$$
$$I_{10} = 10 \cdot I_9 - 1 \approx -6.5360$$

Очевидно, что отрицательные значения при $n = 9, 10$ не имеют смысла. Дело в том, что ошибка, сделанная при округлении I_0 до 6-ти значащих цифр сохранилась при вычислении I_1, умножилась на 2! при вычислении I_2, на 3! — при вычислении I_3, и так далее, т. е. ошибка растет очень быстро, пропорционально $n!$.

Пример 1.10. Рассмотрим методический пример вычислений на модельном компьютере, обеспечивающем точность $\delta_M = 0,0005$. Проанализируем причину происхождения ошибки, например, при вычитании двух чисел, взятых с точностью до третьей цифры после десятичной точки $u = 1,001$, $v = 1,002$, разность которых составляет $\Delta = |v_M - u_M| = 0,001$.

В памяти машины эти же числа представляются в виде

$$u_M = u(1 + \delta_M^u), v_M = v(1 + \delta_M^v), \text{ причем } |\delta_M^u| \leqslant \delta_M \text{ и } |\delta_M^v| \leqslant \delta_M.$$

Тогда $u_M - u \approx u\,\delta_M^u, v_M - v \approx v\,\delta_M^v$.

Относительная ошибка при вычислении разности $u_M - v_M$ будет равна

$$\delta = \frac{(u_M - v_M) - (u - v)}{(u - v)} = \frac{(u_M - u) - (v_M - v)}{(u - v)} = \frac{\delta_M^u - \delta_M^v}{(u - v)}.$$

Очевидно, что $\delta = \left|\frac{\delta_M^u - \delta_M^v}{\Delta}\right| \leqslant \frac{2\delta_M}{0,001} \approx 2000\,\delta_M = 1$, т. е. все значащие цифры могут оказаться неверными.

Пример 1.11. Рассмотрим рекуррентное соотношение $u_{i+1} = qu_i$, $i \geqslant 0, u_0 = a, q > 0, u_i > 0$.

Пусть при выполнении реальных вычислений с конечной длиной мантиссы на i-м шаге возникла погрешность округления, и вычисления проводятся с возмущенным значением $u_i^M = u_i + \delta_i$, тогда вместо u_{i+1} получим $u_{i+1}^M = q(u_i + \delta_i) = u_{i+1} + q\delta_i$, т. е. $\delta_{i+1} = q\delta_i, i = 0, 1, \ldots$.

Следовательно, если $|q| > 1$, то в процессе вычислений погрешность, связанная с возникшей ошибкой округления, будет возрастать (алгоритм неустойчив). В случае $|q| \leqslant 1$ погрешность не возрастает и численный алгоритм устойчив.

1.3. Экономичность вычислительного метода

Пример 1.12. Пусть требуется вычислить сумму $S = 1 + x + x^2 + \ldots + x^{1023}$ при $0 < x < 1$. Для последовательного вычисления $x, x^2 = x \cdot x, \ldots, x^{1023} = x^{1022} \cdot x$ необходимо проделать 1022 умножения, а затем столько же сложений.

Однако если заметить, что $S = \frac{1-x^{1024}}{1-x}$, то количество арифметических действий значительно уменьшается; в частности, для вычисления x^{1024} требуется всего 10 умножений: $x^2 = x \cdot x; x^4 = (x)^2(x)^2, \ldots, x^{1024} = (x)^{512}(x)^{512}$.

Пример 1.13. Вычисления значений многочленов. Если вычислять значение многочлена $P(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n$ «в лоб», т. е. вычислять значения каждого члена и суммировать, то окажется, что необходимо выполнить $\left(n^2 + [n/2]\right)$ умножений и n сложений. Кроме того, такой способ вычислений может привести к накоплению ошибок округления при вычислениях с плавающей точкой.

Его очевидным улучшением является вычисление каждого члена последовательным умножением на x. Такой алгоритм требует $(2n - 1)$ умножение и n сложений.

Еще более экономичным алгоритмом является хорошо известная в алгебре схема Горнера:

$$P(x) = ((\ldots((a_n x + a_{n-1})x + a_{n-2})x + \ldots + a_0),$$

требующая n операций сложения и n операций умножения. Этот метод был известен в средние века в Китае под названием Тянь-Юань и был заново открыт в Европе в начале XIX века англичанином Горнером и итальянцем Руффини.

Пример 1.14. Рассмотрим систему линейных алгебраических уравнений (СЛАУ) вида $\mathbf{Au} = \mathbf{f}, \mathbf{u} = \{u_1, \ldots, u_n\}^{\mathrm{T}}, \mathbf{f} = \{f_1, \ldots, f_n\}^{\mathrm{T}}$, с трехдиагональной матрицей

$$\mathbf{A} = \begin{pmatrix} b_1 & c_1 & 0 & \ldots & 0 & \\ a_2 & b_2 & c_2 & \ldots & 0 & \\ 0 & a_3 & b_3 & c_3 & \ldots & 0 \\ 0 & \ldots & 0 & a_n & b_n & \end{pmatrix}.$$

Если проводить решение такой системы без учета специфической структуры матрицы (например, с помощью метода Гаусса), то количество арифметических действий будет порядка n^3, если же учесть эту структуру, то количество операций можно уменьшить до n.

1.4. Погрешность метода

Оценим погрешность при вычислении первой производной при помощи соотношения: $f'(x) \approx \frac{f(x+h)-f(x)}{h}$:

$$\frac{f(x+h) - f(x)}{h} = \frac{[f(x) + hf'(x) + O(h^2)] - f(x)}{h} = f'(x) + O(h),$$

где $O(h)$ есть погрешность метода. В данном случае под погрешностью метода понимается абсолютная величина разности $\left|f'(x) - \frac{f(x+h)-f(x)}{h}\right|$, которая составляет $O(h)$ (более точно $\frac{h}{2}f''(\xi)$, где $\xi \in [x, x+h]$).

Если же взять другой метод вычисления производной $f'(x) \approx \approx \frac{f(x+h)-f(x-h)}{2h}$, то получим, что его погрешность составляет $O(h^2)$, это оказывается существенным при малых h. Однако уменьшать h до бесконечности не имеет смысла, что видно из следующего примера. Реальная погрешность при вычислении первой производной будет

$$\Delta = \frac{h}{2}\max_{\xi\in[x,x+h]}|f''(\xi)| + \frac{2\delta_M}{h} = O(h) + O(h^{-1}),$$

поскольку абсолютная погрешность вычисления значения функции за счет машинного округления не превосходит $\frac{2\delta_M}{h}$.

В этом случае можно найти оптимальный шаг h. Будем считать полную погрешность в вычислении производной Δ функцией шага h. Отыщем минимум этой функции. Приравняв производную $\Delta'_h(h)$ к нулю, получим оптимальный шаг численного дифференцирования

$$h_{\text{опт}} = 2\sqrt{\frac{\delta_M}{\max\limits_{\xi\in[x,x+h]}|f''(\xi)|}}.$$

Выбирать значение h меньше оптимального не имеет смысла, так как при дальнейшем уменьшении шага суммарная погрешность начинает расти из-за возрастания вклада ошибок округления.

1.5. Элементы теории погрешностей

Определение. Пусть u и u^* — точное и приближенное значение некоторой величины, соответственно. Тогда *абсолютной погрешностью* приближения u^* называется величина $\Delta(u^*)$, удовлетворяющая неравенству

$$|u - u^*| \leqslant \Delta(u^*).$$

Определение. *Относительной погрешностью* называется величина $\delta(u^*)$, удовлетворяющая неравенству

$$\left|\frac{u-u^*}{u^*}\right| \leqslant \delta(u^*).$$

Обычно используется запись $u = u^*(1 \pm \delta(u^*))$.

Определение. Пусть искомая величина u является функцией параметров $t_1, \dots t_n \in \Omega$, u^* — приближенное значение u. Тогда предельной абсолютной погрешностью называется величина

$$D(u^*) = \sup_{(t_1,\dots,t_n)\in\Omega} |u(t_1,\dots,t_n) - u^*|,$$

Предельной относительной погрешностью называется величина $D(u^*)/|u^*|$.

Пусть $|t_j - t_j^*| \leqslant \Delta(t_j^*), j = 1 \div n$ — приближенное значение $u* = u(t_1^*,\dots,t_n^*)$. Предполагаем, что u — непрерывно дифференцируемая функция своих аргументов. Тогда, по формуле Лагранжа,

$$u(t_1,\dots,t_n) - u^* = \sum_{j=1}^{n} \gamma_j(\alpha)\,(t_j - t_j^*),$$

где $\gamma_j(\alpha) = u'_{t_j}(t_1^* + \alpha(t_1 - t_1^*),\dots,t_n^* + \alpha(t_n - t_n^*)), 0 \leqslant \alpha \leqslant 1$.

Отсюда $|u(t_1,\dots,t_n) - u^*| \leqslant D_1(u^*) = \sum\limits_{j=1}^{n} b_j\,\Delta(t_j^*)$, где $b_j = \sup\limits_{\Omega} \left|u'_{t_j}(t_1,\dots,t_n)\right|$.

Можно показать, что при малых $\rho = \sqrt{(\Delta(t_1^*))^2 + \dots + (\Delta(t_n^*))^2}$ эта оценка не может быть существенно улучшена. На практике иногда пользуются грубой (линейной) оценкой

$$|u(t_1,\dots,t_n) - u^*| \leqslant D_2(u^*), \text{где } D_2(u^*) = \sum_{j=1} |\gamma_j(0)|\,\Delta(t^*).$$

Несложно показать, что

a) $\Delta(\pm t_1^* \pm,\dots,\pm t_n^*) = \Delta(t_1^*) + \dots + \Delta(t_n^*)$, предельная погрешность суммы или разности равна сумме предельных погрешностей.

b) Предельная относительная погрешность произведения или частного приближенного равна сумме предельных относительных погрешностей

$$\delta(t_1^* \cdots t_m^* \cdot d_1^{*-1} \cdots d_m^{*-1}) = \delta(t_l^*) + \dots + \delta(t_m^*) + \delta(d_1^*) + \dots + \delta(d_n^*).$$

1.6. Задача численного дифференцирования

В пункте 1.2 уже была введена простейшая формула численного дифференцирования. Рассмотрим задачу приближенного вычисления значения производной подробнее.

Пусть задана таблица значений x_i. В дальнейшем совокупность точек на отрезке, на котором проводятся вычисления, иногда будут называться *сеткой*, каждое значение x_i — *узлом сетки*. Пусть сетка — равномерная, и расстояние между узлами равно h — *шагу сетки*. Пусть узлы сетки пронумерованы в порядке возрастания, т. е.

$$x_0 = a,$$

$$x_j = a + jh, , j = 0, 1, \ldots, N$$

Пусть $f(x_j) = f_j$ — функция, определенная в узлах сетки. Такие функции будут называться табличными, или сеточными функциями. Считаем, кроме того, что рассматриваемая сеточная функция есть проекция (или ограничение) на сетку некоторой гладкой нужное число раз непрерывно дифференцируемой функции $f(x)$. По определению производной

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x},$$

тогда, если шаг сетки достаточно мал, по аналогии можно написать формулу в конечных разностях, дающую приближенное значение производной сеточной функции:

$$f'(x_j) \approx \frac{f(x_j + h) - f(x_j)}{h}. \tag{1.1}$$

Если параметр h достаточно мал, то можно считать полученное значение производной достаточно точным. Погрешность формулы (1.1) оценена в пункте 1.2. Как показано выше, при уменьшении шага сетки h ошибка будет уменьшаться, но при некотором значении h ошибка может возрасти до бесконечности. При оценке погрешности метода обычно считается, что все вычисления были точными. Но существует ошибка округления. При оценке ее большую роль играет машинный ε — мера относительной погрешности машинного округления, возникающей из-за конечной разрядности мантиссы при работе с числами в формате с плавающей точкой. Напомним, что по определению машинным ε называют наибольшее из чисел, для которых в рамках используемой системы вычислений выполнено $1 + \varepsilon = 1$. Тогда абсолютная погрешность при вычислении значения функции (или представлении табличной функции) есть $f(x_j) \cdot \varepsilon$. Максимальный вклад погрешностей округления при вычислении производной по формуле (1.1) будет $\frac{2f(x_j) \cdot \varepsilon}{h}$, тогда, когда члены в знаменателе (1.1) имеют ошибки разных знаков.

Пусть $k = \max |f(x)|$, максимум ищется на отрезке, на котором вычисляются значения производных. Тогда суммарная ошибка, состоящая

из погрешности метода и погрешности округления, есть $\Delta = \frac{2k\varepsilon}{h} + \frac{M_2 h}{2}$, $M_2 = \max|f''(x)|$.

Для вычисления оптимального шага численного дифференцирования найдем минимум суммарной ошибки, как функции шага сетки $\frac{M_2}{2} - \frac{2k\varepsilon}{h^2} = 0$, откуда

$$h_{\text{opt}} = 2\sqrt{\frac{k\varepsilon}{M_2}}.$$

Если требуется повысить точность вычисления производных, необходимо воспользоваться формулами, имеющими меньшие погрешности метода. Так, из курсов математического анализа известно, что

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x-\Delta x)}{2\Delta x}.$$

По аналогии напишем конечно-разностную формулу

$$f'(x_j) \approx \frac{f(x_j + h) - f(x_j - h)}{2h}. \tag{1.2}$$

(1.2) — формула с центральной разностью. Исследуем ее на аппроксимацию, т. е. оценим погрешность метода. Предположим, что функция, которую спроектировали на сетку, трижды непрерывно дифференцируема, тогда

$$f(x_j + h) = f(x_j) + f'(x_j)\,h + f''(x_j)\frac{h^2}{2} + f'''(x_j + \theta_1 h)\frac{h^3}{6},$$

$$f(x_j - h) = f(x_j) - f'(x_j)\,h + f''(x_j)\frac{h^2}{2} - f'''(x_j - \theta_2 h)\frac{h^3}{6}.$$

Погрешность метода определяется 3-й производной функции. Введем $M_3 = \max\limits_{x\in[a,b]} |f'''(x)|$, тогда суммарная погрешность при вычислении по формуле с центральной разностью есть

$$\Delta = \frac{M_3 h^2}{3} + \frac{k\varepsilon}{h},$$

для вычисления оптимального шага, находя минимум погрешности, как функции шага сетки, имеем $\frac{2M_3 h}{6} - \frac{k\varepsilon}{h^2} = 0$, откуда $h_{\text{opt}} = \sqrt[3]{\frac{3k\varepsilon}{M_3}}$.

Для более точного вычисления производной необходимо использовать разложение более высокого порядка, шаг h_{opt} будет увеличиваться.

Формула (1.1) — двухточечная, (1.2) — трехточечная: при вычислении производной используются точки (узлы) x_j (узел входит с нулевым коэффициентом), $x_j + h, x_j - h$ — совокупность узлов, участвующих в

каждом вычислении производной, в дальнейшем будем иногда называть *сеточным шаблоном.*

Введем на рассматриваемом отрезке шаблон из нескольких точек.

Считаем, что сетка равномерная — шаг сетки постоянный, расстояния между любыми двумя соседними узлами равны. Используем для вычисления значения первой производной следующую приближенную (конечно-разностную) формулу:

$$f'(x_j) \approx \frac{1}{h} \sum_{k=-l}^{m} \alpha_k f(x_j + kh), \tag{1.3}$$

шаблон включает l точек слева от рассматриваемой точки x_j и m справа. Коэффициенты α_k — *неопределенные коэффициенты.* Формула дифференцирования может быть и *односторонней* — либо l, либо m могут равняться нулю. В первом случае иногда называют (на наш взгляд, не слишком удачно) такую приближенную формулу формулой дифференцирования вперед, во втором — формулой дифференцирования назад. Потребуем, чтобы (1.3) приближала первую производную с точностью $O(h^{l+m})$. Используем разложения в ряд Тейлора в окрестности точки x_j. Подставляя их в (1.3), получим

$$\frac{1}{h} \sum_{k=-l}^{m} \alpha_k f(x_j + kh) = \frac{1}{h} f(x_j) \sum \alpha_k + f'(x_j) \sum k\alpha_k + f''(x_j) \sum \frac{k^2}{2} \alpha_k h +$$

$$+ f'''(x_j) \sum \frac{k^3}{6} \alpha_k h^2 + \ldots + f^{(n)}(x_j) \sum \alpha_k \frac{k^n}{n!} h^{n-1} + \ldots$$

Потребуем выполнение условий:

$$\sum \alpha_k = 0, \sum k\alpha_k = 1, \sum \alpha_k \frac{k^2}{2} = 0, \ldots \sum \alpha_k \frac{k^n}{n!} = 0, \ldots \tag{1.4}$$

Получаем систему линейных алгебраических уравнений для неопределенных коэффициентов α (1.4). Матрица этой системы есть

$$\begin{pmatrix} 1 & 1 & \ldots & 1 \\ -l & -l+1 & \ldots & m \\ l^2 & (l-1)^2 & \ldots & m^2 \\ (-l)^3 & (-l+1)^3 & \vdots & m^3 \\ \ldots & \ldots & \ldots & \ldots \end{pmatrix}.$$

Вектор правых частей $(0, 1, 0, \ldots, 0)^T$.

Определитель данной матрицы — детерминант Вандермонда. Из курса линейной алгебры следует, что он не равен нулю. Тогда существует единственный набор коэффициентов α, который позволяет найти на шаблоне из $(1 + l + m)$ точек значение первой производной с точностью $O(h^{l+m})$.

Для нахождения второй производной можно использовать ту же самую формулу (1.3) с небольшой модификацией

$$f'(x_j) \approx \frac{1}{h^2} \sum_{k=-l}^{m} \alpha_k f(x_j + kh),$$

только теперь $\sum k\alpha_k = 0, \sum \alpha_k k^2 = \frac{1}{2}$.

Очевидно, что и данная система уравнений для нахождения неопределенных коэффициентов имеет единственное решение. Для получения с той же точностью приближенных значений производных до порядка $l + m$ включительно с точностью $O(h^{l+m})$ модификации формулы (1.3) и условий (1.4) очевидны, набор неопределенных коэффициентов находится единственным образом.

Таким образом, доказано следующее утверждение. На сеточном шаблоне, включающем в себя $N + 1$ точку, с помощью метода неопределенных коэффициентов всегда можно построить единственную формулу для вычисления производной от первого до N порядка включительно с точностью $O(h^N)$.

Утверждение доказано для равномерной сетки, но на случай произвольных расстояний между сеточными узлами обобщение проводится легко.

Так как на практике вычисления проводятся с конечной длиной мантиссы, то получить нулевую ошибку невозможно.

Читателям предлагается оценить значение оптимального шага при вычислениях по формулам типа (1.3) самостоятельно.

1.7. Задачи

1. Найти абсолютную предельную погрешность, погрешность по производной, линейную погрешность для функции $u = t^{10}$, если заданы точка приближения $t^* = 1$, значение функции u^* в этой точке и погрешность $\Delta t^* = 10^{-1}$.

Решение: обозначим $b = \sup\limits_{|t-1|\leqslant 0,1} |u'_t(t)| = \sup\limits_{|t-1|\leqslant 0.1} 10{\cdot}t^9 = 10{\cdot}(1,1)^9 \approx$ $\approx 23, \ldots$

Абсолютная предельная погрешность может быть определена как

$$D(u^*) = \sup_{|t-1|\leqslant 0,1} \left|t^{10} - 1\right| = (1.1)^{10} - 1 \approx 1{,}5\ldots$$

Оценка погрешности u при вычислении значения функции по максимуму производной и линейная оценка соответственно будут $D_1(u^*) = b\,\Delta(t^*) = 2,3\ldots$; $D_2(u^*) = |\gamma(0)|\,\Delta(t^*) = 1$.

2. Дать линейную оценку погрешности при вычислении неявной функции $\varphi(u, t_1, t_2, \ldots, t_n) = 0$, если известны точка приближения $\{t_1^*, \ldots, t_n^*\}$, значение функции в точке приближения u^* и погрешность в определении аргументов $\Delta t_1^*, \ldots, \Delta t_n^*$.

Решение. Дифференцируя по t_j, получим

$$\frac{\partial \varphi}{\partial u}\frac{\partial u}{\partial t_j} + \frac{\partial \varphi}{\partial t_j} = 0,$$

откуда

$$\frac{\partial u}{\partial t_j} = -\left(\frac{\partial \varphi}{\partial t_j}\right)\left(\frac{\partial \varphi}{\partial u}\right)^{-1}.$$

При заданных $\{t_1^*, \ldots, t_n^*\}$, можно найти u^* как корень уравнения $\varphi(u, t_1, t_2, \ldots, t_n) = 0$, а затем — значения

$$b_j(0) = -\left(\frac{\partial \varphi}{\partial t_j}\right)\left(\frac{\partial \varphi}{\partial u}\right)^{-1}\Bigg|_{(u^*, t_1^*, \ldots, t_n^*)},$$

откуда можно получить линейную оценку погрешности функции $D_2(u^*)$.

3. Вычислить относительную погрешность в определении значения функции

$u = xy^2z^3$, если $x^* = 37,1, y^* = 9,87, z^* = 6,052, \Delta x^* = 0,3, \Delta y^* = 0,11, \Delta z^* = 0,016$.

Решение:

$$\delta_x = \frac{0.3}{37.1} \approx 0.81 \cdot 10^{-2}, \delta_y = \frac{0,11}{9,87} \approx 1,12 \cdot 10^{-2}, \delta_z = \frac{0,016}{6,052} \approx 0,26 \cdot 10^{-2},$$

$$\delta(u) = \delta(x^*) + 2\delta(y^*) + 3\delta(z^*) = 3.8 \cdot 10^{-2}.$$

4. Оценить погрешность в определении корней квадратного уравнения $\varphi(u, t_1, t_2) = u^2 + t_1 u + t_2 = 0$, если заданы приближения $t_1^*, t_2^*, \Delta(t_1^*), \Delta(t_2^*)$.

Пусть u^* — решение уравнения

$$u^{*2} + t_1^* u^* + t_2^* = 0.$$

Из формулы

$$b_j(0) = -\left(\frac{d\varphi}{dt_j}\right)\left(\frac{d\varphi}{du}\right)^{-1}\Bigg|_{(u^*, t_1^*, \ldots, t_n^*)}$$

получим

$$b_1(0) = \left.\frac{du}{dt_1}\right|_{(t_1^*,t_2^*)} = -\frac{u^*}{2u^* + t_1^*},$$

$$b_2(0) = \left.\frac{du}{dt_2}\right|_{(t_1^*,t_2^*)} = -\frac{1}{2u^* + t_1^*}.$$

Следовательно, линейная оценка будет

$$D_2(u^*) = \frac{|u^*| \cdot \Delta(t_1^*) + \Delta(t_2^*)}{|2u^* + t_1^*|}.$$

1.8. Задачи для самостоятельного решения

1. Найти абсолютную предельную погрешность, погрешность по производной и линейную оценку погрешности для функций $u = \sin t$, $u = \frac{1}{t^2-5t+6}$.

   Заданы точка приближения $t = t^*$ и погрешность Δt.

2. Определить шаг τ, при котором погрешность вычисления производной $u'(t)$, приближенно вычисляемой в соответствии с формулами

$$u'(t) \approx \frac{f(t+\tau) - f(t)}{\tau},$$

$$u'(t) \approx \frac{f(t+\tau) - f(t-\tau)}{2\tau},$$

   не превосходит 10^3. Известно, что $|u''(t)| \leqslant 1$, $|u'''(t)| \leqslant 1$ для любых t.

3. Пусть для вычисления функции $u = f(t)$ используется частичная сумма ряда Маклорена

$$u(t) \approx u(0) + \frac{u'(0)}{1!}t + \ldots + \frac{u^{(n)}(0)}{n!}t^n,$$

   причем аргумент задан с погрешностью $\Delta t = 10^{-3}$.

   Найти n такое, чтобы погрешность в определении функции $u(y)$ по данной формуле не превышала Δt. Рассмотреть отрезки $t \in [0,\ 1]$, $t \in [10,\ 11]$.

   Предложить более совершенный алгоритм для вычисления функций $u(t) = \sin t$, $u(t) = e^t$ на отрезке $t \in [10,\ 11]$.

4. Определить оптимальный шаг численного дифференцирования τ при использовании для вычисления производной приближенной формулы

$$u'(t) \approx \frac{u(t-2\tau) - 8(t-\tau) + 8(t+\tau) - u(t+2\tau)}{12t},$$

имеющей четвертый порядок точности, если известно, что $\left|u^{(5)}(t)\right| \leqslant M_5$, а значения функций вычисляются с точностью ε.

5. Вычислить относительную погрешность в определении значения функции $u(x, y, z) = x^2y^2/z^4$, если заданы

$$x^* = 37,1,\ y^* = 9,87,\ z^* = 6,052,\ \Delta(x^*) = 0,1;$$

$$\Delta(y^*) = 0,05;\ \Delta(z^*) = 0,02.$$

Литература

[1] *В. С. Рябенький.* Введение в вычислительную математику. — М.: Физматлит, 2000. 294 с.

[2] *Н. В. Бахвалов, Н. П. Жидков, Г. М. Кобельков.* Численные методы. — М: Лаборатория Базовых Знаний, 2002. 632 с.

[3] *В. И. Косарев.* 12 лекций по вычислительной математике. — М.: Изд-во МФТИ, Физматкнига, 2000. 220 с.

[4] *А. А. Самарский.* Введение в численные методы. М.: Наука, 1997. 234 с.

[5] *А. А. Амосов, Ю. А. Дубинский, Н. В. Копченова.* Вычислительные методы для инженеров. — М.: Высшая школа, 1994. 544 с.

[6] *Д. Каханер, К. Моулер, С. Нэш.* Численные методы и программное обеспечение. — М.: Мир, 1998. 575 с.

Лекция 2. Численное решение систем линейных алгебраических уравнений

Рассматриваются наиболее употребительные приближенные методы решения систем линейных алгебраических уравнений. Вводятся согласованные нормы векторов и матриц. Вычисляется число обусловленности в различных нормах. Анализируется влияние ошибок округления на погрешность результата. Дается понятие о спектральных задачах. Для самосопряженной матрицы рассматривается метод вращений поиска собственных значений.

Ключевые слова: системы линейных алгебраических уравнений, согласованные нормы вектора и матрицы, обусловленность системы, число обусловленности матрицы, прямые методы, метод Гаусса, метод Гаусса с выбором главного элемента, LU-разложение, итерационные методы, метод простых итераций, методы Якоби, Зейделя, верхней релаксации, метод наискорейшего спуска, метод сопряженных градиентов, спектральные задачи, метод вращений.

К численному решению систем линейных алгебраических уравнений (СЛАУ) сводятся многие задачи математической физики. Математические модели, представляющие собой СЛАУ большой размерности, встречаются в математической экономике, биологии и т.п. Теория получения приближенных решений СЛАУ — часть вычислительной линейной алгебры. Сама вычислительная линейная алгебра, по-видимому, является наиболее обширной темой во всем курсе вычислительной математики. По прикладной линейной алгебре существует обширная литература (например, [1, 2, 3, 4, 5]), а программы, реализующие наиболее популярные алгоритмы вычислительной линейной алгебры, являются неотъемлемой частью прикладного программного обеспечения, в частности, современных математических пакетов.

2.1. Постановка задачи

Рассмотрим СЛАУ вида

$$\mathbf{Au} = \mathbf{f}, \tag{2.1}$$

где $\mathbf{A}$ — невырожденная ($\det \mathbf{A} \neq 0$) квадратная матрица размером $n \times n$

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix},$$

$\mathbf{u} = \{u_1, \dots, u_n\}^{\mathrm{T}}$ — вектор-столбец решения, $\mathbf{f} = \{f_1, \dots, f_n\}^{\mathrm{T}}$ — вектор-столбец правой части.

Так как матрица системы — невырожденная, $\Delta = \det \mathbf{A} \neq 0$, то решение системы (2.1) существует и единственно.

Из курса линейной алгебры [6] известно правило Крамера нахождения решения. Так, каждый компонент вектора неизвестных может быть вычислен как

$$u_i = \frac{\Delta_i}{\Delta},$$

где Δ_i — определитель матрицы, получаемой из $\mathbf{A}$ заменой i столбца столбцом правых частей. Однако несложные арифметические оценки позволяют понять, что использование этой формулы приводит к неоправданно большим затратам машинного времени [3]. Так, например, если одно слагаемое в Δ вычисляется за 10^{-6} с, то время расчета для $n = 100$ на существующих в момент написания книги компьютерах будет измеряться годами.

На самом деле в настоящее время с помощью компьютеров численно решаются СЛАУ намного более высокого порядка (примерно до $n \approx 10^6$). Такие решения осуществляются при помощи *прямых* или *итерационных* численных методов. *Прямые методы* позволяют в предположении отсутствия ошибок округления (при проведении расчетов на идеальном, т. е. бесконечноразрядном компьютере) получить точное решение задачи за конечное число арифметических действий. *Итерационные* методы, или методы последовательных приближений, позволяют вычислить последовательность $\{\mathbf{u}_k\}$, сходящуюся к решению задач при $k \to \infty$ (на практике, разумеется, ограничиваются конечным k, в зависимости от требуемой точности).

Однако неточность в задании правых частей и элементов матрицы $\mathbf{A}$ может приводить к значительным погрешностям при вычислении решения (2.1). В первой лекции на примере было показано, что такое явление наблюдается в случае плохо обусловленной системы. Остановимся подробнее на важном вопросе оценки погрешности решения СЛАУ.

Для этого напомним некоторые сведения из функционального анализа, которые понадобятся в дальнейшем.

2.2. Согласованные нормы векторов и матриц

В векторном n-мерном линейном нормированном пространстве введем следующие нормы вектора:

кубическая:

$$\|\mathbf{u}\|_1 = \max_{1 \leqslant i \leqslant n} |u_i|\,, \tag{2.2а}$$

октаэдрическая:

$$\|\mathbf{u}\|_2 = \sum_{i=1}^{n} |u_i|, \tag{2.2б}$$

евклидова (в комплексном случае — эрмитова):

$$\|\mathbf{u}\|_3 = \left(\sum_{i=1}^{n} |u_i|^2\right)^{1/2} = (\mathbf{u}, \mathbf{u})^{1/2}. \tag{2.2в}$$

Рассмотрим квадратную матрицу $\mathbf{A}$ и связанное с ней линейное преобразование $\mathbf{v} = \mathbf{A}\mathbf{u}$, где $\mathbf{v}$, $\mathbf{u} \in L^n$ (L^n — n-мерное линейное нормированное пространство). Норма матрицы определяется как действительное неотрицательное число, характеризующее это преобразование и определяющееся как

$$\|\mathbf{A}\| = \sup_{\|\mathbf{u}\| \neq 0} \frac{\|\mathbf{A}\mathbf{u}\|}{\|\mathbf{u}\|} \tag{2.3}$$

Укажем некоторые свойства нормы матрицы:

$$\|\mathbf{A} + \mathbf{B}\| = \|\mathbf{A}\| + \|\mathbf{B}\|\,,$$
$$\|\lambda\mathbf{A}\| = |\lambda|\|\mathbf{A}\|,$$
$$\|\mathbf{A}\mathbf{B}\| \leqslant \|\mathbf{A}\|\,\|\mathbf{B}\|\,,$$
$$\|\mathbf{A}\| = 0 \text{ тогда и только тогда, когда } \mathbf{A} = 0.$$

Заметим, что норму матрицы (2.3) называют подчиненной норме вектора. Говорят, что норма матрицы $\mathbf{A}$ согласована с нормой вектора $\mathbf{u}$, если выполнено условие

$$\|\mathbf{A}\mathbf{u}\| \leqslant \|\mathbf{A}\|\,\|\mathbf{u}\|\,.$$

Нетрудно видеть, что подчиненная норма согласована с соответствующей метрикой векторного пространства. В самом деле

$$\|\mathbf{A}\| = \sup_{\|\mathbf{u}\| \neq 0} \frac{\|\mathbf{A}\mathbf{u}\|}{\|\mathbf{u}\|} \geqslant \frac{\|\mathbf{A}\mathbf{u}\|}{\|\mathbf{u}\|} \text{ откуда } \|\mathbf{A}\mathbf{u}\| \leqslant \|\mathbf{A}\| \cdot \|\mathbf{u}\|\,.$$

Согласованные с введенными выше нормами векторов нормы матриц будут определяться следующим образом:

$$\|\mathbf{A}\|_1 = \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |a_{ij}|,$$

$$\|\mathbf{A}\|_2 = \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{ij}|,$$

$$\|\mathbf{A}\|_3 = \sqrt{\max_{1\leqslant i\leqslant n} \lambda^i(\mathbf{A}^* \cdot \mathbf{A})}.$$

Покажем, как получается выражение для согласованной нормы матрицы $\|\mathbf{A}\|_1$, соответствующей норме вектора $\|\mathbf{u}\|_1$.

Вычислим норму вектора $\|\mathbf{Au}\|_1$:

$$\|\mathbf{Au}\|_1 = \max_i \left| \sum_j a_{ij} u_j \right| \leqslant \max_i \left(\sum_j |a_{ij}|\,|u_j| \right) \leqslant$$

$$\leqslant \left(\max_i \sum_j |a_{ij}|\right) \max_j |u_j| = \left(\max_i \sum_j |a_{ij}|\right) \|\mathbf{u}\|_1,$$

откуда $\frac{\|\mathbf{Au}\|_1}{\|\mathbf{u}\|_1} \leqslant \max_i \sum_j |a_{ij}|$. По определению нормы матрицы как точной верхней грани отношения $\frac{\|\mathbf{Au}\|}{\|\mathbf{u}\|}$, $\max_j \sum_i |a_{ij}| = \|\mathbf{A}\|_1$, если существует вектор, на котором точная верхняя грань достигается.

Покажем, что таким вектором является, например, $\mathbf{v}_k = \{\mathrm{sign} a_{k1}, \ldots, \mathrm{sign} a_{kn}\}^{\mathrm{T}}$, при этом допустим, что максимум в последнем неравенстве достигается при $i = k$.

Поскольку $\|\mathbf{v}_k\| = k$, то $\sum_j a_{kj} v_j = \sum_j |a_{kj}| = \max_i \sum_j |a_{ij}|$.

Тогда, в соответствии с выражением для первой нормы вектора, получаем

$$\|\mathbf{Av}\|_1 = \max_i \sum_j |a_{ij}|.$$

Таким образом, точная верхняя грань в рассмотренном неравенстве достижима и действительно $\|\mathbf{A}\|_1 = \max \sum_j |a_{ij}|$.

Для третьей нормы (2.2в) $\|\mathbf{A}\|_3 = \sup_{\mathbf{u}} \frac{\|\mathbf{Au}\|_3}{\|\mathbf{u}\|_3} = \sup_{\mathbf{u}} \sqrt{\frac{(\mathbf{Au},\mathbf{Au})}{(\mathbf{u},\mathbf{u})}} =$

$= \sup_{\mathbf{u}} \sqrt{\frac{(\mathbf{A}^*\mathbf{Au},\mathbf{u})}{(\mathbf{u},\mathbf{u})}}$.

Заметим, что матрица $\mathbf{B} = \mathbf{A}^*\mathbf{A}$ — симметричная. Без ограничения общности предположим, что все собственные числа матрицы различны. Матрица обладает всеми действительными собственными значениями, и каждому собственному числу соответствует собственный вектор. Все собственные векторы взаимно ортогональны. Можно рассмотреть ортонормированную систему собственных векторов $\omega_1, \ldots, \omega_n$; $\lambda_1, \ldots, \lambda_n$ — соответствующие им собственные значения. Любой вектор $\mathbf{u}$ можно представить в виде своего разложения по базису из собственных векторов: $\sum\limits_i \xi_i \omega_i$. Кроме того, $(\mathbf{A}^*\mathbf{A})\,\omega_i = \lambda_i \omega_i$. Поэтому

$$\sup_{\|\mathbf{u}\|\neq 0} \sqrt{\frac{(\mathbf{A}^*\mathbf{A}\mathbf{u}, \mathbf{A}\mathbf{u})}{(\mathbf{u}, \mathbf{u})}} = \sup_{\|\mathbf{u}\|\neq 0} \sqrt{\sum_i \frac{(\lambda_i \xi_i \omega_i, \xi_i \omega_i)}{(\xi_i \omega_i, \xi_i \omega_i)}} =$$

$$= \sup_{\|\mathbf{u}\|\neq 0} \sqrt{\frac{\sum \lambda_i \xi_i^2}{\sum \xi_i^2}} = \sqrt{\max \lambda_i(\mathbf{A}^*\mathbf{A})},$$

причем точная верхняя грань достигается при $\mathbf{u} = \omega_i$. Действительно,

$$\sup_u \sqrt{\frac{(\mathbf{A}^*\mathbf{A}\omega_i, \omega_i)}{(\omega_i, \omega_i)}} = \sup_u \sqrt{\lambda^i(\mathbf{A}^*\mathbf{A})} = \sqrt{\max_i \lambda_i(\mathbf{A}^*\mathbf{A})},$$

т. к. $\mathbf{A}^*\mathbf{A}\omega_i = \lambda_i \omega_i$, откуда $(\mathbf{A}^*\mathbf{A}\omega_i, \omega_i) = \lambda_i(\omega_i, \omega_i)$,

$$\frac{(\mathbf{A}^*\mathbf{A}\omega_i, \omega_i)}{(\omega_i, \omega_i)} = \lambda_i.$$

В важном частном случае симметричной (самосопряженной) матрицы $\mathbf{A}$ имеем $\lambda^i_{\mathbf{A}^*\mathbf{A}} = \lambda^i_{\mathbf{A}^2} = \left|\lambda^i_{\mathbf{A}}\right|^2$, поэтому $\|\mathbf{A}\|_3 = \max\limits_i \left|\lambda^i_{\mathbf{A}}\right|$.

2.3. Обусловленность СЛАУ. Число обусловленности матрицы

Понятия согласованных норм матриц и векторов позволяют оценить погрешности, возникающие при численном решении СЛАУ. Пусть и матрица, и правая часть системы заданы с некоторой погрешностью, тогда наряду с системой

$$\mathbf{A}\mathbf{u} = \mathbf{f} \tag{2.4}$$

рассматривается система

$$(\mathbf{A} + \Delta\mathbf{A})(\mathbf{u} + \Delta\mathbf{u}) = \mathbf{f} + \Delta\mathbf{f}. \tag{2.5}$$

Теорема. *Пусть правая часть и невырожденная матрица СЛАУ* (2.4) *вида* $\mathbf{Au} = \mathbf{f}$, $\mathbf{u} \in L^n, \mathbf{f} \in L^n$, *получили приращения* $\Delta\mathbf{f}$ *и* $\Delta\mathbf{A}$ *соответственно. Пусть существует обратная матрица* $\mathbf{A}^{-1}$ *и выполнены условия* $\|\mathbf{A}\| \neq 0$, $\mu\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|} < 1$, *где* $\mu = \|\mathbf{A}\| \cdot \left\|\mathbf{A}^{-1}\right\|$. *В этом случае оценка относительной погрешности решения* $\|\Delta\mathbf{u}\| / \|\mathbf{u}\|$ *удовлетворяет неравенству*

$$\frac{\|\Delta\mathbf{u}\|}{\|\mathbf{u}\|} \leqslant \frac{\mu}{1 - \mu\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}} \left(\frac{\|\Delta\mathbf{f}\|}{\|\mathbf{f}\|} + \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}\right).$$

Доказательство.

Из (2.5) следует, что $\Delta\mathbf{u} = \mathbf{A}^{-1}(\Delta\mathbf{f} - \Delta\mathbf{A}\,\mathbf{u} - \Delta\mathbf{A}\,\Delta\mathbf{u})$. Переходя в этом равенстве к норме и использовав неравенство треугольника, получаем

$$\|\Delta\mathbf{u}\| \leqslant \left\|\mathbf{A}^{-1}\right\| \|\Delta\mathbf{f}\| + \left\|\mathbf{A}^{-1}\right\| \|\Delta\mathbf{A}\| \|\mathbf{u}\| + \left\|\mathbf{A}^{-1}\right\| \|\Delta\mathbf{A}\| \|\Delta\mathbf{u}\|, \text{ или}$$

$$\|\Delta\mathbf{u}\| \leqslant \left\|\mathbf{A}^{-1}\right\| \frac{\|\Delta\mathbf{f}\|}{\|\mathbf{f}\|} \|\mathbf{f}\| + \left\|\mathbf{A}^{-1}\right\| \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|} \|\mathbf{A}\| \|\mathbf{u}\| + \left\|\mathbf{A}^{-1}\right\| \|\mathbf{A}\| \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|} \|\Delta\mathbf{u}\|.$$

Вводя обозначение $\mu(\mathbf{A}) = \left\|\mathbf{A}^{-1}\right\| \cdot \|\mathbf{A}\|$, перепишем последнее равенство в виде

$$\|\Delta\mathbf{u}\| \left(1 - \mu\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}\right) \leqslant \mu\frac{\|\Delta\mathbf{f}\|}{\|\mathbf{f}\|} \frac{\|\mathbf{f}\|}{\|\mathbf{A}\|} + \mu \cdot \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|} \|\mathbf{u}\| \leqslant$$

$$\leqslant \mu\frac{\|\Delta\mathbf{f}\|}{\|\mathbf{f}\|} \|\mathbf{u}\| + \mu\frac{\|\Delta\mathbf{A}\|}{\|A\|} \|\mathbf{u}\| = \mu\left(\frac{\|\Delta\mathbf{f}\|}{\|\mathbf{f}\|} + \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}\right) \|\mathbf{u}\|.$$

Заметим, что $\frac{\|\mathbf{f}\|}{\|\mathbf{A}\|} \leqslant \|\mathbf{u}\|$ т.к. $\|\mathbf{f}\| = \|\mathbf{Au}\| \leqslant \|\mathbf{A}\| \cdot \|\mathbf{u}\|$.

Тогда для оценки относительной погрешности решения окончательно получим

$$\frac{\|\Delta\mathbf{u}\|}{\|\mathbf{u}\|} \leqslant \frac{\mu}{1 - \mu\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}} \left(\frac{\|\Delta\mathbf{f}\|}{\|\mathbf{f}\|} + \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}\right). \tag{2.6}$$

При $\Delta A \approx 0$ получаем оценку при наличии погрешности только правых частей

$$\frac{\|\Delta\mathbf{u}\|}{\|\mathbf{u}\|} \leqslant \mu\frac{\|\Delta\mathbf{f}\|}{\|\mathbf{f}\|}, \tag{2.7}$$

если в (2.5) положить $\Delta\mathbf{A} \cdot \Delta\mathbf{u} \approx 0$, то

$$\frac{\|\Delta\mathbf{u}\|}{\|\mathbf{u}\|} \leqslant \mu\left(\frac{\|\Delta\mathbf{f}\|}{\|\mathbf{f}\|} + \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}\right). \tag{2.8}$$

В результате получено важное соотношение, показывающее, на сколько возрастают относительные ошибки решения СЛАУ в случае наличия относительных ошибок при задании правых частей и элементов матриц.

Величина

$$\mu(\mathbf{A}) = \left\|\mathbf{A}^{-1}\right\| \|\mathbf{A}\| \tag{2.9}$$

называется *числом обусловленности* матрицы $\mathbf{A}$. Число обусловленности определяет, насколько погрешность входных данных может повлиять на решение системы (2.1). Почти очевидно, что всегда $\mu \geqslant 1$. Действительно

$$1 = \|\mathbf{E}\| = \left\|\mathbf{A}^{-1}\mathbf{A}\right\| \leqslant \left\|\mathbf{A}^{-1}\right\| \|\mathbf{A}\| = \mu.$$

При $\mu \approx 1 \div 10$ ошибки входных данных слабо сказываются на решении и система (2.1) считается хорошо обусловленной. При $\mu >> 10^2 \div 10^3$ система является плохо обусловленной.

Пример. Решением системы

$$\begin{cases} 100u + 99v = 199 \\ 99u + 98v = 197 \end{cases}$$

будет пара чисел $u = v = 1$.

Внесем возмущение в правые части системы:

$$\begin{cases} 100u + 99v = 198,99 \\ 99u + 98v = 197,01. \end{cases}$$

При этом решение заметно изменится: $u = 2,97; v = -0,99$. Воспользовавшись выбранными согласованными нормами, получим

$$\|\mathbf{f}\|_1 = 199, \quad \|\Delta\mathbf{f}\|_1 = 10^{-2},$$

$$\delta f = \frac{\|\Delta\mathbf{f}\|_1}{\|\mathbf{f}\|_1} \approx 0,5 \cdot 10^{-4} \text{ (это очень малая величина)},$$

$$\|\mathbf{A}\|_1 = \left\|\mathbf{A}^{-1}\right\|_1 = 199, \mu = 199 \cdot 199 \approx 4 \cdot 10^4.$$

Значит, $\delta u = \frac{\|\Delta\mathbf{u}\|}{\|\mathbf{u}\|} \leqslant \mu \frac{\|\Delta\mathbf{f}\|}{\|\mathbf{f}\|} \approx 4 \cdot 10^4 \cdot \frac{10^{-4}}{2} = 2$, что согласуется с результатами решения возмущенной и невозмущенной задач. Для невозмущенной задачи $\|\Delta\mathbf{u}\| \approx 2$, $\|\mathbf{u}\| = 1$.

Рассмотрим еще одно важное свойство. Число обусловленности матрицы, как было показано ранее, можно определить, как $\delta u / \delta f \leqslant \mu(\mathbf{A})$, если $\Delta\mathbf{A} \approx \mathbf{0}$ при $\delta u = \frac{\|\Delta\mathbf{u}\|}{\|\mathbf{u}\|}$. Можно ли найти более тонкую оценку отношения $\delta u / \delta f$, учитывающую зависимость обусловленности СЛАУ

от выбора правых частей? В этом случае параметр обусловленности системы, вообще говоря, зависит и от $\mathbf{f}$, и от $\Delta\mathbf{f}$, и удовлетворяет неравенству $\nu(\mathbf{f}, \Delta\mathbf{f}) \geqslant \frac{\delta u}{\delta f} = \frac{\|\Delta\mathbf{u}\|}{\|\mathbf{u}\|} \cdot \frac{\|\mathbf{f}\|}{\|\Delta\mathbf{f}\|}$. Его можно определить как точную верхнюю грань отношения $\frac{\delta u}{\delta f}$ по $\Delta\mathbf{f}$, что соответствует наихудшей ситуации. Тогда

$$\nu(\mathbf{f}) = \sup_{\Delta\mathbf{f}} \frac{\|\Delta\mathbf{u}\|}{\|\mathbf{u}\|} \frac{\|\mathbf{f}\|}{\|\Delta\mathbf{f}\|} = \sup_{\Delta\mathbf{f}} \frac{\|\mathbf{f}\|}{\|\mathbf{u}\|} \frac{\|\Delta\mathbf{u}\|}{\|\Delta\mathbf{f}\|} =$$

$$= \frac{\|\mathbf{f}\|}{\|\mathbf{u}\|} \sup_{\Delta\mathbf{f}} \frac{\|\mathbf{A}^{-1}\Delta\mathbf{f}\|}{\|\Delta\mathbf{f}\|} = \frac{\|\mathbf{f}\|}{\|\mathbf{u}\|} \|\mathbf{A}^{-1}\|.$$

Далее,

$$\sup_{\mathbf{f}} \|\mathbf{A}^{-1}\| \cdot \frac{\|\mathbf{f}\|}{\|\mathbf{u}\|} = \|\mathbf{A}^{-1}\| \cdot \sup_{\mathbf{f}} \frac{\|A\mathbf{u}\|}{\|\mathbf{u}\|} = \|\mathbf{A}^{-1}\| \cdot \|\mathbf{A}\| = \mu(\mathbf{A}),$$

с другой стороны

$$\inf_{\mathbf{f}} \nu(\mathbf{f}) = \inf_{\mathbf{f}} \|\mathbf{A}^{-1}\| \cdot \frac{\|\mathbf{f}\|}{\|\mathbf{u}\|} = \inf_{\mathbf{f}} \|\mathbf{A}^{-1}\| \sup_{\mathbf{f}} \left(\frac{\|\mathbf{u}\|}{\|\mathbf{f}\|}\right)^{-1} =$$

$$= \|\mathbf{A}^{-1}\| \left(\sup_{\mathbf{f}} \frac{\|\mathbf{A}^{-1}\mathbf{f}\|}{\|\mathbf{f}\|}\right)^{-1} = 1.$$

Параметр $\nu(\mathbf{f})$, характеризующий обусловленность системы, зависит от правых частей. Более тонкая его оценка есть $\nu(\mathbf{f}) = \|\mathbf{A}^{-1}\| \frac{\|\mathbf{f}\|}{\|\mathbf{u}\|}$, причем $1 \leqslant \nu(\mathbf{f}) \leqslant \mu$. Так как такую оценку провести не всегда возможно, то чаще используется точная верхняя грань $\|\mathbf{A}^{-1}\| \|\mathbf{A}\|$. Такая оценка, конечно, может быть существенно завышенной.

Можно также показать, что для симметричной матрицы $\mathbf{A}$ имеет место $\mu = \left|\max_k \lambda_{\mathbf{A}}^k\right| / \left|\min_k \lambda_{\mathbf{A}}^k\right|$, т. е. обусловленность СЛАУ зависит от ее спектральных свойств. Это следует из определения третьей нормы матрицы $\|\mathbf{A}\|_3 = \left|\max_k \lambda_A^k\right|$ и соотношения $\|\mathbf{A}^{-1}\|_3 = \left|\min_k \lambda_A^k\right|^{-1}$, которое предлагается доказать самостоятельно. ■

2.4. Прямые методы решения СЛАУ

Трудность численного решения рассматриваемых СЛАУ определяется видом матрицы $\mathbf{A}$. Легко получается решение системы с диагональной матрицей, в этом случае система распадается на n линейных уравнений, каждое из которых содержит лишь одну неизвестную величину. Для

диагональной системы очевидны явные формулы

$$u_k = f_k/a_{kk}, k = 1 \div n.$$

В случае треугольной матрицы

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ 0 & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_{nn} \end{pmatrix}$$

из последнего уравнения получаем $u_n = f_n/a_{nn}, (a_{ii} \neq 0$, т. к. $\Delta = \det \mathbf{A} \neq 0)$.

Решая систему линейных уравнений с треугольной матрицей «снизу вверх», для u_k имеем

$$u_k = \frac{1}{a_{kk}}(f_k - a_{kn}u_n - a_{k,n-1}u_{n-1} - \dots - a_{k,k+1}u_{k+1}),$$

$$\text{или } u_k = a_{kk}^{-1}(f_k - \sum_{j=k+1}^{n} a_{kj}u_j), k = n-1, n-2, \dots, 1.$$

Можно оценить количество арифметических действий, затрачиваемых на решение такой системы. Оно составляет $O(n^2)$.

Пусть теперь система уравнений имеет матрицу общего вида. Стандартная схема такого решения разделяется на два этапа: прямой ход — приведение матрицы к треугольному виду, и обратный — вычисление решения системы.

2.4.1. Метод исключения Гаусса

Рассматривается система уравнений

$$\begin{cases} a_{11}u_1 + a_{12}u_2 + \dots + a_{1n}u_n = f_1, \\ a_{21}u_1 + a_{22}u_2 + \dots + a_{2n}u_n = f_2, \\ \dots \\ a_{n1}u_1 + a_{n2}u_2 + \dots + a_{nn}u_n = f_n. \end{cases} \tag{2.10}$$

Прямой ход метода Гаусса состоит в следующем. Положим, что $a_{11} \neq 0$ и исключим u_1 из всех уравнений, начиная со второго, для чего ко второму уравнению прибавим первое, умноженное на $-a_{21}/a_{11} = -\eta_{21}$, к третьему прибавим первое, умноженное на $-a_{31}/a_{11} = -\eta_{31}$ и т. д. После этих преобразований получим эквивалентную систему:

$$\begin{cases} a_{11}u_1 + a_{12}u_2 + \dots + a_{1n}u_n = f_1, \\ a_{22}^1u_2 + \dots + a_{2n}^1u_n = f_2^1, \\ \dots \\ a_{n2}^1u_2 + \dots + a_{nn}^1u_n = f_n^1, \end{cases} \tag{2.11}$$

в которой коэффициенты и правые части определяются следующим образом:

$$a_{ij}^1 = a_{ij} - \eta_{i1}a_{1j}; f_i^1 = f_i - \eta_{i1}f_1; i,j = 2,\ldots,n.$$

Теперь положим $a_{22}^1 \neq 0$. Аналогично, вычислив множители второго шага $-a_{i2}^1/a_{22}^1 = -\eta_{i2}$ $(i = 3,\ldots,n)$, исключаем u_2 из последних $(n-2)$ уравнений системы (2.17). В результате преобразований получим новую эквивалентную систему уравнений

$$\begin{cases} a_{11}u_1 + a_{12}u_2 + a_{13}u_3 + \ldots + a_{1n}u_n = f_1 \\ a_{22}^1 u_2 + a_{23}^1 u_3 + \ldots + a_{2n}^1 u_n = f_2^1 \\ a_{33}^2 u_3 + \ldots + a_{3n}^2 u_n = f_3^2 \\ \ldots \\ a_{n3}^2 u_3 + \ldots + a_{nn}^2 u_n = f_n^2 \end{cases},$$

в которой $a_{ij}^2 = a_{ij}^1 - \eta_{i2}a_{2j}^1$; $f_i^2 = f_i^1 - \eta_{i2}f_2^1$; $i,j = 3,\ldots,n$. Продолжая алгоритм, т.е. исключая u_i $(i = k+1,\ldots,n)$, приходим на $n-1$ шаге к системе с треугольной матрицей

$$\begin{cases} a_{11}u_1 + a_{12}u_2 + a_{13}u_3 + \ldots + a_{1n}u_n = f_1 \\ a_{22}^1 u_2 + a_{23}^1 u_3 + \ldots + a_{2n}^1 u_n = f_2^1 \\ a_{33}^2 u_3 + \ldots + a_{3n}^2 u_n = f_3^2 \\ \qquad\qquad \ldots \\ a_{nn}^{(n-1)} u_n = f_n^{(n-1)}. \end{cases} \tag{2.12}$$

Обратный ход метода Гаусса позволяет определить решение системы линейных уравнений. Из последнего уравнения системы находим u_n; подставляем это значение в предпоследнее уравнение, получим u_{n-1}. Поступая так и далее, последовательно находим $u_{n-2}, u_{n-3}, \ldots, u_1$. Вычисления компонент вектора решения проводятся по формулам

$$u_n = f_n^{(n-1)}/a_{nn}^{(n-1)},$$

$$\ldots$$

$$u_k = \frac{1}{a_{kk}^{(k-1)}}(f_k^{(k-1)} - a_{k,k+1}^{(k-1)}u_{k+1} - \ldots - a_{kn}^{(k-1)}u_n), \quad k = n-1, n-2, \ldots, 1,$$

$$\ldots$$

$$u_2 = \frac{1}{a_{22}^1}(f_2^1 - a_{23}^1 u_3 - \ldots - a_{2n}^1 u_n),$$

$$u_1 = \frac{1}{a_{11}}(f_1 - a_{12}u_2 - \ldots - a_{1n}u_n).$$

Этот алгоритм прост и легко реализуем при условии, что $a_{11} \neq 0$, $a_{22} \neq 0$ и т. д. Количество арифметических действий прямого хода $\approx 2/3n^3$, обратного $\approx n^2$. Это уже приемлемая для современных компьютеров величина.

Рассмотрим метод Гаусса с позиции операций с матрицами. Пусть $\mathbf{A}_1$ — матрица системы после исключения первого неизвестного

$$\mathbf{A}_{11} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ 0 & a_{22}^1 & a_{23}^1 & \dots & a_{2n}^1 \\ 0 & a_{32}^1 & a_{33}^1 & \dots & a_{3n}^1 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & a_{n2}^1 & a_{n3}^1 & \dots & a_n^1 n \end{pmatrix}, \quad \mathbf{f}_1 = \{f_1, f_2^1, \dots, f_n^1\}^{\mathrm{T}}.$$

Введем новую матрицу

$$\mathbf{N}_1 = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ -\eta_{21} & 1 & 0 & \dots & 0 \\ -\eta_{31} & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ -\eta_{n1} & 0 & 0 & \dots & 1 \end{pmatrix}.$$

Очевидно, $\mathbf{A}_1 = \mathbf{N}_1\mathbf{A}$, $\mathbf{f}_1 = \mathbf{N}_1\mathbf{f}$. Аналогично, после второго шага система приводится к виду $\mathbf{A}_2\mathbf{u} = \mathbf{f}_2$, где $\mathbf{A}_2 = \mathbf{N}_2\mathbf{A}_1$, $\mathbf{f}_2 = \mathbf{N}_2\mathbf{f}_1$,

$$\mathbf{A}_1 = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ 0 & a_{22}^1 & a_{23}^1 & \dots & a_{2n}^1 \\ 0 & 0 & a_{33}^2 & \dots & a_{3n}^2 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & a_{n3}^2 & \dots & a_{nn}^2 \end{pmatrix}, \mathbf{N}_2 = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & -\eta_{32} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & -\eta_{n2} & 0 & \dots & 1 \end{pmatrix},$$

$$\mathbf{f}_2 = \left\{f_1, f_2^1, f_3^2, \dots, f_n^2\right\}^{\mathrm{T}}.$$

После $n-1$ шага получим $\mathbf{A}_{n-1}\mathbf{u} = \mathbf{f}_{n-1}$, $\mathbf{A}_{n-1} = \mathbf{N}_{n-1} \cdot \mathbf{A}_{n-2}$, $\mathbf{f}_{n-1} = \mathbf{N}_{n-1}\mathbf{f}_{n-2}$,

$$\mathbf{A}_{n-1} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ 0 & a_{22}^1 & a_{23}^1 & \dots & a_{2n}^1 \\ 0 & 0 & a_{33}^2 & \dots & a_{3n}^2 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & a_{nn}^{(n-1)} \end{pmatrix},$$

$$\mathbf{N}_{n-1} = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & -\eta_{n,n-1} & 1 \end{pmatrix}, \quad \mathbf{f}_{n-1} = \{f_1, f_2^1, f_3^2, \dots, f_n^{n-1}\}^{\mathrm{T}}.$$

В итоге получаются матрица и вектор $\mathbf{A}_{n-1} = \mathbf{N}_{n-1} \dots \mathbf{N}_2\mathbf{N}_1\mathbf{A}$, $\mathbf{f}_{(n-1)} = \mathbf{N}_{n-1} \dots \mathbf{N}_2\mathbf{N}_1\mathbf{f}$, откуда $\mathbf{A} = \mathbf{N}_1^{-1}\mathbf{N}_2^{-1} \dots \mathbf{N}_{n-1}^{-1} \cdot \mathbf{A}_{n-1}$. При этом

$$\mathbf{N}_1^{-1} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ \eta_{21} & 1 & 0 & \dots & 0 \\ \eta_{31} & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \eta_{n1} & 0 & 0 & \dots & 1 \end{pmatrix}, \quad \mathbf{N}_2^{-1} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & \eta_{32} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \eta_{n2} & 0 & \dots & 1 \end{pmatrix},$$

$$\mathbf{N}_{n-1}^{-1} = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & \eta_{n,n-1} & 1 \end{pmatrix}.$$

После введения обозначений $\mathbf{U} = \mathbf{A}_{n-1}$, $\mathbf{L} = \mathbf{N}_1^{-1}\mathbf{N}_2^{-1} \dots \mathbf{N}_{n-1}^{-1}$, где

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ \eta_{21} & 1 & 0 & \dots & 0 \\ \eta_{31} & \eta_{32} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \eta_{n1} & \eta_{n2} & \eta_{n3} & \dots & 1 \end{pmatrix},$$

получим $\mathbf{A} = \mathbf{LU}$.

Это представление матрицы $\mathbf{A}$ называется LU-разложением (на произведение нижней и верхней треугольных матриц $\mathbf{L}$ и $\mathbf{U}$). Прямой ход метода Гаусса можно рассматривать как один из вариантов представления матрицы в виде произведения двух треугольных матриц, или LU-разложения. Его можно провести и другими способами.

Вспомним «отрицательный» пример из лекции 1. Пусть необходимо решить систему

$$-10^{-7}u_1 + u_2 = 1,$$
$$u_1 + 2u_2 = 4.$$

Исключая u_1 из первого уравнения и подставляя во второе, получим $u_2 = (10^7 + 4)/(10^7 + 2)$. После вычислений с семью значащими цифрами получаем $u_1 = 0,000000, u_2 = 1,000000$, что неверно (см. второе уравнение). Теперь исключим u_1 из второго уравнения и подставим в первое. При этом получим $u_2 = \frac{1+4\cdot10^{-7}}{1+2\cdot10^{-7}}$. После вычислений с той же точностью имеем: $u_2 = 1,000000, u_1 = 2,000000$, что является правильным решением (с заданным количеством значащих цифр).

В реальных вычислениях используются методы с *выбором главного* (или *ведущего*) *элемента*. Выбор главного элемента *по столбцам* реализуется следующим образом: перед исключением u_1 отыскивается $\max\limits_i |a_{i1}|$.

Пусть максимум достигается при $i = k$. В этом случае меняются местами первое и k уравнения (или в матрице меняются местами две строки) и реализуется процедура исключения.

Затем отыскивается $\max\limits_i |a_{i2}^1|$, и процедура поиска главного элемента в столбцах повторяется. Так же реализуется выбор главного элемента по строкам: перед исключением u_1 отыскивается $\max\limits_j |a_{kj}|$. Если максимум достигается при $i = k$, то у u_1 и u_k меняются номера, то есть максимальный элемент из коэффициентов первого уравнения окажется на месте a_{11}, и т. д. Наиболее эффективным является метод Гаусса с выбором главного элемента по всей матрице. Во многих методах важным является условие *диагонального преобладания* $|a_{ii}| \geqslant \sum\limits_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|$ для $i = 1, \ldots, n$, при выполнении которого проблемы, появляющиеся в методе Гаусса, не возникают. Если для всех строк матрицы выполняются строгие неравенства, то говорят о *строгом диагональном преобладании.*

Полученное решение можно улучшить следующим образом. Пусть $\mathbf{r}^1 = \mathbf{f} - \mathbf{A}\mathbf{u}^1$ есть невязка, допущенная при решении рассматриваемой системы ($\mathbf{u}^1$ — полученное численное решение) за счет ошибки округлений. Очевидно, что погрешность $\boldsymbol{\varepsilon}^1 = \mathbf{u} - \mathbf{u}^1$ удовлетворяет СЛАУ $\mathbf{A}\boldsymbol{\varepsilon}^1 = \mathbf{r}^1$, так как $\mathbf{A}\boldsymbol{\varepsilon}^1 = \mathbf{A}\mathbf{u} - \mathbf{A}\mathbf{u}^1 = \mathbf{f} - \mathbf{A}\mathbf{u}^1$.

Решив последнюю систему, получаем $\boldsymbol{\varepsilon}^1$, после чего уточняем решение:

$$\mathbf{u}^2 = \mathbf{u}^1 + \boldsymbol{\varepsilon}^1.$$

Эту процедуру можно продолжить.

2.4.2. Модификация метода Гаусса для случая линейных систем с трехдиагональными матрицами — метод прогонки

В приложениях часто возникают системы линейных уравнений с матрицами специального вида. В дальнейшем, например, при интерполяции функции сплайнами, при решении задачи Штурма–Лиувилля, при численном решении уравнений теплопроводности, уравнений Лапласа и Пуассона придется иметь дело с системами, матрицы которых имеют ненулевые элементы лишь на главной диагонали и еще на двух диагоналях — одной под главной, одной над главной. Такие матрицы называются трехдиагональными. Для решения систем с трехдиагональными матрицами существует экономичный (требующий малого количества памяти и арифметических действий) вариант метода Гаусса — *прогонка.* В англоязычной литературе метод прогонки называется алгоритмом Томаса. Подробнее о свойствах метода прогонки речь пойдет в лекции 10 в связи с решением разностными методами задачи Штурма–Лиувилля.

2.4.3. LU-разложение

Среди прямых методов численного решения СЛАУ широко используется также LU-разложение матрицы **A** и метод Холецкого (или метод квадратного корня).

Если матрица **A** представима в виде произведений матриц **LU**, то СЛАУ может быть представлена в виде

$$(\mathbf{LU})\mathbf{u} = \mathbf{f}. \tag{2.13}$$

Перепишем (2.13), вводя вспомогательный вектор **v**, в следующем виде

$$\mathbf{Lv} = \mathbf{f}, \mathbf{Uu} = \mathbf{v}. \tag{2.14}$$

Решение СЛАУ свелось к последовательному решению двух систем с треугольными матрицами. Первый этап решения системы $\mathbf{Lv} = \mathbf{f}$:

$$\begin{cases} v_1 = f_1, \\ l_{21}v_1 + v_2 = f_2, \\ \dots \\ l_{n1}v_1 + l_{n2}v_2 + \dots + l_{n,n-1}v_{n-1} + v_n = f_n, \end{cases}$$

откуда можно вычислить все v_k последовательно по формулам

$$v_k = f_k - \sum_{j=1}^{k-1} l_{kj}\, v_j; k = 2, \dots, n.$$

Далее рассмотрим систему $\mathbf{Uu} = \mathbf{v}$ или

$$\begin{cases} d_{11}u_1 + d_2u_2 + \dots + d_{1n}u_n = v_1, \\ d_{22}u_2 + \dots + d_{2n}u_n = v_2, \\ \qquad\qquad \dots \\ d_{nn}u_n = v_n, \end{cases}$$

решение которой находится в обратном порядке, т. е. при $k = n-1, \dots, 1$ по очевидным формулам $u_k = d_{kk}^{-1}(v_k - \sum\limits_{j=k+1}^{n} d_{kj}u_j)$. Условия существования такого разложения даются следующей теоремой [5] (без доказательства).

Теорема. *Если все главные миноры квадратной матрицы* **A** *отличны от нуля, то существуют единственные нижняя и верхняя треугольные матрицы* $\mathbf{L} = l_{ij}$ *и* $\mathbf{U} = d_{ij}$ *такие, что* $\mathbf{A} = \mathbf{LU}$. *При этом все диагональные коэффициенты матрицы* **L** *фиксированы и равны единице.*

Опишем алгоритм нахождения элементов $l_{ij}d_{ij}$ матриц $\mathbf{L}$, $\mathbf{U}$. Выписав равенство $\mathbf{A} = \mathbf{LU}$ в компонентах, получим

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ l_{21} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ l_{n1} & l_{n2} & \dots & 1 \end{pmatrix} \begin{pmatrix} d_{11} & d_{12} & \dots & d_{1n} \\ 0 & d_{22} & \dots & d_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & d_{nn} \end{pmatrix}.$$

Выполнив умножение матриц, приходим к системе линейных уравнений размером $n \times n$:

$$d_{11} = a_{11}, d_{12} = a_{12}, \dots, d_{1n} = a_{1n},$$
$$l_{21}d_{11} = a_{21}, l_{21}d_{12} + d_{22} = a_{22}, \dots, l_{21}d_{1n} + d_{2n} = a_{2n},$$
$$\dots$$
$$l_{n1}d_{11}{=}a_{n1}, l_{n1}d_{12}{+}l_{n2}d_{22}{=}a_{n2}, \dots, l_{n1}d_{1n}{+}\dots{+}l_{n,n-1}d_{n-1,n}{+}d_{nn}{=}a_{nn}$$

относительно неизвестных $d_{11}, d_{12}, \dots, d_{1n}, l_{21}, d_{22}, \dots, d_{2n}, l_{n1}, l_{n2}, \dots, d_{nn}$.

Специфика этой системы позволяет решить ее последовательно. Из первой строки находим $d_{1j} = a_{1j} (j = 1, \dots, n)$.

Из уравнений, входящих в первый столбец приведенной выше системы, находим $l_{i1} = a_{i1}/d_{11}, i = 1, \dots, n$. Теперь можно из уравнений второй строки найти $d_{2j} = a_{2j} - l_{21}d_{1j}, j = 2, \dots, n$, а из уравнений, входящих во второй столбец, получим $l_{i2} = d_{22}^{-1}(a_{i2} - l_{i1}d_{12}), i = 2, \dots, n$ и так далее. Последним вычисляется элемент

$$d_{nn} = a_{nn} - \sum_{k=1}^{n-1} l_{nk}d_{kn}.$$

Можно выписать общий вид этих формул:

$$d_{ij} = a_{ij} - \sum_{k=1}^{j-1} l_{ik}d_{kj}, \quad i \leqslant j,$$
$$l_{ij} = d_{ij}^{-1}\left(a_{ij} - \sum_{k=1}^{j-1} l_{ik}d_{kj}\right), i > j.$$

Приведение матриц к треугольному виду аналогично приведению матрицы в методе Гаусса и также требует количества арифметических действий порядка $O(n^3)$, точнее, $\approx 2n^3$.

2.4.4. Метод Холецкого (метод квадратного корня)

Пусть матрица рассматриваемой линейной системы $\mathbf{A}$ — симметричная, т. е. $a_{ij} = a_{ji}$, положительная матрица. Тогда она представима в

виде $\mathbf{A} = \mathbf{L}\mathbf{L}^T$, где

$$\mathbf{L}^{\mathrm{T}} = \begin{pmatrix} l_{11} & l_{12} & \dots & l_{1n} \\ 0 & l_{22} & \dots & l_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & l_{nn} \end{pmatrix}, \mathbf{L} = \begin{pmatrix} l_{11} & 0 & \cdots & 0 \\ l_{12} & l_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ l_{1n} & l_{2n} & \dots & l_{nn} \end{pmatrix}.$$

Далее, как и в случае LU-разложения, решение СЛАУ $\mathbf{Au} = \mathbf{f}$ сводится к последовательному решению двух линейных систем с треугольными матрицами $\mathbf{Lv} = \mathbf{f}$, $\mathbf{L}^{\mathrm{T}}\mathbf{u} = \mathbf{v}$, для решения которых требуется примерно $2n^2$ арифметических действий.

Первая из этих линейных систем

$$\begin{gathered} l_{11}v_1 = f_1, \\ l_{12}v_1 + l_{22}v_2 = f_2, \\ \dots \\ l_{1n}v_1 + l_{2n}v_2 + \ldots + l_{nn}v_n = f_n, \end{gathered}$$

она легко решается. Для решения получаем очевидные формулы

$$v_i = l_{ii}^{-1}(f_i - \sum_{k=1}^{i} l_{ki}v_k), i = 1, \ldots, n.$$

Вторая система уравнений есть

$$l_{11}u_1 + l_{12}u_2 + \ldots + l_{1n}u_n = v_1, l_{22}u_2 + \ldots + l_{2n}u_n = v_2, \ldots l_{nn}u_n = v_n.$$

Из нее находим значения переменных u_i в обратном порядке по формуле

$$u_k = l_{ii}^{-1}(v_k - \sum_{j=k+1}^{n} l_{kj}u_j).$$

Определенной опасностью при реализации этого метода являются возможная близость к нулю l_{ii} и отрицательность подкоренных выражений при вычислении l_{ii} (последнего не должно быть при симметричной положительной матрице A)

Элементы матрицы L находим из уравнения $\mathbf{L}\mathbf{L}^T = \mathbf{A}$, приравнивая соответствующие элементы матриц $\mathbf{L}\mathbf{L}^T$ и A. В результате получим

систему уравнений

$$l_{11}^2 = a_{11},$$
$$l_{i1}l_{11} = a_{i1}, i = 2, \ldots, n,$$
$$l_{21}^2 + l_{22}^2 = a_{22},$$
$$l_{i1}l_{21} + l_{i2}l_{22} = a_{i2}, i = 3, \ldots, n,$$
$$\ldots$$
$$l_{k1}^2 + l_{k2}^2 + \ldots + l_{kk}^2 = a_{kk},$$
$$l_{i1}l_{k1} + l_{i2}l_{k2} + \ldots + l_{ik}l_{kk} = a_{kk}, i = k+1, \ldots, n.$$

Решение этой системы легко находится:

$$l_{11} = \sqrt{a_{11}},$$
$$l_{i1} = a_{i1}/l_{11}, i = 2, \ldots, n,$$
$$l_{22} = \sqrt{a_{22} - l_{21}^2},$$
$$l_{i2} = (a_{i2} - l_{i1}l_{21})/l_{22}, i = 3, \ldots, n,$$
$$\ldots$$
$$l_{kk} = \sqrt{a_{kk} - l_{k1}^2 - l_{k2}^2 - \ldots - l_{k,k-1}^2},$$
$$l_{ik} = (a_{ik} - l_{i1}l_{k1} - l_{i2}l_{k2} - \ldots - l_{i,k-1}l_{k,k-1})/l_{kk}, i = k+1, \ldots, n,$$

Метод также называется методом квадратного корня.

Внимание! Не следует путать матрицу (оператор) $\mathbf{L}$ с оператором $\mathbf{A}^{1/2}$ — квадратным корнем из самосопряженного положительного оператора.

2.5. Итерационные методы решения СЛАУ

2.5.1. Метод простой итерации

Рассмотрим систему линейных алгебраических уравнений

$$\mathbf{Au} = \mathbf{f}.$$

Проведем несколько равносильных преобразований. Умножим обе части системы на один и тот же скалярный множитель τ, затем прибавим к правой и левой частям системы вектор $\mathbf{u}$. Систему уравнений можно теперь записать в виде, удобном для итераций:

$$\mathbf{u} = \mathbf{Bu} + \mathbf{F}, \tag{2.15}$$

где $\mathbf{B} = \mathbf{E} - \tau\mathbf{A}, \mathbf{F} = \tau\mathbf{f}$.

Теперь построим последовательность приближений к решению системы. Выберем произвольный вектор $\mathbf{u}_0$ — начальное приближение к решению. Чаще всего его просто полагают нулевым вектором. Скорее всего, начальное приближение не удовлетворяет (2.15) и, следовательно, исходной системе. При подстановке его в исходное уравнение возникает невязка $\mathbf{r}_0 = \mathbf{f} - \mathbf{A}\mathbf{u}_0$. Вычислив невязку, с помощью (2.15) можно уточнить приближение к решению, считая, что

$$\mathbf{u}_1 = \mathbf{u}_0 + \tau\mathbf{r}_0.$$

По первому приближению снова вычисляется невязка, процесс продолжается. В ходе итерации получаем $\mathbf{u}_{k+1} = \mathbf{u}_k + \tau\mathbf{r}_k$, $\mathbf{r}_k = \mathbf{f} - \mathbf{A}\mathbf{u}_k$. Эквивалентная формулировка метода, называемого методом простых итераций, заключается в следующем. Решение (2.15) находится как предел последовательности $\{\mathbf{u}_0, \mathbf{u}_1, \mathbf{u}_2, \ldots\}$ приближений, члены которой связаны рекуррентным соотношением (оно эквивалентно приведенному выше, из записи исключен вектор невязки):

$$\mathbf{u}_{k+1} = \mathbf{B}\mathbf{u}_k + \mathbf{F}, \tag{2.16}$$

$\mathbf{u}_0 = 0$ (или любому произвольному вектору). Если предел такой последовательности существует, то говорят о *сходимости* итерационного процесса к решению СЛАУ.

Существуют другие формы записи метода итераций, например

$$\mathbf{u}_{k+1} = (\mathbf{E} - \tau\mathbf{A})\mathbf{u}_k + \tau\mathbf{f}. \tag{2.17}$$

Канонической формой записи двухслойного итерационного процесса называется следующая:

$$\mathbf{D}_{k+1}\frac{\mathbf{u}_{k+1} - \mathbf{u}_k}{\tau_{k+1}} + \mathbf{A}\mathbf{u}_k = \mathbf{f}. \tag{2.18}$$

При $\mathbf{D}_k = \mathbf{E}$, $\tau_k = \tau$ последняя формула соответствует однопараметрическому итерационному процессу — рассмотренному выше *методу простых итераций*. При $\mathbf{D}_k = \mathbf{E}$, $\tau_k = \{\tau_k, k = 1, \ldots, n\}$ — n-шаговому явному итерационному процессу, при $\mathbf{D}_k = \mathbf{D}'$, $\tau_k = 1$ — методу простой итерации без итерационного параметра. В случае, когда $\mathbf{D} \neq \mathbf{E}$, итерационный метод называется *неявным* — для вычисления следующего приближения к решению придется решать (как правило, более простую, чем исходную) систему линейных уравнений.

Теорема (достаточное условие сходимости метода простой итерации). *Итерационный процесс* (2.16) *сходится к решению* $\mathbf{U}$ *СЛАУ* $\mathbf{A}\mathbf{u} = \mathbf{F}$ *со скоростью геометрической прогрессии при выполнении условия:* $\|\mathbf{B}\| \leqslant q < 1$.

Доказательство.

Пусть $\mathbf{U}$ — точное решение системы (2). Вычитая из (2.16)-(2.15), получим $\mathbf{u}_k - \mathbf{U} = \mathbf{B}(\mathbf{u}_{k-1} - \mathbf{U})$, или, обозначив погрешность $\boldsymbol{\varepsilon}_k = \mathbf{u}_k - \mathbf{U}$, получим для эволюции погрешности уравнение $\boldsymbol{\varepsilon}_k = \mathbf{B}\boldsymbol{\varepsilon}_{k-1}$. Справедлива цепочка неравенств: $\|\mathbf{u}_k - \mathbf{U}\| = \|\boldsymbol{\varepsilon}_k\| \leqslant \|\mathbf{B}\| \cdot \|\boldsymbol{\varepsilon}_{k-1}\| \leqslant q\|\boldsymbol{\varepsilon}_{k-1}\| \leqslant \ldots \leqslant q^k \|\boldsymbol{\varepsilon}_0\| = q^k \|\mathbf{u}_0 - \mathbf{U}\|$, где $0 < q \leqslant \|\mathbf{B}\|$.

Отсюда следует, что при $q < 1$ $\lim\limits_{k\to\infty} \mathbf{u}_k = \mathbf{U}$.

Из неравенства $\|\boldsymbol{\varepsilon}_k\| \leqslant q^k \|\boldsymbol{\varepsilon}_0\|$ можно получить оценку количества итераций, необходимых для достижения точности ε, т. е. для выполнения условия $\|\mathbf{u}_k - \mathbf{U}\| = \|\boldsymbol{\varepsilon}_k\| \leqslant \varepsilon$. Эта оценка имеет вид $k \geqslant \left(\ln \frac{\varepsilon}{\|\boldsymbol{\varepsilon}_0\|}\right) / \ln q$.

■

Теорема (критерий сходимости метода простой итерации (без доказательства)). *Пусть СЛАУ* (2.15) *имеет единственное решение. Тогда для сходимости итерационного процесса* (2.16) *необходимо и достаточно, чтобы все собственные значения матрицы* $\mathbf{B}$ *по абсолютной величине были меньше единицы.*

Сравним по количеству арифметических действий прямые и итерационные методы. Метод Гаусса без выбора главного элемента при $n \gg 1$ требует $\approx \left(\frac{2}{3}n^3\right)$ арифметических действий; метод простой итерации (2.16) $\approx (2n^2 \cdot I)$, где I — число приближений, необходимое для достижения заданной точности. Значит, при $I < n/3$ метод итераций становится предпочтительнее. В реальных задачах, в основном, $I \ll n$. Кроме того, итерационные методы можно делать более эффективными, изменяя итерационные параметры. В ряде случаев итерационные методы оказываются более устойчивыми по отношению к накоплению ошибок округления, чем прямые.

2.5.2. Влияние ошибок округления на результат численного решения

Будем трактовать суммарный эффект ошибок округления при выполнении одного итерационного шага как возмущение правой части в итерационном процессе

$$\mathbf{u}_k = \mathbf{B}\mathbf{u}_{k-1} + \mathbf{F}. \tag{2.19}$$

Результат вычислений на каждой итерации при наличии ошибок округления представим в виде

$$\mathbf{u}_k^M = \mathbf{B}\mathbf{u}_{k-1}^M + \mathbf{F} + \boldsymbol{\delta}_k, \tag{2.20}$$

где $\boldsymbol{\delta}_k$ — суммарная погрешность округления. Норму разности между реальным и идеальным (т. е. в отсутствии ошибки округления) результатами

расчетов получим, вычитая (2.19) из (2.20). Учтем, что $\|\mathbf{B}\| < q < 1$,

$$\|\mathbf{u}_k^M - \mathbf{u}_k\| \leqslant q\|\mathbf{u}_{k-1}^M - \mathbf{u}_{k-1}\| + \\ + \|\boldsymbol{\delta}_k\| \leqslant q^2\|\mathbf{u}_{k-2}^M - \mathbf{u}_{k-2}\| + q\|\boldsymbol{\delta}_{k-1}\| + \|\boldsymbol{\delta}_k\| \leqslant \ldots \leqslant q^k\|\mathbf{u}_0^M - \mathbf{u}_0\| + \\ + (\max_i \|\boldsymbol{\delta}_i\|)(1 + q + \ldots + q^{k-1}), i = 1, \ldots, k.$$

Так как начальное приближение задано точно $\|\mathbf{u}_0^M - \mathbf{u}_0\| = 0$. Обозначим $\delta = \max_i \|\boldsymbol{\delta}_i\|$ и вычислим сумму членов геометрической прогрессии. Получим $\|\mathbf{u}_k^M - \mathbf{u}_k\| \leqslant \delta\frac{q^k - 1}{q - 1} \leqslant \frac{\delta}{1-q}$, то есть погрешность, вносимая в решение из-за конечной разрядности мантиссы, не зависит от количества итераций. Этот результат является характеристикой устойчивости рассматриваемого вычислительного процесса.

2.5.3. Методы Якоби, Зейделя, верхней релаксации

Представим матрицу $\mathbf{A}$ в виде

$$\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{U}, \tag{2.21}$$

где $\mathbf{L}$ и $\mathbf{U}$ — нижняя и верхняя треугольные матрицы с нулевыми элементами на главной диагонали, $\mathbf{D}$ — диагональная матрица. Рассматриваемая СЛАУ может быть переписана в следующем эквивалентном виде:

$$\mathbf{Lu} + \mathbf{Du} + \mathbf{Uu} = \mathbf{f}.$$

Построим два итерационных метода

$$\mathbf{Lu}_k + \mathbf{Du}_{k+1} + \mathbf{Uu}_k = \mathbf{f}$$

и

$$\mathbf{Lu}_{k+1} + \mathbf{Du}_{k+1} + \mathbf{Uu}_k = \mathbf{f},$$

или, соответственно,

$$\mathbf{u}_{k+1} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})\mathbf{u}_k + \mathbf{D}^{-1}\mathbf{f} \tag{2.22}$$

и

$$\mathbf{u}_{k+1} = -(\mathbf{L} + \mathbf{D})^{-1}\mathbf{U}\mathbf{u}_k + (\mathbf{L} + \mathbf{D})^{-1}\mathbf{f}. \tag{2.23}$$

Очевидно, что эти формулы описывают итерационные процессы вида (2.16), если положить в (2.22)

$$\mathbf{B} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U}), \mathbf{F} = \mathbf{D}^{-1}\mathbf{f}$$

или

$$\mathbf{B} = -(\mathbf{L}+\mathbf{D})^{-1}\mathbf{U}, \mathbf{F} = (\mathbf{L}+\mathbf{D})^{-1}\mathbf{f}.$$

Эти итерационные процессы называются *методами Якоби и Зейделя*. Представим их в компонентной записи. Метод Якоби будет иметь вид (перенесем итерационный индекс k вверх):

$$u_1^{k+1} = -(a_{12}u_2^k + a_{13}u_3^k + \ldots + a_{1n}u_n^k - f_1)/a_{11},$$

$$u_2^{k+1} = -(a_{21}u_1^k + a_{23}u_3^k + \ldots + a_{2n}u_n^k - f_2)/a_{22},$$

$$\ldots$$

$$u_n^{k+1} = -(a_{n1}u_1^k + a_{n2}u_2^k + \ldots + a_{n,n-1}u_{n-1}^k - f_n)/a_{nn}.$$

Метод Зейделя можно представить следующим образом:

$$u_1^{k+1} = -(a_{12}u_2^k + a_{13}u_3^k + \ldots + a_{1n}u_n^k - f_1)/a_{11},$$

$$u_2^{k+1} = -(a_{21}u_1^{k+1} + a_{23}u_3^k + \ldots + a_{2n}u_n^k - f_2)/a_{22},$$

$$\ldots$$

$$u_n^{k+1} = -(a_{n1}u_1^{k+1} + a_{n2}u_2^{k+1} + \ldots + a_{n,n-1}u_{n-1}^{k+1} - f_n)/a_{nn}.$$

Эти формулы легко выводятся, если учесть, что элементами матрицы D^{-1} являются $d_{ii} = a_{ii}^{-1}$.

Теорема (достаточное условие сходимости метода Якоби). *Итерационный метод Якоби сходится к решению соответствующей СЛАУ, если выполнено условие диагонального преобладания*

$$|a_{ii}| > \sum_{\substack{j=1 \\ (j\neq i)}}^{n} |a_{ij}|, i = 1, \ldots, n. \tag{2.24}$$

Доказательство.

Выполненные условия (2.24) означает, что в любой строке матрицы перехода

$$\mathbf{B} = \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} & -\frac{a_{13}}{a_{11}} & \ldots & -\frac{a_{1n-1}}{a_{11}} & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & -\frac{a_{23}}{a_{22}} & \ldots & -\frac{a_{2n-1}}{a_{22}} & -\frac{a_{2n}}{a_{22}} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ -\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & -\frac{a_{n3}}{a_{nn}} & \ldots & -\frac{a_{n,n-1}}{a_{nn}} & 0 \end{pmatrix},$$

сумма модулей элементов меньше единицы. В этом случае по крайней мере одна из норм матрицы $\mathbf{B}$ меньше единицы. Тогда выполняется достаточное условие сходимости метода простых итераций. ■

Теорема (критерий сходимости итерационного метода Якоби). *Для сходимости итерационного метода Якоби необходимо и достаточно, чтобы все корни уравнения*

$$\begin{vmatrix} \lambda a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & \lambda a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & \lambda a_{nn} \end{vmatrix} = 0$$

по модулю не превосходили единицы.

Доказательство.

Легко проверить, что в силу диагональности $\mathbf{D}$ имеет место

$$\det(\mathbf{B} - \lambda\mathbf{E}) = \det\left[-\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U}) - \lambda\mathbf{E}\right] = \det(-\mathbf{D}^{-1}) \cdot \det\left[(\mathbf{L}+\mathbf{U}) + \mathbf{D}\lambda\right].$$

Собственными значениями матрицы $\mathbf{B} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})$ являются корни уравнения

$$\det\left[(\mathbf{L} + \mathbf{U}) + \mathbf{D}\lambda\right] = 0,$$

которые в соответствии с критерием сходимости метода простой итерации должны быть по модулю меньше единицы.

Аналогичную теорему можно доказать и для метода Зейделя, однако матрица в этой теореме будет иметь другой вид:

$$\begin{pmatrix} \lambda a_{11} & a_{12} & \dots & a_{1n} \\ \lambda a_{12} & \lambda a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ \lambda a_{n1} & \lambda a_{n2} & \dots & \lambda a_{nn} \end{pmatrix}.$$

■

Теорема (достаточное условие сходимости метода Зейделя (без доказательства)). *Пусть* $\mathbf{A}$ *— вещественная, симметричная, положительно определенная матрица. В этом случае итерационный метод Зейделя сходится.*

Доказательство этой теоремы сводится к проверке того, что выполнение условий теоремы для матрицы $\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{L}^T$ влечет выполнение условия сходимости итерационного метода с матрицей перехода — $(\mathbf{L} + \mathbf{D})^{-1}\mathbf{L}^T$. СЛАУ с вещественной матрицей $\mathbf{A}$ такой, что $\det \mathbf{A} \neq 0$ может быть симметризована умножением на матрицу $\mathbf{A}^\mathbf{T}$:

$$(\mathbf{A}^\mathbf{T}\mathbf{A})\mathbf{u} = \mathbf{A}^\mathbf{T}\mathbf{f}$$

(симметризация Гаусса).

Развитием метода Зейделя является метод релаксации. В этом методе вводится итерационный параметр τ, называемый параметром релаксации. Представим метод релаксации в матричной форме:

$$(\tau\mathbf{L}\mathbf{u}_{k+1} + \mathbf{D}\mathbf{u}_{k+1}) + (\tau - 1)\mathbf{D}\mathbf{u}_k + \tau\mathbf{U}\mathbf{u}_k = \tau\mathbf{f}.$$

Выбирая τ, можно существенно изменять скорость сходимости итерационного метода. Выразим $\mathbf{u}_{k+1}$

$$\mathbf{u}_{k+1} = -(\mathbf{D} + \tau\mathbf{L})^{-1}\left[(\tau - 1)\mathbf{D} + \tau\mathbf{L}\right]\mathbf{u}_k + \tau(\mathbf{D} + \tau\mathbf{L})^{-1}\mathbf{f}.$$

В общем случае задача вычисления $\tau_{\text{опт}}$ (оптимального итерационного параметра) не решена, однако известно, что $1 < \tau_{\text{опт}} < 2$. В этом случае итерационный метод называется методом последовательной верхней релаксации или SOR — Successive Over Relaxation. Иногда встречается термин «сверхрелаксация» при $1 < \tau_{\text{опт}} < 2$. При $0 < \tau < 1$ имеем метод нижней релаксации.

2.6. Вариационные итерационные методы

2.6.1. Связь между вариационной задачей и задачей решения СЛАУ

Пусть $\mathbf{u} \in L^n$, где L^n есть n-мерное евклидово пространство. Рассмотрим квадратичный функционал от $\mathbf{u}$, называемый *функционалом энергии*:

$$\Phi(\mathbf{u}) = (\mathbf{A}\mathbf{u}, \mathbf{u}) - 2(\mathbf{f}, \mathbf{u}) + c,$$

где $\mathbf{A}$ — линейный оператор, $\mathbf{f} \in L^n$, c — константа. Этот функционал совпадает с квадратичным функционалом $\Phi(\mathbf{u}) = (\mathbf{A}^*\mathbf{u}, \mathbf{u}) - 2(\mathbf{f}, \mathbf{u}) + c$, где $\mathbf{A}^*$ — сопряженный к $\mathbf{A}$ оператор. Действительно, $(\mathbf{A}\mathbf{u}, \mathbf{u}) \equiv (\mathbf{u}, \mathbf{A}^*\mathbf{u})$ по определению сопряженного оператора и $(\mathbf{u}, \mathbf{A}^*\mathbf{u}) = (\mathbf{A}^*\mathbf{u}, \mathbf{u})$ в силу коммутативности скалярного произведения. Тогда

$$\Phi(\mathbf{u}) = \left(\frac{\mathbf{A} + \mathbf{A}^*}{2}\mathbf{u}, \mathbf{u}\right) - 2(\mathbf{f}, \mathbf{u}) + c,$$

так как $\frac{1}{2}(\mathbf{A}\mathbf{u}, \mathbf{u}) + \frac{1}{2}(\mathbf{A}^*\mathbf{u}, \mathbf{u}) = \left(\frac{\mathbf{A} + \mathbf{A}^*}{2}\mathbf{u}, \mathbf{u}\right)$.

Без ограничения общности предположим, что оператор $\mathbf{A}$ — самосопряженный, $\mathbf{A} = \mathbf{A}^*$. В противном случае будем рассматривать задачу с оператором $\frac{1}{2}(\mathbf{A} + \mathbf{A}^*)$ при решении вариационной задачи.

Будем также считать, что $\mathbf{A}$ — положительный оператор, т. е. $\mathbf{A} > 0$, это означает, что для любого ненулевого вектора $\mathbf{u}$ выполнено $(\mathbf{A}\mathbf{u}, \mathbf{u}) > 0$.

Поставим задачу об отыскании элемента $\mathbf{v}$, придающего наименьшее значение функционалу $\Phi(\mathbf{u})$:

$$\Phi(\mathbf{v}) = \min_{\mathbf{u}\in L^n} \Phi(\mathbf{u}).$$

Теорема. *Пусть* $\mathbf{A} = \mathbf{A}* > 0$. *В этом случае существует единственный элемент* $\mathbf{v} \in L^n$, *придающий наименьшее значение квадратичному функционалу* $\Phi(\mathbf{u}) = (\mathbf{Au}, \mathbf{u}) - (2\mathbf{f}, \mathbf{u}) + c$, *являющийся решением СЛАУ* $\mathbf{Au} = \mathbf{f}$.

Доказательство.

СЛАУ $\mathbf{Au} = \mathbf{f}$ имеет единственное решение $\mathbf{v}$, поскольку $\mathbf{A}$ является невырожденным оператором в силу его положительной определенности. Покажем, что в этом случае при $\mathbf{Av} - \mathbf{f} = 0$ для любого вектора Δ имеет место $\Phi(\mathbf{v} + \Delta) > \Phi(\mathbf{v})$, т. е. при $\mathbf{u} = \mathbf{v}$ достигается минимум квадратичного функционала $\Phi(\mathbf{u})$.

Действительно,

$$\Phi(\mathbf{v}+\Delta)=(\mathbf{A}(\mathbf{v}+\Delta),\mathbf{v}+\Delta)-2(\mathbf{f},\mathbf{v}+\Delta)+c =$$

$$= (\mathbf{Av}+\mathbf{A}\Delta,\mathbf{v}+\Delta)-2(\mathbf{f},\mathbf{v}+\Delta)+c=$$

$$= (\mathbf{Av},\mathbf{v}) + (\mathbf{Av},\Delta) + (\mathbf{A}\Delta,\mathbf{v}) + (\mathbf{A}\Delta,\Delta) - 2(\mathbf{f},\mathbf{v}) - 2(\mathbf{f},\Delta) + c =$$

$$= (\mathbf{Av},\mathbf{v}) + 2(\mathbf{Av},\Delta) + (\mathbf{A}\Delta,\Delta) - 2(\mathbf{f},\mathbf{v}) - 2(\mathbf{f},\Delta) + c =$$

$$= [(\mathbf{Av},\mathbf{v}) - 2(\mathbf{f},\mathbf{v}) + c] + 2(\mathbf{Av},\Delta) - 2(\mathbf{f},\Delta) + (\mathbf{A}\Delta,\Delta) =$$

$$= \Phi(\mathbf{v}) + 2(\mathbf{Av}-\mathbf{f},\Delta) + (\mathbf{A}\Delta,\Delta) = \Phi(\mathbf{v}) + (\mathbf{A}\Delta,\Delta) > \Phi(\mathbf{v}),$$

т. е. при $\mathbf{Av} = \mathbf{f}$ и любом Δ имеет место $\min\limits_{\mathbf{u}} \Phi(\mathbf{u})$. Докажем, что верно и обратное утверждение. Если элемент доставляет минимальное значение функционалу энергии, то он является решением системы линейных уравнений $\mathbf{Av} = \mathbf{f}$. Из курса математического анализа известно, что в точке минимума должно выполняться условие $\operatorname{grad}\Phi(\mathbf{u}) = 0$, $\mathbf{A} > 0$. Вычисляя градиент, приходим к условию минимума функционала $\operatorname{grad}\Phi(\mathbf{u}) = 2\mathbf{Au} - 2\mathbf{f} = 0$. Таким образом установлена эквивалентность вариационной задачи (отыскание элемента, придающего минимум $\Phi(\mathbf{u})$) и задачи о нахождении решения СЛАУ. ■

Заметим, что СЛАУ с самосопряженным и положительно определенным оператором $\mathbf{A}$ представляют собой важный класс задач в математической физике, в частности, они возникают при решении краевых задач для эллиптических уравнений. При необходимости можно произвести симметризацию по Гауссу исходной системы.

2.6.2. Методы градиентного и наискорейшего спуска

Метод градиентного спуска состоит в нахождении следующего приближения в итерационном процессе из предыдущего, путем смещения в направлении градиента функционала

$$\Phi(\mathbf{u}) = (\mathbf{Au}, \mathbf{u}) - 2(\mathbf{f}, \mathbf{u}) \tag{2.25}$$

$$\mathbf{u}_{k+1} = \mathbf{u}_k - \alpha_k \cdot \operatorname{grad} \Phi(\mathbf{u}_k), \tag{2.26}$$

где $\mathbf{A}$ — положительно определенная симметричная матрица; α_k — параметр, определяемый из заданных условий; например, из условия минимума величины

$$\Phi\left[\mathbf{u}_k - \alpha_k \cdot \operatorname{grad} \Phi(\mathbf{u}_k)\right].$$

В этом случае итерационный метод называется методом наискорейшего спуска. Так как $\operatorname{grad} \Phi(\mathbf{u}) = 2(\mathbf{Au} - \mathbf{f})$, то (2.26) приобретает вид

$$\mathbf{u}_{k+1} = \mathbf{u}_k - \tau_k(\mathbf{Au}_k - \mathbf{f}), \quad \text{где} \quad \tau_k = 2\alpha_k, \tag{2.27}$$

что соответствует записи итерационного метода в форме (2.17). Здесь τ_k является итерационным параметром, который в методе наискорейшего спуска определяется из условия минимума функции $\Phi(\tau_k, \mathbf{u}_{k+1})$ по τ_k. Найдем условие этого минимума:

$$0 = 2(\mathbf{Au}_{k+1} - \mathbf{f}, (\mathbf{u}_{k+1})'_{\tau_k}) = -2(\mathbf{Au}_{k+1} - \mathbf{f}, \mathbf{Au}_k - \mathbf{f}).$$

Здесь учтено соотношение: $(\mathbf{Au}', \mathbf{u}) = (\mathbf{u}', \mathbf{A}^*\mathbf{u}) = (\mathbf{Au}, \mathbf{u}')$, поскольку $\mathbf{A} = \mathbf{A}^*$ и $(\mathbf{v}, \mathbf{Aw}) = (\mathbf{Aw}, \mathbf{v})$ в силу самосопряженности оператора $\mathbf{A}$. Подставим в последние равенства $\mathbf{u}_{k+1}$ из (2.27), получим $(\mathbf{Au}_k - \mathbf{f} - \tau_k\mathbf{A}(\mathbf{Au}_k - \mathbf{f}), \mathbf{Au}_k - \mathbf{f}) = 0$, откуда следует

$$(\mathbf{Au}_k - \mathbf{f}, \mathbf{Au}_k - \mathbf{f}) - \tau_k(\mathbf{A}(\mathbf{Au}_k - \mathbf{f}), \mathbf{Au}_k - \mathbf{f}) = 0,$$

$\tau_k = (\mathbf{Au}_k - \mathbf{f}, \mathbf{Au}_k - \mathbf{f})/(\mathbf{A}(\mathbf{Au}_k - \mathbf{f}), \mathbf{Au}_k - \mathbf{f})$, или $\tau_k = \frac{(\mathbf{r}_k, \mathbf{r}_k)}{(\mathbf{Ar}_k, \mathbf{r}_k)}$. где $\mathbf{r}_k = \mathbf{Au}_k - \mathbf{f}$.

Вектор $\mathbf{r}_k$ называют *вектором невязки.*

2.6.3. Метод минимальных невязок

Этот итерационный метод определяется следующим образом. Пусть $\mathbf{u}_{k+1} = \mathbf{u}_k - \tau_k\mathbf{r}_k$, как и ранее, $\mathbf{r}_k = \mathbf{Au}_k - \mathbf{f}$. Итерационный параметр τ_k на каждой итерации выбирается так, чтобы минимизировать евклидову норму невязки $\mathbf{r}_{k+1}$. Заметим, что итерационный процесс $\mathbf{u}_{n+1} = \mathbf{u}_n + \tau_n\mathbf{r}_n$ может быть представлен в равносильном виде в терминах невязки

$\mathbf{r}_{n+1} = \mathbf{r}_n + \tau_n \mathbf{A}\mathbf{r}_n$. Тогда для квадрата евклидовой (третьей) нормы невязки получаем условие

$$(\mathbf{r}_{k+1}, \mathbf{r}_{k+1}) = (\mathbf{r}_k, \mathbf{r}_k) - 2\tau_k(\mathbf{A}\mathbf{r}_k, \mathbf{r}_k) + \tau_k^2(\mathbf{A}\mathbf{r}_k, \mathbf{A}\mathbf{r}_k).$$

Для отыскания минимума невязки на следующей итерации приравняем нулю производную последнего выражения по итерационному параметру τ_k. Получим равенство

$$-2(\mathbf{A}\mathbf{r}_k, \mathbf{r}_k) + 2\tau_k(\mathbf{A}\mathbf{r}_k, \mathbf{A}\mathbf{r}_k) = 0.$$

Из последнего соотношения находим значение итерационного параметра

$$\tau_k = \frac{(\mathbf{A}\mathbf{r}_k, \mathbf{r}_k)}{(\mathbf{A}\mathbf{r}_k, \mathbf{A}\mathbf{r}_k)}.$$

2.6.4. Метод сопряженных градиентов

Этот метод применяется для решения систем уравнений с самосопряженной положительной матрицей $\mathbf{A} = \mathbf{A}^* > 0$. Оптимизируем градиентный метод, выбирая параметры τ таким образом, чтобы на последующем шаге невязка была ортогональна всем предыдущим. На первом шаге невязку ищем аналогично методу наискорейшего спуска. Получим невязки, образующие ортогональный базис. На последнем шаге невязка становится равна нулю, так как пространство конечномерно, и единственный элемент, ортогональный всем базисным векторам конечномерного пространства — нулевой. Получаем точное решение за конечное число шагов (прямой метод). Однако этот метод работает не всегда, так как при плохой обусловленности матрицы он становится вычислительно неустойчивым.

Идея метода состоит в следующем. Выбираем произвольное начальное приближение и вычисляем по нему вектор невязки $\mathbf{r}^0 = \mathbf{A}\mathbf{u}^0 - \mathbf{f}$, тогда первое приближение $\mathbf{u}^1 = \mathbf{u}^0 + \tau_0\mathbf{r}^0$.

Из условия ортогональности невязок на двух первых шагах находим значение итерационного параметра $\tau_0 = -\frac{(\mathbf{r}^0, \mathbf{r}^0)}{(\mathbf{A}\mathbf{r}^0, \mathbf{r}^0)}$.

Построим такое приближение, чтобы учитывались две предыдущие — *трехслойный итерационный метод*. Фактически, при построении его применяется процесс ортогонализации Грамма-Шмидта. Если $\left(\mathbf{A}\mathbf{r}^{n-2}, \mathbf{r}^{n-1}\right) = 0$, то $\mathbf{r}^{n-2}, \mathbf{r}^{n-1}$ — $\mathbf{A}$-*сопряженные невязки*.

Имеем *метод сопряженных градиентов*.

Все невязки уменьшаются по норме, поэтому данный метод эффективен даже для плохо обусловленных задач, т.к. на определенном шаге можно оборвать вычисления и получить приближение к решению с заданной точностью. Тогда этот метод становится итерационным.

Приведем последовательность расчетных формул одного из вариантов метода сопряженных градиентов.

$$\mathbf{u}_1 = (\mathbf{E} - \tau_1 \mathbf{A})\mathbf{u}_0 + \tau_1 \mathbf{f},$$

$$\dots$$

$$\mathbf{u}_{k+1} = \alpha_{k+1}(\mathbf{E} - \tau_{k+1}\mathbf{A})\mathbf{u}_k + (1 - \alpha_{k+1})\mathbf{u}_{k-1} + \alpha_{k+1}\tau_{k+1}\mathbf{f}, \qquad (2.28)$$

где

$$\tau_{k+1} = \frac{(\mathbf{r}_k, \mathbf{r}_k)}{(\mathbf{A}\mathbf{r}_k, \mathbf{r}_k)},$$

$$\alpha_1 = 1; \alpha_{k+1} = \left[1 - \frac{1}{\alpha_k} \cdot \frac{\tau_{k+1}}{\tau_k} \cdot \frac{(\mathbf{r}_k, \mathbf{r}_k)}{(\mathbf{r}_{k-1}, \mathbf{r}_{k-1})}\right]^{-1}, \quad k = 1, 2, \dots$$

В вычислительной практике этот метод используется при умеренном числе обусловленности, больших n и неизвестных границах спектра матрицы $\mathbf{A}$, как итерационный метод, поскольку k_0 обычно достаточно большое число. Подробнее о методах сопряженных градиентов можно прочитать в [2].

2.7. О спектральных задачах

Спектральные задачи — вычислительно наиболее трудоемкие задачи в прикладной линейной алгебре. Различают полную и частичную проблемы собственных значений. В первом случае необходимо отыскать ВСЕ собственные числа матрицы, во втором — лишь максимальное по абсолютной величине собственное число. Различают также самосопряженную спектральную задачу и задачу для произвольной матрицы. Очевидно, самосопряженная проблема решается проще — спектр самосопряженной матрицы всегда действительный.

Рассмотрим два алгоритма для самосопряженных матриц. Первый — *степенной* алгоритм, для вычисления наибольшего по абсолютной величине собственного числа. Выбираем произвольный ненулевой вектор $\mathbf{u}_0$ и строим последовательность векторов

$$\mathbf{u}_{k+1} = \mathbf{A}\mathbf{u}_k.$$

Легко показать, что выражение

$$\lambda \approx \frac{(\mathbf{A}\mathbf{u}_k, \mathbf{u}_k)}{(\mathbf{u}_k, \mathbf{u}_k)} = \frac{(\mathbf{u}_{k+1}, \mathbf{u}_k)}{(\mathbf{u}_k, \mathbf{u}_k)}$$

приближает максимальное по абсолютной величине собственное значение с точностью $O(\lambda_N/\lambda_{N-1})^k$. Здесь λ_N/λ_{N-1} — отношение самого

большого по модулю собственного числа матрицы к следующему по абсолютной величине.

Для решения полной самосопряженной проблемы собственных значений применяется *метод вращений.*

Определение собственных значений самосопряженной матрицы **A** эквивалентно отысканию такой ортогональной матрицы **T**, что

$$\Lambda = \mathbf{T}'\mathbf{A}\mathbf{T},$$

матрица Λ — диагональная. Среди всех ортогональных преобразований данное минимизирует сумму квадратов внедиагональных элементов исходной матрицы. Построим итерационный метод, минимизирующий эту сумму на каждой итерации. Пусть каждое преобразование подобия на каждой итерации содержит лишь одну матрицу вращения $\hat{\mathbf{A}} = \mathbf{T}'_{ij}\mathbf{A}\mathbf{T}_{ij}$, где матрица $\mathbf{T}_{ij}$ есть матрица поворота в плоскости $\mathbf{u}_i\mathbf{u}_j$ на угол α. Эта матрица отличается от матрицы **A** только двумя строками и двумя столбцами (с номерами i и j). Так как евклидова норма матрицы не изменяется при ортогональных преобразованиях, то легко получить соотношение между суммами квадратов внедиагональных элементов старой и новой матриц:

$$\sum_{i\neq j}\hat{a}_{ij}^2 = \sum_{i\neq j}a_{ij}^2 - 2a_{ij}^2 + \frac{1}{2}\left((a_{jj}-a_{ii})\sin 2\alpha + 2a_{ij}\cos 2\alpha\right)^2.$$

Очевидны условия минимизации суммы в левой части последнего равенства. Следует на текущей итерации выбирать индексы так, чтобы выполнялось условие $|a_{ij}| = \max\limits_{k\neq l}|a_{kl}|$, а угол поворота выбирается из условия $0 = \left((a_{jj}-a_{ii})\sin 2\alpha + 2a_{ij}\cos 2\alpha\right)^2$. Тогда он удовлетворяет условию $\operatorname{tg} 2\alpha = \frac{2a_{ij}}{a_{ii}-a_{jj}}$, $|\alpha| \leqslant \frac{\pi}{4}$.

Независимо от наличия кратных собственных значений метод вращений обладает квадратичной сходимостью.

Выбор максимального по модулю внедиагонального элемента — затратная операция, поэтому часто реализуется метод вращений с барьерами. Его идея состоит в следующем. При переборе внедиагональных значений вращение производится тогда, когда значение элемента по абсолютной величине превосходит некоторую величину (барьер). Если все элементы меньше барьера, его значение уменьшается, например, на порядок, и снова начинается циклический перебор внедиагональных элементов. Подробнее о методе вращений смотри в [1].

Другие алгоритмы решения спектральных задач описаны в специальной литературе [1, 3, 10].

Задачи

1. Методом Гаусса решить систему линейных уравнений $\mathbf{Ax} = \mathbf{f}$, где

$$\mathbf{A} = \begin{pmatrix} 5 & 0 & 1 \\ 2 & 6 & -2 \\ -3 & 2 & 10 \end{pmatrix}, \mathbf{f} = (11, 8, 6)^{\mathbf{T}}$$

или

$$5x_1 + 0 \cdot x_2 + x_3 = 11,$$

$$2x_1 + 6 \cdot x_2 + 2x_3 = 8,$$

$$-3x_1 + 2 \cdot x_2 + 10x_3 = 6.$$

Решение. Расширенная матрица $\tilde{\mathbf{A}}$ имеет вид

$$\tilde{\mathbf{A}} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & f_1 \\ a_{21} & a_{22} & a_{23} & f_2 \\ a_{31} & a_{32} & a_{33} & f_3 \end{pmatrix} = \begin{pmatrix} 5 & 0 & 1 & 11 \\ 2 & 6 & -2 & 8 \\ -3 & 2 & 10 & 6 \end{pmatrix}$$

Разделив элементы первой строки на ведущий элемент $a_{11} = 5$ получаем первую опорную строку $(1, 0, 0.2, 2.2)$.

Далее умножим ее на $a_{21} = 2$ и вычтем из второй строки, после чего умножим опорную строку на $a_{31} = 3$ и вычтем из третьей. Получаем матрицу $\tilde{\mathbf{A}}_1$

$$\tilde{\mathbf{A}} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & f_1 \\ 0 & \tilde{a}_{22} & \tilde{a}_{23} & \tilde{f}_2 \\ 0 & \tilde{a}_{32} & \tilde{a}_{33} & \tilde{f}_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0.2 & 2.2 \\ 0 & 6 & -2.4 & 3.6 \\ 0 & 2 & 10.6 & 12.6 \end{pmatrix}$$

Вторая опорная строка — результат деления второй строки матрицы $\tilde{\mathbf{A}}_1$ на

$$a_{22}^1 = 6:$$

$$(0,\ 1,\ -0.4,\ 0.6)$$

Матрицы $\tilde{\mathbf{A}}_2$ после умножения опорной строки на 2 и вычитания ее из третьей

$$\tilde{\mathbf{A}} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & f_1 \\ 0 & \tilde{a}_{22} & \tilde{a}_{23} & \tilde{f}_2 \\ 0 & 0 & \tilde{\tilde{a}}_{33} & \tilde{\tilde{f}}_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0.2 & 2.2 \\ 0 & 6 & -0.4 & 0.6 \\ 0 & 0 & 11.4 & 11.4 \end{pmatrix}$$

Третья опорная строка (результат деления третьей строки на 11,4) есть $(0, 0, 1, 1)$.

Матрица $\tilde{\mathbf{A}}_3$ будет

$$\tilde{\mathbf{A}}_1 = \begin{pmatrix} 1 & 0 & 0,2 & 2,2 \\ 0 & 1 & -0,4 & 0,6 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

Обратный ход метода Гаусса.

Система уравнений с матрицей $\tilde{\mathbf{A}}_3$

$$x_1 + 0 \cdot x_2 + 0,2x_3 = 2,2,$$

$$x_2 - 0,4x_3 = 0,6,$$

$$x_3 = 1.$$

Разрешая эту систему, начиная с последнего уравнения, получим

$$x_3 = 1, \quad x_2 = 1, \quad x_1 = 2.$$

Определитель матрицы $\det \mathbf{A}$ можно вычислить как произведение ведущих элементов

$$\det \mathbf{A} = 5 \cdot 6 \cdot 11,4 = 342.$$

2. Показать, что решение системы линейных уравнений

$$x_1 + 2x_2 + 3x_3 + 4x_4 = 2,$$

$$x_1 + 3x_2 + x_3 + 2x_4 = -1,$$

$$2x_1 + 3x_2 + 8x_3 + 7x_4 = 10,$$

$$2x_1 + 5x_2 + 3x_3 + 7x_4 = 3$$

методом Гаусса невозможно.

Решение. Преобразование матрицы рассматриваемой системы будет

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 2 \\ 1 & 3 & 1 & 2 & -1 \\ 2 & 3 & 8 & 7 & 10 \\ 2 & 5 & 3 & 7 & 3 \end{pmatrix} \to \begin{pmatrix} 1 & 2 & 3 & 4 & 2 \\ 0 & 1 & -2 & -2 & -3 \\ 0 & -1 & 2 & -1 & 6 \\ 0 & 1 & -3 & -1 & 1 \end{pmatrix} \to \begin{pmatrix} 1 & 2 & 3 & 4 & 2 \\ 0 & 1 & -2 & -2 & -3 \\ 0 & 0 & 0 & -3 & -3 \\ 0 & 0 & -1 & 1 & 2 \end{pmatrix}$$

Поскольку $a_{33}^2 = 0$ то вычисление третьей опорной строки невозможно.

3. Показать, что

$$\mu(\mathbf{A}) \geqslant \frac{\left|\max\limits_i \lambda_i(\mathbf{A})\right|}{\left|\min\limits_i \lambda_i(\mathbf{A})\right|}.$$

Рассмотреть случай симметричной матрицы $\mathbf{A}$.

Решение. Для собственного вектора ω, соответствующего наибольшему по модулю собственному значению матрицы, выполняется равенство $\mathbf{A}\omega = \lambda\omega$, откуда

$$\|\mathbf{A}\omega\| = \left|\max_i \lambda_i\right| \|\omega\|.$$

Учитывая, что $\|\mathbf{A}\omega\| \leqslant \|\mathbf{A}\|\,\|\omega\|$, получим $\|\mathbf{A}\| \geqslant \left|\max\limits_i \lambda_i(\mathbf{A})\right|$. Для обратной матрицы $\mathbf{A}^{-1}$ максимальным по модулю является собственное число $\min\limits_i \lambda_i^{-1}$, откуда $\left\|\mathbf{A}^{-1}\right\| \geqslant \left|\min\limits_i \lambda_i(\mathbf{A})\right|^{-1}$. Объединяя два последних неравенства, получим

$$\mu(\mathbf{A}) = \left\|\mathbf{A}^{-1}\right\| \|\mathbf{A}\| \geqslant \frac{\left|\max\limits_i \lambda_i(\mathbf{A})\right|}{\left|\min\limits_i \lambda_i(\mathbf{A})\right|}.$$

В случае симметричной матрицы $\mathbf{A}(\mathbf{A}^* = \mathbf{A})$ имеем

$$\|\mathbf{A}\|_3 = \sqrt{\lambda_{\max}(\mathbf{A}^*\mathbf{A})} = \sqrt{\lambda_{\max}(\mathbf{A}^2)} = \sqrt{\lambda^2_{\max}(\mathbf{A})} = |\lambda_{\max}(\mathbf{A})|,$$

т. к. из $\mathbf{A}\omega_i = \lambda_i\omega_i$ следует $(\mathbf{A}\omega_i)^2 = \lambda_i^2\omega_i^2$. Аналогично

$$\left\|\mathbf{A}^{-1}\right\|_3 = \sqrt{\lambda_{\max}[(\mathbf{A}^{-1})^*(\mathbf{A}^{-1})]} = \sqrt{\lambda_{\max}[(\mathbf{A}^{-1})]^2} =$$

$$= \sqrt{\lambda^{-1}_{\min}(\mathbf{A}^2)} = \left|\lambda^{-1}_{\min}(\mathbf{A})\right|.$$

Тогда $\mu(\mathbf{A}) = \|\mathbf{A}\|_3 \left\|\mathbf{A}^{-1}\right\|_3 = |\lambda_{\max}(\mathbf{A})| \,/\, |\lambda_{\min}(\mathbf{A})|$.

4. Найти число обусловленности матрицы $\mathbf{A}$, выразив его через число обусловленности матрицы $\mathbf{B}$, если $\mathbf{A} = \mathbf{B}^*\mathbf{B} > 0$.

Решение. Для самосопряженной положительной матрицы $\mathbf{A}$ имеем

$$\|\mathbf{A}\| = \sup_{\|\mathbf{u}\|\neq 0} \frac{(\mathbf{u}, \mathbf{A}\mathbf{u})}{(\mathbf{u}, \mathbf{u})}.$$

Тогда

$$\mu(\mathbf{A}) = \|\mathbf{A}\|_3 \left\|\mathbf{A}^{-1}\right\|_3 = \sup_{\|\mathbf{u}\|\neq 0} \frac{(\mathbf{u}, \mathbf{A}\mathbf{u})}{(\mathbf{u}, \mathbf{u})} \sup_{\|\mathbf{u}\|\neq 0} \frac{(\mathbf{u}, \mathbf{A}^{-1}\mathbf{u})}{(\mathbf{u}, \mathbf{u})} =$$

$$= \sup_{\|\mathbf{u}\|\neq 0} \frac{(\mathbf{B}\mathbf{u}, \mathbf{B}\mathbf{u})}{(\mathbf{u}, \mathbf{u})} \sup_{\|\mathbf{u}\|\neq 0} \frac{(\mathbf{B}^{-1}\mathbf{u}, \mathbf{B}^{-1}\mathbf{u})}{(\mathbf{u}, \mathbf{u})} = \|\mathbf{B}\|_3^2 \left\|\mathbf{B}^{-1}\right\|_3^2,$$

откуда

$$\mu(\mathbf{A}) = \mu^2(\mathbf{B}).$$

5. Показать, что норма матрицы

$$\|\mathbf{A}\|_2 = \max_{1\leqslant j\leqslant n} \sum_{j=1}^{n} |a_{ij}|$$

согласована с нормой вектора

$$\|\mathbf{u}\|_2 = \sum_{i=1}^{n} |u_i|.$$

Решение.

$$\|\mathbf{A}\mathbf{u}\|_2 = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij} u_j \right| \leqslant \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}| \cdot |u_j| \leqslant \sum_{j=1}^{n} |u_j| \cdot \sum_{i=1}^{n} |a_{ij}| \leqslant$$

$$\leqslant \left(\max_{1\leqslant i\leqslant n} \sum_{j} |a_{ij}|\right) \cdot \|\mathbf{u}\|_2 = \|\mathbf{A}\|_2 \|\mathbf{u}\|_2 .$$

Положим

$$\max \sum_{i=1}^{n} |a_{ij}| = \sum_{i=1}^{n} |a_{ik}|.$$

Покажем, что существует вектор $\mathbf{v}$, для которого достигается равенство. В качестве такового можно взять вектор $\mathbf{v}$ с компонентами $v_i = 0, i \neq k, v_k = 1$.

Таким образом, норма матрицы $\|\mathbf{A}\|_2 = \max\limits_{1\leqslant j\leqslant n} \sum |u_{ij}|$ согласована с нормой вектора $\|\mathbf{u}\|_2 = \sum\limits_{i=1}^{n} |u_i|$.

6. Дана жорданова клетка порядка n

$$\mathbf{A}=\begin{pmatrix} 1 & d & 0 & \dots & \dots & 0 & 0 \\ 0 & 1 & d & \dots & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & d & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \dots & \dots & 1 & d \\ 0 & 0 & \dots & \dots & \dots & \dots & 1 \end{pmatrix}.$$

Найти $\mu(\mathbf{A})$ и оценить возмущение в компоненте u_1 решения системы $\mathbf{Au}=\mathbf{f}$, если компонент f_n вектора f возмущен на величину ε.

Решение. Из $\mathbf{Au}=\mathbf{f}$ следует, что $\mathbf{u}=\mathbf{A}^{-1}\mathbf{f}$.

С помощью обратной подстановки $u_n=1, u_{n-1}=\dots$ находим компоненты матрицы

$$\mathbf{A}^{-1}=\begin{pmatrix} 1 & -d & d^2 & \dots & \dots & (-d)^{n-2} & (-d)^{n-1} \\ 0 & 1 & -d & \dots & \dots & (-d)^{n-3} & (-d)^{n-2} \\ 0 & 0 & 1 & \dots & \dots & (-d)^{n-4} & (-d)^{n-3} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \dots & 1 & -d \\ 0 & 0 & 0 & \dots & \dots & 0 & 1 \end{pmatrix}.$$

В этом случае $\|\mathbf{A}\|_1=1+|d|$, $\left\|\mathbf{A}^{-1}\right\|_1=1+|d|+d^2+\dots+|d|^{n-1}=\frac{|d|^n-1}{|d|-1}$;

Видно, что при $|d|>1$ матрица $\mathbf{A}$ плохо обусловлена, при $|d|<1$ - хорошо. При $n=20$ и $d=5$ имеем $\mu(\mathbf{A})\approx 10^{14}$.

Компонент $\tilde{u}_1$ решения возмущенной системы $\tilde{\mathbf{u}}=\mathbf{A}^{-1}\tilde{\mathbf{f}}$ будет

$$\tilde{u}_1=f_1-df_2+d^2f_3+\dots+(-d)^{n-2}f_{n-1}+(-d)^{n-1}(f_n+\varepsilon)=u_1+(-d)^{n-1}\varepsilon,$$

где u_1 — компонент решения невозмущенной системы $\mathbf{Au}=\mathbf{f}$

Отсюда видно, что при $|d|>1$ возмущение в n компоненте вектора $\mathbf{f}$ увеличивается в компоненте u_1 вектора $\mathbf{u}$ в $|d|^{n-1}$ раз, а при $|d|<1$ — в $|d|^{n-1}$ раз убывает.

7. Пусть в системе линейных уравнений

$$u_1+0,99u_2=f_1,$$

$$0,99u_1+u_2=f_2$$

вектор $\mathbf{f} = (f_1, f_2)^T$ получает приращение $\Delta\mathbf{f} = (\delta f_1, \delta f_2)^T$, а решение получает приращение $\Delta\mathbf{u} = (\delta u_1, \delta u_2)^T$, Найти наименьшее число μ, при котором независимо от $\mathbf{f}$ и $\Delta\mathbf{f}$ выполняется оценка

$$\frac{\|\Delta\mathbf{u}\|}{\|\mathbf{u}\|} \leqslant \mu \frac{\|\Delta\mathbf{f}\|}{\|\mathbf{f}\|}.$$

Решить задачу, используя нормы матриц $\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|_3$.

Решение. Для возмущенной задачи $\mathbf{A}(\mathbf{u}+\Delta\mathbf{u}) = \mathbf{f}+\Delta\mathbf{f}$, из линейности системы следует $\mathbf{A}\Delta\mathbf{u} = \Delta\mathbf{f}$. Для возмущения решения выполняется равенство $\Delta\mathbf{u} = \mathbf{A}^{-1}\Delta\mathbf{f}$. Тогда $\|\Delta\mathbf{u}\| \leqslant \|\mathbf{A}^{-1}\|\cdot\|\mathbf{A}\|\cdot\frac{\|\Delta\mathbf{f}\|}{\|\mathbf{f}\|}\cdot\frac{\|\mathbf{f}\|}{\|\mathbf{A}\|}$. Отсюда сразу следует $\|\Delta\mathbf{u}\| \leqslant \|\mathbf{A}^{-1}\| \cdot \|\mathbf{A}\| \frac{\|\Delta\mathbf{f}\|}{\|\mathbf{f}\|} \|\mathbf{u}\|$, так как $\frac{\|\mathbf{f}\|}{\|\mathbf{A}\|} \leqslant \|\Delta\mathbf{u}\|$.

Тогда искомая оценка будет $\frac{\|\Delta\mathbf{u}\|}{\|\mathbf{u}\|} \leqslant \|\mathbf{A}^{-1}\| \cdot \|\mathbf{A}\| \frac{\|\Delta\mathbf{f}\|}{\|\mathbf{f}\|}$. Обозначим $\mu = \|\mathbf{A}^{-1}\| \cdot \|\mathbf{A}\|$. В этом случае наименьшим числом, при котором выполняется оценка $\frac{\|\Delta\mathbf{u}\|}{\|\mathbf{u}\|} \leqslant \mu\frac{\|\Delta\mathbf{f}\|}{\|\mathbf{f}\|}$, является $\mu = \|\mathbf{A}^{-1}\| \cdot \|\mathbf{A}\|$. Это — число обусловленности системы уравнений. Численное решение в соответствующих нормах получается легко.

8. При заданном фиксированном $\mathbf{f}$ найти наименьшее число ν, при котором независимо от $\Delta\mathbf{f}$ выполняется оценка

$$\frac{\|\Delta\mathbf{u}\|}{\|\mathbf{u}\|} \leqslant \nu(f) \frac{\|\Delta\mathbf{f}\|}{\|\mathbf{f}\|}.$$

Найти такую правую часть системы $\mathbf{f}$, которой соответствует наименьшее ν, а также само это значение при использовании третьей нормы матрицы.

Решение.

По условию задачи $\nu \geqslant \frac{\|\mathbf{f}\|}{\|\mathbf{u}\|} \cdot \frac{\|\Delta\mathbf{u}\|}{\|\Delta\mathbf{f}\|}$. Рассмотрим, какие значения может принимать это число. Точная нижняя грань для такой оценки, очевидно, $\inf\limits_{\|\Delta\mathbf{f}\|\neq 0} \nu = \frac{\|\mathbf{f}\|}{\|\mathbf{u}\|} \cdot \sup\limits_{\|\Delta\mathbf{f}\|\neq 0} \frac{\|\mathbf{A}^{-1}\cdot\Delta\mathbf{f}\|}{\|\Delta\mathbf{f}\|}$. Так как надо найти оценку, не зависящую от начального возмущения (при решении конкретной задачи оно, очевидно, неизвестно), получим

$$\nu = \inf_{\|\Delta\mathbf{f}\|\neq 0} \nu = \|\mathbf{A}^{-1}\| \cdot \frac{\|\mathbf{f}\|}{\|\mathbf{u}\|}.$$

Для точной нижней грани выполнено $\inf\limits_{\mathbf{f}} \nu(\mathbf{f}) = \|\mathbf{A}^{-1}\| \times$ $\times \left(\sup \frac{\|\mathbf{A}^{-1}\mathbf{f}\|}{\|\mathbf{f}\|}\right)^{-1} = 1$. Можно оценить и точную верхнюю грань:

$$\sup_{\mathbf{f}} \nu = \|\mathbf{A}^{-1}\| \sup_{\mathbf{u}} \frac{\|\mathbf{A}\mathbf{u}\|}{\|\mathbf{u}\|} = \|\mathbf{A}^{-1}\| \cdot \|\mathbf{A}\| = \mu$$

Таким образом, $1 \leqslant \nu \leqslant \mu$.

Ответим на вопрос, при каких $\mathbf{f}$ достигается $\sup\limits_{\mathbf{f}} \nu$ и $\inf\limits_{\mathbf{f}} \nu$. Для этого используем разложение вектора правой части системы по базису из собственных векторов матрицы $\mathbf{A}^2$ (без ограничения общности полагаем, что такой базис существует). В этом базисе $\frac{\|\mathbf{A}\mathbf{u}\|}{\|\mathbf{u}\|} =$ $= \sqrt{\frac{(\mathbf{A}^*\mathbf{A}\mathbf{u},\mathbf{u})}{(\mathbf{u},\mathbf{u})}} - \sqrt{\frac{(\sum \lambda_i \xi_i \omega_i, \xi_i \omega_i)}{(\xi_i\omega_i, \xi_i\omega_i)}} = \sqrt{\frac{\sum \lambda_i \xi_i^2}{\sum \xi_i^2}}$, $\sqrt{\lambda_{\min}(\mathbf{A}^*\mathbf{A})} \leqslant$ $\leqslant \sqrt{\frac{\sum \lambda_i \xi_i^2}{\sum \xi_i^2}} = \frac{\|\mathbf{A}\mathbf{u}\|}{\|\mathbf{u}\|} \leqslant \sqrt{\lambda_{\max}(\mathbf{A}^*\mathbf{A})}$.

Для самосопряженной положительной матрицы $\mathbf{A}^* = \mathbf{A} > 0$ получаем

$$|\lambda_{\min}(\mathbf{A})| \leqslant \sqrt{\frac{\sum\limits_i \lambda_i \xi_i^2}{\sum\limits_i \xi_i^2}} = \frac{\|\mathbf{A}\mathbf{u}\|}{\|\mathbf{u}\|} \leqslant |\lambda_{\max}(\mathbf{A})|\,.$$

В то же время

$$\frac{\|\mathbf{A}\omega_i\|}{\|\omega_i\|} = \sqrt{\frac{(\mathbf{A} * \mathbf{A}\omega_i, \omega_i)}{(\omega_i, \omega_i)}} = \sqrt{\frac{\lambda_i(\omega_i, \omega_i)}{(\omega_i, \omega_i)}} = \sqrt{\lambda_i(\mathbf{A} * \mathbf{A})}\,,$$

для самосопряженной положительной матрицы $\frac{\|\mathbf{A}\omega_i\|}{\|\omega_i\|} = |\lambda_i(\mathbf{A})|$. $\sup\limits_{\omega} \frac{\|\mathbf{A}\omega\|}{\|\omega\|} = \sqrt{\lambda_{\max}(\mathbf{A}^*\mathbf{A})}$; $\inf\limits_{\omega} \frac{\|\mathbf{A}\omega\|}{\|\omega\|} = \sqrt{\lambda_{\min}(\mathbf{A}^*\mathbf{A})}$.

В случае $\mathbf{A}^* = \mathbf{A} > 0$ получаем $\sup\limits_{\omega} \frac{\|\mathbf{A}\omega\|}{\|\omega\|} = |\lambda_{\max}(\mathbf{A})|$ $\inf\limits_{\omega} \frac{\|\mathbf{A}\omega\|}{\|\omega\|} =$ $= |\lambda_{\min}(\mathbf{A})|$.

Таким образом,

$$\sup_{\mathbf{u}} \frac{\|\mathbf{A}\mathbf{u}\|}{\|\mathbf{u}\|} = \sqrt{\lambda_{\max}(\mathbf{A}^*\mathbf{A})}; \sup_{\omega} \frac{\|\mathbf{A}\omega\|}{\|\omega\|} = \sqrt{\lambda_{\max}(\mathbf{A}^*\mathbf{A})},$$

$$\inf_{\mathbf{u}} \frac{\|\mathbf{A}\mathbf{u}\|}{\|\mathbf{u}\|} = \sqrt{\lambda_{\min}(\mathbf{A}^*\mathbf{A})}; \inf_{\omega} \frac{\|\mathbf{A}\omega\|}{\|\omega\|} = \sqrt{\lambda_{\min}(\mathbf{A}^*\mathbf{A})}.$$

9. Выписать формулы итерационных методов Якоби, Зейделя, верхней релаксации для СЛАУ

$$\mathbf{Ax} = \mathbf{f}, \mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \mathbf{x} = (1, -1)^T, \mathbf{f} = (1, -1)^T$$

или

$$2u + v = 1,$$

$$u + 2v = -1.$$

Оценить количество итераций для метода Якоби.

(Решение системы: $u = 1; v = -1$).

Решение.

Итерационные методы Якоби, Зейделя, релаксации соответственно записываются

$$\begin{cases} u_{k+1} = -\dfrac{1}{2}v_k + \dfrac{1}{2}, \\ v_{k+1} = -\dfrac{1}{2}u_k - \dfrac{1}{2}, \end{cases}$$

$$(u_0, v_0) = (u_0^0, v_0^0),$$

или $\mathbf{x}_{k+1} = \mathbf{B}\mathbf{x}_k + \mathbf{f}$,

$$\mathbf{B} = \begin{pmatrix} 0 & -0,5 \\ -0,5 & 0 \end{pmatrix},$$

$$u_{k+1} = -\frac{1}{2}v_k + \frac{1}{2},$$

$$v_{k+1} = -\frac{1}{2}u_{k+1} - \frac{1}{2};$$

$$u_{k+1} = (1 - \tau)u_k + \frac{\tau}{2}(1 - v_k),$$

$$v_{k+1} = (1 - \tau)v_k - \frac{\tau}{2}(1 + u_k).$$

Оценка количества итераций проводится по формуле

$$k \approx \ln \frac{\varepsilon}{\varepsilon_0} / \ln \|\mathbf{B}\| = \ln 10^{-3} / \ln \frac{1}{2}.$$

10. Представить графическую интерпретацию итерационного метода Якоби для СЛАУ

$$a_{11}u + a_{12}v = f_1,$$

$$a_{21}u + a_{22}v = f_2,$$

$$a_{11} \neq 0, a_{22} \neq 0$$

Решение. Итерационный процесс Якоби записывается как

$$u_{k+1} = -\frac{a_{12}}{a_{11}}v_k + \frac{f_1}{a_{11}},$$

$$v_{k+1} = -\frac{a_{21}}{a_{22}}u_k + \frac{f_2}{a_{22}}.$$

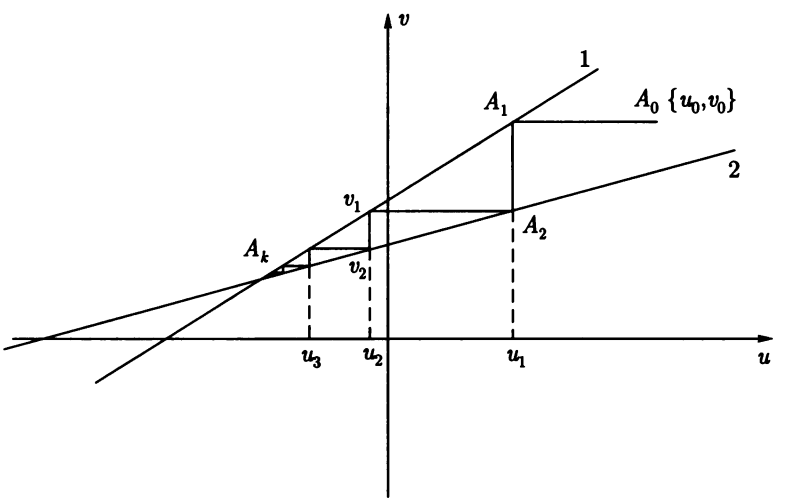


Рис. 2.1

Первое уравнение соответствует прямой 1, второе — прямой 2. Вычисление u_1 соответствует проведению отрезка, параллельного оси $0u$ и (при $v = v_0$) до пересечения с прямой 1; точка пересечения даст первое приближение u_1. Вычислению v_1 соответствует проведение из точки A_1 прямой, параллельной оси $0v$ до пересечения с прямой 2 и т. д. до сходимости итераций к точке пересечения прямых 1 и 2 (A_k) с заданной точностью.

11. При каких a, b сходится метод простой итерации $\mathbf{u}_{k+1} = \mathbf{B}\mathbf{u}_k + \mathbf{f}$, где

$$\mathbf{B} = \begin{pmatrix} a & b & 0 \\ b & a & b \\ 0 & b & a \end{pmatrix}.$$

Решение. Для того, чтобы метод простой итерации сходился к решению соответствующей СЛАУ, необходимо и достаточно, чтобы все собственные значения матрицы $\mathbf{B}$ по модулю были меньше единицы: $|\lambda_i| < 1$. Решаем характеристическое уравнение

$$det(\mathbf{B} - \lambda\mathbf{E}) = \begin{vmatrix} a-\lambda & b & 0 \\ b & a-\lambda & b \\ 0 & b & a-\lambda \end{vmatrix} =$$

$$= (a-\lambda)\begin{vmatrix} a-\lambda & b \\ b & a-\lambda \end{vmatrix} - b\begin{vmatrix} b & a \\ 0 & a-\lambda \end{vmatrix} =$$

$$= (a-\lambda)\left[(a-\lambda)^2 - b^2\right] - b^2(a-\lambda) =$$

$$= (a-\lambda)(a-\lambda-\sqrt{2b})(a-\lambda+\sqrt{2b}) = 0,$$

откуда получим условие сходимости итерационного метода

$$|a| < 1, \left|a \pm \sqrt{2b}\right| < 1.$$

12. Найти условие сходимости итерационных методов Якоби и Зейделя для СЛАУ $\mathbf{A}\mathbf{u} = \mathbf{f}$ с матрицей $\mathbf{A}$ вида

$$\mathbf{A} = \begin{pmatrix} a & b & 0 \\ b & a & b \\ 0 & b & a \end{pmatrix}.$$

Решение. Для метода Якоби

$$\mathbf{u}_{k+1} = \mathbf{B}\mathbf{u}_k + \mathbf{f}, \quad \text{где} \quad \mathbf{B} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U}).$$

Имеет место уравнение:

$\mathbf{B}\omega = \lambda\omega$, где λ и ω — собственное число и собственный вектор, соответственно. В таком случае $-\mathbf{D}(\mathbf{L} + \mathbf{U})\omega = \lambda\omega$, или: $(\mathbf{L} + \mathbf{U} + \lambda\mathbf{D})\,\omega = 0$, откуда (предполагаем наличие нетривиальных решений у последней СЛАУ):

$$\det(\mathbf{L} + \mathbf{U} + \lambda\mathbf{D}) = 0.$$

Решим это уравнение:

$$\det \begin{vmatrix} \lambda a & b & 0 \\ b & \lambda a & b \\ 0 & b & \lambda a \end{vmatrix} = 0,$$

откуда получим условия сходимости итерационного метода Якоби:

$$\left| \frac{b}{a} \right| < 2^{-\frac{1}{2}}.$$

Для метода Зейделя имеем

$$\cdot \mathbf{B} = -(\mathbf{D} + \mathbf{L})^{-1}\mathbf{U}, \quad \mathbf{B}\boldsymbol{\omega} = \lambda\boldsymbol{\omega}.$$

В таком случае

$$-(\mathbf{D} + \mathbf{L})^{-1}\mathbf{U}\boldsymbol{\omega} = \lambda\boldsymbol{\omega},$$

откуда следует уравнение

$$\det(\lambda\mathbf{L} + \lambda\mathbf{D} + \mathbf{U}) = 0,$$

Вычислив детерминант, придем к алгебраическому уравнению

$$\det \begin{vmatrix} \lambda a & b & 0 \\ \lambda b & \lambda a & b \\ 0 & \lambda b & \lambda a \end{vmatrix} = a\lambda^2(a^2\lambda - 2b^2) = 0.$$

В таком случае, поскольку $\lambda_{1,2} = 0, \lambda_3 = 2\frac{b^2}{a^2}$, получим условие сходимости метода Зейделя: $\frac{b}{a} < 2^{-\frac{1}{2}}$. Видно, что в данном случае условия сходимости для обоих методов совпадают.

Литература

[1] *Воеводин В.В.* Вычислительные основы линейной алгебры. М.: Наука, 1977. 303 с.

[2] *Голуб Дж., Ван Лоун Ч.* Матричные вычисления. М.: Мир, 1999. 548 с.

[3] *Деммель Дж.* Вычислительная линейная алгебра. Теория и приложения. М., Мир, 2001. 429 с.

[4] *Фадеев А.К., Фадеева В.Н.* Вычислительные методы линейной алгебры. СПб.: Лань, 2002. 736 с.

[5] *Воеводин В.В., Кузнецов Ю.А.* Матрицы и вычисления. М.: Наука, 1984. 320 с.

[6] *Беклемишев Д.В.* Курс аналитической геометрии и линейной алгебры. М.: Наука, 1980. 240 с.

[7] *Рябенький В.С.* Введение в вычислительную математику. М.: Физматлит, 2000. 294 с.

[8] *Бахвалов Н.В., Жидков Н.П., Кобельков Г.М.* Численные методы. М.: Лабо-ратория Базовых Знаний, 2002. 632 с.

[9] *Косарев В.И.* 12 лекций по вычислительной математике. М.: Изд-во МФТИ, Физматкнига, 2000. 220 с.

[10] *Коновалов А.Н.* Введение в вычислительные методы линейной алгебры. Новосибирск, Наука, 1993. 158 с.

[11] *Амосов А.А., Дубинский Ю.А., Копченова Н.В.* Вычислительные методы для инженеров. М.: Высшая школа, 1994. 544 с.

[12] *Вержбицкий В.М.* Численные методы. Линейная алгебра и нелинейные уравнения. М.: Высшая школа, 2000. 266 с.

Лекция 3. Численное решение переопределенных СЛАУ. Метод наименьших квадратов

В лекции рассматриваются методы решения переопределенных систем уравнений. Обсуждается вопрос о выборе базиса на погрешность результата. Вкратце описываются итерационные методы решения плохо обусловленных систем линейных уравнений.

Ключевые слова: переопределенная система, метод наименьших квадратов, обобщенный многочлен, конечный ряд Фурье.

3.1. Пример использования метода наименьших квадратов (МНК)

Приведем простой пример получения переопределенной системы линейных уравнений. Такого рода задачи часто встречаются, например, при обработке результатов экспериментов.

Пусть f — линейная (или близкая к линейной) функция аргумента x: $f(x) = u_1 x + u_0$. В точках x_k известны значения функции $f(x_k)$. Тогда u_0, u_1 — коэффициенты, которые необходимо подобрать так, чтобы выполнялись условия $u_1 x_k + u_0 = f_k$, $k = 0, 1, 2, 3, 4$, $f_k = f(x_k)$.

Получим систему пяти уравнений относительно двух неизвестных. Это — переопределенная система. Она не имеет классического решения, так как в общем случае не существует прямой, проходящей через все 5 точек (это возможно только тогда, когда какие-либо три уравнения полученной системы линейными преобразованиями сводятся к двум другим — система линейно зависима).

Рассмотрим общий случай. Пусть коэффициенты $\{u_0, u_1\}$ необходимо определить по результатам $n+1$ измерения. Введем функцию, равную сумме квадратов невязок $r_k = u_1 x_k + u_0 - f_k$

$$\Phi(u_1, u_0) = \sum_{k=0}^{n} r_k^2 = \sum_{k=0}^{n} (u_1 x_k + u_0 - f_k)^2. \tag{3.1}$$

Примем за обобщенное решение переопределенной СЛАУ такие $\{u_0, u_1\}$ для которых $\Phi(u_0, u_1)$ принимает наименьшие значение. Для определения обобщенного решения из условия минимума суммы квадратов невязки получаем систему двух уравнений, имеющую классическое

решение:

$$\frac{\partial \Phi}{\partial u_0} = 0, \frac{\partial \Phi}{\partial u_1} = 0.$$

Выбор функции $\Phi(u_0, u_1)$ имеет некоторый произвол. Например, возможно каждому измерению придать некоторый вес b_k. От набора таких весовых множителей зависело бы решение системы. В этом случае функция Φ будет

$$\Phi(u_0, u_1) = \sum_{k=0}^{n} b_k (u_1 x_k - u_0 - f_k)^2.$$

Если в качестве невязки выбрать $r_k = |u_1 x_k + u_0 - f_k|$, то получим задачу линейного программирования на отыскании минимума функции

$$\Phi(u_1 u_2) = \sum_{k=0}^{3} |u_1 x_k + u_0 - f_k|.$$

Получившийся таким образом функционал, вообще говоря, недифференцируем. Для решения задачи нельзя использовать метод наименьших квадратов.

Произвол имеется и в выборе базисных функций. Вообще говоря, можно было бы записать невязку r_k в виде

$$r_k = \sum_{j=0}^{p} u_j \varphi_j(x_k) - f(x_k), k = 1, \ldots, n,$$

где $\varphi_j(x)$ — некоторые функции, образующие базис, например, тригонометрические: $\varphi_j(x) = \sin(jx)$. Выражение $\sum_{j=0}^{p} u_j \varphi_j(x)$ называется *обобщенным полиномом*. В приведенном выше примере в качестве базисных функций были выбраны степенные функции $\varphi_j(x) = x^j$. Обобщенный полином превратился в алгебраический.

В случае выбора произвольной системы базисных функций переопределенная СЛАУ и функционал $\Phi(u_0, \ldots, u_p)$ будут

$$u_0 \varphi_0(x_0) + \ldots + u_p \varphi_p(x_0) = f_0,$$

$$\ldots$$

$$u_0 \varphi_0(x_n) + \ldots + u_p \varphi_p(x_n) = f_n,$$

$$\Phi(u_0, \ldots, u_n) = \sum_{i=0}^{n} (\sum_{j=0}^{p} u_j \varphi_j(x_i) - f_i)^2$$

Отыщем обобщенное решение методом наименьших квадратов. Приравнивая все частные производные по компонентам обобщенного решения к нулю $\frac{\partial \Phi}{\partial u_k} = 0$ (условия минимума) и изменяя порядок суммирования, получаем СЛАУ

$$\sum_{j=0}^{p} \left(\sum_{i=0}^{n} \varphi_j(x_i)\varphi_k(x_i) \right) u_j = \sum_{i=0}^{n} f_i \varphi_k(x_i), k = 0, \ldots, p,$$

или

$$(\varphi_0, \varphi_0)u_0 + (\varphi_0, \varphi_1)u_1 + \ldots + (\varphi_0, \varphi_p)u_p = (\varphi_0, f),$$
$$(\varphi_1, \varphi_0)u_0 + (\varphi_1, \varphi_1)u_1 + \ldots + (\varphi_1, \varphi_p)u_p = (\varphi_1, f),$$
$$(\varphi_p, \varphi_0)u_0 + (\varphi_p, \varphi_1)u_1 + \ldots + (\varphi_p, \varphi_p)u_p = (\varphi_p, f),$$

Система метода наименьших квадратов имеет вид $\mathbf{Du} = \mathbf{f}$ с матрицей $\mathbf{D}$, элементами которой являются скалярные произведения $(\varphi_i, \varphi_j) = \sum\limits_{i=0}^{n} \varphi_j(x)\varphi_k(x_i)$ Это — матрица Грамма. Ее свойства известны из курса линейной алгебры, эта матрица симметричная и положительно определенная. Таким образом, решение исследуемой СЛАУ существует и единственно. В правой части системы стоят проекции свободного члена исходной задачи на подпространство базисных функций $(\varphi, f) = \sum\limits_{i=0}^{n} \varphi_j(x_i) f_i$.

Здесь учтено, что $\frac{\partial \Phi}{\partial u_k} = 2 \sum\limits_{i=0}^{n} \varphi_k(x_i) \left(\sum\limits_{j=0}^{p} u_j \varphi_j(x_i) - f_i \right)$, или, в развернутом виде

$$\sum_{i=0}^{n} \varphi_0(x_i) \left(u_0 \varphi_0(x_i) + u_1 \varphi_0(x_i) + \ldots + u_p \varphi_p(x_i) - f(x_i) \right) = 0,$$

$$\sum_{i=0}^{n} \varphi_1(x_i) \left(u_0 \varphi_0(x_i) + u_1 \varphi_1(x_i) + \ldots + u_p \varphi_p(x_i) - f(x_i) \right) = 0,$$

$$\ldots,$$

$$\sum_{i=0}^{n} \varphi_s(x_i) \left(u_0 \varphi_0(x_i) + u_1 \varphi_1(x_i) + \ldots + u_p \varphi_p(x_i) - f(x_i) \right) = 0.$$

Часто выбирают $\varphi_k(x) = x^k$, в этом случае система уравнений принимает следующую форму:

$$\sum_{j=0}^{p} \left(\sum_{i=0}^{n} x_i^{j+k} \right) u_j = \sum_{i=0}^{n} f_i x_i^k, k = 1, \ldots, p.$$

Эта система может быть легко выписана в компонентах:

$$u_0 + \left(\sum_{i=0}^{n} x_i\right) u_1 + \ldots + \left(\sum_{i=0}^{n} x_i^p\right) u_p = \sum_{i=0}^{n} f(x_i),$$

$$\left(\sum_{i=0}^{n} x_i\right) u_0 + \left(\sum_{i=0}^{n} x_i^2\right) u_1 + \ldots + \left(\sum_{i=0}^{n} x_i^{p+1}\right) = \sum_{i=0}^{n} x_i\, f(x_i),$$

$$\ldots$$

$$\left(\sum_{i=0}^{n} x_i^p\right) u_0 + \left(\sum_{i=0}^{n} x_i^{p+1}\right) + \ldots + \left(\sum_{i=0}^{n} x_i^{2p+1}\right) u_p = \sum_{i=0}^{n} x_i^p f(x_i).$$

В случае использования ортонормированных систем базисных функций $\varphi_j(x)$, т. е., при выполнении условия $(\varphi_k, \varphi_j) = \delta_{kj}$ решение принимает простой вид $u_0 = \frac{(\varphi_0, f)}{(\varphi_0, \varphi_0)} = \|\varphi_0\|^{-2}(\varphi_0, f); u_1 = \|\varphi_1\|^{-2}(\varphi_1, f); \ldots u_p = \|\varphi_p\|^{-2}(\varphi_p, f)$. Поскольку система функции ортонормирована, то $\|\varphi_j\| = 1$ и $u_k = (\varphi_k, f), k = 1, \ldots, n$. Эти коэффициенты называются *коэффициентами Фурье*, а обобщенный многочлен с этими коэффициентами — *обобщенным многочленом Фурье*. В частности, в качестве базисных можно использовать ортогональную систему функций на $[-\pi, \pi]$: $\{\sin kx, \cos kx\}$ $k = 1, \ldots, n$. Такие представления называют отрезками тригонометрических рядов Фурье или конечными рядами Фурье.

Докажем теорему о методе наименьших квадратов, обобщающую изложенную информацию.

Запишем переопределенную СЛАУ

$$a_{11}u_1 + \ldots + a_{1p}u_p = f_1,$$

$$\ldots$$

$$a_{n1}u_1 + \ldots + a_{np}u_p = f_n, n > p. \tag{3.2}$$

$$\mathbf{u} = \{u_1, \ldots, u_p\}^{\mathrm{T}} \in L^p, \mathbf{f} = \{f_1, \ldots, f_n\}^{\mathrm{T}} \in L^n,$$

где линейные нормированные пространства L_p и L_n имеют размерности p и n соответственно. Перепишем (3.2) в матричной форме:

$$\mathbf{Au} = \mathbf{f},$$

где $\mathbf{A} = \begin{pmatrix} a_{11} & \ldots & a_{1p} \\ \ldots & \ldots & \ldots \\ a_{n1} & \ldots & a_{np} \end{pmatrix}$.

Наряду с основным скалярным умножением в L_n

$$(\mathbf{x},\mathbf{y})^n = \sum_{k=1}^{n} x_k y_k \tag{3.3}$$

введем скалярное умножение с весовой матрицей $\mathbf{B}$:

$$[\mathbf{x},\mathbf{y}]^n = (\mathbf{Bx},\mathbf{y})^n, \mathbf{B} = \mathbf{B}^* > 0, \mathbf{x},\ \mathbf{y} \in L^n \tag{3.4}$$

Оба этих умножения удовлетворяют аксиомам скалярного умножения элементов линейного пространства. Матрица $\mathbf{B}$ является весовой и определяет вклад невязки каждого слагаемого суммы (3.1). Система (3.2) не имеет классического решения. Определим обобщенное решение этой системы как элемент линейного пространства $\mathbf{v}$, придающий наименьшее значение квадратичной форме:

$$\Phi(\mathbf{u}) = [\mathbf{Au} - \mathbf{f}, \mathbf{Au} - \mathbf{f}]^n.$$

Теорема. *Пусть столбцы матрицы* $\mathbf{A}$ *линейно независимы, т.е. ранг* $\mathbf{A}$ *равен* p. *Тогда существует единственный элемент* p*-мерного евклидова пространства* $\mathbf{v} \in L^P$, *являющийся обобщенным решением системы (3.2), и решением СЛАУ*

$$\mathbf{A^*BAu} = \mathbf{A^*Bf}, \tag{3.5}$$

состоящей из p *скалярных уравнений относительно неизвестных* $\{\mathbf{u}_k\}_{k=1}^{k=p}$, *доставляющий минимум квадратичной форме*

$$\Phi(\mathbf{u}) = [\mathbf{Au} - \mathbf{f}, \mathbf{Au} - \mathbf{f}]^n$$

Доказательство.

Покажем, что решение СЛАУ $\mathbf{A^*BAu} = \mathbf{A^*Bf}$ существует и единственно. Введем обозначение $\mathbf{q}_k \in L^n$ для вектора — k столбца матрицы системы $\mathbf{A}$:

$$\mathbf{q}_k = \{a_{1k}, \ldots, a_{nk}\}^{\mathrm{T}}; k = 1, \ldots, p.$$

Несложно показать, что матрица $\mathbf{D} = \mathbf{A^*BA}$ системы (3.5) есть квадратная матрица $p \times p$. Элемент d_{ij} этой матрицы, стоящий на пересечении i строки и j столбца, есть $d_{ij} = (\mathbf{q}_i, \mathbf{Bq}_j)^n = (\mathbf{Bq}_i, \mathbf{q}_j)^n = [\mathbf{q}_i, \mathbf{q}_j]^n$, в силу коммутативности скалярного произведения $d_{ij} = d_{ji}$, что означает самосопряженность матрицы $\mathbf{D}$: $\mathbf{D} = \mathbf{D}^*$.

Покажем, что матрица $\mathbf{D}$ невырождена и положительно определена. Напомним, что $(\mathbf{f}, \mathbf{A}\xi)^n = (\mathbf{A^*f}, \xi)^P; \mathbf{f}_1 \in L^n,\ \xi \in L^P$. Это равенство проверяется непосредственно, если записать обе его части в развернутом

виде. Невырожденность матрицы $\mathbf{D}$ следует из того, что ранг матрицы $\mathbf{A}$ равен p.

В таком случае $0 < [\mathbf{Ax}, \mathbf{x}]^n = (\mathbf{BAx}, \mathbf{Ax})^n = (\mathbf{A^*BAx}, \mathbf{x})^p = (\mathbf{Dx}, \mathbf{x})^p$. Поскольку $\mathbf{D}$ невырождена и положительно определена, то (3.5) имеет единственное решение $\mathbf{v} \in L^p$. Теперь покажем, что $\mathbf{v}$ — единственное обобщенное решение системы. Для любого вектора $\Delta \neq 0$ выполнено $\Phi(\mathbf{v} + \Delta) > \Phi(\mathbf{v})$:

$$\Phi(\mathbf{v} + \Delta) = [\mathbf{A}(\mathbf{v} + \Delta) - \mathbf{f}, \mathbf{A}(\mathbf{v} + \Delta) - \mathbf{f}]^n =$$

$$= [\mathbf{Av} - \mathbf{f}, \mathbf{Av} - \mathbf{f}]^n - 2[\mathbf{Av} - \mathbf{f}, \mathbf{A}\Delta]^n + [\mathbf{A}\Delta, \mathbf{A}\Delta]^n =$$

$$= \Phi(\mathbf{v}) + 2[\mathbf{Av} - \mathbf{f}, \mathbf{A}\Delta]^n + (\mathbf{D}\Delta, \Delta)^p = \Phi(\mathbf{v}) + (\mathbf{D}\Delta, \Delta)^p > \Phi(\mathbf{v}),$$

что и требовалось доказать. ■

При доказательстве использовалось

$$[\mathbf{Av} - \mathbf{f}, \mathbf{A}\Delta]^n = (\mathbf{B}(\mathbf{Av} - \mathbf{f}), \mathbf{A}\Delta)^n = (\mathbf{A^*BAv} - \mathbf{A^*Bf}, \Delta)^p = 0,$$

поскольку $\mathbf{A^*BAv} = \mathbf{A^*Bf}$.

Так как матрица $\mathbf{D} = \mathbf{A^*BA}$ — симметричная и положительно определенная, то для численного решения полученной СЛАУ можно воспользоваться итерационными методами.

Если система векторов $\{\mathbf{q}_k\}_{k=1}^p$ оказывается ортонормированной, т. е. $[\mathbf{q}_i, \mathbf{q}_j] = \delta_{ij} i, j = 1, \ldots, p$, то матрица $\mathbf{D}$ оказывается единичной. Ее элементы и есть скалярные произведения $[\mathbf{q}_i, \mathbf{q}_j]$. В этом случае решением системы будет

$$\mathbf{u} = \mathbf{A^*Bf}.$$

Следует отметить, что если базисные функции $\varphi_i(x), i = 0, \ldots, p$ не выбираются специальным образом, то при достаточно больших p ($p \geqslant 5$) полученная СЛАУ оказывается плохо обусловленной. Строки матрицы $\mathbf{D} = \mathbf{A^*BA}$ могут оказаться почти линейно зависимыми. Простейшим примером такого почти линейно зависимого базиса является система функций $x^i, i = 1, \ldots, p$ при больших p. В этом случае желательно использовать ортогональные функциональные базисы, однако такой выбор не всегда возможен и удобен.

Для примера возьмем в качестве базисных функций степенные, обобщенный полином в этом случае будет

$$f(x) = \sum_{j=0}^{p} u_j \, x^j.$$

Скалярные произведения на отрезке [0, 1], записанные в интегральной форме (т. е. при $n \to \infty$), будут иметь вид

$$(\varphi_i, \varphi_j) = \int_0^1 x^i x^j dx = \int_0^1 x^{i+j} dx = \frac{1}{i+j+1}.$$

В таком случае СЛАУ после применения МНК, т. е. минимизации функционала $\Phi(u) = \int_0^1 (F(x) - f(x))^2 dx$, где $F(x)$ — заданная функция, будет:

$$u_0 + \frac{1}{2}u_1 + \ldots + \frac{1}{p+1}u_p = \int_0^1 F(x)dx,$$

$$\frac{1}{2}u_0 + \frac{1}{3}u_1 + \ldots + \frac{1}{p+2}u_p = \int_0^1 xF(x)dx,$$

$$\ldots$$

$$\frac{1}{p+1}u_0 + \frac{1}{p+2}u_1 + \ldots + \frac{1}{2p+1}u_p = \int_0^1 pF(x)dx,$$

или

$$\mathbf{H}_{p+1}\mathbf{u} = \mathbf{F}',$$

где

$$\mathbf{u} = \{u_0, \ldots, u_p\}^{\mathrm{T}}, \quad \mathbf{F}' = \left\{ \int_0^1 F(x)dx, \ldots, \int_0^1 x^p F(x)dx \right\}^{\mathrm{T}},$$

$$\mathbf{H}_{p+1} = \left\{ \frac{1}{i+j-1} \right\}_{i,j=1}^{p+1}$$

Матрица $\mathbf{H}_{p+1}$ называется матрицей Гильберта. Это классический пример плохо обусловленной матрицы. Число обусловленности очень быстро растет с ростом p. Так при $p = 1$ $\mu = \|\mathbf{H}\| \cdot \|\mathbf{H}^{-1}\| \approx 20$, при $p = 9\mu \approx 10^{13}$. Если получим СЛАУ для дискретной системы точек, т. е. для $(\varphi_j, \varphi_k) = \sum_{i=0}^{n} x_i^{k+j}, (\varphi_j, f) = \sum_{i=0}^{n} x_i^k f(x_i)\, x_i = \frac{i}{n}$, то ее матрица будет асимптотически приближаться к матрице Гильберта $\mathbf{H}_{p+1}$ при $n \to \infty$.

3.2. Понятие о методах решения плохо обусловленных СЛАУ

Улучшить качество численного решения СЛАУ метода наименьших квадратов возможно, если использовать различные преобразования матрицы $\mathbf{A}$.

Большинство прямых методов решения линейных систем основано либо на замене исходной системы $\mathbf{Au} = \mathbf{f}$ ($\mathbf{A}$ — квадратная матрица) на эквивалентную $\mathbf{CAu} = \mathbf{Cf}$, либо на представлении матрицы $\mathbf{A}$ в виде произведения других матриц, таких, чтобы новая система либо решалась более просто, либо была лучше (по крайней мере, не хуже) обусловлена, чем исходная.

Подход, использующий спектральную эквивалентность матриц $\mathbf{A}$ и $\mathbf{C}^{-1}$ (в смысле границ спектра собственных значений), основан на умножении $\mathbf{A}$ на близкую в некотором смысле матрицу $\mathbf{C}$. Последняя матрица выбирается таким образом, чтобы произведение было близким к единичной матрице (при этом $\mathbf{C} \neq \mathbf{A}^{-1}$, так как обращение плохо обусловленной матрицы приводит к накоплению вычислительных ошибок. Число обусловленности матрицы $\mathbf{CA}$ $\mu = \lambda_{\max}/\lambda_{\min}$ будет близко к единице. Метод энергетически эквивалентных операторов оказался эффективным при численном решении сеточных уравнений при разностной аппроксимации уравнений в частных производных эллиптического типа.

Идея предобусловливания СЛАУ состоит в том, чтобы вместо исходной системы $\mathbf{Au} = \mathbf{f}$ решать систему $\mathbf{A'u'} = \mathbf{f'}$, где $\mathbf{A'} = \mathbf{C^{-1}AC^{-1}}, \mathbf{u'} = \mathbf{Cu}, \mathbf{f'} = \mathbf{C^{-1}f}$, матрица $\mathbf{C}$ выбирается так, чтобы она была симметричной положительной, хорошо обусловленной.

Определение. Матрица $\mathbf{Q}$ с вещественными элементами q_{ij} является ортогональной, если

$$\mathbf{Q}^* = \mathbf{Q}^{-1}.$$

Пусть матрица $\mathbf{C}$ невырождена. Тогда она представима в виде

$$\mathbf{C} = \mathbf{QR},$$

где $\mathbf{Q}$ — ортогональная, а $\mathbf{R}$ — верхняя треугольная матрицы. В качестве матрицы $\mathbf{Q}$ часто используется симметричная ортогональная матрица $\mathbf{H}$:

$$\mathbf{H} = \mathbf{E} - 2\mathbf{ww}^{\mathrm{T}},$$

где $\mathbf{w}$ - произвольный вектор-столбец, такой, что $(\mathbf{w}^{\mathrm{T}}\mathbf{w}) = 1$. Матрица $(\mathbf{w}, \mathbf{wT})$ есть произведение вектора-столбца $\mathbf{w}$ на вектор-строку $\mathbf{wT}$ (преобразование Хаусхолдера или *метод отражений*).

Заметим, что симметричность матрицы $\mathbf{H}$ (матрицы отражений) показывается непосредственной проверкой; ортогональность $\mathbf{H}$ можно показать следующим образом:

$$\mathbf{H}\mathbf{H}^{\mathrm{T}} = (\mathbf{E} - 2\mathbf{w}\mathbf{w}^{\mathrm{T}})(\mathbf{E} - 2\mathbf{w}\mathbf{w}^{\mathrm{T}}) =$$

$$= \mathbf{E} - 2\mathbf{w}\mathbf{w}^{\mathrm{T}} - 2\mathbf{w}\mathbf{w}^{\mathrm{T}} + 4\mathbf{w}\mathbf{w}^{\mathrm{T}} \cdot \mathbf{w}\mathbf{w}^{\mathrm{T}} =$$

$$= \mathbf{E} - 4\mathbf{w}\mathbf{w}^{\mathrm{T}} - 4\mathbf{w}(\mathbf{w}^{\mathrm{T}}\mathbf{w})\mathbf{w}^{\mathrm{T}} = \mathbf{E},$$

так как $\mathbf{w}^{\mathrm{T}}\mathbf{w} = 1$.

Матрица СЛАУ представляется в виде $\mathbf{A} = \mathbf{H}^{\mathrm{T}}\mathbf{R}$, после чего решается эквивалентная система уравнений $\mathbf{R}\mathbf{u} = \mathbf{H}\mathbf{f}$.

3.3. Задачи

1. Для функции $f(x) = \sqrt{x}$ на отрезке [0,1] построить многочлен $F(x) = u_0 + u_1 x$ среднеквадратичного приближения со скалярными произведениями:

$$(\varphi_i, \varphi_j) = \int_0^1 x^i x^j dx, \quad f, \varphi_i) = \int_0^1 f(x) x^i dx.$$

Решение. Введем базисные функции $\varphi_0(x) = 1, \varphi_1(x) = x$, и вычислим скалярные произведения

$$(\varphi_0, \varphi_0) = \int_0^1 1^2 dx = 1, (\varphi_1, \varphi_1) = \int_0^1 x^2 dx = \frac{1}{3},$$

$$(\varphi_0, \varphi_1) = \int_0^1 x dx = \frac{1}{2}, (f, \varphi_0) = \int_0^1 \sqrt{1}\, x\, dx = \frac{2}{3},$$

$$(f, \varphi_1) = \int_0^1 \sqrt{x}\, x\, dx = \frac{2}{5},$$

Для вычисления коэффициентов получим СЛАУ

$$u_0 + \frac{1}{2}u_1 = \frac{2}{3},$$

$$\frac{1}{2}u_0 + \frac{1}{3}u_1 = \frac{2}{5},$$

откуда $u_0 = \frac{4}{15}, c_1 = \frac{4}{5}, F(x) = \frac{4}{15} + \frac{4}{5}x$.

2. Получить СЛАУ

$$c_{11}x + c_{12}y = f_1,$$
$$c_{22}x + c_{21}y = f_2,$$

если она задана в форме метода наименьших квадратов: $\mathbf{A^*BAu} = \mathbf{A^*Bf}$, где

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix}, \mathbf{B} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \mathbf{u} = (u_1, u_2)^{\mathrm{T}}, \mathbf{f} = (f_1, f_2, f_3)^{\mathrm{T}}.$$

Решить эту систему для случая $a_{11} = a_{32} = 1, a_{21} = a_{12} = 2, a_{31} = 0, a_{22} = 1, f_1 = 1, f_2 = 2, f_3 = 1$.

Решение. Проведем необходимые вычисления:

$$\begin{pmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix} =$$

$$= \begin{pmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix},$$

$$\begin{pmatrix} a_{11}^2 + a_{21}^2 + a_{31}^2 & a_{11}a_{12} + a_{21}a_{22} + a_{31}a_{32} \\ a_{11}a_{12} + a_{21}a_{22} + a_{31}a_{32} & a_{12}^2 + a_{22}^2 + a_{32}^2 \end{pmatrix} =$$

$$= \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} a_{11}f_1 + a_{21}f_2 + a_{31}f_3 \\ a_{12}f_1 + a_{22}f_2 + a_{32}f_3 \end{pmatrix},$$

$$(a_{11}^2+a_{21}^2+a_{31}^2)u_1+(a_{11}a_{12}+a_{21}a_{22}+a_{31}a_{32})u_2 = a_{11}f_1+a_{21}f_2+a_{31}f_3,$$
$$(a_{11}a_{12}+a_{21}a_{22}+a_{31}a_{32})u_1+(a_{12}^2+a_{22}^2+a_{32}^2)u_2 = a_{12}f_1+a_{22}f_2+a_{32}f_3.$$

Решение в числах предлагается найти читателям.

3. С помощью метода наименьших квадратов найти коэффициенты полинома второй степени $f(x) = u_0 + u_1x + u_2x^2$, если таблица измерений задана $\{x_i,\ f_i\}|_{i=0}^{n}$.

Решение. Переопределенная система уравнений:

$$u_0 + u_1x_k + u_2x_k^2 = f_k, k = 0, 1, \ldots, n,$$

определяет функционал

$$\Phi(u_0, \ldots, u_n) = \sum_{k=0}^{n} r_k^2 = \sum_{k=0}^{n} (u_0 + u_1x_k + u_2x_k^2 - f_k)^2.$$

Условия минимума последнего

$$\frac{\partial \Phi}{\partial u_k} = 0.$$

После проведения алгебраических преобразований получим для коэффициентов систему линейных уравнений

$$u_0 + \left(\sum_{k=0}^{n} x_k\right) u_1 + \left(\sum_{k=0}^{n} x_k^2\right) u_2 = \sum_{k=0}^{n} f_k,$$

$$\left(\sum_{k=0}^{n} x_k\right) u_0 + \left(\sum_{k=0}^{n} x_k^2\right) u_1 + \left(\sum_{k=0}^{n} x_k^3\right) u_2 = \sum_{k=0}^{n} f_k x_k,$$

$$\left(\sum_{k=0}^{n} x_k^2\right) u_0 + \left(\sum_{k=0}^{n} x_k^3\right) u_1 + \left(\sum_{k=0}^{n} x_k^4\right) u_2 = \sum_{k=0}^{n} f_k x_k^2.$$

3.4. Задачи для самостоятельного решения

1. Задана таблица $\{x_k, f_k\}$, $k = 0, 1, \ldots, n, n > 1$.

   Найти линейную функцию $f(x) = \alpha x + \beta$, минимизирующую функционал $\Phi(\alpha, \beta) = \sum\limits_{k=0}^{n} (f_k - \alpha x_k - \beta)^2$.

   Используя этот результат, решить переопределенную СЛАУ

   $$x + y = 3, x + 3y = 7, 2x - y = 0.2, 3x + y = 5.$$

2. Пусть $\mathbf{B}$ — квадратная матрица размером $n \times n$, $\mathbf{u}$ — n-мерный вектор, $\Phi(x)$ — функционал, $\Phi(x) = (\mathbf{Bu} - \mathbf{u}x, \mathbf{Bu} - \mathbf{u}x)$. Доказать, что $\Phi(x)$ достигает минимума при $x = (\mathbf{Bu}, \mathbf{u})/(\mathbf{u}, \mathbf{u})$.

3. Барометрическое давление изменяется с высотой по закону $p = ae^{bh}$.

   Определить коэффициенты a, b по результатам наблюдений, приведенных в таблице (h — высота в метрах над уровнем моря; p — давление в мм рт. ст.):

h	0	270	840	1452	2116	3203
p	760	737	686	636	584	508

4. Скорость корабля связана с мощностью его двигателя формулой $P = a_0 + a_3 v^3$ (P — мощность в лошадиных силах, v — скорость в узлах).

   Определить a_0, a_3 по табличным данным.

v	6	8	10	12	13
P	423	805	1378	2357	2893

5. Найти приближение многочленом третьей степени методом наименьших квадратов для функции $f(x) = \sin(\pi x)$ по значениям в точках $x_0 = -1, x_1 = -0,5, x_2 = 0, x_4 = 1$.

6. Вычислить матрицы $\mathbf{A^*BA}$ и $\mathbf{A^*B}$ и найти решение системы уравнений вида $\mathbf{A^*BAu} = \mathbf{A^*Bf}$, если заданы матрицы $\mathbf{A}, \mathbf{B}$ и вектор $\mathbf{f}$.

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 2 & -1 \\ 1 & 3 \\ 3 & 1 \end{pmatrix}, \mathbf{B} = \begin{pmatrix} 1 & 0,5 \\ 0,5 & 1 \end{pmatrix}, \mathbf{f} = (3,\ 0,2,\ 7,\ 5)^{\mathrm{T}}.$$

7. Функция $f(x) = \sqrt{1 + \sin^2(x-1)}$ приближенно заменяется тригонометрическим многочленом $P(x) = a_0 + a_1 \sin x + b_1 \cos x + a_2 \sin 2x + b_2 \cos 2x$ по десяти точкам $x_0, \ldots, x_9$ с помощью метода наименьших квадратов.

   Опишите алгоритм вычисления коэффициентов $a_0,\ a_1,\ b_1,\ a_2,\ b_2$.

8. Доказать, что прямая, проведенная по методу наименьших квадратов, проходит через точку с координатами $x' = \frac{\sum\limits_i y_i x_i}{\sum\limits_i y_i}, y' = \frac{\sum\limits_i x_i y_i}{\sum\limits_i x_i}$.

Литература

[1] *Рябенький В.С.* Введение в вычислительную математику. М.: Физматлит, 2000. 294 с.

[2] *Бахвалов Н.В., Жидков Н.П., Кобельков Г.М.* Численные методы. М.: Лаборатория Базовых Знаний, 2002. 632 с.

[3] *Каханер Д., Моулер К., Нэш С.* Численные методы и программное обеспечение. М.: Мир, 1998. 575 с.

[4] *Голуб Дж., Ван Лоун Ч.* Матричные вычисления. М.: Мир, 1999. 548 с.

[5] *Форсайт Дж., Малькольм М., Моулер К.* Машинные методы математических вычислений. М.: Мир, 1980. 279 с.

[6] *Коновалов А.Н.* Введение в вычислительные методы линейной алгебры: Новосибирск. М.: Наука, 1993. 158 с.

Лекция 4. Численные методы решения экстремальных задач

Рассматриваются наиболее употребительные методы поиска минимума функций нескольких переменных.

Ключевые слова: функционал, целевая функция, стационарная точка, метод перебора, метод золотого сечения, метод парабол, метод покоординатного спуска, градиентного спуска, метод наискорейшего спуска, математическое программирование.

4.1. Поиск безусловного минимума функции

Определение. Пусть на множестве U, состоящем из элементов u линейного метрического пространства определена скалярная функция $\Phi(u)$.

1. Говорят, что $\Phi(u)$ имеет локальный минимум на элементе u^*, если существует его конечная ε-окрестность, в которой выполнено

$$\Phi(u^*) \leqslant \Phi(u), \|u - u^*\| \leqslant \varepsilon \tag{4.1}$$

2. $\Phi(u)$ достигает глобального минимума в U на элементе u^* (строгий, абсолютный минимум), если имеет место равенство

$$\Phi(u^*) = \inf_U \Phi(u) \tag{4.2}$$

Замечание. Если U — числовая ось, решается задача на нахождение минимума функции одного переменного, если U — n-мерное векторное пространство, имеется задача на нахождение минимума функции n переменных, если U — функциональное пространство, то решается задача на отыскание функции, доставляющей минимум функционалу (задача оптимального управления или динамического программирования).

Если к (4.1) или (4.2) добавляются условия

$$u_k^0 \leqslant u_k \leqslant u_k^1, k = 1, \ldots, K$$

$$F_i^0 \leqslant \Phi_i(u) \leqslant F_i^1, i = 1, \ldots, I,$$

($u_k^{\pm}, F_i^{\pm}$ — числа, а Φ_i — заданные функции), то это задача поиска условного минимума, если подобные ограничения отсутствуют, то это задача

поиска безусловного минимума. Причем, если функции $\Phi_i(u)$ линейны, задача поиска условного минимума называется задачей линейного программирования, если хотя бы одна из этих функций нелинейна, то имеется задача нелинейного программирования. Обе эти задачи вместе с задачей динамического программирования в теории оптимального управления называются задачами математического программирования.

Говорится о поиске минимума функции, не ограничивая общности, так как максимум функции $\Phi(u)$ является минимумом функции $-\Phi(u)$.

$\Phi(u)$ называют целевой функцией.

Отметим связь между задачами вычисления корней системы нелинейных алгебраических уравнений (СНАУ) и задачи минимизации.

Пусть на множестве $U \in L^n$ решается система нелинейных уравнений

$$f_1(u_1, \ldots, u_n) = 0,$$

$$\ldots$$

$$f_n(u_1, \ldots, u_n) = 0.$$

Определим целевую функцию следующим образом:

$$\Phi(u_1, \ldots, u_n) = \sum_{k=1}^{n} f_k^2(u_1, \ldots, u_n).$$

В области U справедливо $\Phi(u) \geqslant 0$, причем минимальное значение $\Phi(u)$ имеет при $u = u^*$, где u^* — корень рассмотренной системы. Поэтому ее решение эквивалентно поиску минимума $\Phi(u)$ в U. Если $\Phi(u)$ строго больше нуля, то система решений не имеет.

Теперь положим, что необходимо найти минимум целевой функции $\Phi(u)$, у которой существуют первые производные. В этом случае задача сводится к решению СНАУ

$$\frac{\partial \Phi(u_1, \ldots, u_n)}{\partial u_1} = 0,$$

$$\ldots$$

$$\frac{\partial \Phi(u_1, \ldots, u_n)}{\partial u_n} = 0.$$

Точка, являющаяся решением указанной СНАУ, называется стационарной. Однако не всякая стационарная точка может быть точкой локального минимума целевой функции.

Следующую теорему приведем без доказательства.

Теорема. *Пусть функция $\Phi(u)$ дважды непрерывно дифференцируема. Тогда достаточным условием того, чтобы стационарная точка u^* была точкой локального минимума, является положительная определенность матрицы Гессе*

$$\mathbf{G}(u^*) = \left\{ \begin{array}{ccc} \frac{\partial^2 \Phi}{\partial u_1^2} & \cdots & \frac{\partial^2 \Phi}{\partial u_1 \partial u_m} \\ \cdots & \cdots & \cdots \\ \frac{\partial^2 \Phi}{\partial u_m \partial u_1} & \cdots & \frac{\partial^2 \Phi}{\partial u_m^2} \end{array} \right\}.$$

Отметим, что методы отыскания минимума $\Phi(u)$ нередко оказываются более эффективными, чем методы численного решения СНАУ.

Метод перебора.

Пусть $U = [a, b]$, т. е. отрезок числовой оси. Разобьем его на n равных частей с узлами в точках $u_i = a + i(b-a)/n; i = 0, \ldots, n$.

Вычислив значение $\Phi(u)$ в этих точках, найдем путем сравнения точку u^*, в которой

$$\Phi(u^*) = \min_{0 \leqslant i \leqslant n} \Phi(u_i).$$

Далее полагаем: $u^* \approx u_{\min}, \Phi^* \approx \Phi(u^*)$. Погрешность в определении u^* этого простейшего метода не превосходит числа

$$\varepsilon_n = \frac{b-a}{n}.$$

Этот метод прост, но неэкономичен, особенно когда ищется минимум функции многих переменных. Например, в гиперкубе $U = \{0 \leqslant u_i \leqslant 1, 1 \leqslant i \leqslant 10\}$ с разбиением каждого из отрезков (по каждой из координат) на 10 частей, с быстродействием 10^6 операций в секунду потребуется около 10^7 с (примерно 4 месяца) для нахождения $\min\limits_U \Phi(u)$, если предположить, что количество арифметических действий, необходимое для вычисления значений $\Phi(u)$ в каждой точке требует тысячи арифметических операций. Этот метод можно сделать более эффективным, если сначала определить минимум с грубым шагом, затем уже искать минимум с меньшим шагом на том из отрезков $[x_i, x_{i+1}]$, на котором предполагается наличие минимума; можно и далее уточнять решение задачи таким же образом.

Усовершенствованием этого метода являются методы исключения отрезков, дихотомии (деления отрезка пополам) и золотого сечения. В них отрезок $[a, b]$ делится на 4 части выбором внутри отрезка точек u_1, u_2, в которых вычисляются значения целевой функции. Сравнив ее значения в этих точках, можно сократить отрезок поиска точки минимума, перейдя к отрезку $[a, u_2]$, если $\Phi(u_1) \leqslant \Phi(u_2)$ или $[u_1, b]$, если $\Phi(u_1) \geqslant \Phi(u_2)$. Эту процедуру можно продолжить.

В методе дихотомии точки u_1, u_2 выбираются близко к середине отрезка $u_1 = \frac{b+a-\Delta}{2}, u_2 = \frac{b+a+\Delta}{2}$, где Δ достаточно мало. Поскольку отношение $\frac{b-u_1}{b-a}, \frac{u_2-a}{b-a}$ близко к $1/2$, такой выбор объясняется стремлением обеспечить максимальное относительное уменьшение отрезков.

В конце вычисления в качестве приближенного значения u^* берется середина последнего отрезка. В результате n итераций длина отрезка будет $\Delta_n = \frac{b-a}{2^n} + (\frac{1}{2^n} + \frac{1}{2^{n-1}} + \ldots + \frac{1}{2})\Delta = \frac{b-a}{2^n} + (1 - \frac{1}{2^n})\Delta$, т. е. точность определения u^* составляет $\varepsilon_n = \Delta_n/2$.

Находя n из условия $\varepsilon_n \leqslant \varepsilon$, получим количество итераций, необходимое для достижения данной точности

$$n \geqslant \log_2 \frac{b-a-\Delta}{2\varepsilon - \Delta}.$$

Если в предыдущем неравенстве положить Δ малой, то

$$\varepsilon_n \approx \frac{b-a}{2^{n+1}}.$$

Метод золотого сечения.

Расположим точки u_1, u_2 на $[a, b]$ так, чтобы одна из них стала бы также пробной, но уже на новом отрезке, после исключения части исходного отрезка. Это позволит уменьшить количество вычислений, поскольку необходимо будет вычислить значение $\Phi(u)$ лишь в одной из пробных точек, так как во второй оно уже известно.

Найдем расположение таких точек, для чего рассмотрим отрезок [0, 1] и, для определенности, положим, что при его уменьшении исключается его правая часть.

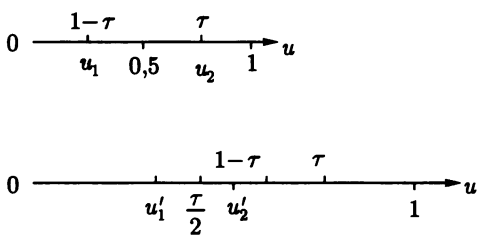


Рис. 4.1

Пусть $u_2 = \tau$, тогда симметрично расположенная относительно центра отрезка точка имеет координату $u_1 = 1 - \tau$ (рис. 4.1).

Пробная точка u_1 отрезка [0, 1] перейдет в пробную точку $u_2^1 = 1 - \tau$ нового отрезка $[0, \tau]$. Условием деления отрезков [0, 1] и $[0, \tau]$ в одном и том же отношении точками $u_2 = \tau$ и $u_2^1 = 1 - \tau$ является равенство

$$\frac{1}{\tau} = \frac{\tau}{1-\tau}, \quad \text{или} \quad \tau^2 + \tau - 1 = 0,$$

откуда находим положительный корень

$$\tau = \frac{\sqrt{5}-1}{2} \approx 0,61803\ldots,$$

т. е. $u_1 = 1 - \tau = \frac{3-\sqrt{5}}{2}$, $u_2 = \tau = \frac{\sqrt{5}-1}{2}$.

Для отрезка $[a, b]$

$$u_1 = a + \frac{3-\sqrt{5}}{2}(b-a); u_2 = a + \frac{\sqrt{5}-1}{2}(b-a)$$

Замечания.

1. Точки u_1, u_2 обладают следующим свойством: каждая из них делит отрезок $[a, b]$ на две неравные части так, что отношение длины всего отрезка к длине его большей части равно отношению длин большей и меньшей части. Точки, обладающие таким свойством, называются точками золотого сечения, введенного Леонардо да Винчи.
2. На каждой итерации отрезок поиска минимума уменьшается в одном и том же отношении

$$\tau = \frac{\sqrt{5}-1}{2},$$

поэтому в результате n итераций длина становиться равной

$$\Delta_n = \tau^n (b-a).$$

Следовательно, точность ε_n определения точки u^* после n итераций равна

$$\varepsilon_n = \frac{\Delta_n}{2} = \frac{1}{2}(\frac{\sqrt{5}-1}{2})^n (b-a);$$

а условие окончания вычислительного процесса будет $\varepsilon_n \leqslant \varepsilon$.

Метод парабол.

Методы, использующие исключение отрезков, основаны на сравнении функций в двух точках пробного отрезка, учитываются лишь значения функции в этих точках.

Учесть информацию о значениях функции между точками позволяют методы полиномиальной аппроксимации. Их основная идея заключена в том, что функция $\Phi(u)$ аппроксимируется полиномом, а точка его минимума служит приближением к u^*. Разумеется, в этом случае кроме свойства унимодальности (т. е. наличия единственного минимума на рассматриваемом отрезке), необходимо на $\Phi(u)$ наложить и требования достаточной гладкости для ее полиномиальной аппроксимации.

Для повышения точности поиска u^* можно как увеличивать степень полинома, так и уменьшать пробный отрезок. Поскольку первый прием приводит к заметному увеличению вычислительной работы и появлению дополнительных экстремумов, обычно пользуются полиномами второй (метод парабол) или третьей (метод кубической интерполяции) степени.

Алгоритм поиска минимума состоит в следующем.

Выбираем на пробном отрезке три точки u_1, u_2, u_3 такие, что $u_1 < u_2 < u_3$ и $u_1 \leqslant u^* \leqslant u_3$.

Построим параболу (квадратичный полином)

$$Q(u) = a_0 + a_1(u - u_1) + a_2(u - u_1)(u - u_2),$$

график которой проходит через точки $(u_1, f(u_1)), (u_2, f(u_2)), (u_3, f(u_3))$.

Коэффициенты $a_k, k = 1, 2, 3$ находим из системы уравнений

$$Q(u_1) = f(u_1),$$
$$Q(u_2) = f(u_2),$$
$$Q(u_3) = f(u_3),$$

откуда

$$a_0 = f(u_1), a_1 = \frac{f(u_2) - f(u_1)}{u_2 - u_1},$$
$$a_2 = \frac{1}{u_3 - u_2}\left[\frac{f(u_3) - f(u_1)}{u_3 - u_1} - \frac{f(u_2) - f(u_1)}{u_2 - u_1}\right].$$

Точку $\bar{u}$ минимума $Q(u)$ находим, приравнивания его производную к нулю:

$$\bar{u} = \frac{1}{2}(u_1 + u_2 - \frac{a_1}{a_2}) =$$
$$= \frac{1}{2}\left[(u_1 + u_2) - \frac{(f_2 - f_1)(u_3 - u_2)}{u_2 - u_1}/(\frac{f_3 - f_1}{u_3 - u_1} - \frac{f_2 - f_1}{u_2 - u_1})\right].$$

Далее полагаем: $u^* \approx \bar{u}$ (очередное приближение точки минимума). Эту процедуру можно продолжить до достижения необходимой точности, выбирая новые точки $u_k, k = 1, 2, 3$. Для этого можно использовать методы исключения отрезков, используя в качестве двух пробных точек u_2 и $\bar{u}$, таких, что $u_2, \bar{u} \in [u_1, u_3]$.

4.2. Методы спуска

Основная идея методов спуска состоит в том, чтобы построить алгоритм, позволяющий перейти из точки начального приближения $u_0 = \{u_0^1, \ldots, u_0^n\}$ в следующую точку $u_1 = \{u_1^1, \ldots, u_1^n\}$ таким образом, чтобы значение целевой функции приблизилось к минимальному.

4.2.1. Метод покоординатного спуска

Этот метод является редукцией поиска функции многих переменных к последовательности поиска минимумов функции одной переменной. Пусть $u^0 \in U$ — начальное приближение к минимуму $\Phi(u)$.

Рассмотрим $\Phi(u_0) = \Phi(u_0^1, \ldots, u_0^n)$ как функцию одной переменной u_1 при фиксированных $u_2^0, \ldots, u_n^0$ и находим одним из приведенных методов поиска минимума функции одной переменной

$$\min_{u_1 \in U} \Phi(u^1, u_0^2, \ldots, u_0^n).$$

Полученное значение u_1, доставляющее минимум $\Phi(u_1)$, обозначим u_1^1; при этом

$$\Phi(u_1^1, u_0^2, \ldots, u_0^n) \leqslant \Phi(u_0^1, \ldots, u_0^n).$$

Далее, при фиксированных значениях $u_1^1, u_3^0, \ldots, u_n^0$ ищем

$$\min_{u_2 \in U} \Phi(u_1^1, u_1^2, x_0^3, \ldots, u_0^n),$$

как функции от u_2; соответствующее значение u_2 обозначим u_2^1; при этом

$$\Phi(u_1^1, u_1^2, \ldots, u_0^n) \leqslant \Phi(u_1^1, u_0^2, \ldots, u_0^n).$$

Этот процесс продолжаем аналогичным образом и для оставшихся координат; в результате получим

$$\Phi(u_1^1, \ldots, u_1^n) \leqslant \Phi(u_1^1, \ldots, u_0^n).$$

Таким образом, переходим из точки u_0 в точку u_1. Этот процесс повторяется до тех пор, пока не будет выполнено условие выхода из итераций, например:

$$|\Phi(u_{k+1}) - \Phi(u_k)| \leqslant \varepsilon,$$

где $\varepsilon > 0$ — заданная точность.

Пример. Найти минимум функции двух переменных

$$\Phi(u) = u_1^2 + u_2^2$$

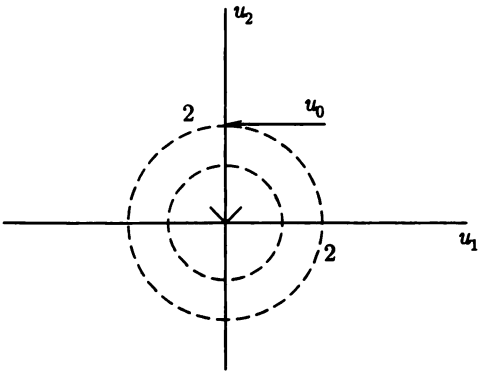


Рис. 4.2

Выбрав некоторую точку начального приближения, например, $u_0 = (2, 2)$, получим минимум целевой функции за два шага, так как ее линии уровня — окружности с центром в начале координат (рис. 4.2).

Если же целевой функцией является, например

$$\Phi(u) = 5u_1^2 + 5u_2^2 + 8u_1u_2,$$

которая поворотом системы координат на угол $-45°$ и преобразованием

$$u_1 = \frac{v_1 + v_2}{\sqrt{2}}; u_2 = \frac{(-v_1 + v_2)}{\sqrt{2}}$$

приводится к виду $\Phi'(v) = v_1^2 + 9v_2^2$, то ее линиями уровня являются эллипсы $v_1^2/9 + v_2^2 = c^2$ поэтому спуск будет иметь иной характер (рис. 4.3).

Можно показать, что покоординатный спуск реализуется (сходится к точке минимума) при условии существования вторых производных $\Phi''_{u_1}, \Phi''_{u_2}, \Phi''_{u_1u_2}$, причем $\Phi''_{u_1} \geqslant a_1 > 0, \Phi''_{u_2} \geqslant a_2 > 0, \left|\Phi''_{u_1u_2}\right| \leqslant a_3, a_1a_2 > a_3^2$. Изломы приводят к подъему. Этот метод сходится достаточно медленно, а при наличии так называемых «оврагов», очень медленно. Разделим «рельефы», образуемые линиями уровня — на два типа: «котловинный» и «овражный». В первом случае линии уровня похожи на эллипсы, а функция вблизи своего минимума практически не изменяется при изменении переменных. Этот случай можно назвать простым (рис. 4.4).

Рельеф овражного типа имеет либо точки излома (рис. 4.4), либо участки с большей кривизной («разрешимый овраг»). Если линии уровня — кусочно-гладкие, то выделим на них точки излома, геометрическое место которых назовем истинным «оврагом», если угол направлен

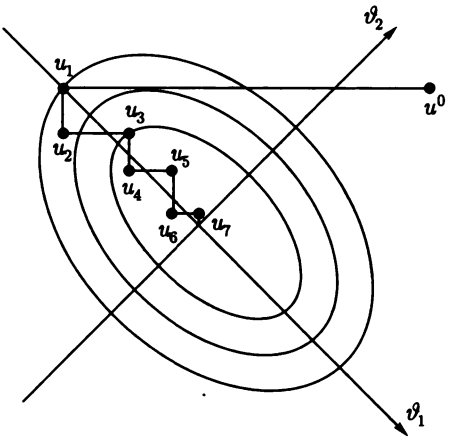


Рис. 4.3

в сторону возрастания функции — и «гребнем», если в сторону убывания (рис. 4.5).

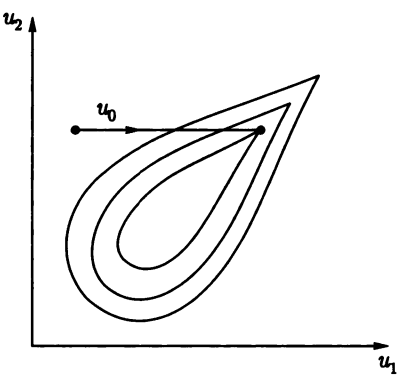


Рис. 4.4

Примером разрешимого оврага является функция $\Phi(u_1, u_2) = 10(u - \sin u_1)^2 + 0,1u_1^2$ (рис. 4.6).

Неупорядоченный тип рельефа характеризуется наличием многих экстремумов; примером может служить функция $\Phi(u_1, u_2) = (1+\sin 2u_1)\times (1 + \sin 2u_2)$ (рис. 4.7).

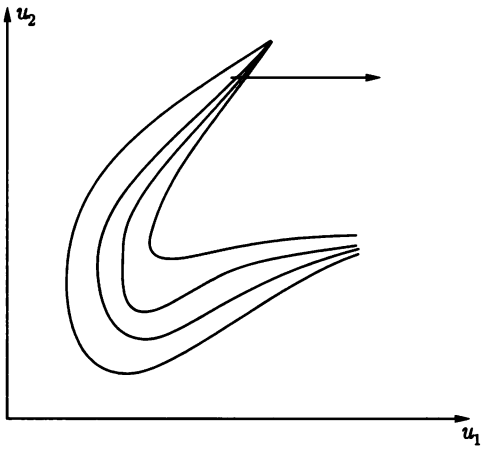


Рис. 4.5

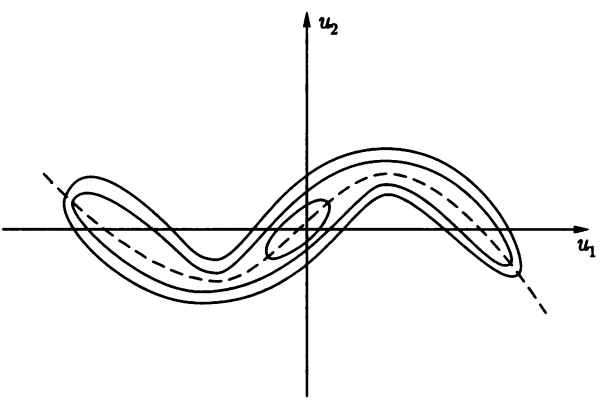


Рис. 4.6

Метод оврагов используется в случае, если «дно» оврага узкое, а «склоны» крутые. В этом случае спустимся из двух точек P_0 и P_1, например, с помощью метода координатного или градиентного спуска на «дно» оврага (или в его окрестность) в точки с координатами r_0 или r_1, не требуя высокой точности сходимости. Проведем через эти две точки прямую и выберем на ней новую точку

$$P_2 = r_1 \pm h(r_1 - r_0),$$

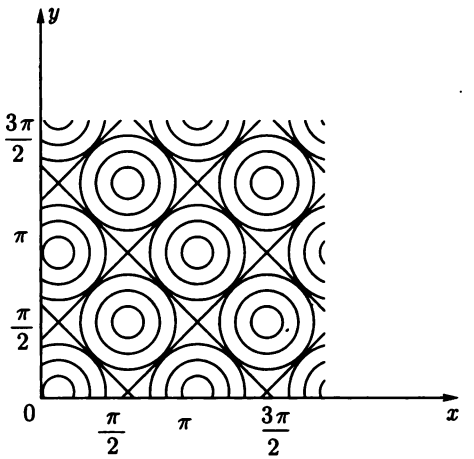


Рис. 4.7

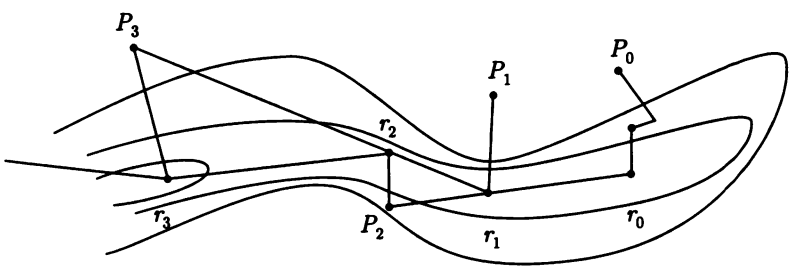


Рис. 4.8

где $h = \text{const} > 0$ — «овражный шаг», который выбирается для каждой функции путем расчета (рис. 4.8). Точка лежит на «склоне» оврага. Из нее спускаемся на «дно» и попадаем в некую точку r_2, через точки r_1 и r_2 проводим прямую и находим точку P_3, из которой возможно опуститься в точку r_3.

Процесс продолжается до тех пор, пока значения целевой функции на «дне» оврага убывают, т. е. пока

$$\Phi(r_{n+1}) > \Phi(r_n).$$

4.2.2. Метод градиентного спуска

Напомним, что градиент функции $\operatorname{grad}\Phi(u) = \left(\frac{d\Phi}{du^1}, \ldots, \frac{d\Phi}{du^n}\right)$ есть вектор, ортогональный линиям уровня целевой функции, а его направление совпадает с направлением наибольшего роста $\Phi(u)$ в данной точке. В точке минимума $\operatorname{grad}\Phi(u) = 0$.

Построим итерационный процесс следующим образом:

$$u_{k+1} = u_k - \tau \cdot \operatorname{grad}\Phi, u_0 = a,$$

где τ — шаг спуска (итерационный параметр). Итерации продолжим до выполнения заданного условия окончания процесса поиска минимума, например

$$\|\operatorname{grad}\Phi(u_{k+1})\| \leqslant \varepsilon > 0.$$

Пример. Рассмотрим функцию двух переменных $\Phi(u_1, u_2) = \frac{(u^1)^2}{4} + + (u^2)^2$.

В соответствии с методом градиентного спуска получим

$$u^1_{k+1} = u^1_k - \tau\frac{u^1_k}{2},$$
$$u^2_{k+1} = u^2_k - \tau \cdot 2u^2_k.$$

Пусть начальное приближение $u_0 = \{1;\ 1\}; \tau = 0,1$.

Тогда $u_1 = \{0{,}95; 0{,}80\}$; $u_2 = \{0{,}9025; 0{,}6400\}$; $u_3 = \{0{,}8574; 0{,}5120\}$; $\Phi(u_1) = 1,25; \Phi(u_3) = 0,446$.

Если взять $\tau = 2$, то $u_1 = \{0; -3\}$ и $\Phi(u_1) = 9$, в то время как $\min\limits_U \Phi(u) = 0$. Выбор шага оказывается существенным в этом методе, поэтому чаще используются методы с переменным шагом.

4.2.3. Метод наискорейшего спуска

В методе градиентного спуска выберем шаг τ так, чтобы функция $\Phi(u)$ максимально уменьшала свое значение:

$$\Phi(u_{k+1}) = \min\Phi(u_k - \tau \cdot \operatorname{grad}\Phi(u_k)).$$

В предыдущем примере выбор шага в точке u_0 сводится к задаче о поиске минимума функции

$$\frac{1}{4}\left(1 - \frac{\tau}{2}\right)^2 + (1 - 2\tau)^2,$$

откуда $\tau = 10/9$, поскольку

$$\Phi(u^1) = \frac{1}{4}(u^1_0 - \tau\frac{u^1_0}{2})^2 + (u^2_0 - 2\tau u^2_0)^2 = \frac{1}{4}(1 - \frac{\tau}{2})^2 + (1 - 2\tau)^2, u^1_0 = u^2_0 = 1.$$

На следующих шагах τ будет зависеть от $u_k^i, k > 0, i = 1, 2$.

Общий случай этого метода, а также метод сопряженных градиентов рассмотрены в лекции, посвященной численным методам решения систем линейных алгебраических уравнений.

Отметим следующее важное обстоятельство. Решение экстремальных задач в L^n зачастую сопряжено со значительными трудностями, особенно для многоэкстремальных задач. Некоторые из этих трудностей исчезают, если ограничиться рассмотрением только выпуклых функций на выпуклых множествах.

Определение. Функция $\Phi(u)$, заданная на выпуклом множестве $U \in L^n$, называется выпуклой, если для любых точек $u, v \in U$ и любого $\alpha \in [0, 1]$ выполнено:

$$\Phi[\alpha u + (1-\alpha)v] \leqslant \alpha\Phi(u) + (1-\alpha)\Phi(v).$$

Определение. Функция $\Phi(u)$ называется строго выпуклой, если для всех $\alpha \in (0, 1)$ выполнено строгое неравенство

$$\Phi[\alpha u + (1-\alpha)v] < \alpha\Phi(u) + (1-\alpha)\Phi(v).$$

Это определение имеет наглядный геометрический смысл: график функции $\Phi(u)$ на интервале, соединяющем точки u, v лежит ниже хорды, проходящей через точки $\{u, \Phi(u)\}$ и $\{v, \Phi(v)\}$ (рис. 4.11).

Для дважды непрерывно дифференцируемой функции $\Phi(u)$ положительная определенность матрицы Гессе $\Phi''_u(u)$ есть достаточное условие строгой выпуклости.

Теорема. *Пусть $\Phi(u)$ — выпуклая функция на выпуклом множестве U, $u \in U$. Тогда любой ее локальный минимум на U является одновременно и глобальным.*

Глобальный минимум строго выпуклой функции $\Phi(u)$ на выпуклом множестве U достигается в единственной точке.

Доказательство.

Предположим противное, т. е. u_0 — точка локального, а u^* — глобального минимума $\Phi(u)$ на U, $u^* \neq u_0$ и $\Phi(u_0) > \Phi(u^*)$. Отсюда, с учетом выпуклости $\Phi(u)$ имеем

$$\Phi[\alpha u* + (1-\alpha)u_0] \leqslant \alpha\Phi(u*) + (1-\alpha)\Phi(u_0) < \Phi(u_0).$$

При $\alpha \to +0$ точка $u = \alpha u^* + (1-\alpha)u^0$ попадает в сколь угодно малую окрестность u_0. Поэтому полученное неравенство $\Phi(u) < \Phi(u_0)$ противоречит предположению о том, что u_0 — точка локального минимума (первая часть теоремы доказана).

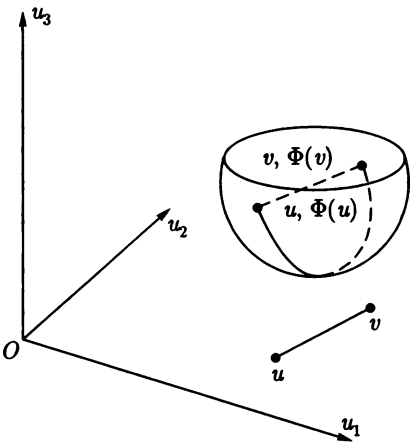


Рис. 4.9

Пусть $u^{(1)}, u^{(2)}$ — две различные точки глобального минимума. Из строгой выпуклости $\Phi(u)$ следует, что для всех $\alpha \in [0,1]$ выполняется строгое неравенство $\Phi\left[\alpha u^{(1)} + (1-\alpha)u^{(2)}\right] < \alpha\Phi(u^{(1)}) + (1-\alpha)\Phi(u^{(2)}) = \Phi* = \min\limits_{U} \Phi(u)$, что противоречит предположению о том, что $u^{(1)}, u^{(2)}$ — точки глобального минимума. ■

4.3. Задачи математического программирования

Под линейным программированием понимают часть экстремальных задач, рассматривающую минимизацию линейных функций и переменных при наличии дополнительных линейных условий трех типов:

$$\begin{aligned} &\min \Phi(u); \Phi(u) = \sum_{i=1}^{n} c_i u_i; \\ &u_i \geqslant 0; 1 \leqslant i \leqslant n; \\ &\sum_{i=1}^{n} a_{ij} u_i = b_j, 1 \leqslant j \leqslant J_1 \\ &\sum_{i=1}^{n} c_{ij} u_i \leqslant b_j, 1_1 < j \leqslant J_2 \end{aligned} \tag{4.3}$$

Каждое из этих условий определяет полупространство, ограниченное гиперплоскостью; вместе эти условия определяют выпуклый n-мер-

ный многогранник J', являющейся пересечением полупространств. Условия типа равенств выделяют из n-мерного пространства $(n-m)$-мерную плоскость. Ее пересечение с M дает выпуклый $(n-m)$-мерный многогранник G. Таким образом, задача состоит в том, чтобы найти минимум линейной функции $\Phi(u)$ в многограннике G.

G — выпуклый многогранник (возможно и неограниченный), поэтому внутри него линейная функция $\Phi(u)$ не может достигать минимума. Показывается, что ее минимум, если он существует, достигается в какой-то из его вершин. Теоретически задача линейного программирования достаточно проста: необходимо вычислить значение функций в конечном числе точек — вершинах многогранника, сравнить их между собой и найти среди них наименьшее. Однако трудность заключается в том, что в экономических задачах количество переменных порядка $10^2 \div 10^4$, поэтому решение оказывается достаточно сложным.

Пример. Рассмотрим следующую простую задачу линейной оптимизации на плоскости: найти $\min \Phi(u) = c_1u_1 + c_2u_2$ при ограничениях

$$\begin{cases} a_{11}u_1 + a_{12}u_2 = b_1 \\ a_{21}u_1 + a_{22}u_2 = b_2 \\ u_1 \geqslant 0, u_2 \geqslant 0. \end{cases}$$

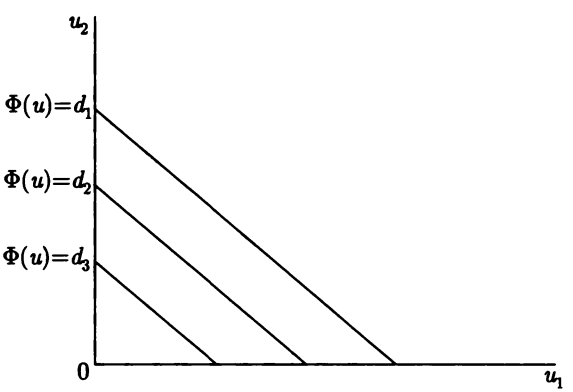


Рис. 4.10

Неравенства $u_1 \geqslant 0$ и $u_2 \geqslant 0$ выделяют I квадрант плоскости (u_1, u_2), а линейная функция $\Phi(u) = c_1u_1 + c_2u_2$ при определенных c_1, c_2 задает в этом квадранте семейство прямых уравнением $c_1u_1 + c_2u_2 = d$ (d неизвестно) (рис. 4.10).

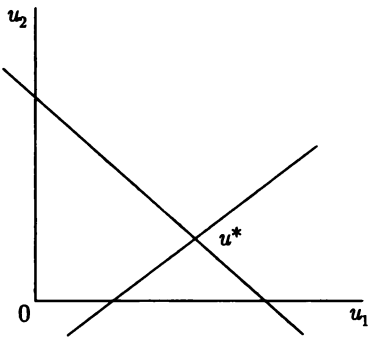


Рис. 4.11

Вообще говоря, ограничения в форме равенств могут быть следующего типа: отсутствует, одно ограничение, два ограничения (последний случай соответствуют рассматриваемому). Пусть СЛАУ в данном примере имеет единственное решение $u^* = (u_1^*, u_2^*)$.

Если выполнены неравенства $u_1 \geqslant 0, u_2 \geqslant 0$, то u^* и есть решение задачи линейной оптимизации, а $\min \Phi(u) = c_1 u_1^* + c_2 u_2^*$ (рис. 4.11).)

Если же, например, $u_1^* < 0$ и $u_2^* < 0$ то задача линейного программирования неразрешима, так как по условию решения находится внутри первого квадранта. В данном случае ограничения определяют единственное решение рассматриваемой задачи, а целевая функция принимает «навязанное» значение $\Phi(u^*)$. Процесс решения оказался весьма простым. Если бы имелось больше условий типа неравенств и меньше равенств, рассмотрение оказалось бы более сложным.

Пример. Графическое решение задачи

$$\min \Phi(u) = -3u_1 - 3u_2,$$

$$u_1 + 2u_2 \leqslant 7,$$

$$2u_1 + u_2 \leqslant 8,$$

$$u_2 \leqslant 3,$$

$$u_2 \leqslant 3,$$

$$u_1 \geqslant 0,$$

$$u_2 \geqslant 0$$

приводит к решению $u^* = (3, 2)$ и $\Phi(u^*) = -15$. Для решения задачи рассматриваются линии уровня функции $\Phi(u)$, т. е. семейство параллельных прямых $-3u_1 - 3u_2 = d = \text{const}$.

Нормальный к этим прямым вектор $-\Phi'_u(u)$ указывает направление убывания целевой функции. Решением задачи u^* является одна из вершин многогранника, образованного прямыми системы неравенств.

4.4. Задачи

1. Свести задачу о нахождении решения системы нелинейных уравнений

$$\begin{cases} u - 5 \cdot 10^{-2} e^{uv} = 0, \\ v - 5 \cdot 10^{-2} e^{-(u+v)} = 0, \end{cases}$$

к вариационной задаче в области $\Omega = \{|u - 0,1| \leqslant 0,1; |v - 0,1| \leqslant 0,1\}$.

Решение. Решение системы сводится к нахождению условий минимума функционала

$$\Phi(u, v) = (u - 5 \cdot 10^{-2} e^{uv})^2 + (u - 5 \cdot 10^{-2} e^{-(u+v)})^2.$$

2. Свести задачу о нахождении минимума функции

$$\Phi(u, v) = u^4 + v^4 - u^2 - v^2 \text{ в области } \Omega > \{|u| \leqslant 1; |v| \leqslant 1\}.$$

к решению системы алгебраических уравнений.

Решение. Задача о нахождении минимума функции $\Phi(u, v)$ сводится к решению системы уравнений $\frac{\partial \Phi}{\partial u} = 0$, $\frac{\partial \Phi}{\partial v} = 0$.

3. Найти значения $\{x, y\}$, при которых достигается минимум функции $f(x, y) = x^3 + y^3 - 3xy$.

Решение. Вычислим частные производные $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$ и приравняем их к нулю. Получим систему двух нелинейных уравнений $3x^2 - 3y = 0, -3x + 3y^2 = 0$. Решениями этой системы являются пары $\{0, 0\}, \{1, 1\}$. Подстановкой убеждаемся, что вторая точка является точкой глобального минимума.

4. Методом деления отрезка пополам найти точку локального минимума для функции $f(x) = x^3 + e^{-x} - x$ на отрезке [0, 1] с точностью $\varepsilon = 10^{-2}$.

Решение. Обозначим границы отрезка $a_0 = 0, b_0 = 1$ и зададим $\frac{\Delta}{2} = \delta = 10^{-3}$.

Вычислим $f\left(\frac{a_0+b_0-\Delta}{2}\right) \approx 0,2324$ и $f\left(\frac{a_0+b_0+\Delta}{2}\right) \approx 0,2307$. Так как второе значение меньше первого, то положим $\{a_1, b_1\} = \{0.4990, 1\}$. Продолжая далее, получим $\{a_2, b_2\} = \{0,4990; 0,7505\}$, $\{a_3, b_3\} = \{0,6238; 0,7505\}$, $\{a_4, b_4\} = \{0,6861; 0,7505\}$, $\{a_5, b_5\} = \{0,6861; 0,7191\}$, $\{a_6, b_6\} = \{0,7016; 0,7191\}$, $\{a_7, b_7\} = \{0,7101; 0,7198\}$.

5. С помощью метода Ньютона найти минимум функции $F(t) = \sin t - \cos t$, $t_0 = -0,5$.

Решение. Найдем точку минимума функции $F(t)$ как корень уравнения $F'(t) = 0$. Для этого построим итерационный процесс Ньютона:

$$t_{n+1} = t_n - \frac{F'(t_n)}{F''(t_n)}, t_0 = -0,5.$$

При $t_0 = -0,5$ имеем $F'(t_0) = \cos t + \sin t \approx 0,3982$, $F''(t_0) = -\sin t_0 + \cos t_0 \approx 1,3570$, $t_1 = t_0 - \frac{0,3982}{0,1357} \approx -0,7934$. Дальнейшие вычисления дают $t_2 = -0,7854$, $t_3 = -0,7854$.

4.5. Задачи для самостоятельного решения

1. Найти точку локального минимума функций:

$$f(x,y) = 3x^2 - 2x\sqrt{y} + y - 8x + 8,$$

$$f(x,y) = x^3 + 8y^3 - 6xy + 1,$$

$$f(x,y) = x^2 + y^2 + xy + x - y + 1,$$

$$f(x,y) = 2x^3 - xy^2 + 5x^2 + y^2.$$

2. Найти точку локального минимума функций:

$$f(t) = 2x^2 - \ln x,$$

$$f(t) = \frac{t^3}{3} + t^2,$$

$$f(t) = \frac{t^4}{4} - 2t^2,$$

$$f(t) = te^{-\frac{t^2}{2}},$$

$$f(t) = 3t^4 - 8t^3 + 6t^2,$$

$$f(t) = (t-5)e^t,$$

$$f(t) = \frac{t^2-3}{t+2}.$$

используя методы дихотомии и сведения вариационной задачи к решению алгебраического уравнения.

3. Найти точки локального минимума функций

$$f(x,y) = (x^2+y-11)^2 + (x+y^2-7)^2,$$

$$f(x,y) = (x^2+y^2-1)^2 + (y-x\sin x)^2,$$

$$f(x,y) = x^3 + 8y^3 - 6xy + 1,$$

$$f(x,y) = (x-3)^2 + (y-2)^2 + (x-y-4)^2,$$

$$f(x,y) = 2x^3 - xy^2 + 5x^2 + y^2$$

методом покоординатного спуска.

Литература

[1] *Федоренко Р.П.* Введение в вычислительную физику. — М.: Изд-во МФТИ, 1994. 526 с.

[2] *Калиткин Н.Н.* Численные методы. — М.: Наука, 1978. 512 с.

[3] *Амосов А.А., Дубинский Ю.А., Копченова Н.В.* Вычислительные методы для инженеров. М.: Высшая школа, 1994. 544 с.

[4] *Бирюков С.И.* Оптимизация. Введение в теорию. Численные методы. М.: МЗ-пресс, 2003. 244 с.

Лекция 5. Численное решение нелинейных алгебраических уравнений и систем

Рассматриваются численные методы решения нелинейных уравнений и систем. На основе принципа сжимающих отображений рассматриваются условия сходимости итерационных методов. Доказывается квадратичная сходимость метода Ньютона. Рассматривается задача о динамике простейшего нелинейного дискретного отображения — логистического. Дается понятие о бифуркациях дискретного отображения.

Ключевые слова: итерации, метод простой итерации. Терема о неподвижной точке. Метод Ньютона, метод хорд, дискретные отображения, бифуркации.

5.1. Сжимающие отображения. Итерации. Метод простых итераций(МПИ)

Рассмотрим системы нелинейных алгебраических уравнений, записанные в векторном виде.

Система нелинейных алгебраических уравнений

$$\mathbf{f}(\mathbf{u}) = 0 \tag{5.1}$$

может быть также представлена в равносильном виде

$$\mathbf{u} = \mathbf{F}(\mathbf{u}), \tag{5.2}$$

где $\mathbf{u} \in L^n$ L_n — n-мерное евклидово пространство. Как правило, для нелинейной системы переход от формы записи (5.1) к равносильному виду (5.2) осуществляется не единственным образом.

Поставим в соответствие системе (5.2) итерационный процесс, определяющий последовательность итераций (последовательных приближений к решению). Соответствующий итерационный процесс записывается в форме

$$\mathbf{u}_{k+1} = \mathbf{F}(\mathbf{u}_k), \quad \mathbf{u}_0 = a, k = 0, 1, \ldots. \tag{5.3}$$

Для дальнейшего изложения потребуется понятие отображения. *Отображением* называется закон, по которому каждому элементу x некоторого множества X однозначно сопоставляется определенный элемент

y множества Y (X может совпадать с Y). Это соотношение между элементами $x \in X$ и $y \in Y$ записывается как $y = f(x)$ или $f : x \to y$. Говорят, что отображение f действует из X в $Y (f : X \to Y)$. Отображение $f : X \to X$ называют преобразованием множества X, это отображение f преобразует множество X в себя. В функциональном анализе и линейной алгебре вместо термина «отображение» часто употребляется термин «оператор», в случае, если X и Y — числовые множества, употребляется термин «функция».

Определение. Область $\Omega \in L^N$ называется выпуклой, если наряду с любыми двумя точками $a \in \Omega$ и $b \in \Omega$ она включает все точки отрезка $[a, b]$, т. е. точки с координатами $u = a + t(b - a)$, где $0 \leqslant t \leqslant 1$.

Определение. Отображение $\mathbf{v} = \mathbf{F}(\mathbf{u})$ называется *сжимающим* в замкнутой выпуклой области Ω, если существует такое число $0 < q < 1$, что

$$\rho[\mathbf{F}(\mathbf{u}_1), \mathbf{F}(\mathbf{u}_2)] \leqslant q\rho(\mathbf{u}_1, \mathbf{u}_2)$$

при любых u_1, u_2, принадлежащих области Ω, здесь $\rho(\mathbf{u}_1, \mathbf{u}_2)$ — расстояние между элементами. В линейном нормированном пространстве $\rho(\mathbf{u}_1, \mathbf{u}_2) = \|\mathbf{u}_1 - \mathbf{u}_2\|$.

Приведем без доказательства одну из основных теорем функционального анализа.

Теорема (принцип сжимающих отображений). *Всякое сжимающее отображение имеет в Ω одну и только одну неподвижную точку $u^* \in \Omega$.*

Более подробно о сжимающих отображениях и другие теоремы о неподвижных точках можно найти, например, в [1, 2, 3].

Теорема (о сжимающем отображении [1, 5]).
Последовательность $\{\mathbf{u}_k\}, k = 0, 1, \ldots$ элементов n-мерного евклидова пространства, порожденная итерационным процессом

$$\mathbf{u}_{k+1} = \mathbf{F}(\mathbf{u}_k), \mathbf{u}_0 = \mathbf{a},$$

сходится к решению $\mathbf{U}$ системы нелинейных алгебраических уравнений $\mathbf{u} = \mathbf{F}(\mathbf{u})$, если отображение

$$\mathbf{v} = \mathbf{F}(\mathbf{u})$$

является сжимающим; при этом выполнено

$$\rho(\mathbf{U}, \mathbf{u}_k) \leqslant \frac{q^k}{1 - q}\rho(\mathbf{u}_0, \mathbf{u}_1).$$

Доказательство.

По определению сжимающего отображения

$$\rho(\mathbf{u}_{k+1}, \mathbf{u}_k) = \rho\left[\mathbf{F}(\mathbf{u}_k), \mathbf{F}(\mathbf{u}_{k-1})\right] \leqslant$$

$$\leqslant q\,\rho(\mathbf{u}_k, \mathbf{u}_{k-1}) \leqslant \ldots \leqslant q^k \rho(\mathbf{u}_0, \mathbf{u}_1) = q^k \rho_0.$$

В таком случае получим цепочку неравенств при $p > k$:

$$\rho(\mathbf{u}_p, \mathbf{u}_k) \leqslant \rho(\mathbf{u}_p, \mathbf{u}_{p-1}) + \ldots + \rho(\mathbf{u}_{k+1}, \mathbf{u}_k) \leqslant$$

$$\leqslant q^{p-1}\rho_0 + \ldots + q^k \rho_0 \leqslant q^k \rho_0 \sum_{i=0}^{\infty} q^i = \rho_0 \frac{q^k}{1-q}.$$

В соответствии с критерием Коши существования предела последовательности, последовательность $\mathbf{u}_k$ стремится к пределу $\mathbf{U}$, поскольку правая часть неравенства стремиться к нулю при $k \to \infty$.

Напомним критерий Коши сходимости числовой последовательности: последовательность $\{u_k\}$, $k = 0, 1, \ldots$ является сходящейся, если для любого положительного числа ε существует номер N такой, что при всех $k > N$ и любых натуральных p расстояние между членами последовательности u_k и u_{k+1} меньше ε, т. е. $|u_k - u_{k+p}| < \varepsilon$.

Напомним критерий Коши для последовательности элементов метрического пространства: последовательность $\{\mathbf{u}_k\}$, $k = 0, 1, \ldots$ является сходящейся, если для любого $\varepsilon > 0$ существует номер N такой, что при всех $k > N$ и любом натуральном p расстояние $\rho(\mathbf{u}_k, \mathbf{u}_{k+p}) < \varepsilon$.

Продолжим доказательство. Переходя в последнем неравенстве к пределу при $p \to \infty$, получим

$$\rho(\mathbf{U}, \mathbf{u}_k) \leqslant \rho_0 \frac{q^k}{1-q}.$$

Покажем, что $\mathbf{U}$ есть корень уравнения (5.2)

$$\rho\left[\mathbf{U}, \mathbf{F}(\mathbf{U})\right] \leqslant \rho(\mathbf{U}, \mathbf{u}_{k+1}) + \rho\left[\mathbf{u}_{k+1}, \mathbf{F}(\mathbf{U})\right] = \rho(\mathbf{U}, \mathbf{u}_{k+1}) + \rho\left[\mathbf{F}(\mathbf{u}_k), \mathbf{F}(\mathbf{U})\right] \leqslant$$

$$\leqslant \rho_0 \frac{q^{k+1}}{1-q} + q\rho(\mathbf{u}_k, \mathbf{U}) \leqslant \rho_0 \frac{q^{k+1}}{1-q} + q\rho_0 \frac{q^k}{1-q} = 2\rho_0 \frac{q^{k+1}}{1-q}$$

Поскольку k выбрано произвольно, а левая часть от k не зависит, то $\rho\left[\mathbf{U}, \mathbf{F}(\mathbf{U})\right] = 0$, или $\mathbf{U} = \mathbf{F}(\mathbf{U})$.

В случае скалярного уравнения имеем $\theta = u_k + t(u_{k+1} - u_k)$,

$$|u_{k+1} - u_k| = |F(u_k) - F(u_{k-1})| \leqslant$$

$$\leqslant \max_{\Delta} |F'(\theta)|\,|u_k - u_{k-1}| \leqslant \ldots \leqslant (\max_{\Delta} |F'(\theta)|)^k\,|u_1 - u_0|,$$

откуда следует условие сходимости итерационного процесса $\max |F'(u)| \leqslant q < 1$. Отрезок Δ включает в себя всю последовательность u_k, $u_k \in \Delta$, $k = 0, 1, 2, \ldots$

В случае решения системы нелинейных уравнений достаточным условием сходимости итерационного процесса будет $\|\mathbf{F}'(\mathbf{u})\| < 1$, где $\mathbf{F}'(\mathbf{u})$ — матрица Якоби. ■

Теорема (без доказательства). *Пусть область $G \in L^n$ выпуклая, $\mathbf{u} \in G$ а компоненты $F_i(\mathbf{u})$ вектор-функции $\mathbf{F}(\mathbf{u}) = (F_1, \ldots, F_N)^T$ имеют равномерно непрерывные производные первого порядка. Положим, что норма матрицы Якоби*

$$\mathbf{Y} = \frac{d\mathbf{F}(\mathbf{u})}{d\mathbf{u}} = \begin{pmatrix} \dfrac{\partial F_1}{\partial u_1} \cdots \dfrac{\partial F_1}{\partial u_n} \\ \cdots \quad \cdots \quad \cdots \\ \dfrac{\partial F_n}{\partial u_1} \cdots \dfrac{\partial F_n}{\partial u_n} \end{pmatrix}$$

не превосходит некоторого числа $0 \leqslant q < 1$, т. е. $\|\mathbf{Y}\| \leqslant q < 1$ для всех $\mathbf{u} \in G$.

В этом случае отображение $\mathbf{v} = \mathbf{F}(\mathbf{u})$ является сжимающим в области G, т. е. $\rho(\mathbf{F}(\mathbf{u}_1), \mathbf{F}(\mathbf{u}_2)) \leqslant q\,\rho(\mathbf{u}_1, \mathbf{u}_2)$, или $\|\mathbf{F}(\mathbf{u}_1) - \mathbf{F}(\mathbf{u}_2)\| \leqslant q\,\|\mathbf{u}_1 - \mathbf{u}_2\|$.

Геометрическая интерпретация метода простой итерации для скалярного случая $u_{k+1} = F(u_k)$ приведена на рис. 5.1. Алгоритм метода простых итераций таков.

1. Локализуем корень, приближенно определяем, на каком отрезке он находится. Вопрос локализации корня не решается алгоритмически, это, скорее, вопрос искусства вычислителя, хотя во многих случаях локализовать корень достаточно легко.

2. Выбираем точку u_0 на оси $0u$.

3. Вычисляем $F(u_0)$.

4. Определяем точку u_1 по значению $F(u_0)$:

   4.1. Пересечение горизонтальной прямой AA' с прямой $v = u$ есть точка C ($OA = v_1, AC = u_1$)

   4.2. Очевидно, что горизонтальная координата точки C и есть u_1 (так как $F(u_0) = u_1$).

   4.3. Опустим перпендикуляр из C на u. Поскольку $OA = u_1$, то u_1 — значение на первой итерации.

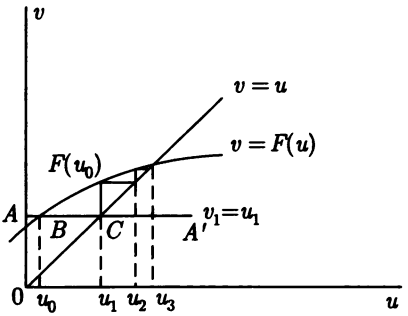


Рис. 5.1. Геометрическая интерпретация метода простых итераций

5. Аналогично строим точки $u_2, u_3,$. Получившаяся диаграмма носит название *лесенка Ламерея.*

Метод релаксации. Без ограничения общности рассмотрим скалярный случай. Положим $F(u) = u + \tau f(u)$ и построим итерационный процесс $u_{k+1} = u_k + \tau f(u_k), u_0 = a.$

Тогда $F'(u) = 1 + \tau f'(u)$ и τ выбирается из условия $|F'(u)| < 1$, причем, чем меньше значение $|F'(u)|$, тем быстрее будет сходиться итерационный процесс. В частности, если положить $F'(u) = 0$, то, $\tau = -[f'(u)]^{-1}$, а формулы итерационного процесса будут $u_{k+1} = u_k - [f'(u_k)]^{-1} f(u_k), u_0 = a.$

5.2. Метод Ньютона

Как и выше, необходимо найти решение уравнения $f(u) = 0$. Пусть u_k есть k приближение решения (k итерация). Следующее приближение ищем в виде $u_{k+1} = u_k + \Delta u_k$, разложив функцию $f(u)$ в ряд Тейлора с точностью до членов первого порядка: $f(u_k + \Delta u_k) = f(u_k) + f'_u(u_k) \cdot \Delta u_k + O(\Delta^2 u_k)$.

Пренебрегая членами $O(\Delta^2 u_k)$, получим линеаризованное уравнение для определения Δu_k:

$$f(u_k) + f'_u(u_k)\Delta u_k = 0,$$
$$u_{k+1} - u_k = \Delta u_k = -f_u^{-1}(u_k) f(u_k),$$
$$u_{k+1} = u_k - f_u^{-1}(u_k) f(u_k), u_0 = a.$$

Это уже знакомая формула, полученная в результате оптимизации релаксационного варианта метода простой итерации.

Для системы уравнений матрица Якоби $\mathbf{f}'_{\mathbf{u}}(\mathbf{u}_k)$ будет

$$\mathbf{A} = \left\{ \frac{\partial f_i}{\partial x_j} \right\} = \begin{pmatrix} \frac{\partial f_1}{\partial u_1} & \dots & \frac{\partial f_1}{\partial u_n} \\ \dots & \dots & \dots \\ \frac{\partial f_n}{\partial u_1} & \dots & \frac{\partial f_n}{\partial u_n} \end{pmatrix},$$

а метод Ньютона выглядит следующим образом:

$$f_1^{k+1} = f_1^k + (\frac{\partial f_1}{\partial u_1})^k \Delta u_1^k + \dots + (\frac{\partial f_1}{\partial u_n}) \Delta u_n^k,$$

$$\dots$$

$$f_n^{k+1} = f_n^k + (\frac{\partial f_n}{\partial u_1})^k \Delta u_1^k + \dots + (\frac{\partial f_n}{\partial u_n})^k \Delta u_n^k,$$

где $f_i = f_i(u_1, \dots, u_n)$. Приходим к СЛАУ вида $\mathbf{A}(\Delta \mathbf{u}) = -\mathbf{f}$, где $\Delta \mathbf{u} = (\Delta u_1, \Delta u_2, \dots, \Delta u_n)^{\mathrm{T}}$, $\mathbf{f} = (f_1, f_2, \dots, f_n)^{\mathrm{T}}$, $\Delta \mathbf{u} \in L^n, \mathbf{f} \in L^n$, $\Delta u_i = u_i^{k+1} - u_i^k$, $i = 1, \dots, n, k = 0, 1, \dots$.

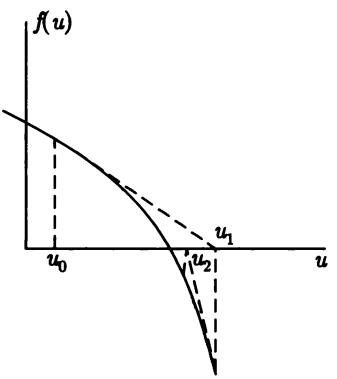


Рис. 5.2. Геометрическая интерпретация метода Ньютона в одномерном случае

Геометрический смысл метода Ньютона в одномерном случае проиллюстрирован на рис. 5.2. Заменим $f(u)$ в точках u_k — каждом приближении к корню — касательными. За следующее приближение по методу Ньютона примем значение u точки пересечения касательной с осью абсцисс. Метод Ньютона называют также методом линеаризации или методом касательных.

Теорема о квадратичной сходимости метода Ньютона [2, 4].

Сформулируем и докажем теорему для одномерного (скалярного) случая. Аналогичная теорема будет справедлива и для систем нелинейных уравнений.

Теорема. *Пусть существуют первые две ограниченные производные* $f(u)$ *и, кроме того, существует* $[f'_u(u)]^{-1}$*; причем имеют место оценки* $|f''_{uu}| \leqslant C_2, \left|[f'(u)]^{-1}\right| \leqslant C_1$ *(отображение* $f(u)$ *равномерно невырождено), а начальное приближение выбирается из условия*

$$C_1^2 C_2 \, |f(u_0)| \leqslant q < 1.$$

Тогда метод Ньютона сходится с квадратичной скоростью сходимости.

Доказательство.

Разложим $f(u_{k+1})$ в ряд Тейлора в окрестности $f(u_k)$, ограничившись квадратичными членами разложения:

$$f(u_{k+1}) = f(u_k) + f'_u(u_k)\,\Delta u_k + O(\Delta^2 u_k).$$

Здесь введено обозначение $\Delta u_k = u_{k+1} - u_k$. Переходя к абсолютной величине и учитывая, что для метода Ньютона $f(u_k) + f'_u(u_k)\,\Delta u_k = 0$, или $\Delta u = -[f'_u(u)]^{-1} f(u_k)$, получим

$$|f(u_{k+1})| = O(\Delta^2 u_k) \leqslant C_2 \cdot \Delta^2 u_k = C_2 \left|[f'_u(u)]^{-1} f(u_k)\right|^2 \leqslant C_2 C_1^2 \, |f(u_k)|^2,$$

так как $\left|[f'_u(u)]^{-1}\right| < C_1$.

Введем в рассмотрение *невязку* как $r_k = |f(u_k)|$, получим $r_{k+1} \leqslant Cr_k^2$, где $C = C_2 C_1^2$.

Можно выписать цепочку соотношений $r_1 \leqslant Cr_0^2, r_2 \leqslant Cr_1^2 \leqslant C^3 r_0^4, r_3 \leqslant Cr_2^2 \leqslant C^7 r_0^8$, и т.д., в результате для невязки на k итерации получается выражение $r_k \leqslant C^{-1}(Cr_0)^{2^k}$.

Неравенства $r_{k+1} \leqslant Cr_k^2$ и $r_k \leqslant C^{-1}(Cr_0)^{2^k}$ являются определением квадратичной скорости сходимости.

Для сходимости итерационного процесса Ньютона достаточно, чтобы было выполнено условие, следующее из последнего неравенства: $Cr_0 = C\,|f(u_0)| \leqslant q < 1$. Отсюда следуют ограничения на начальное приближение, в частности, $|f(u_0)| \leqslant 1/C$. Теорема доказана. ■

Замечание. Несложно показать, что погрешность, определяемая, как $\varepsilon_k = \rho(\mathrm{u}_k - \mathrm{U})$, или, в скалярном случае, $\varepsilon_k = |u_k - U|$, убывает

квадратично. Для этого разложим $f(u_{k+1})$ в окрестности u_k в ряд Тейлора до первого члена (или линеаризуем $f(u_{k+1})$)

$$f(u_{k+1}) \approx f(u_k) + f(u_k)(u_{k+1} - u_k) + \frac{f''}{2}(u_{k+1} - u_k)^2.$$

Так как в методе Ньютона приближения находятся достаточно близко к корню уравнения и $u_{k+1} \approx U$, получим

$$0 = f(U) = f(u_k) + f'(u_k)(U - u_k) + \frac{f''}{2}[(U - u_k)]^2.$$

Разделив полученное равенство на $f'_u(u_k)$, приходим к оценке $U - \left[u_k - \frac{f(u_k)}{f'(u_k)}\right] = \frac{f''_u(\theta)}{2f'_u(u_k)}(U - u_k)^2$, откуда следует $\left|U - (u_k - \frac{f(u_k)}{f'_u(u_k)})\right| \leqslant \left|\frac{\max|f''_u(\Theta)|}{2f'_u(u_k)}\right| |U - u_k|^2, u_k \in \Delta, k = 0, 1, \ldots$. Левая часть последнего неравенства по формуле Ньютона равна $\varepsilon_{k+1} = |U - u_{k+1}|$. В таком случае $\varepsilon_{k+1} \leqslant \bar{C}\varepsilon_k^2$, где $\bar{C} = \frac{\max\limits_{\Delta}|f''(\theta)|}{2|f'(u_k)|}$, откуда последовательно находим $q^{(k)}[f(u)]$.

Итерационные процессы, имеющие третий и четвертый порядок сходимости, представляются формулами

$$u_{k+1} = u_k - \frac{f_k}{f'_k} - \frac{f''_k \cdot f_k^2}{2(f'_k)^3},$$

$$u_{k+1} = u_k - \frac{f_k}{f'_k} - \frac{f''_k f_k^2}{2(f')_k^3} - \frac{(f'')^2 f_k^3}{2(f')_k^5} + \frac{f_k^{(3)} f^3}{6(f')_k^7}, \quad \text{где} \quad f_k = f(u_k).$$

Отметим, что итерационные методы высших порядков используются достаточно редко вследствие повышенных требований к гладкости функций и необходимости вычисления ее производных и обратных к ним величин.

Иногда для численного решения нелинейных алгебраических систем уравнений, чтобы не вычислять на каждой итерации обратную матрицу, используют упрощенный метод Ньютона

$$\mathbf{u}_{k+1} = \mathbf{u}_k - \left[\mathbf{f}_{\mathbf{u}}^{-1}(\mathbf{u}_0)\right]^{-1} \mathbf{f}(\mathbf{u}_k), \mathbf{u}_0 = \mathbf{a}.$$

Этот метод оказывается приемлемым, поскольку начальное приближение в методе Ньютона обычно выбирается достаточно близким к корню уравнения.

Методом секущих (или разностным методом Ньютона) называется итерационный метод, в котором вместо производной вычисляется

разностное выражение $f'(u_k) \approx \frac{f(u_k)-f(u_{k-1})}{\tau_k}$, откуда $u_{k+1} = u_k - \frac{f(u_k)\tau_k}{f(u_k)-f(u_{k-1})}, \tau_k = u_k - u_{k-1}$.

5.3. О вариационных подходах к решению нелинейных систем уравнений

Рассмотрим систему нелинейных уравнений $f(u,v) = 0, g(u,v) = 0$, $u, v \in \Omega$.

Рассмотрим функционал $\Phi(u,v) = f^2(u,v) + g^2(u,v)$.

Так как Φ неотрицателен, то найдется точка $\{\bar{u}, \bar{v}\}$ такая, что $\{\bar{u}, \bar{v}\} = \arg \min_{u,v\in\Omega} \Phi(u,v)$, но $\min \Phi(u,v)$, очевидно, достигается при $f(u,v) = 0$, $g(u,v) = 0$, т. е. на решении исходной системы уравнений.

Построим итерационный процесс, соответствующий методу *градиентного спуска*

$$\begin{pmatrix} u_{k+1} \\ v_{k+1} \end{pmatrix} = \begin{pmatrix} u_k \\ v_k \end{pmatrix} - \tau_k \begin{pmatrix} \Phi'_u(u_k, v_k) \\ \Phi'_v(u_k, v_k) \end{pmatrix},$$

где τ_k — параметр, который выбирается, например, из условия минимальности $\Phi(u_{k+1}, v_{k+1})$ в данном направлении (метод наискорейшего спуска), $\{p_k, q_k\}$ — вектор, определяющий направление минимизации. На каждом шаге итераций решается задача минимизации Φ по одному аргументу.

5.4. Метод Чебышёва построения итерационных процессов высшего порядка

Предположим, что существует функция $g(u)$, обратная к $f(u)$. При этом $u = g[f(u)], U = g(0)$. Пусть, кроме того, $f(u)$ непрерывна и имеет необходимое число непрерывных производных на отрезке, внутри которого лежат все члены последовательности $\{u_k\}, k = 0, 1, \ldots$ Обратная функция имеет такое же количество непрерывных производных, как и $f(u)$. Разложим функцию $g(f[v]) = g(h)$ в ряд Тейлора в окрестности корня — точки $w = f(u)$.

$$g(h) \approx g(w) + \sum_{i=1}^{n} \frac{g^{(i)}(w)}{i!}(h-w)^i.$$

Тогда, учитывая, что $u = g[f(u)], w = f(u), h = f(v)$, получим

$$g(0) = U \approx u + \sum_{i=1}^{n} \frac{g^i[f(u)]}{i!}[-f(u)]^i + \ldots.$$

Можно показать, что итерационный метод

$$u_{k+1} = u_k + \sum_{i=1}^{n} (-1)^i \frac{g^{(i)}[f(u_k)]}{i!} [f(u_k)]^i, u^0 = a$$

имеет порядок сходимости $n+1$. Для вычисления производных обратной функции $u = g[f(u)]$ воспользуемся правилом дифференцирования сложной функции:

$$1 = g^{(1)}[f(u)] \cdot f^{(1)}(u),$$

$$0 = g^{(2)}[f(u)] \cdot [f^{(1)}(u)]^2 + g^{(1)}[f(u)] \cdot f^{(2)}_{(u)},$$

$$0 = g^{(3)}[f(u)] \cdot [f^{(1)}_{(u)}]^3 + 3g^{(2)}[f(u)] \cdot f^{(2)}_{(u)} \cdot f^{(1)}_{(u)} + g^{(1)}[f(u)] \cdot f^{(3)}_{(u)},$$

$$\ldots$$

5.5. Разностные отображения в нелинейной динамике

Рассмотрим последовательность чисел $u_{k+1} \in R$ (R — множество вещественных чисел), каждый член которой связан с предыдущим рекуррентным соотношением

$$u_{k+1} = f(u_k, u_{k-1}, \ldots, u_1, k), \tag{5.4}$$

где $k \in N$ (N — множество натуральных чисел). Соотношения (5.4) называются разностными отображениями (уравнениями) с дискретным аргументом.

Такие уравнения появляются при моделировании процессов, в которых величина u рассматривается через определенные промежутки времени. Например, еще в середине XIX века Ферхюльст для описания динамики популяционной системы предложил измерять ежегодно численность особей u_k, где k — номер года. Относительная численность u_{k+1} полагалась пропорциональной численности в k год, однако она начинает убывать, когда животных становится много (u_k сравнимо с 1):

$$u_{k+1} = f(u_k), \tag{5.5}$$

где

$$f(u_k) = \lambda u_k(1 - u_k), u_0 = a. \tag{5.6}$$

Другой пример из экономической области — задача о банковских сбережениях. Пусть u_0 — денежный вклад, растущий в соответствии с постоянным процентом δ, по закону:

$$u_{k+1} = (1+\delta)u_k = \ldots = (1+\delta)^k u_0.$$

Пусть далее законодательный орган, желая воспрепятствовать такому обогащению вкладчика, издает закон о том, чтобы процент убывал пропорционально u_k, т. е.

$$\delta_k = \delta_0 \left(1 - \frac{u_k}{u_{\max}}\right).$$

Тогда счет в банке изменился бы по закону

$$u_{k+1} = \left[1 + \delta_0 \left(1 - \frac{u_k}{u_{\max}}\right)\right] u_k, \tag{5.7}$$

т. е. в соответствии с моделью (5.6).

Так как $u_k \in [0,1]$, то $\lambda \in [0,4]$. Отображение (5.6) называется логистическим. К нему можно также придти, применив простейший из численных методов решения обыкновенных дифференциальных уравнений — явный метод Эйлера (лекция 8) для решения дифференциального уравнения динамики популяции (уравнения Ферхюльста)

$$\dot{u} = \lambda u(1-u), u(0) = a \tag{5.8}$$

где u — численность популяции. Вводя шаг по времени, получим разностный аналог уравнения (5.8):

$$\frac{u_{k+1} - u_k}{\tau} = \lambda u_k (1 - u_k), u_0 = a, \tag{5.9}$$

откуда получаем

$$u_{k+1} = \alpha u_k - \beta u_k^2, \tag{5.10}$$

$$\alpha = \lambda\tau + 1, \quad \beta = \lambda\tau.$$

И отображение (5.7), и отображение (5.10) легко приводится к виду (5.6). Достаточно произвести очевидную замену переменных. Для (5.10) эта замена будет $\mu = \alpha$, $z_k = \frac{\alpha}{\beta} u_k$. Как двумерное обобщение логистического отображения можно рассматривать отображение Хенона

$$u_{k+1} = 1 - \alpha u_k^2 + y_k, \quad y_{k+1} = \beta u_k, \quad |\beta| \leqslant 1, \tag{5.11}$$

или

$$u_{k+1} = 1 - \alpha u_k^2 + \beta u_{k-1}. \tag{5.12}$$

К довольно известным двумерным дискретным моделям относится также отображение Чирикова, предложенное для моделирования поведения незатухающего ротатора, возбуждаемого внешними толчками:

$$u_{k+1} = u_k - \alpha \sin y_k, y_{k+1} = y_k + u_{k+1}. \tag{5.13}$$

Рассмотрим подробнее свойства отображения:

$$u_{k+1} = \lambda u_k(1 - u_k), u_0 = a.$$

Заметим, что $f(0) = f(1) = 0$ и $\max f(u) = f(0,5) = \lambda/4$, то при $0 < \lambda < 4$ интервал $X = [0,1]$ отображается в себя, $u \in X$.

Введем обозначения

$$f^2 = f(f(u)), f^3 = f(f(f(u))), f^k = \underbrace{f(f \dots f(u) \dots)}_{k}.$$

Последовательность $f, f_2, \dots, f_k, \dots$ называется траекторией отображения и обозначается $\{f^k(u_0)\}_0^\infty$.

Определение. Точка $a \in X$ (X — множество, включающее в себя все значения отображения (5.4)) называется предельной точкой траектории $\{f^k(u_0)\}_{k=0}^\infty$, если существует последовательность $k_1 < k_2 < \dots < k_n \underset{n\to\infty}{\to} \infty$ такая, что $f^{k_n} \to a, n = 1, 2, \dots$.

Рассмотрим вначале случай $0 < \lambda < 1$. На $X = [0,1]$ существует только одна предельная (или неподвижная) точка $x = 0$. Любая последовательность, $\{f^k(u_0)\}_{k=0}^\infty$ сходится к предельной точке рассматриваемого отображения $x = 0$. Если рассматривается популяционная модель, то это означает, что рассматриваемая популяция не может выжить.

Из теоремы о сжимающем отображении следует, что последовательность $\{u_k\}_{n=0}^\infty$ сходится к своей предельной точке, если $|f'_u| \leqslant 1$.

В этом случае точка называется притягивающей. При выполнении условия $|f'_u| > 1$ точка называется отталкивающей.

Графическое изображение траектории (лесенка Ламерея) представлено на рис. 5.3.

Теперь рассмотрим случай $1 < \lambda < 3$.

В случае, когда $\lambda > 1$ — неподвижная точка, $u = 0$ становится отталкивающей, поскольку $|f'(0)| > 1$, а на отрезке [0, 1] появляется другая неподвижная точка $u_1 = 1 - \lambda^{-1}$.

Производная для рассматриваемого отображения $|f'(u_1)| = |2 - \lambda| < 1$. Точка u_1 при $1 < \lambda \leqslant 3$ является притягивающей.

Отметим, что при $1 < \lambda \leqslant 2$ производная $f'(u_1) > 0$ и траектория $\{f^k(u_0)\}_{k=1}^\infty$ стремится монотонно к u_1 (рис. 5.4); при $2 < \lambda \leqslant 3$ производная $f'(u_1) < 0$ и траектория приближается к u_1 немонотонно, попеременно принимая значения то меньше, то больше этого значения.

При $\lambda = 3$ точка u_1 остается притягивающей, но значение производной в этой точке является предельным: $|f'(u_1)| = 1$.

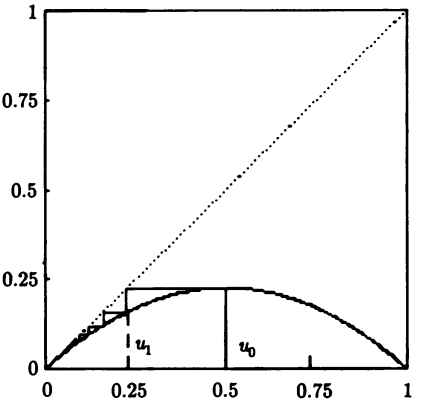


Рис. 5.3

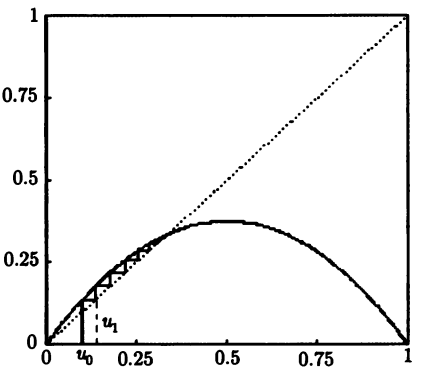


Рис. 5.4

При значениях параметра логистического отображения $\lambda = 1$ и $\lambda = 3$ неподвижная точка этого отображения теряет устойчивость и появляется либо другая устойчивая неподвижная точка, как это произошло в первом случае, либо притягивающий *цикл*; определение цикла будет дано ниже. Качественное изменение поведения решения (траектории отображения) при изменении параметра называется *бифуркацией*.

Пусть теперь $3 < \lambda \leqslant 1 + \sqrt{6}$. Как уже отмечалось, при значении параметра $\lambda = 3$ происходит *бифуркация*: неподвижная точка $u_2 = 1 -$

$-\lambda^{-1}$ из притягивающей превращается в отталкивающую: $|f'(u)| > 1$ при $\lambda > 3$. После того как точка стала отталкивающей, рассмотрим корни u_3, u_4 уравнения $f^2(u) = u$, или $\lambda^2 u^2 - \lambda(\lambda+1)u + (\lambda+1) = 0$.

Заметим, что если u_1 — предельная точка отображения $f(u) = u$, то она является также и предельной точкой отображения $f^2(u) = u$. Действительно, $f^2(u_1) = f(f(u_1)) = f(u_1) = u_1$, где u_1 — любая предельная точка рассматриваемого отображения, отличная от корней уравнения $f^2(u) = u$. Тогда, зная два корня уравнения $f^2(u) = u$ точки u_3, u_4 легко находятся как корни квадратного уравнения, они есть

$$u_{3,4} = \frac{(\lambda+1) \pm \sqrt{2\lambda - 3\lambda^2 - 3}}{2\lambda}.$$

Эти корни связаны соотношениями

$$f(u_3) = u_4, f(u_4) = u_3.$$

В данном случае говорят, что отображение имеет *цикл периода* 2, который будем обозначать P_2. Его наличие, например, в популяционной модели говорит об изменении численности особей с периодом в 2 единицы времени. Траектория для случая такого цикла изображена на рис. 5.5. Можно считать, что неподвижная (предельная) точка отображения есть цикл периода 1.

Переход от цикла P_1 (предельная точка логистического отображения) к циклу P_2 называют *бифуркацией рождения цикла (удвоения периода)*.

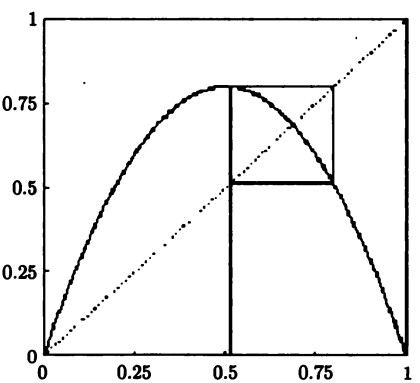


Рис. 5.5

Определение. Точка $a \in X$ называется периодической периода m, если $f^m(a) = a$ и $f^i(a) \neq a$ при $0 < i < m$.

Отметим, что каковы бы ни были попарно различные точки u_1, $u_2, \ldots, u_m$, если положить $f(u_i) = u_{i+1}$, $i = 1, 2, \ldots, m-1$ и $f(u_m) = u_1$, то рассматриваемое отображение будет иметь периодическую траекторию периода m: $u_1, u_2, \ldots, u_m, u_1, u_2, \ldots, u_m, \ldots$

Если к тому же $f(u)$ имеет первую производную, то в окрестности каждой из точек u_i выполнено

$$|f(u) - f(u_i)| \approx |f'(u_i)| \cdot |u - u_i|,$$

или

$$|f(u) - u_{i+1}| \approx |f'(u_i)| \cdot |u - u_i|.$$

Будем рассматривать $f^m(u)$ как сложную функцию. Пользуясь правилом дифференцирования сложной функции, получим

$$|f^m(u) - u_{i+1}| \approx \left| \prod_{k=1}^{m} f'(u_k) \right| \cdot |u - u_i|.$$

Если $\left| \prod_{k=1}^{m} f'(u_k) \right| < 1$, то траектория $\{f_k(u_0)\}_{k=0}^{\infty}$ приближается к циклу $\{u_1, \ldots, u_k\}$, или $\{u^k\}_{k=1}^{m}$. Такой цикл называется притягивающим циклом, а величина $\left| \prod_{k=1}^{m} f'(u_k) \right|$ — мультипликатором цикла. Цикл может быть как притягивающим, так и отталкивающим.

Определение. Цикл $P_m = \{u_1, \ldots, u_m\}$ отображения $f : X \to X$, переводящего множество X в себя, называется притягивающим, если существует число k_0, такое, что для любого $k > k_0$ траектория $\{f_k(u_0)\}_{k=0}^{\infty}$ распадается на m последовательностей, каждая из которых сходится к точкам $u_1, \ldots, u_m$ соответственно.

Достаточным условием существования притягивающего (отталкивающего) цикла является выполнение неравенства $\mu(P_m) > 1$, где $\mu(P_m) = \prod_{k=1}^{m} f'(u_k)$, $u \in P_m$ — мультипликатор цикла.

Отметим интересные свойства функции $f^2(u)$, в частности, ее график пересекается с прямой $y = u$ не только в неподвижных точках рассматриваемого отображения, u_1, u_2, но и в точках цикла P_2. Таким образом, можно сказать, что бифуркация рождения цикла обусловлена потерей устойчивости одной предельной точки и появлением двух устойчивых предельных точек отображения $f^2(u)$. На рис. 5.6а, в показано поведение функции $f^2(u)$ при разных значениях параметра λ ($\lambda = 2{,}8$; $\lambda = 1 + \sqrt{5}$).

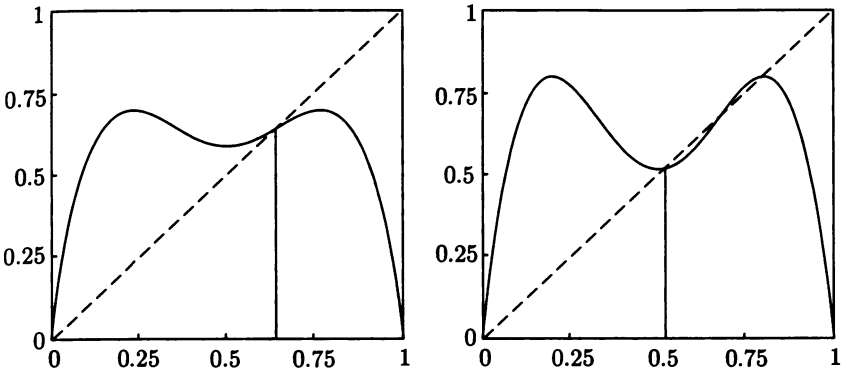


Рис. 5.6

При увеличении λ у отображения появляются новые неподвижные точки. Мультипликатор цикла P_2 вычисляется следующим образом:

$$\mu(u_3, u_4) = f'(u_3)f'(u_4) = \lambda^2(1-2u_3)(1-2u_4) = 4 + 2\lambda + \lambda^2.$$

Очевидно, что $|\mu(u_3, u_4)| < 1$, если $3 < \lambda < 1 + \sqrt{6}$, тогда цикл P_2 — притягивающий. Траектория $\{f_k(u_0)\}_{k=0}^{\infty}$ притягивается циклом $\{u_3, u_4\}$ и подпоследовательность $\{f^{2k}(u_0)\}_{k=0}^{\infty}$ сходится к одной точке цикла, а $\{f^{2k+1}(u_0)\}_{k=0}^{\infty}$ — к другой.

Знак мультипликатора дает информацию о характере приближения траектории к циклу. В частности, если $3 < \lambda < 1 + \sqrt{5}$, то подпоследовательности $\{f^{2k}(u_0)\}_{k=0}^{\infty}$ и $\{f^{2k+1}(u_0)\}_{k=0}^{\infty}$, начиная с некоторого u, являются монотонными, одна из них возрастающая, а другая — убывающая, что зависит от знаков $f'(u_3)$ и $f'(u_4)$.

При $1 + \sqrt{5} < \lambda \leqslant 1 + \sqrt{6}$ значение мультипликатора $\mu < 0$, и подпоследовательности $\{f^{2k}(u_0)\}_{k=0}^{\infty}$ и $\{f^{2k+1}(u_0)\}_{k=0}^{\infty}$ приближаются к точкам $\{u_3, u_4\}$ немонотонно.

Рассмотрим теперь случай $1 + \sqrt{6} \leqslant \lambda < 3,54\ldots$

При $\lambda = 1 + \sqrt{6}$ происходит вторая бифуркация удвоения периода.

Цикл $\{u_3, u_4\}$ из притягивающего превращается в отталкивающий, $|\mu(u_3, u_4)| > 1$ при $\lambda > 1 + \sqrt{6}$. Появляется новый притягивающий цикл P_4:

$$u_{4m} \to u_5, u_{4m+1} \to u_6, u_{4m+2} \to u_7, u_{4m+3} \to u_8,$$
$$u_6 = f(u_5), u_7 = f(u_6), u_8 = f(u_7), u_5 = f(u_8).$$

Для популяционной динамики это означает, что численность особей колеблется с периодом 4 единицы времени. Соответствующий график приведен на рис. 5.7.

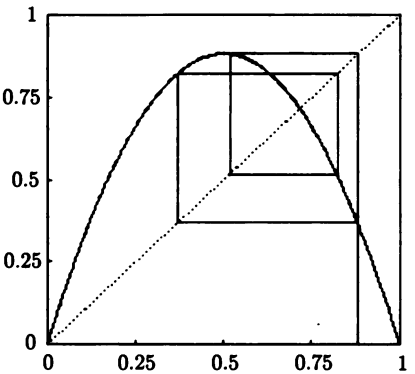


Рис. 5.7

При $\lambda \approx 3,54$ цикл P_4 периода 4 становится отталкивающим, $|\mu(u_5, \ldots, u_8)| > 1$; при этом появляются притягивающий цикл P_8 периода 8. Дальнейшее увеличение параметра λ будет приводить к появлению циклов P_{16}, P_{32} и т.д. Происходит *каскад бифуркаций удвоения периода.*

Заметим, что рассмотренный простой процесс имеет сложное поведение. Наблюдается каскад бифуркаций при увеличении величины λ; кроме того, все циклы, которые при этом встречаются, имеют период 2^p. Эта важнейшая закономерность, которая прослеживается не только в расчетах, но и в природе! Рассмотренные бифуркации при увеличении λ можно наглядно представить на бифуркационной диаграмме (рис. 5.8). Диаграмма получается, если обозначить через Λ_1, Λ_2 те значения λ, в которых происходят бифуркации, а через $\lambda_1, \lambda_2, \ldots$ при которых $u = 0,5$ является элементом циклов $P_2, P_4, \ldots$; по вертикальной оси откладываются значения предельных точек отображения. Обозначим за $d_1, d_2, \ldots$ величины, равные расстоянию между $x = 0,5$ и ближайшим к нему элементом цикла P_2 при $\lambda = \lambda_k$. Численный эксперимент показал, что Λ_k и λ_k при достаточно больших k ведут себя, как геометрическая прогрессия со знаменателем $\delta = 4,66920\ldots$, т. е.

$$\lim_{k\to\infty} \frac{\Lambda_{k+1} - \Lambda_k}{\Lambda_{k+2} - \Lambda_{k+1}} = \delta.$$

Отношение d_k/d_{k+1} имеет предел, равный $\alpha = 2,50290\ldots$.

Эти закономерности были замечены американским математиком Фейгенбаумом.

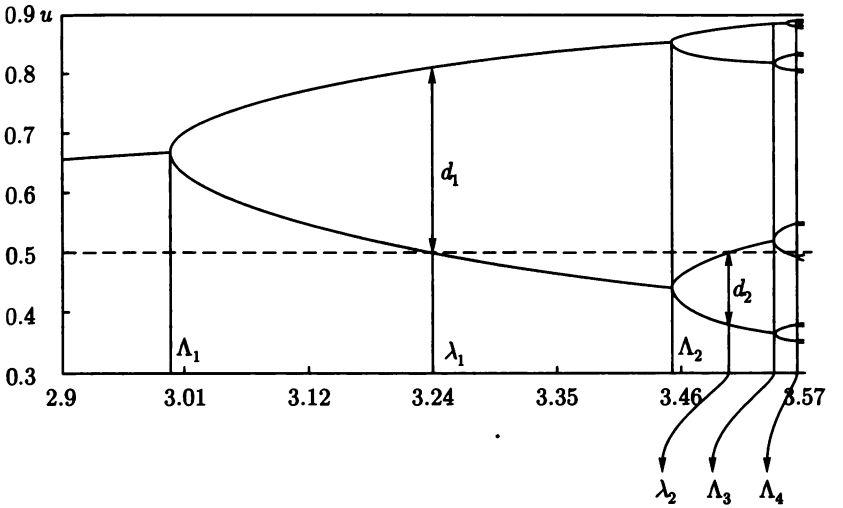


Рис. 5.8

При дальнейшем увеличении λ последовательность $\{u_k\}_{k=0}^{\infty}$ приобретает хаотический характер ($\lambda = \lambda_\infty \approx 3,569$), что видно на рис. 5.9.

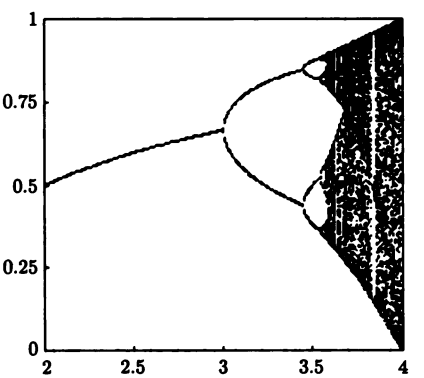


Рис. 5.9

Примечательно, что каскады Фейгенбаума имеют фрактальный характер (т. е. сохраняют подобие при изменении масштабов, рис. 5.10а, в).

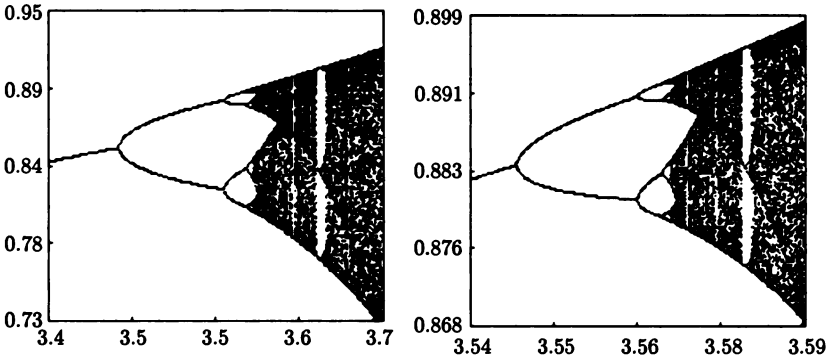


Рис. 5.10

Изучение графиков функций $f^2(u)$ и $f^1(u)$ показывает, что их фрагменты вблизи максимумов близки друг к другу, более того, они отличаются лишь масштабами. Оказывается, что такое же подобие имеет место для функции $f^{2^k}, n > 1$ при $\lambda = \lambda_k$, и выполняется тем точнее, чем больше n. Если положить $u' = u - 1/2$ (в дальнейшем штрих будем опускать) и считать α коэффициентом растяжения вдоль осей, то для некой симметричной функции $g(u)$, определенной на отрезке $[-1, 1]$, можно получить следующее функциональное уравнение:

$$g(u) = -\alpha g\left[g\left(-\frac{u}{\alpha}\right)\right],$$

которое универсально определяет α:

$$g(0) = -\alpha g(g(0)).$$

Вблизи максимума $g(x)$ должна быть близка к квадратичной параболе, причем $g(0) = 1$. В *теории универсальности* показывается, что эта функция вычисляется с помощью ряда

$$g(u) = 1 - 1,52763u^2 + 0,104815u^4 - 0,0267057u^6 + \ldots$$

Пусть теперь $\lambda = 3,83$. В этом случае из хаотической области, изображенной на рис. 5.10, появляется устойчивый цикл P_3 (рис. 5.11a, b представляют циклы в последовательные моменты времени).

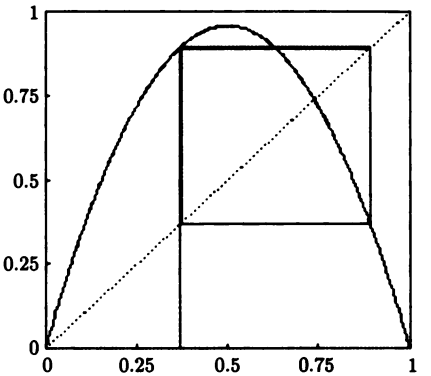


Рис. 5.11 a

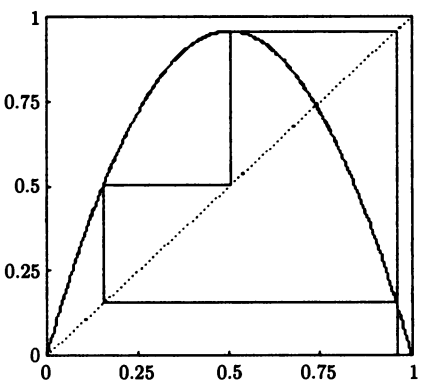


Рис. 5.11 b

Циклу на рис. 5.11 соответствует самое большое окно устойчивых циклов $P_{3\cdot 2^k}$. Чередование хаотических и регулярных зон — называется перемежаемостью. Возможно, нечто подобное наблюдается в гидродинамических потоках, где ламинарные зоны чередуются с турбулентными.

5.6. Задачи

1. Методами простых итераций и Ньютона решить уравнение

$$x^2 - e^{-x} = 0, x_0 \in [0.5,\ 1]\,.$$

Решение. Метод простых итераций будет $x_{k+1} = e^{-\frac{x_k}{2}}, x_0 = 0,75$, $F(x) = e^{-\frac{x_k}{2}}, |F'_x(x)| < 1$ при $x \in [0.5,\ 1]$.

Приведем таблицу приближений до точности $|x_5 - x_4| < 10^{-4}$.

k	0	1	2	3	4	5
x_k	0,75	0,6873	0,7091	0,7015	0,7042	0,7032

Метод Ньютона запишется как

$$x_{k+1} = x_k - \frac{(x_k)^2 - e^{-x_k}}{2x_k - e^{-x_k}}, x_0 = 1.$$

Результаты расчетов:

k	0	1	2	3
x_k	1	0,7330	0,7038	0,7035

2. Построить итерационные методы для вычисления корней уравнения

$$x^3 + 3x^2 - 1 = 0.$$

Использовать метод простых итераций.

Решение. Все три корня данного уравнения лежат на отрезках $[-3, -2]$, $[-1, 0]$, $[0, 1]$.

Построим итерационный процесс для вычисления первого корня, лежащего на отрезке $[-3, -2]$:

$$x_{k+1} = x_k^{-2} - 3, F(x) = x^{-2} - 3, |F'_x(x)| = \left|-2x^{-3}\right| \leqslant \frac{1}{4} < 1,$$

т. е. для начального приближения $x_0 \in [-3, -2]$ метод простых итераций сходится.

Для вычисления двух оставшихся корней, лежащих на отрезках $[-1, 0]$ и $[0, 1]$, построим итерационный метод

$$x_{k+1} = \pm(x_k + 3)^{-\frac{1}{2}}, F(x) = \pm(x_k + 3)^{-\frac{1}{2}}.$$

Поскольку для рассматриваемых отрезков $|F'_x(x)| = \left|2(x+3)^{-\frac{1}{2}}\right|^3 < 1$, то этот итерационный процесс сходится к корням рассматриваемого уравнения.

3. Исследовать сходимость итерационного метода релаксации для численного решения уравнения вида $f(x) = 0$:

$$x_{k+1} = x_k - \tau \quad f(x_k), \quad x_0 = a, \quad f'(x) > 0.$$

Решение. Итерационный метод релаксации записывается как

$$x_{k+1} = F(x_k), x_0 = 0,$$

$$F(x) = x - \tau f(x).$$

Из условия того, что отображение является сжимающим, получим $|F(x) - F(y)| = |x - y - \tau(f(x) - f(y))| \leqslant q|x - y|$, где $q = \max|1 - \tau f'(x + \alpha(x - y))|, 0 \leqslant \alpha \leqslant 1$.

Положим $0 < f'_{\min} \leqslant f' \leqslant f'_{\max}$, тогда $q(\tau) = \max\{|1 - \tau f'_{\min}|, |1 - \tau f'_{\max}|\}$.

Отсюда видно, что итерационный процесс будет сходиться, если $|1 - \tau f'_{\max}| < 1$, или $\tau < 2/f'_{\max}$, а оптимальное значение параметра, при котором $q = q_{\min}$ достигается при $|1 - \tau f'_{\min}| = |1 - \tau f'_{\max}|$, или $\tau = \tau_0 = 2(f'_{\min} + f'_{\max})^{-1}, q(\tau_0) = (1 - f'_{\min}/f'_{\max})(1 + f'_{\min}/f'_{\max})$.

4. Построить метод простых итераций и итерационный процесс Ньютона для системы уравнений

$$f(x, y) = x + 3\lg x - y^2 = 0,$$

$$g(x, y) = 2x^2 - xy - 5x + 1 = 0.$$

Решение. Графики функций $f(x, y)$ и $g(x, y)$ приведены на рис. 5.12.

Метод простых итераций запишем как

$$x_{k+1} = y_k^2 - 3\lg x_k,$$

$$y_{k+1} = 2x_k + \frac{1}{x_k} - 5,$$

где $x_0 = 3,4, y_0 = 2,2, F(x, y) = y^2 - 3\lg x; G(x, y) = 2x + \frac{1}{x} - 5$.

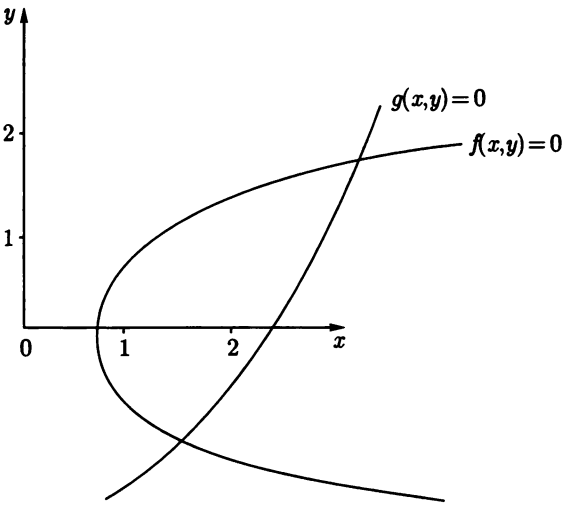


Рис. 5.12

Матрица Якоби для такого процесса запишется как

$$\mathbf{J} = \begin{pmatrix} \frac{\partial F}{\partial x} & \frac{\partial F}{\partial y} \\ \frac{\partial G}{\partial x} & \frac{\partial G}{\partial y} \end{pmatrix} = \begin{pmatrix} -\frac{3\lg e}{x} & 2y \\ 2 - x^{-2} & 0 \end{pmatrix}.$$

Несложно проверить, вычислив норму матрицы Якоби, что для приведенного начального приближения достаточное условие сходимости не выполняется.

Рассмотрим другой итерационный процесс:

$$x_{k+1} = \left(\frac{x_k (y_k + 5) - 1}{2} \right)^{1/2},$$

$$y_{k+1} = (x_k + 3\lg x_k)^{1/2},$$

$$F(x,y) = \left(\frac{x_k (y_k + 5) - 1}{2} \right)^{1/2}, G(x,y) = (x + 3\lg x)^{1/2}.$$

Матрица Якоби для этого итерационного метода будет

$$\mathbf{J} = \begin{pmatrix} \frac{5+y}{2\sqrt{2}\sqrt{x(y+5)-1}} & \frac{x}{2\sqrt{2}\sqrt{x(y+5)-1}} \\ \frac{1+\frac{3\lg e}{x}}{2\sqrt{x+3\lg x}} & 0 \end{pmatrix}.$$

В окрестности начального приближения условие сходимости выполнено. Таблица первых пяти приближений будет

k	0	1	2	3	4	5
x	3,4	3,426	3,451	3,466	3,475	3,480
y	2,2	2,243	2,2505	2,255	2,258	2,259

5. Построить итерационный метод Ньютона для вычисления $\sqrt[n]{a}$, $a>0$, $n \in R$.

Решение. Найдем корень уравнения

$$f(x) = x^n - a = 0.$$

Метод Ньютона для этого уравнения запишется

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = x_k - \frac{x_n^k - a}{n x_m^{p-1}} =$$

$$= \frac{n-1}{n} x_k + \frac{a}{n x_k^{n-1}} = \frac{1}{n}\left[(n-1)\, x_k + \frac{a}{x_k^{n-1}}\right].$$

В частности, при $n = 2$ имеем

$$x_{k+1} = \frac{1}{2}\left(x_k + \frac{a}{x_k}\right).$$

6. Получить расчетные формулы метода Ньютона для численного решения системы двух нелинейных уравнений $f(x, y) = 0, g(x, y) = 0$.

Решение. Положим $x_{k+1} = \Delta x_k$, $y_{k+1} = y_k + \Delta y_k$, $f(x_k, y_k) = f^k$, получим

$$f(x_k, y_k) + \left(\frac{\partial f}{\partial x}\right)^k \Delta x_k + \left(\frac{\partial f}{\partial y}\right)^k \Delta y_k = 0,$$

$$g(x_k, y_k) + \left(\frac{\partial g}{\partial x}\right) \Delta x_k + \left(\frac{\partial g}{\partial y}\right) \Delta y_k = 0,$$

откуда

$$\Delta x_k = \frac{|\mathbf{X}_k|}{|\mathbf{J}_k|} = \begin{vmatrix} \left(\frac{\partial f}{\partial x}\right)_k & \left(\frac{\partial f}{\partial y}\right)_k \\ \left(\frac{\partial g}{\partial x}\right)_k & \left(\frac{\partial g}{\partial y}\right)_k \end{vmatrix}^{-1}.$$

$$\cdot \begin{vmatrix} -f_k & \left(\frac{\partial f}{\partial y}\right)_k \\ -g_k & \left(\frac{\partial g}{\partial y}\right)_k \end{vmatrix} = \frac{-f_k\left(\frac{\partial g}{\partial y}\right)_k + g_k\left(\frac{\partial f}{\partial y}\right)_k}{\left(\frac{\partial f}{\partial x}\right)_k\left(\frac{\partial g}{\partial y}\right)_k - \left(\frac{\partial g}{\partial x}\right)_k\left(\frac{\partial f}{\partial y}\right)_k},$$

$$\Delta y_k = \frac{|\mathbf{Y}_k|}{|\mathbf{J}_k|} = \begin{vmatrix} \left(\frac{\partial f}{\partial x}\right)_k & \left(\frac{\partial f}{\partial y}\right)_k \\ \left(\frac{\partial g}{\partial x}\right)_k & \left(\frac{\partial g}{\partial y}\right)_k \end{vmatrix}^{-1} \cdot$$

$$\cdot \begin{vmatrix} \left(\frac{\partial f}{\partial x}\right)_k & -f_k \\ \left(\frac{\partial g}{\partial x}\right)_k & -g_k \end{vmatrix} = \frac{-g_k\left(\frac{\partial g}{\partial x}\right)_k + f_k\left(\frac{\partial f}{\partial x}\right)_k}{\left(\frac{\partial f}{\partial x}\right)_k\left(\frac{\partial g}{\partial y}\right)_k - \left(\frac{\partial g}{\partial x}\right)_k\left(\frac{\partial f}{\partial y}\right)_k},$$

где $\mathbf{J}_k = \begin{pmatrix} \left(\frac{\partial f}{\partial x}\right)_k & \left(\frac{\partial f}{\partial y}\right)_k \\ \left(\frac{\partial g}{\partial x}\right)_k & \left(\frac{\partial g}{\partial y}\right)_k \end{pmatrix}$ — матрица Якоби, $\mathbf{X}_k$, $\mathbf{Y}_k$ — матрицы

$$\mathbf{X}_k = \begin{pmatrix} -f_k & \left(\frac{\partial f}{\partial x}\right)_k \\ -g_k & \left(\frac{\partial f}{\partial y}\right)_k \end{pmatrix}, \mathbf{Y}_k = \begin{pmatrix} \left(\frac{\partial f}{\partial x}\right)_k & -f_k \\ \left(\frac{\partial f}{\partial y}\right)_k & -g_k \end{pmatrix}.$$

Тогда запишем расчетные формулы для итерационного метода Ньютона:

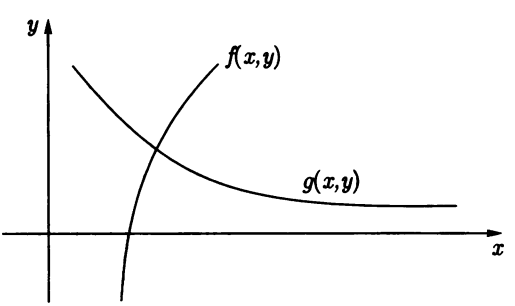
$$x_{k+1} = x_k + \Delta x_k, y_{k+1} = y_k + \Delta y_k.$$

7. Найти приближенное решение системы двух нелинейных алгебраических уравнений $x^3 - y^3 - 1 = 0, xy^3 - y - 4 = 0$, используя метод Ньютона.

Решение.

Пусть $f(x,y) = x^3 - y^3 - 1, g(x,y) = xy^3 - y - 4$.

Начальное приближение можно найти графически: $x_0 = y_0 = 1,5$.



Матрицы $\mathbf{J}_k, \mathbf{X}_k, \mathbf{Y}_k$ в этом случае будут

$$\mathbf{J}_k = -\begin{pmatrix} 3x_k^2 & -2y_k \\ y_k^3 & 3x_k y_k^2 - 1 \end{pmatrix}, \mathbf{X}_k = -\begin{pmatrix} x_k^3 - y_k^3 - 1 & -2y_k \\ x_k y_k^3 - y_k - 4 & 3x_k y_k^2 - 1 \end{pmatrix},$$

$$\mathbf{Y}_k = -\begin{pmatrix} 3x_k^2 & x_k^3 - y_k^3 - 1 \\ y_n^3 & x_n y_n^3 - y_n - 4 \end{pmatrix}.$$

Следующее приближение вычисляется по формуле Ньютона:

$$x_{k+1} = x_k + \frac{|\mathbf{X}_k|}{|\mathbf{J}_k|}, y_{k+1} = y_k + \frac{|\mathbf{Y}_k|}{|\mathbf{J}_k|}.$$

Результаты вычислений первых двух итераций приведены в таблице (точность $\varepsilon \approx 10^{-3}$).

k	$x_k; y_k$	$f_k; g_k$	$\lvert\mathbf{J}_k\rvert$	$\lvert\mathbf{X}_k\rvert$	$\lvert\mathbf{Y}_k\rvert$
0	1,5; 1,5	0,12500 -4,33750	71,71875	-0,171875	-3,3750
1	1,502397 1,547059	-0,002170 0,015844	77,73277	0,0277988	0,1153255
2	1,5020396 1,545570	0,0000017 0,000019			

8. . Найти приближенное решение системы трех нелинейных уравнений $x^2 + y^2 + z^2 = 1, 2x^2 + y^2 - 4z = 0, 3x^2 - 4y + z^2 = 0$, используя метод Ньютона. За начальное приближение решения выбрать точку $x_0 = y_0 = z_0 = 0,5$.

Решение. Обозначим $\mathbf{F} = (x^2+y^2+z^2-1, 2x^2+y^2-4z, 3x^2-4y+z^2)^{\mathrm{T}}$, тогда матрица Якоби будет $\mathbf{J} = \begin{pmatrix} 2x & 2y & 2z \\ 4x & 2y & -4 \\ 6x & -4 & 2z \end{pmatrix}$.

Соответственно, для точки начального приближения вычислим

$$\mathbf{F}_0 = \begin{pmatrix} 0,25 + 0,25 + 0,25 - 1 \\ 0,5 + 0,25 - 2,00 \\ 0,75 - 2,00 + 0,25 \end{pmatrix} = \begin{pmatrix} -0,25 \\ -1,25 \\ -1,00 \end{pmatrix},$$

$$\mathbf{J}_0 = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & -4 \\ 3 & -4 & 1 \end{pmatrix}, \det \mathbf{J}_0 = -40 \neq 0,$$

$$\mathbf{J}_0^{-1} = -\frac{1}{40}\begin{pmatrix} -15 & -5 & -5 \\ -14 & -2 & 6 \\ -11 & 7 & -1 \end{pmatrix}.$$

Вычислим первое приближение

$$\mathbf{x}_1 = \mathbf{x}_0 - \mathbf{J}_0^{-1}\mathbf{F}_0 = \begin{pmatrix} 0,5 \\ 0,5 \\ 0,5 \end{pmatrix} -$$

$$-\frac{1}{40}\begin{pmatrix} -15 & -5 & -5 \\ -14 & -2 & 6 \\ -11 & 7 & -1 \end{pmatrix}\begin{pmatrix} -0,25 \\ -1,25 \\ -1,00 \end{pmatrix} = \begin{pmatrix} 0,875 \\ 0,500 \\ 0375 \end{pmatrix}.$$

Для второго приближения

$$\mathbf{F}_1 = \begin{pmatrix} 0,875^2 + 0,500^2 + 0,375^2 - 1 \\ 2\cdot 0,875^2 + 0,500^2 - 4\cdot 0,375 \\ 3\cdot 0,875^2 - 4\cdot 0,500 + 0,375^2 \end{pmatrix} = \begin{pmatrix} 0,15625 \\ 0,28125 \\ 0,43750 \end{pmatrix},$$

$$\mathbf{J}_1 = \begin{pmatrix} 2\cdot 0,875 & 2\cdot 0,500 & 2\cdot 0,375 \\ 4\cdot 0,875 & 2\cdot 0,500 & -4 \\ 6\cdot 0,875 & -4 & 2\cdot 0,375 \end{pmatrix} = \begin{pmatrix} 1,750 & 1 & 0,750 \\ 3,500 & 1 & -4 \\ 5,250 & -4 & 0,75 \end{pmatrix},$$

$$\det \mathbf{J}_1 = -64,75,$$

$$\mathbf{J}_1^{-1} = -\frac{1}{64,75}\begin{pmatrix} -15,25 & -3,75 & -4,75 \\ -23,625 & -2,625 & 9,625 \\ -19,25 & 12,25 & -1,75 \end{pmatrix}.$$

По формуле Ньютона получим

$$\mathbf{x}_2 = \mathbf{x}_1 - \mathbf{J}_1^{-1}\mathbf{F}_1 = \begin{pmatrix} 0,875 \\ 0,500 \\ 0,375 \end{pmatrix} +$$

$$+\frac{1}{64,75}\begin{pmatrix} -15,25 & -3,75 & -4,75 \\ -23,625 & -2,625 & 9,625 \\ -19,25 & 12,25 & -1,75 \end{pmatrix}\cdot\begin{pmatrix} 0,15625 \\ 0,28125 \\ 0,43750 \end{pmatrix} = \begin{pmatrix} 0,78981 \\ 0,49662 \\ 0,36993 \end{pmatrix}.$$

Аналогично находим третье приближение, которым и ограничимся:

$$\mathbf{x}_3 = \begin{pmatrix} 0,78521 \\ 0,49662 \\ 0,36992 \end{pmatrix}.$$

5.7. Задачи для самостоятельного решения

1. Исследовать возможность применения трех итерационных процессов

$$x_{k+1} = -e^{-x_k},$$

$$x_{k+1} = -\ln x_k,$$

$$x_{k+1} = \frac{1}{2}(x_k + e^{-x_k}).$$

для численного решения уравнения $x + \ln x = 0$, имеющего корень $\bar{x} \approx 0,6$.

2. Исследовать сходимость процесса простой итерации в зависимости от начального приближения для численного решения уравнения

$$x - 2^{x-1} = 0,$$

имеющего корни $x_1 = 1, x_2 = 2$.

3. Предложить процессы простых итераций для решения уравнений

$$x - \frac{\cos x}{2} = 0,$$

$$x = \ln(x + 2) \quad \text{(уравнение имеет два корня)},$$

$$e^{-x} = \cos x \quad \text{(для поиска ближайшего к нулю корня)}.$$

4. Предложить итерационный процесс Ньютона для вычисления решений систем уравнений

$$\begin{cases} x^{10} + y^{10} = 1024, \\ e^x - e^y = 1. \end{cases}$$

$$\begin{cases} \sin(x + 1) - y = 1,2, \\ 2x + \cos y = 2. \end{cases}$$

5. Предложить методы простых итераций и Ньютона для численного решения нелинейных уравнений:

$$e^x - \frac{1}{x} = 0,$$

$$x^2 - 20\sin x = 0,$$

$$x\,2^x - 1 = 0,$$

$$\sqrt{x+1} - \frac{1}{x} = 0,$$

$$\operatorname{arctg}(x - 1) + 2x = 0.$$

6. С помощью методов дихотомии и Ньютона найти точку локального минимума функции $f(t) = e^{-t} + t^3 - t$, с точностью $\varepsilon = 10^{-6}$.

Литература

[1] *Колмогоров А.Н., Фомин С.В.* Элементы теории функций и функционального анализа. М.: Наука, 1981.

[2] *Бабенко К.И.* Основы численного анализа. М.: Наука, 1986. 744 с.

[3] *Коллатц Л.* Функциональный анализ и вычислительная математика. М.: Мир, 1969. 448 с.

[4] *Федоренко Р.П.* Введение в вычислительную физику. М.: Изд-во МФТИ, 1994. 526 с.

[5] *Н. В. Бахвалов, Н. П. Жидков, Г. М. Кобельков.* Численные методы - М: Лаборатория Базовых Знаний, 2002. - 632 с.

[6] *Лобанов А.И., Петров И.Б.* Вычислительные методы для анализа моделей сложных динамических систем. Ч. 1. М.: МФТИ, 2004. 168 с.

[7] *Вержбицкий В.М.* Численные методы. М.: Высшая школа, 2005. 866 с.

[8] *Шарковский А.Н., Майстренко Ю.А., Романенко Е.Ю.* Разностные уравнения и их приложения. Киев: Наукова думка, 1986. 279 с.

[9] *Ахромеева Т.С., Курдюмов С.П., Малинецкий Г.Г., Самарский А.А.* Нестационарные структуры и диффузионный хаос. М.: Наука, 1992. 541 с.

Лекция 6. Интерполяция функций

Рассматривается задача алгебраической интерполяции. Обусловленность задачи исследуется на основе рассмотрения константы Лебега. Доказывается теорема об остаточном члене интерполяции. Выводятся формулы алгебраической интерполяции с кратными узлами. Рассматривается задача гладкого восполнения функции (локальными и нелокальными сплайнами), а также естественный базис в пространстве сплайн-функций — B-сплайны.

Ключевые слова: интерполяция, обобщенный полином, метод неопределенных коэффициентов, интерполяционный полином в формах Лагранжа и Ньютона, постоянная Лебега, кусочно-линейная интерполяция, сплайны, B-сплайны.

6.1. Постановка задачи интерполяции

Пусть задана совокупность *узлов интерполяции* или *сетка* на некотором отрезке $[a, b]$. В простейшем случае сетка — равномерная, т. е. расстояние между соседними узлами одинаково. В дальнейшем также рассмотрим неравномерные сетки.

1. Совокупность узлов $\{t_n\}_{n=0}^{N}$, $t_n = a + n\tau, \tau = (b-a)/N, t \in [a,\ b]$.

2. Сеточная проекция функции $f(t)$ на $[a, b]$, т. е. таблица $f_n = \{f(t_n)\}_{n=0}^{N}$; эту таблицу задает оператор ограничения на сетку или рестрикции (от английского *restriction*) R.

Задача состоит в том, чтобы по таблице f_n восстановить непрерывную функцию. Обозначим ее через $F(t)$. Разумеется, она отличается от исходной функции $f(t)$, причем такое восстановление неоднозначно и осуществляется оператором интерполяции I. Сама функция $F(t)$ называется интерполирующей или интерполянтом. Необходимо оценить потерю информации при действии этого оператора, т. е. величину $|f(t) - F(t)|$, зависящую от типа оператора интерполяции и свойств $f(t)$, в частности, ее гладкости. Таким образом, имеем схему:

$$f(t) \underset{\mathrm{R}}{\to} \{f_n\}_{n=0}^{N} \underset{\mathrm{I}}{\to} F(t).$$

6.2. Кусочно-линейная интерполяция

Простейший способ интерполяции — кусочно-линейная, требующая минимальных требований на гладкость функции $f(t)$. При таком способе интерполяции соседние точки (t_n, f_n) и (t_{n+1}, f_{n+1}) соединяют отрезками прямых

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

Теорема. *Пусть $f(t)$ — Липшиц непрерывная функция, т. е. $|f(t_1) - f(t_2)| \leqslant c\,|t_1 - t_2|$, тогда $|f(t) - F(t)| \leqslant c\frac{\tau}{2}$.*

Примечание. Если сетка неравномерная и $\tau = \max\limits_n(t_{n+1} - t_n)$, то теорема верна и для этого случая.

Доказательство.

Пусть $t \in [t_n, t_{n+1}]$, обозначим $\tau = t_{n+1} - t_n$. Тогда $t = t_n + \alpha \cdot \tau$; $0 \leqslant \alpha \leqslant 1$. В силу линейности $F(t)$ имеем равенство $F(t) = \alpha f_{n+1} + (1 - \alpha) f_n$.

Оценим разность

$$|F(t) - f(t)| = |\alpha f_{n+1} + (1 - \alpha) f_n - \alpha f(t) - (1 - \alpha) f(t)| \leqslant$$

$$\leqslant \alpha\,|f_{n+1} - f(t)\,| + (1 - \alpha)|\,f_n - f(t)\,|.$$

Поскольку $f_{n+1} = f(t_n + \tau)$, имеем

$$|f_{n+1} - f(t)| = |f(t_n + \tau) - f(t_n + \alpha\tau)| \leqslant |c(1 - \alpha)\tau| =$$

$$= c(1 - \alpha)\tau, \quad \text{т. к.} \quad 0 \leqslant \alpha \leqslant 1.$$

Аналогично $|f_n - f(t)| \leqslant c\alpha\tau$. В таком случае $|f(t) - F(t)| \leqslant 2\alpha(1 - \alpha)c\tau \leqslant c\tau/2$. ■

Замечание. Простой аппарат кусочно-линейной интерполяции позволяет ввести объекты, на которых базируется один из наиболее известных современных численных методов — метод конечных элементов. Сетке t_n ставится в соответствие набор базисных функций $\varphi_n(t)$, каждая из которых сопоставляется своему узлу t_n, причем $\varphi_n(t_k) = \delta_k^n, \varphi_n(t_{n-1}) = \varphi_n(t_{n+1}) = 0, \varphi_n(t_n) = 1$, а в остальных точках она вычисляется с помощью кусочно-линейной интерполяции.

Функция $F(t)$ в этом случае представляется в виде

$$F(t) = \sum_{n=0}^{N} f_n \varphi_n(t).$$

В вычислительной математике часто используется кусочно - полиномиальная интерполяция. Так, эрмитовым кубическим интерполянтом называется кусочно-кубический интерполянт с непрерывной производной, кубическим сплайном называется кусочно-кубический интерполянт с двумя непрерывными производными. О сплайнах речь пойдет ниже.

6.3. Интерполяция обобщенными полиномами

Для того чтобы функция (обобщенный полином) $F(t) = \sum_{n=0}^{N} u_n \varphi_n(t)$ была интерполирующей, необходимо выполнение условий: $F(t_k) = f_k$, $k = 0 \div N$, где f_k — значения функции в точках интерполяции. Для коэффициентов обобщенного полинома получаем систему уравнений:

$$\begin{cases} u_0 \cdot \varphi_0(t_0) + u_1 \cdot \varphi_1(t_0) + \ldots + u_N \cdot \varphi_N(t_0) = f_0, \\ u_0 \cdot \varphi_0(t_1) + u_1 \cdot \varphi_1(t_1) + \ldots + u_N \cdot \varphi_N(t_1) = f_1, \\ \ldots \\ u_0 \cdot \varphi_0(t_N) + u_1 \cdot \varphi_1(t_N) + \ldots + u_N \cdot \varphi_N(t_N) = f_N, \end{cases}$$

или в векторной форме

$$\mathbf{Au} = \mathbf{f},$$

где

$$\mathbf{A} = \begin{pmatrix} \varphi_0(t_0) & \varphi_1(t_0) & \ldots & \varphi_N(t_0) \\ \varphi_0(t_1) & \varphi(t_1) & \ldots & \varphi_N(t_1) \\ \ldots & \ldots & \ldots & \ldots \\ \varphi_0(t_N) & \varphi_1(t_N) & \ldots & \varphi_N(t_N) \end{pmatrix}, \mathbf{u} = \begin{pmatrix} u_0 & u_1 & \ldots & u_N \end{pmatrix}^T,$$

$$\mathbf{f} = \begin{pmatrix} f_0 & f_1 & \ldots & f_N \end{pmatrix}^T.$$

Теорема (доказывается в курсе линейной алгебры). *Для того чтобы решение задачи интерполяции существовало и было единственным, необходимо и достаточно, чтобы система базисных функций $\varphi_n(t_k)$ была линейно независима.*

Теорема (доказывается в курсе линейной алгебры). *Для того чтобы система функций $\varphi_n(t_k)$ была линейной независимой в точках $t_0, \ldots, t_n$, необходимо и достаточно, чтобы определитель матрицы Грамма*

$$\mathbf{C} = \mathbf{A}^*\mathbf{A} = \begin{pmatrix} (\varphi_0, \varphi_0) & (\varphi_0, \varphi_1) & \ldots & (\varphi_0, \varphi_N) \\ (\varphi_1, \varphi_0) & (\varphi_1, \varphi_1) & \ldots & (\varphi_1, \varphi_N) \\ \ldots & \ldots & \ldots & \ldots \\ (\varphi_N, \varphi_0) & (\varphi_N, \varphi_1) & \ldots & (\varphi_N, \varphi_N) \end{pmatrix},$$

был отличен от нуля. Здесь каждый элемент матрицы Грамма имеет вид

$$\gamma_{jk} = (\varphi_k, \varphi_j) = \sum_{i=0}^{N} \varphi_k(t_i) \cdot \varphi_j(t_i).$$

В случае, если система функций $\{\varphi_j\}_0^N$ ортогональна на множестве точек $\{t_j\}_0^N$, решение задачи интерполяции значительно упрощается (напомним, что система функций $\{\varphi_j\}_0^N$ является ортогональной на множестве точек $\{t_j\}_0^N$, если $(\varphi_k, \varphi_j) = 0$ при $k \neq j$ и $(\varphi_k, \varphi_j) \neq 0$ при $k = j$ для всех $k = 0,\ 1, \ldots, N;\ j = 0, 1, \ldots, n$).

Дело в том, что матрица Грамма для ортогональной системы функций диагональна, и ее определитель отличен от нуля (всякая ортогональная система функций заведомо линейно независима). Линейная система уравнений представляется как $\mathbf{A} * \mathbf{Au} = \mathbf{A} * \mathbf{f}$, или $\mathbf{Cu} = \mathbf{b}$, где $\mathbf{C} = \mathbf{A} * \mathbf{A}$, $\mathbf{b} = \mathbf{A} * \mathbf{f}$ — вектор, а ее решение в случае $\mathbf{A} * \mathbf{A} = \mathbf{E}$ есть $\mathbf{u} = \mathbf{A} * \mathbf{f}$.

Примером ортогональной системы являются показательные функции $e^{2\pi i k t_j}$ на множестве точек $t_j = j/N, j = 0, 1, \ldots, N$ (на отрезке $[0, 1]$).

6.4. Полиномиальная (алгебраическая) интерполяция

В этом случае ($u_k(t) = t^k$) СЛАУ для определения коэффициентов имеет вид

$$\begin{cases} u_0 + u_1 t_0 + \ldots + u_N t_0^N = f_0, \\ u_0 + u_1 t_1 + \ldots + u_N t_1^N = f_1, \\ \ldots \\ u_0 + u_1 t_N + \ldots + u_N t_N^N = f_N, \end{cases}$$

а ее определитель

$$\det \begin{pmatrix} 1 & t_0 & t_0^2 & \ldots \\ 1 & t_1 & t_1^2 & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ 1 & t_N & t_N^2 & \ldots \end{pmatrix} = \prod_{i \neq j}^{N} (t_i - t_j), 0 \leqslant j < i \leqslant N,$$

отличен от нуля, если узлы интерполяции попарно различны. Это известный из курса линейной алгебры определитель Вандермонда.

Ответ на вопрос о существовании и единственности решения СЛАУ оказывается — утвердительным - решение задачи алгебраической интерполяции всегда существует и единственно, но при больших N система

оказывается плохо обусловленной. Однако решение этой задачи можно выписать в явном виде

$$L_N(t) = \sum_{n=0}^{N} f_n \cdot \varphi_n^N(t),$$

где $\varphi_n^N(t) = \prod\limits_{\substack{i=0 \\ i \neq n}}^{N} \frac{t-t_j}{t_n-t_i}$ — базисные функции, являющиеся полиномами степени N, каждый из которых сопоставлен со своим узлом сетки так, что $\varphi_n^N(t_k) = \delta_k^n$. Заметим, что правильнее было бы писать $L_N(t, \{t_n\}, \{f_n\})$, т. е. интерполянт зависит от t, сетки и сеточной функции. Такой вид записи алгебраического интерполяционного полинома не единственен. Выписанный полином называется интерполяционным полиномом *в форме Лагранжа*. Он удобен для теоретического рассмотрения, но на практике часто оказывается более удобной другая форма представления — полином *в форме Ньютона*, о котором речь пойдет ниже.

6.5. Теорема об остаточном члене интерполяции

Введем понятие остаточного члена интерполяции для оценки погрешности

$$R_N(t) = f(t) - L_N(t). \tag{6.1}$$

Теорема. *Пусть функция $f(t)$ имеет на отрезке $[a,b]$ $---$ $N+1$ ограниченную производную. Тогда $R_N(t) = \frac{1}{(N+1)!} \prod\limits_{j=0}^{N} (t-t_j) \cdot f^{(N+1)}(\xi)$, где $\xi \in [a,b]$.*

Доказательство.

Рассмотрим функцию

$$\psi(x) = f(x) - L_N(x) - R_N(t)\frac{(x-t_0)(x-t_1)\dots(x-t_N)}{(t-t_0)(t-t_1)\dots(t-t_N)},$$

имеющую, по крайней мере, $N+1$ производную. По условию, эту производную имеет $f(x)$, а два остальных члена — полиномы.

Кроме того, $\psi(x)$ на $[a,b]$ имеет, по крайней мере, $N+2$ нуля.

Их можно указать. Точки $x = t_n (n = 0, \dots, N)$ — нули, поскольку $f(t_n) = L(t_n)$, а последнее слагаемое обращается в них в нуль. $N+2$ нулем является точка $x = t$ в силу определения остаточного члена. Далее, поскольку между каждыми двумя нулями непрерывно дифференцируемой функции имеется хотя бы один нуль ее производной, на $[a,b]$ имеется хотя бы $N+1$ нуль ψ'. Применяя это рассуждение к $\psi'', \psi''', \dots$ можно показать, что существует точка $\xi \in [a,b]$ такая, что $\psi^{(N+1)}(\xi) = 0$.

Вычислим $N+1$ производную правой части выражения для $F(x)$ с учетом того, что $L^{(N+1)}=0$. Кроме того, в точке ξ

$$\psi^{(N+1)}(\xi)=f^{(N+1)}(\xi)-L^{(N+1)}(\xi)-\frac{d^{N+1}}{dx^{N+1}}\left[R_N(t)\cdot\frac{(x-t_0)\ldots(x-t_N)}{(t-t_0)\ldots(t-t_N)}\right]_{\xi},$$

$$L^{(N+1)}(\xi)=0;\Psi^{(N+1)}(\xi)=0;$$

$\frac{d^{N+1}}{dx^{N+1}}\left[\frac{(x-t_0)\ldots(x-t_N)}{(t-t_N)\ldots(t-t_N)}\right]\Big|_{x=\xi}=\frac{(N+1)!}{\prod\limits_{j=0}^{N}(t-t_j)}$. Тогда $f^{(N+1)}(\xi)-R_N(t)\cdot\frac{(N+1)!}{\prod\limits_{j=0}^{N}(t-t_j)}=$

$=0$, откуда получим выражение для $R_N(t)$: $R_N(t)=\frac{f^{(N+1)}(\xi)}{(N+1)!}\prod\limits_{j=0}^{N}(t-t_j)$

■

Рассмотрим некоторые важные следствия этой теоремы.

Следствие (точность интерполяции на равномерной сетке). *Положим, что $t_n=n\tau,\tau=(b-a)/N,t\in[a,\ b]$, — сетка равномерная. В этом случае имеет место оценка*

$$|R_N(t)|\leqslant\frac{\tau^{N+1}}{N+1}C,C=\max_{t\in[a,b]}\left|f^{(N+1)}(t)\right|.$$

Доказательство.

Пусть $t=t_k+\alpha\tau,\alpha\in[0,1],k=0,1,\ldots,N-1$.

Тогда $t-t_n=k\tau+\alpha\tau-n\tau=(k+\alpha-n)\tau$; откуда $\prod\limits_{n=0}^{N}(t-t_n)=$

$=\tau^{N+1}\prod\limits_{n=0}^{N}(k+\alpha-n)$. Можно показать, что $\prod\limits_{n=0}^{N}|k+\alpha-n|\leqslant N!$. Остаточный член оценивается следующим образом: $R_N(t)=\frac{f^{(N+1)}(\xi)}{(N+1)!}\prod\limits_{n=0}^{N}(t-$ $-t_n)$, поэтому с учетом приведенных оценок получим

$$|R_N(t)|\leqslant\frac{\tau^{N+1}}{N+1}\max_{\xi\in[a,b]}\left|f^{(N+1)}(\xi)\right|.$$

Рассмотрим, как ведет себя оценка в задаче экстраполяции при удалении точки t от интервала $[t_0,t_N]$. При $t\in[t_N,t_N+\tau]$ имеем $|R_N(t)|\leqslant$ $\tau^{N+1}\cdot\max\limits_{\xi\in[t_0,\,t_N+\tau]}\left|f^{(N+1)}(\xi)\right|$, поскольку $\prod\limits_{n=0}^{N+1}|k+\alpha-n|\leqslant(N+1)!$. При $t\in[t_N+\tau,t_N+2\tau]\,|R_N(t)|\leqslant(N+2)\tau^{N+1}\cdot\max\limits_{\xi\in[t_0,\,t_N+2\tau]}\left|f^{(N+1)}(\xi)\right|$, так как $\prod\limits_{n=0}^{N+2}|(k+\alpha-n)|\sim(N+2)!$

При $t\in[t_N+2\tau,t_N+3\tau]\,|R_N(t)|\leqslant\frac{(N+2)(N+3)}{2!}\tau^{N+1}\max\limits_{\xi\in[t_0,\,t_N+3\tau]}\left|f^{(N+1)}(\xi)\right|$ и так далее. ■

Видно, что ошибка экстраполяции растет быстро, но не сразу: экстраполяция допустима на интервалах $\sim O(\tau)$.

6.6. Интерполяционный полином в форме Ньютона

6.6.1. Разделенные и конечные разности

Определение. Пусть задана система узлов $\{t_n\}_{n=0}^{N}$, $t_n \in [a,b]$, $t_0 = a$, $t_N = b$.

Разделенные разности нулевого порядка в точке t_i совпадают со значениями функции $f(t_i)$;

Разности первого порядка определяются для двух точек t_i, t_{i+1} равенством

$$f(t_i, t_{i+1}) = \frac{f(t_{i+1}) - f(t_i)}{t_{i+1} - t_i},$$

разности второго порядка — для трех точек t_i, t_{i+1}, t_{i+2}

$$f(t_i, t_{i+1}, t_{i+2}) = \frac{f(t_{i+1}, t_{i+2}) - f(t_i, t_{i+1})}{t_{i+2} - t_i},$$

разности порядка k — для $k+1$ точки по рекуррентной формуле

$$f(t_i, t_{i+1}, \ldots, t_{i+k}) = \frac{f(t_{i+1}, \ldots, t_{i+k}) - f(t_i, \ldots, t_{i+k-1})}{t_{i+k} - t_i}.$$

Методом математической индукции можно показать, что

a) $f(t_i, t_{i+1}, \ldots, t_{i+k}) = \sum\limits_{j=0}^{k} \dfrac{f(t_{i+j})}{\prod\limits_{\substack{r=0 \\ r \neq i+j}}^{k} (t_{i+j} - t_{i+r})}$;

b) существует точка $\xi \in [a,b]$ такая, что $k! f(t_i, t_{i+1}, \ldots, t_{i+k}) = f^{(k)}(\xi)$.

Отсюда следует, что разделенная разность есть симметричная функция своих аргументов $t_i, \ldots, t_{i+k}$ и она не изменяется при их перестановке.

Для удобства введем таблицу разделенных разностей:

t_0	$f(t_0)$			$\ldots$	
		$f(t_0, t_1)$		$\ldots$	
t_1	$f(t_1)$		$f(t_0, t_1, t_2)$	$\ldots$	
		$f(t_1, t_2)$		$\ldots$	
t_2	$f(t_2)$		$\ldots$	$\ldots$	$f(t_0, \ldots, t_n)$
$\ldots$	$\ldots$	$\ldots$	$f(t_{n-2}, t_{n-1}, t_n)$	$\ldots$	
		$f(t_{n-1}, t_n)$		$\ldots$	
t_n	$f(t_n)$			$\ldots$	

Пусть сетка — равномерная. Тогда конечной разностью первого порядка функции $f(t)$ в точке t_k с шагом τ называют величину $\Delta f_k = f_{k+1} - f_k$, где $f_k = f(t_k)$, второго порядка — величину

$$\Delta^2 f_k = \Delta f_{k+1} - \Delta f_k = f_{k+2} - 2f_{k+1} + f_k,$$

третьего

$$\Delta^3 f_k = \Delta f_{k+3} - 3f_{k+2} + 3f_{k+1} - f_k = f_{k+3} - 3f_{k+2} + 3f_{k+1} - f_k,$$

четвертого

$$\Delta^4 f_k = f_{k+4} - 4f_{k+3} + 6f_{k+2} - 4f_{k+1} + f_k,$$

причем

$$\Delta^n f_k = \Delta^{n-1} f_{k+1} - \Delta^{n-1} f_k, k \geqslant 1, \Delta^0 f_k = f_k.$$

Методом математической индукции доказывается формула

$$\Delta^n f_k = \sum_{s=0}^{n} (-1)^{n-1} \cdot C_n^s f_{k+s},$$

где $C_i^s = \frac{i!}{s!(i-s)!}$ - биномиальные коэффициенты.

Нетрудно показать, например, используя формулу Лагранжа, что существует точка $\xi \in [a, b]$ такая, что $\tau^n f^{(n)}(\xi) = \Delta^n f_k, t_k \in [a, b]$, поэтому в вычислительных методах используется приближенная формула

$$f^{(n)}(t) \approx \frac{\Delta^n f_k}{\tau^n};$$

аналогичная тем формулам численного дифференцирования, что были получены в первой лекции методом неопределенных коэффициентов.

Заметим, что введенные конечные разности называют «разностями вперед». Аналогично можно ввести «разности назад»:

$$\Delta f_k = f_k - f_{k-1}, \Delta^2 f_k = \Delta f_k - \Delta f_{k-1} = f_k - 2f_{k-1} + f_{k-2}, \ldots$$

$$\Delta^n f_k = \Delta^{n-1} f_k - \Delta^{n-1} f_{k-1}$$

и центральные разности

$$\Delta f_k = f_{k+1/2} - \Delta f_{k-1/2}, \Delta^2 f_k = \Delta f_{k+1/2} - \Delta f_{k-1/2} = f_{k+1} - 2f_k + f_{k-1}, \ldots,$$

$$\Delta^n f_k = \Delta^{n-1} f_{k+1/2} - \Delta^{n-1} f_{k-1/2}.$$

Иногда для обозначения первых конечных разностей вперед и назад используют обозначения $\Delta^+ f_k, \Delta^- f_k$.

6.6.2. Интерполяционный полином в форме Ньютона

Интерполяционный полином может быть записан с использованием введенных выше разделенных разностей. Такая форма его записи называется интерполяционным полиномом в форме Ньютона. Полином имеет вид

$$N_n(t) = f(t_1) + f(t_1, t_2)(t - t_1) + \ldots + f(t_1, \ldots, t_{n+1})(t - t_1)\ldots(t - t_n).$$

То, что это интерполяционный полином, может быть доказано, например, методом математической индукции. Отметим, что полином в форме Ньютона напоминает ряд Тейлора, а остаточный член интерполяционного полинома — остаточный член этого ряда. Достоинством записи интерполянта в форме Ньютона является то, что для повышения порядка полинома нет необходимости в его полной перестройке; достаточно лишь добавить к уже полученному выражению еще одно или несколько слагаемых. С помощью разделенных разностей можно оценивать погрешность интерполяции. Читатели могут предложить способ контроля точности вычислений, основанный на использовании разделенных разностей.

6.7. Многочлены Чебышёва и минимизация остаточного члена интерполяции

Многочленом Чебышева первого рода называется функция $T_n(t) = \cos(n \arccos t)$, где $t \in [-1, 1], n = 0, 1, \ldots$

Убедимся в том, что функция $T_n(t)$ действительно является многочленом. При $n = 0$ и $n = 1$ имеем $T_0(t) = 1, T_1(t) = t$.

Положив $\theta = \arccos t$, получим $T_1(t) = \cos\theta, T_n(t) = \cos n\theta$, $T_{n-1}(t) = \cos(n-1)\theta, T_{n+1}(t) = \cos(n+1)\theta$. По формуле суммы косинусов $\cos(n+1)\theta + \cos(n-1)\theta = 2\cos\theta\cos n\theta$, и справедливо рекуррентное соотношение $T_{n+1}(t) + T_{n-1}(t) = 2T_1(t)T_n(t)$, или $T_{n+1}(t) = 2tT_n(t) - T_{n-1}(t)$. Отсюда следует вид записи полиномов Чебышева: $T_2 = 2t^2 - 1, T_3(t) = 4t^3 - 3t, T_4(t) = 8t^4 - 8t^2 + 1$ и так далее. Функции $T_n(t)$ являются многочленами степени n со старшим членом $2^{n-1}t^n$.

Введем также нормированные многочлены Чебышева $\bar{T}_n(t) = \frac{T_n(t)}{2^{n-1}}$.

Нули многочлена Чебышева находятся из очевидного уравнения $T_n(t) = \cos(n \arccos t) = 0$, откуда $t_m = \cos\left(\frac{2m-1}{n}\pi\right), m = 1, 2, \ldots n, t \in [-1, 1]$. Для произвольного отрезка $[a, b]$ нули полинома Чебышева получаются очевидным линейным преобразованием, выражения для них будут $t_m = \frac{a+b}{2} + \frac{b-a}{2}\cos\left(\frac{2m-1}{n}\pi\right), m = 1, 2, \ldots n$. Легко отыскиваются также точки экстремумов полинома Чебышева, для них $|T_n(t)| = 1$ и на отрезке $t \in [-1, 1]$ точки экстремумов есть $t_m = \cos(\frac{m}{n}\pi), m = 1, 2, \ldots n$.

Нас интересует решение следующей задачи на минимакс: найти

$$\min_{\{t_n\}_{n=0}^N} \left\{ \max_{t\in[-1,1]} \left| \prod_{n=0}^{N} (t - t_n) \right| \right\}$$

чтобы путем выбора узлов сетки минимизировать остаточный член интерполяции. Эта задача была решена П. Л. Чебышевым.

Теорема Чебышева (без доказательства). *Среди всех многочленов степени $n \geqslant 1$, со старшим коэффициентом a_n равным единице, наименьшее уклонение от нуля, равное 2^{1-n}, имеет нормированный полином Чебышева* $\bar{T}_n(t) = 2^{1-n}T_n(t), t \in [-1,\ 1]$.

Это свойство полиномов Чебышева, наименьшее уклонение от нуля, можно сформулировать по-другому: для любого полинома $P_n(t) = t^n + a_{n-1}t^{n-1} + \ldots + a_0$, отличного от $\bar{T}_n(t)$ справедливо $2^{1-n} = \max\limits_{[-1,1]} |\bar{T}_n(t)| < \max\limits_{[-1,1]} |P_n(t)|, t \in [-1,\ 1]$.

Если в качестве интерполяционных узлов выбрать нули полинома Чебышева, то произведение $\prod\limits_{j=0}^{N+1}(t - t_j)$, а также $R_n(t)$ будут наименее уклоняющимися от нуля.

6.8. Обусловленность задачи интерполяции. Постоянная Лебега

В процессе вычислений значения интерполируемой функции известны с некоторой погрешностью. При работе на вычислительной машине ошибки округления неизбежны. Возникает вопрос о чувствительности интерполяционного полинома к ошибкам начальных данных (обусловленности задачи интерполяции) и к ошибкам округления (вопрос вычислительной устойчивости). Интерполяционный полином — оператор, линейный по отношению к значениям интерполируемой функции. С учетом погрешности начальных данных полином в форме Лагранжа может быть записан следующим образом:

$$L_N(t) = \sum_{n=0}^{N} f_n\varphi_n^N(t) + \sum_{n=0}^{N} \delta f_n\varphi_n^N(t),$$

причем слагаемое

$$\Delta_N(t, \delta f) = \sum_{n=0}^{N} \delta f_n\varphi_n^N(t)$$

характеризует чувствительность к ошибкам начальных данных и ошибкам вычислений. Нас интересует оценка

$$\max_{t\in[a,b]}|\Delta_N(t,\delta f)|\leqslant l_N\delta;\delta=\max_{t\in[a,b]}|\delta f_n|,l_N=\max_{t\in[a,b]}\sum_{n=0}^{N}\left|\varphi_n^N(t)\right|,$$

коэффициент l_N называется постоянной Лебега.

Введем в рассмотрение еще один объект. Пусть $\sum\left|\varphi_i^N(x)\right|$ — сумма модулей всех базисных функций. Обозначим ее $L(x)=\sum\left|l_i^N(x)\right|$ — функция Лебега (сетки). Тогда константа Лебега $l_N=\sup\limits_{x\in[a,b]}L(x)$.

Так как функция Лебега зависит лишь от расположения узлов сетки, то и константа Лебега зависит лишь от введенной сетки. Обусловленность и устойчивость задачи интерполяции зависят от константы Лебега.

Если рассматривать оператор интерполяции как оператор проекции (проектор), переводящий элемент одного пространства (пространства сеточных функций) в другое (пространство непрерывно дифференцируемых функций), то постоянная Лебега есть норма такого оператора проекции. Подробнее об этом в [7].

Конечно, реальная погрешность при интерполяции будет заведомо меньше, чем приведенная выше оценка. Тем не менее, оценка является достижимой (это свойство нормы оператора). Наихудшим распределением погрешности будет такое распределение, когда погрешности максимальны и меняют знак от точки к точке. То, что при этом будет достижима приведенная выше оценка, следует из вида функции Лебега и каждой из базисных функций. Предлагаем читателям соответствующие построения провести самостоятельно.

Приведем (без доказательства) примерные оценки роста постоянной Лебега в зависимости от числа узлов сетки. Константа Лебега растет примерно как $l_N\sim 2^N$ для равномерной сетки и $l_N\sim\ln(N)$ для сетки с чебышевским набором узлов. Доказано, что рост константы Лебега для последней сетки асимптотически стремится к минимально возможному, и сетка с чебышевскими узлами близка к оптимальной для задач интерполяции.

6.9. Интерполяция с кратными узлами

Определение. Пусть в узлах сетки $\{t_n\}_{n=0}^M$ заданы не только значения функции $f(t_n)$, но и значения ее производных $f'(t_n), f''(t_n),\ldots,$ $f^{(k_n-1)}(t_n)$. В этом случае узел t_n называется кратным, а число k_n, равное количеству заданных значений производных в n узле — кратностью узла.

Доказывается теорема о существовании единственного полинома $P_N(t)$, удовлетворяющего условиям

$$P_N(t_n) = f_n, P'_N(t_n) = f'_n, \ldots, P_N^{(k_n-1)}(t_n) = f_n^{k_n-1},$$
$$N = k_0 + k_1 + \ldots + k_M - 1.$$

Такой полином называется полиномом с кратными узлами. Отметим два частных случая.

а) в точке $t = t_0$ заданы $f_0, f'_0, \ldots, f_0^{(N)} (M = 0,\ k_0 = N + 1)$.

Тогда многочлен $P_N(t)$, удовлетворяющий этим условиям, может быть записан как

$$P_N(t) = \sum_{i=0}^{N} f^{(i)}(t_0) \frac{(t - t_0)^i}{i!}.$$

Это — ряд Тейлора, который является интерполянтом с кратным узлом в точке $t = t_0$ кратности $N + 1$.

б) Пусть на концах отрезка $[t_0, t1]$ заданы значения $f_0, f_1, f'_0, f'_1 (M = 1, k_0 = 2, k_1 = 2, N = 3)$. Тогда $P_3(t_0) = f_0, P'_3(t_0) = f'_0, P_3(t_1) = f_1, P'_3(t_1) = f'_1$, а интерполянт имеет вид

$$P_3(t) = f_0 \frac{(t_1 - t)^2 [2(t - t_0) + \tau]}{\tau^3} + f'_0 \frac{(t_1 - t)^2 (t - t_0)}{\tau^2} +$$
$$+ f_1 \frac{(t - t_0)^2 [2(t_1 - t) + \tau]}{\tau^3} + f'_1 \frac{(t - t_0)^2 (t - t_1)}{\tau^2},$$

здесь $\tau = t_1 - t_0$.

Такой многочлен называется кубическим интерполяционным многочленом Эрмита.

Теорема без доказательства. . *Пусть $f(t)$ имеет $N + 1$ ограниченную производную на отрезке $[a, b]$. Тогда погрешность интерполяционного многочлена Эрмита степени N выражается формулой*

$$R_N(t) = \frac{f^{(N+1)}(\xi)}{(N+1)!} (t - t_0)^{k_0} (t - t_1)^{k_1} \ldots (t - t_M)^{k_M},$$

где t_n — интерполяционные узлы, $n = 0, \ldots, M, k_i$ — кратность i узла, $\xi \in [a, b]$, $N = k_0 + k_1 + \ldots + k_M - 1$.

Поставим теперь следующую задачу: построить кусочно-кубическую интерполирующую функцию, непрерывную на отрезке $[a, b]$ со своими двумя первыми производными.

Обозначим такую функцию $S(t)$; значения производных в узлах t_n обозначим $m_n = S'(t_n)$. Если задать в узлах t_n, t_{n+1} значение функции и ее первой производной, то получим эрмитов кусочно-кубический полином

$$S(t) = \frac{(t_{n+1}-t)^2\,[2(t-t_n)+\tau]}{\tau^3} f_n + \frac{(t-t_n)^2\,[2(t_{n+1}-t)+\tau]}{\tau^3} f_{n+1} +$$

$$+ \frac{(t_{n+1}-t)^2(t-t_n)}{\tau^2} m_n + \frac{(t-t_n)^2(t-t_{n+1})}{\tau^2} m_{n+1},$$

или

$$S(z) = f_n(1-z)^2(1+2z) + f_{n+1}z^2(3-2z) + m_n\tau_n z(1-\tau)^2 - m_{n+1}\tau_n z^2(1-z),$$

где $\tau_n = t_{n+1} - t_n, z = (t-t_n)/\tau_n, m_n = S'(t_n)$.

6.9.1. Замечание о тригонометрической интерполяции

Для периодической функции $f(t)$ с периодом T естественно строить приближение с использованием функций $\varphi_n(t) = a_n \cos\frac{\pi nt}{T} + b_n \sin\frac{\pi nt}{T}$ Тригонометрическая интерполяция состоит в замене $f(t)$ тригонометрическим многочленом $F_N(t) = \sum\limits_{n=0}^{N} \varphi_n(t) = a_0 + \sum\limits_{n=1}^{N} (a_n \cos\frac{\pi nt}{T} + b_n \sin\frac{\pi nt}{T})$, коэффициенты которого находятся при решении СЛАУ $F_N(t_k) = f(t_k)$, $k = 1, \ldots, 2N+1, t_{2N+1} - t_0 = T$, здесь $\{t_k\}_{k=0}^{2N+1}$ — последовательность узлов интерполяции.

6.10. Кусочно-многочленная глобальная интерполяция(сплайны)

Определение. Пусть на отрезке $[a,b]$ задана система узловых точек $\{t_n\}_{n=0}^{N-1}$. Сплайном $S_m(t)$ называется определенная на $[a,b]$ функция, имеющая l непрерывных производных и являющаяся на каждом интервале (t_{n-1}, t_n) многочленом степени m.

Определение. Дефектом сплайна называется разность $d = m - l$ между степенью сплайна и показателем его гладкости l.

Замечание. Для сплайнов также используется обозначение $S_{m,d}(t)$. Если сплайн строится так, чтобы выполнялись условия $S_m(t_n) = f(t_n)$, где $f(t)$ — интерполируемая функция, то он называется интерполяционным сплайном. В соответствии с определением, кусочно-линейная функция является интерполяционным сплайном первой степени дефекта 1, кусочно-квадратичная функция с первой непрерывной производной —

интерполяционным сплайном второй степени дефекта 1. Наиболее известным в приложениях является интерполяционный кубический сплайн дефекта 1 (естественный сплайн), который будем обозначать $S(t)$.

Определение. Кубическим сплайном дефекта 1, интерполирующим на отрезке $[a,b]$ заданную функцию $f(t)$, называется функция $S(t)$, удовлетворяющая следующим условиям:

1. $S(t_n) = f(t_n)$ — условие интерполяции в узлах сетки $\{t_n\}_{n=0}^{N-1}$.
2. $S(t) \in C^2[a,b]$, т. е. является непрерывной вместе с двумя первыми производными.
3. На каждом отрезке $[t_n, t_{n+1}]$, $S(t)$ является кубическим многочленом; $n = 0, \ldots, N-1$.
4. На краях отрезка $[a,b]$ заданы краевые условия. Наиболее часто употребляются следующие:
   (a) $S'(a) = f'(a)$, $S'(b) = f'(b)$;
   (b) $S''(a) = f''(a)$, $S''(b) = f''(b)$; часто полагают $S''(a) = S''(b) = 0$;
   (c) $S(a) = S(b)$, $S'(a) = S'(b)$; эти условия называются периодическими, т. е. интерполируемая функция является периодической с периодом $b-a$.

Покажем, что эта задача имеет единственное решение.

Теорема. *Интерполяционный кубический сплайн $S(t)$, удовлетворяющий условиям* 1–3 *и одному из краевых условий* 4, *существует и единственен.*

Доказательство.

Пусть $S(z)$ — эрмитов кубический многочлен, который на каждом отрезке $[t_n, t_{n+1}]$, $n = 0, \ldots, N-1$, представлен как

$$S(z) = f_n(1-z)^2(1+2z) + f_{n+1}\cdot z^2(3-2z) + m_n\tau_n z(1-z)^2 - m_{n+1}\tau_n z^2(1-z),$$

где $\tau_n = t_{n+1} - t_n$, $z = (t-t_n)/\tau_n$, $m_n = S'(t_n)$. Тогда

$$S''(t) = \frac{(f_{n+1}-f_n)(6-12z)}{\tau_n^2} + m_n\frac{6z-4}{\tau_n} + m_{n+1}\frac{6z-2}{\tau_n},$$

$$S''(t_n+0) = 6\frac{f_{n+1}-f_n}{\tau_n^2} - \frac{4m_n}{\tau_n} - \frac{2m_{n+1}}{\tau_n},$$

$$S''(t_n-0) = -6\frac{f_n-f_{n-1}}{\tau_{n-1}^2} + \frac{2m_{n-1}}{\tau_{n-1}} + m_{n+1}\frac{4m_n}{\tau_{n-1}}.$$

Условие непрерывности второй производной $S''(t_n+0) = S''(t_n-0)$ будет

$$r_n m_{n-1} + 2m_n + s_n m_{n+1} = c_n,$$

$$c_n = 3\left(s_n \frac{f_{n+1}-f_n}{\tau_n} + r_n \frac{f_n - f_{n-1}}{\tau_{n-1}}\right), s_n = \frac{\tau_{n-1}}{\tau_{n-1}+\tau_n}, r_n = 1 - s_n,$$

$$n = 1, \ldots, N-1.$$

После добавления краевых условий получаем систему из $N+1$ уравнение с $N+1$ неизвестным m_n. Для краевых условий первого типа (заданы первые производные) система выглядит как

$$m_0 = f'_0,$$

$$r_n m_{n-1} + 2m_n + s_n m_{n+1} = c_n,$$

$$m_N = f'_N.$$

Для условий второго типа (заданы вторые производные)

$$2m_0 + m_1 = 3\frac{f_1 - f_0}{\tau_0} - \frac{\tau_0}{2} f''_0,$$

$$r_n m_{n-1} + 2m_n + s_n m_{n+1} = c_n,$$

$$m_{N-1} + 2m_N = 3\frac{f_N - f_{N-1}}{\tau_{N-1}} + \frac{\tau_{N-1}}{2} f''_N.$$

Аналогично получается СЛАУ для третьего типа краевых условий.

Во всех случаях матрицы СЛАУ оказываются трехдиагональными симметричными, со строгим диагональным преобладанием и, как показывается, положительно определенными, а, следовательно, и неособенными. Следовательно, решение СЛАУ существует и единственно. Отсюда следует существование и единственность решения задачи о построении кубического сплайна. ■

Приведем еще одно доказательство этой же теоремы.

Доказательство.

Рассмотрим неравномерную сетку: $t_n - t_{n-1} = \tau_{n-1}$, $t_{n+1} - t_n = \tau_n$. В узлах сетки определены значения функции: f_{n-1}, f_n, f_{n+1}. Пусть m_n — значение второй производной в точке t_n (пока неизвестное!). На отрезке $[t_n, t_{n+1}]$ для второй производной кусочно-кубического сплайна имеем

$$S''_{tt} = \frac{1}{\tau_n}(m_n(t_{n+1} - t) + m_{n+1}(t - t_n)). \tag{6.2}$$

Так как сплайн — полином третьей степени, то его вторая производная — линейная функция. Интегрируем (6.2) по t, получаем (на отрезке $[t_n, t_{n+1}]$)

$$S'_t = \frac{1}{\tau_n}\left(m_{n+1}\frac{(t_{n+1}-t)^2}{2} - m_n\frac{(t-t_n)^2}{2}\right) + A_n.$$

Интегрируя последнее соотношение еще раз, получаем:

$$S(t) = \frac{1}{6\tau_n}(m_n(t_{n+1}-t)^3 + m_{n+1}(t-t_n)^3) +$$

$$+\alpha_n\,(t_{n+1}-t) + \beta_{n+1}\,(t-\,t_n).$$

A_n — константа интегрирования. После второго интегрирования положим $A_n t + B_n = (\beta_n - \alpha_n)t + \alpha_n t_{n+1} - \beta_n t_n$, т. е. вместо двух констант A_n, B_n введем две новые константы, более удобные для дальнейших выкладок.

Из условий $S(t_n) = f_n, S(t_{n+1}) = f_{n+1}$, получаем:

$$f_n = \frac{m_n\tau_n^2}{6} + \alpha_n\tau_n \Rightarrow \alpha_n = \frac{f_n}{\tau_n} - \frac{m_n\tau_n}{6}.$$

$$f_{n+1} = \frac{m_{n+1}\tau_n^2}{6} + \beta_n\tau_n \Rightarrow \beta_n = \frac{f_{n+1}}{\tau_n} - \frac{m_{n+1}\tau_n}{6}.$$

$$A_n = \frac{f_{n+1}-f_n}{\tau_n} - \frac{(m_{n+1}-m_n)\tau_n}{6}.$$

Приравняем первые производные в t_n справа и слева $S'_t(t_n+0) = S'_t(t_n-0)$, получим систему уравнений для определения коэффициентов сплайна:

$$\frac{m_n\tau_{n-1}}{2} - \frac{m_{n-1}\tau_{n-1}}{2} + \frac{f_n - f_{n-1}}{\tau_{n-1}} - \frac{(m_n - m_{n-1})\tau_{n-1}}{6} =$$

$$= \frac{m_{n+1}\tau_n}{2} - \frac{m_n\tau_n}{2} + \frac{f_{n+1}-f_n}{\tau_n} - \frac{(m_{n+1}-m_n)\tau_n}{6}, \qquad (6.3)$$

которая дополняется соответствующими граничными условиями. В случае свободного сплайна $m_0 = m_N = 0$.

Систему для определения коэффициентов, называемых *моментами* кубического *сплайна*, можно записать в матричной форме

$$\mathbf{AM} = \mathbf{F},$$

где $\mathbf{A}$ — квадратная матрица:

$$\mathbf{A} = \begin{pmatrix} \frac{\tau_1+\tau_2}{3} & \frac{\tau_2}{6} & 0 & \dots & 0 & 0 \\ \frac{\tau_2}{6} & \frac{\tau_2+\tau_3}{3} & \frac{\tau_3}{6} & \dots & 0 & 0 \\ 0 & \frac{\tau_3}{6} & \frac{\tau_3+\tau_4}{3} & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \frac{\tau_{N-1}}{6} & \frac{\tau_N+\tau_{N-1}}{3} \end{pmatrix}$$

$\mathbf{M}$ и $\mathbf{F}$ — векторы-столбцы:

$$\mathbf{M} = (m_1, m_2, \dots, m_{N-1})^{\mathrm{T}},$$

$$\mathbf{F} = \left(\frac{f_2-f_1}{\tau_1} - \frac{f_1-f_0}{\tau_0}, \frac{f_3-f_2}{\tau_2} - \frac{f_2-f_1}{\tau_1}, \dots, \frac{f_N-f_{N-1}}{\tau_N} - \frac{f_{N-1}-f_{N-2}}{\tau_{N-1}}\right)^{\mathrm{T}}.$$

Матрица $\mathbf{A}$ симметрична, имеет свойство диагонального преобладания и, как можно показать, положительно определена, а следовательно, неособенная. Значит, решение рассматриваемой СЛАУ существует и единственно. Следовательно, и задача о построении кубического сплайна имеет единственное решение. Для других типов краевых условий доказательство проводится аналогично. Метод решения такой СЛАУ, который будет рассмотрен в лекции 10 — прогонка.

Теорема (без доказательства). *Для функции $f(t) \in C^4[a,b]$ и интерполирующего ее сплайна $S(t)$, построенного на сетке $\{t_n\}_{n=0}^N$, имеют место следующие неравенства:*

$$\|f(t) - S(t)\|_{[a,b]} \leqslant M_4 \tau^4,$$

$$\|f'(t) - S'(t)\|_{[a,b]} \leqslant M_4 \tau^3,$$

$$\|f''(t) - S''(t)\|_{[a,b]} \leqslant M_4 \tau^2,$$

где $M_4 = \left\|f^{(4)}(t)\right\|_{[a,b]}$, $\tau = \max\limits_n (t_{n+1} - t_n)$.

Отсюда следует, что при $\tau \to 0$ последовательность функций $S^{(k)}(t)$, $i = 0, 1, 2$ (кубический сплайн и первые две его производные) сходится, соответственно, к $f^{(k)}(t)$.

Теорема (экстремальное свойство кубических сплайнов) (без доказательства). *Пусть сплайн $S(t)$ интерполирует функцию $f(t)$ на системе узлов*

$$\{t_n\}_{n=0}^N ; t_0 = a, t_N = b.$$

Тогда $S(t)$ с краевыми условиями $S''(a) = S''(b) = 0$ доставляет минимум функционалу

$$\int_a^b [F''(t)]^2 dt$$

среди всех функций $F(t) \in C_2^2[a,b]$, т. е. функций, имеющих интегрируемые с квадратом вторые производные ($\int_a^b [F''(t)]^2 dt$) сходится на отрезке $[a,b]$) и интерполирующих $f(t)$ на отрезке $[a,b]$.

Локальный сплайн. Локальная форма сплайн-интерполяции предложена В. С. Рябеньким [9, 10]. Рассмотрим неравномерную сетку: $t_n - t_{n-1} = h_{n-1}, t_{n+1} - t_n = h_n$. В узлах сетки определены значения функции: f_{n-1}, f_n, f_{n+1}. Не вдаваясь в детали, приведем важные для практического использования формулы в случае постоянного шага сетки $h = \text{const}$.

Построим интерполяционный полином второго порядка $P_2(x)$ в форме Ньютона $(d - f_n)/h$:

t_{n-1}	f_{n-1}	$\frac{f_n - f_{n-1}}{h}$	$\frac{f_{n-1} - 2f_n + f_{n+1}}{h^2}$
t_n	f_n	$\frac{f_{n+1} - f_n}{h}$	
t_{n+1}	f_{n+1}		

$$P_2(t, f_n) = f_{n-1} + \frac{f_n - f_{n-1}}{h}(t - t_{n-1}) + \\ + \frac{f_{n-1} - 2f_n + f_{n+1}}{h^2}(t - t_{n-1})(t - t_n), \qquad (6.4)$$

Этот полином приближает f на отрезке $[t_{n-1}, t_{n+1}]$ с точностью до $o(h^2)$. Рассмотрим теперь полином

$$Q_5(t, f_n) = P_2(t, f_n) + \frac{h^3}{2}\left\{\frac{f_{n+2} - 3f_{n+1} + 3f_n - f_{n-1}}{h^3} \times \right. \\ \left. \times \left(\frac{t - t_n}{h}\right)^3 \left(\frac{t - t_{n+1}}{h}\right)\left(3 - \frac{2(t - t_n)}{h}\right)\right\},$$

представляющий собой аппроксимацию функции f на отрезке $[t_{n-1}, t_{n+1}]$ с непрерывными первой и второй производными. В [1] доказано, что выражение (6.4) аппроксимирует $f_x^{(m)}$ с порядком $o(h^{3-m})$ во всех точках отрезка. Так как коэффициенты сплайна зависят от значений функции лишь в 4-х соседних точках и для определения коэффициентов (6.4) не требуется решать систему линейных уравнений, такая кусочно-гладкая интерполяция называется *локальным сплайном.*

Замечание. $Q_5(t, f_n)$ уже не обладает экстремальным свойством.

6.11. B-сплайны

Сплайны с локальным носителем. (B-сплайны). В последнее время в вычислительной практике широкое распространение получили B-сплайны (от английского слова bell — колокол), сосредоточенные на конечном носителе. Они используются как для интерполяции функций, так и в качестве базисных функций при построении методов типа конечных элементов.

Для подробного ознакомления с приложениями B-сплайнов и B-сплайнами произвольной степени рекомендуется обратиться к [5, 6] и современным публикациям, например, в журнале «Математическое моделирование» [14]. В данной лекции ограничимся наиболее распространенными случаями B-сплайнов порядка 2 и 3, см. также [11].

Определение. B-сплайном, или базисным сплайном степени $N-1$ дефекта 1 относительно узлов $\{t_i\}_{i=n}^{n+N}$ называется функция

$$B_{N-1,n}(t) = B_{N-1}(t_n, t_{n+1}, \ldots, t_{n+N}, t) = N \sum_{i=n}^{n+N} \frac{(t_i - t)_{\max}^{N-1}}{\prod\limits_{\substack{=n \\ j \neq i}}^{n+N} (t_i - t_j)}$$

$$(t_i - t)_{\max}^{N-1} = \begin{cases} (t_{i-t})^{N-1}, & t \leqslant t_i, \\ 0, & t > t_i, \end{cases}$$

Пусть $t_{n+i} = t_n + i\tau$, т. е. рассматривается случай равномерной сетки. Рассмотрим несколько частных случаев B-сплайнов.

1. $N = 2$. В этом случае сплайн строится наиболее просто.

$$B_{1,n}(t) = B_1(t_n, t_{n+1}, t_{n+2}, t) = 2\left[\frac{(t_n - t)_{\max}}{(t_n - t_{n+1})(t_n - t_{n+2})} + \right.$$

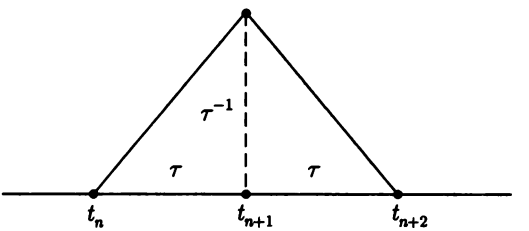
$$\left. \frac{(t_{n+1} - t)_{\max}}{(t_{n+1} - t_n)(t_{n+1} - t_{n+2})} + \frac{(t_{n+2} - t)_{\max}}{(t_{n+2} - t_n)(t_{n+2} - t_{n+1})}\right] =$$

$$= \frac{1}{\tau^2}\left[(t_n - t)_{\max} - 2(t_{n+1} - t)_{\max} + (t_{n+2} - t)_{\max}\right],$$

или

$$B(t) = \begin{cases} \frac{1}{\tau^2}(t_n - t - 2t_{n+1} + 2t + t_{n+2} - t) = 0, & t \leqslant t_n \\ \frac{1}{\tau^2}(0 - 2t_{n+1} + 2t + t_{n+2} - t) = \frac{1}{\tau} + \frac{t - t_{N+1}}{\tau^2}, & t_n \leqslant t \leqslant t_{n+1} \\ \frac{1}{\tau^2}(0 - 0 + t_{n+2} - t) = \frac{1}{\tau} - \frac{t - t_{N+1}}{\tau^2}, & t_{n+1} \leqslant t \leqslant t_{n+2} \\ 0, & t \geqslant t_{n+2} \end{cases}$$

Это функция «крышка» или «крышечка». Она часто используется в качестве базисной функции в методах конечных элементов.



Рассмотрим случай B-сплайна 2-го порядка, задаваемого формулой

$$S_k(x) = \begin{cases} x^2; & x = \frac{t - t_{k-2}}{t_{k-1} - t_{k-2}}, t \in [t_{k-2}, t_{k-1}]; \\ 1 + 2x - x^2; & x = \frac{t - t_{k-1}}{t_k - t_{k-1}}, \quad t \in [t_{k-1}, t_k]; \\ 2 - x^2; & x = \frac{t - t_k}{t_{k+1} - t_k}, \quad t \in [t_k, t_{k+1}]; \\ (1 - x)^2; & x = \frac{t - t_{k+1}}{t_{k+2} - t_{k+1}}, t \in [t_{k+1}, t_{k+2}]. \end{cases}$$

При $t < t_{k-2}, t > t_{k-2}, S_k(x) \equiv 0$. Построенный сплайн обладает следующими свойствами:

(a) $S'_t(t_{k-2}) = S'_t(t_{k+2}) = 0$;
(b) $S(t_{k-1}) = S(t_{k+1}) = 1$;
(c) $S(t_{k-2}) = S(t_{k+2}) = 0$.

При интерполяции функций можно поступить таким способом. Заметим, что для интерполяции с помощью сплайна необходимо потребовать выполнения условия

$$b_{i-1}S_{i-1} + b_iS_i + b_{i+1}S_{i+1} = f_i,$$

где b — коэффициенты интерполяции, S — B-сплайн, индекс указывает на точку носителя, в которой сплайн достигает своего максимума. Система таких соотношений, естественно, дополняется граничными условиями. Известно [5], что получившаяся система для определения коэффициентов разложения будет иметь трехдиагональную матрицу с диагональным преобладанием при выполнении ограничения на длины соседних шагов: они должны различаться не более чем в $\frac{1+\sqrt{13}}{2}$ раза.

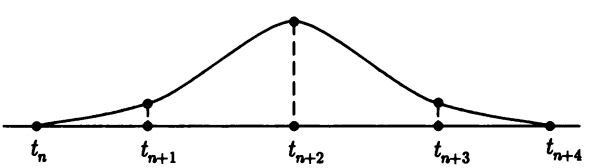
2. $N = 4$ (кубический B-сплайн) имеет вид

$$B_{3,\,n}(t) = \frac{1}{6\tau^4}\left[(t_n - t)^3_{\max} - 4(t_{n+1} - t)^3_{\max} + 6(t_{n+2} - t)^3_{\max} - \right.$$

$$\left. -4(t_{n+3} - t)^3_{\max} + (t_{n+4} - t)^3_{\max}\right],$$

или, после несложных упрощений:

$$\begin{cases} 0, & t \geqslant t_n, \\ \frac{1}{6\tau^4}(t - t_n)^3, & t_n \leqslant t \leqslant t_{n+1}, \\ \frac{1}{6\tau} + \frac{1}{2\tau^2}(t - t_{n+1}) + \frac{1}{2\tau^3}(t - t_{n+1})^2 - \frac{1}{2\tau^4}(t - t_{n+1})^3, & t_{n+1} \leqslant t \leqslant t_{n+2}, \\ \frac{1}{6\tau} + \frac{1}{2\tau^2}(t_{n+3} - t) + \frac{1}{2\tau^3}(t_{n+3} - t)^2 - \frac{1}{2\tau^4}(t_{n+3} - t)^3, & t_{n+2} \leqslant t \leqslant t_{n+3}, \\ \frac{1}{6\tau^4}(t_{n+4} - t)^3, & t_{n+3} \leqslant t \leqslant t_{n+4} \\ 0, & t \geqslant t_{n+4} \end{cases}$$



Базисные сплайны заданной степени являются линейно независимыми функциями и образуют базис в функциональных пространствах, что можно использовать для представления с их помощью других функций этих же пространств. Любая, например, кусочно-постоянная функция на отрезке, составленном из равных интервалов, может быть единственным образом представлена как линейная комбинация B-сплайнов нулевой степени, любая кусочно-линейная функция — B-сплайнов первой степени и т.д. Базисные

сплайны играют существенную роль при построении численных методов решения задач математической физики, например, метода конечных элементов в теории приближения функций, при решении задач компьютерной графики.

Для последнего класса задач также используются функции Бернштейна:

$$B_n^N(t) = \frac{N!}{n!(N-n)!}\frac{(b-t)^{N-n}(t-a)^n}{(b-a)^N},$$
$$n = 0, \ldots, N, t \in [a, b].$$

Функции Бернштейна иногда записывают в форме рекуррентного соотношения:

$$B_{-1}^N(t) = 0, B_0^0(t) = 1, B_i^N(t) = \frac{(b-t)B_i^{N-1}(t) + (t-a)B_{i-1}^{N-1}(t)}{b-a},$$
$$i = 0, \ldots, n, B_{N+1}^N(t) = 0.$$

Такие рекуррентные последовательности применяются с целью уменьшения ошибок округления.

Функции Бернштейна являются базисными для построения кривых Безье, активно использующихся в компьютерной графике и техническом дизайне, появившихся в результате работ Безье и де Кастильо над формами автомобилей фирм Рено и Ситроен. Подробнее о функциях Бернштейна в [12].

6.12. Интерполяция функций двух переменных

Пусть сетка образована пересечением прямых $x = x_n, n = 0, \ldots, N$ и $y = y_m, m = 0, \ldots, M, f_{nm} = f(x_n, y_m)$ — значение функции в узле x_n, y_m. Воспользуемся, например, аппаратом кусочно-многочленной интерполяции. Для этого сначала реализуется кусочно-многочленная интерполяция заданной степени по x на каждой прямой $y = y_m$. Затем при каждом значении $x = x_n$ реализуется кусочно-многочленная интерполяция по y с учетом значений функции, полученных на первом шаге. Так, в случае кусочно-линейной интерполяции по обоим переменным этот метод приводит (для случая прямоугольника $x \in [x_n, x_{n+1}], y \in [y_n, y_{n+1}]$) к интерполяционному многочлену

$$F(x,y) = f_{nm}\frac{(x-x_{n+1})(y-y_{m+1})}{(x_n-x_{n+1})(y_m-y_{m+1})} + f_{n+1,m}\frac{(x-x_n)(y-y_{m+1})}{(x_{n+1}-x_n)(y_m-y_{m+1})} +$$
$$+ f_{n+1,m+1}\frac{(x-x_n)(y-y_m)}{(x_{n+1}-x_n)(y_{m+1}-y_m)} + f_{n,m+1}\frac{(x-x_{n+1})(y-y_m)}{(x_n-x_{n+1})(y_{m+1}-y_m)}.$$

Сходным образом можно провести последовательную лагранжеву интерполяцию, но при каждом фиксированном значении m, затем — при каждом фиксированном значении n с учетом первого шага интерполяции. Общая формула такого интерполянта аналогична одномерной формуле для интерполяционного полинома в форме Лагранжа:

$$L_{NM}(x,\ y) = \sum_{n=0}^{N}\sum_{m=0}^{M} f_{nm} \prod_{\substack{i=0\\ i\neq n}}^{N}\prod_{\substack{j=0\\ j\neq m}}^{M} \frac{(x-x_i)(y-y_j)}{(x_n-x_i)(y_m-y_j)}.$$

Если f_A, f_B, f_D — значение функции $f(x,y)$ в вершинах A, B, некоторого треугольника на треугольной расчетной сетке, то вычислить приближенное значение функции внутри этого треугольника можно с помощью билинейной функции $f(x,\ y) \approx F(x,y) = ax + by + c$, находя коэффициенты a, b, c из условий

$$ax_A + by_A + c = f_A,$$

$$ax_B + by_B + c = f_B,$$

$$ax_D + by_D + c = f_D,$$

где $\{x_A, y_A\}, \{x_B, y_B\}, \{x_D, y_D\}$ — координаты вершин A, B, D. Погрешность такой интерполяции для функции $f(x,y)$ с непрерывными вторыми производными будет $O(h^2)$, где h — длина наибольшей стороны треугольника ABD.

6.13. Задачи

1. Выписать интерполяционные полиномы первой и второй степени в форме Лагранжа и Ньютона.

   Решение. Интерполяционные полиномы первой и второй степени в форме Лагранжа:

$$L_1(t) = f(t_0)\frac{t-t_1}{t_0-t_1} + f(t_1)\frac{t-t_0}{t_1-t_0},$$

$$L_2(t) = f(t_0)\frac{(t-t_1)(t-t_2)}{(t_0-t_1)(t_0-t_2)} + f(t_1)\frac{(t-t_0)(t-t_2)}{(t_1-t_0)(t_1-t_2)} + f(t_2)\frac{(t-t_0)(t-t_2)}{(t_2-t_0)(t_2-t_1)},$$

где t_0, t_1, t_2 — узлы интерполяции, $f(t_1), f(t_2), f(t_3)$ — значения интерполируемой функции.

Интерполяционные полиномы первой и второй степени в форме Ньютона:

$$N_1(t) = f(t_1) + \frac{f(t_2) - f(t_1)}{t_2 - t_1}(t - t_1),$$

$$N_2(t) = f(t_1) + \frac{f(t_2) - f(t_1)}{t_2 - t_1}(t - t_1) +$$

$$\frac{1}{t_3 - t_1}\left[\frac{f(t_3) - f(t_2)}{t_3 - t_3} - \frac{f(t_2) - f(t_1)}{t_2 - t_1}\right](t - t_1)(t - t_2).$$

2. Сравните количество арифметических действий, требуемое для вычисления интерполяционного полинома, записанного в двух формах:

$$L_N(x) = a_0 + a_1 t + \ldots + a_n t^n,$$

$$L_N(x) = a_0 + t(a_1 + t(a_2 + t(\ldots(a_{n-1} + c_n t)\ldots))) \text{ (схема Горнера)}$$

Решение. В первом случае для вычисления значения в одной точке требуется $\frac{n}{2}(n+1)$ умножений и n сложений. Во втором — n умножений и n сложений.

3. Задана система узлов интерполяции:

$$t_i = t_0 + \frac{t_n - t_0}{n - 1}(i - 1), i = 1, \ldots, n.$$

Какова погрешность интерполяции, если $n = 3$?

Решение. Сделаем замену переменных в выражении для остаточного члена

$$R_3(t) = (t - t_0)(t - \frac{t_0 + t_n}{2})(x - t_n), t = \frac{t_0 + t_n}{2} + \frac{t_n - t_0}{2}z, z \in [-1; 1].$$

Получим

$$R_3(t) = \frac{(t_n - t_0)^3}{2}(z^3 - z).$$

Полученный кубический полином имеет на $[-1; 1]$ экстремумы в точках

$$z_{1,2} = \pm\frac{1}{\sqrt{3}}.$$

В таком случае $\max|R_3(t)| = \|R_3(t)\| = \frac{(t_n - t_0)^3}{12\sqrt{3}}$.

4. Предложите простой рекуррентный алгоритм вычисления коэффициентов $a_i (i = 0 \div N-1)$ интерполяционного полинома

$$P_N(t) = a_0 + a_1(t - t_0) + \ldots + a_n(t - t_0)\ldots(t - t_{N-1}).$$

Решение. Для коэффициентов полинома получаем систему линейных уравнений с треугольной матрицей:

$$a_0 = f_0,$$

$$a_0 + a_1(t_1 - t_0) = f_1,$$

$$a_0 + a_1(t_2 - t_0) + a_2(t_2 - t_0)(t_2 - t_1) = f_2,$$

$$\ldots$$

$$a_0 + a_1(t_N - t_0) + \ldots + a_N(t_N - t_0)\ldots(t_N - t_{N-1}) = f_N,$$

которая легко решается от первого уравнения к последнему:

$$a_0 = f_0, a_1 = \frac{f_1 - f_0}{t_1 - t_0}, a = \frac{1}{t_2 - t_1}\left(\frac{f_2 - f_0}{t_2 - t_0} + \frac{f_1 - f_0}{t_1 - t_0}\right), \ldots$$

5. Оценить погрешность приближения функции $\ln t' (t' = 1,23)$ при помощи интерполяционного полинома второй степени по точкам $1,1; 1,2; 1,3$.

Решение. Остаточный член интерполяции будет

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

при $N = 2$ имеем:

$$\varepsilon = |f(x) - L_2(t)| \leqslant \frac{\max\limits_{[1,1;1,3]} |f^{(3)}(t)|}{6} |(t - t_0)(t - t_1)(t - t_2)|,$$

$$\max_{[1,3;1,2]} \left|f^{(3)}(t)\right| = \frac{2}{1 \cdot 1,3} \approx 1,5.$$

В таком случае

$$\varepsilon = |\ln(t') - L_2(t)| \leqslant \frac{1,5}{6} |(1,23 - 1,1)(1,23 - 1,2)(1,23 - 1,3)| \approx$$

$$\approx 6,9 \cdot 10^{-5}.$$

6. Показать, что погрешность интерполяции может быть выражена следующим образом:

$$f(t) - L_N(t) = (t - t_0)\dots(t - t_N) f(t, t_0, \dots, t_N),$$

где $f(t, t_0, \dots, t_N)$ — разделенная разность порядка N.

Решение. Из выражения для разделенной разности порядка $N + 1$

$$f(t, t_0, \dots, t_N) = \frac{f(t)}{(t-t_0)\dots(t-t_N)} + \frac{f(t_0)}{(t_0-t)(t_N-t_0)\dots(t_0-t_N)} + \dots +$$

$$+ \frac{f(t_N)}{(t_N-t)(t_N-t_0)\dots(t_N-t_{N-1})},$$

получим выражение для $f(t)$:

$$f(t) = f(t_0)\frac{(t-t_1)\dots(t-t_N)}{(t_0-t_1)\dots(t_0-t_N)} + f(t_N)\frac{(t-t_0)\dots(t-t_{N-1})}{(t_N-t_0)\dots(t_N-t_{N-1})} +$$
$$+(t-t_0)\dots(t-t_N)\cdot f(t, t_0, \dots, t_N) =$$
$$= L_N(t) + (t-t_0)\dots(t-t_N)\cdot f(t, t_0, \dots, t_N).$$

Тогда

$$f(t) - L_N(t)\cdot(t-t_0)\dots(t-t_N)\cdot f(t, t_0, \dots, t_N).$$

Сравнивая полученное выражение с выражением для остаточного члена интерполяции $R_N(t) = f(t) - L(t) = \frac{f^{(N+1)}(\xi)}{(N+1)!}(t-t_0)\dots(t-t_N)$, приходим к выводу, что для некоторой точки $\xi \in [t_0, t_N]$ имеет место соотношение между разделенной разностью и производной порядка $N + 1$:

$$f(t, t_0, \dots, t_N) = \frac{f^{(N+1)}(\xi)}{(N+1)!}.$$

7. Пусть значения функции $f(t)$ заданы в узлах интерполяции t_1, t_2, t_3. Построить функцию $g(t) = \frac{a_0+a_1 t}{d_0+t}$, для которой выполнялось бы условие интерполяции: $g(t_i) = f(t_i), i = 1,2,3$ (задача дробно-линейной интерполяции).

Решение. Из условий интерполяции $g(t_i) = \frac{a_0+a_1 t_i}{d_0+t_i} = f(t_i), i = 1, 2, 3$ получаем систему трех линейных уравнений:

$$a_0 + a_1 t_1 - d_0 f_1 = t_1 f_1,$$
$$a_0 + a_1 t_2 - d_0 f_2 = t_2 f_2,$$
$$a_0 + a_1 t_3 - d_0 f_2 = t_3 f_3.$$

Вводя обозначения

$$\tau_n = t_n - t_{i-1}, \bar{\tau} = \frac{\tau_n + \tau_{n+1}}{2}, \quad \Delta_n f = \frac{f_n - f_{n-1}}{\tau_n},$$

$$\Delta_{n+1} f = \frac{f_{n+1} - f_n}{\tau_{n+1}}, \quad \delta_n f = \frac{\Delta_{n+1} f - \Delta_n f}{\bar{\tau}},$$

и применив метод последовательного исключения неизвестных, получим

$$a_0 = (2t_n\, \Delta_n f\, \Delta_{n+1} f - f_n\, t_n\, \delta_n f)/\delta_n(tf),$$

$$a_1 = f_n - 2\Delta_n f\, \Delta_{n+1} f, b_0 = -t\, \delta_n(t\, f)/\delta_n f.$$

Здесь учтено, что $\delta_n(tf) = t_n \delta_n f + \frac{f_{n+1}+f_{n-1}}{\bar{\tau}_n}$.

Разумеется, знаменатель дробно-рационального выражения не должен обращаться в нуль на рассматриваемом отрезке.

6.14. Задачи для самостоятельного решения

1. Покажите, что интерполяционный полином в форме Лагранжа может быть построен в соответствии со следующими рекуррентными формулами: $L_0(t) = f(t_0)$,

$$L_n(t) = L_{n-1}(t) + [f(t_n) - L_{n-1}(t)] + \frac{P_n(t)}{P_n(t_n)},$$
$$P_1(t) = t - t_0, P_{n+1}(t) = P_n(t)(t - t_n).$$

2. Построить интерполяционный кубический полином $P_3(t) = \sum_{i=0}^{3} a_i t^i$, для которого выполнено $P_3(1) = 1, P_3(2) = 2, P_3(3) = 2, a_3 = 1$.

3. Построить интерполяционный полином в форме Лагранжа для функций $f(t) = |t|$, $f(t) = t^2$ по узлам $-1; 0; 1$; и $-2; -1; 0; 1; 2$.

4. Оценить погрешность интерполяции функции $f(t) = \sin t$ на отрезке $[0; \pi/4\,]$ по трем равноотстоящим узлам.

5. Оценить, какое количество узлов интерполяции потребуется на отрезке $[0; \pi/4\,]$ для обеспечения точности $\varepsilon = 10^{-3}$ при интерполяции функции $\sin t$.

6. Привести примеры непрерывных функций, для которых расходится последовательность интерполяционных полиномов (на равномерной сетке).

7. Какой величины необходимо выбрать шаг интерполирования τ для обеспечения точности $\varepsilon \leqslant 10^{-4}$ интерполяции функции $f(t) = \sqrt[3]{t}, t \in [1; 10^3]$ при линейной и квадратичной интерполяции?

8. Пусть имеется таблица функции $f(t) = sint$ в равноотстоящих точка $\{t_n\}_0^N$, причем $\max\limits_n(t_{n+1} - t_n) = \tau$. При каком τ линейная интерполяция позволит восстановить $f(t)$ с точностью $\varepsilon \leqslant 10^{-4}$? Тот же вопрос для квадратичной интерполяции. Решить задачу для случаев равномерной и неравномерной сеток $\{t_n\}_0^N$.

9. Как оценить погрешность интерполяционного процесса, если интерполируемая функция задана таблично?

10. По заданным значениям функции

t	1	2	2,5	3
f	-6	-1	15,6	16

найти значение t, при котором $f(t) = 0$ (решение уравнения $f(t) = 0$ для заданной функции методом обратной интерполяции).

11. Построить линейный и кубический сплайны по значениям функции $f(t)$ в точках $\{0, 1\}$и $\{0, 1, 2\}$ соответственно.

12. Показать, что система линейных алгебраических уравнений для определения коэффициентов сплайна (6.3) всегда имеет единственное решение. Показать, что для этой системы устойчив метод прогонки. Оценить число обусловленности системы (6.3) в случае, когда $h_n \equiv h$ не зависит от номера узла.

Литература

[1] *Федоренко Р.П.* Введение в вычислительную физику. М.: Изд-во МФТИ, 1994. 528 с.

[2] *Рябенький В.С.* Введение в вычислительную математику. М.: Физматлит, 2000. 294 с.

[3] *Марчук Г.И.* Методы вычислительной математики. М.: Наука, 1989. 608 с.

[4] *Вержбицкий В.М.* Численные методы математический анализ и обыкновенные дифференциальные уравнения. М.: Высшая школа, 2001. 382 с.

[5] *Завьялов Ю.С., Квасов Б.И., Мирошниченко В.Л.* Методы сплайн-функций. М.: Наука, 1980. 352 с.

[6] *Завьялов Ю.С., Леус В.А., Скороспелов В.А.* Сплайны в инженерной геометрии. М.: Машиностроение, 1985. 224 с.

[7] *Вершинин В.В., Завьялов Ю.С., Павлов Н.Н.* Экстремальные свойства сплайнов и задача сглаживания. Новосибирск: Наука, 1988. 104 с.

[8] *Самарский А.А., Гулин А.В.* Численные методы. М.: Наука, 1989. 430 с.

[9] *Бабенко К.И.* Основы численного анализа. М.: Наука, 1986. 744 с.

[10] *Рябенький В.С., Филиппов А.Ф.* Об устойчивости разностных уравнений. М.: Гостехиздат, 1956. 160 с.

[11] *Рябенький В.С.* Метод разностных потенциалов для некоторых задач механики сплошной среды. М.: Наука, 1987. 320 с.

[12] *Ши Д.* Численные методы в задачах теплообмена. М.: Мир, 1988. 544 с.

[13] *Каханер Д., Моулер К., Нэш С.* Численные методы и программное обеспечение. М.: Мир, 1998. 575 с.

[14] *Калиткин Н.Н. и др.*// Математическое моделирование. 1994, т. 6, №4, с. 77-110, 1997, т. 9, №6, с. 67–81, 1997, т. 9, №9, с. 107–116.

Лекция 7. Численное интегрирование

Исследуются простейшие квадратурные формулы интерполяционного типа — прямоугольников, трапеций, Симпсона. Для оценки реальной погрешности формул используется правило Рунге. Дается понятие о квадратурных формулах Гаусса. Рассматриваются методы вычисления многомерных интегралов.

Ключевые слова: квадратурные формулы интерполяционного типа. Правильные квадратурные формулы. Кубатурные формулы. Квадратуры Гаусса.

Введение. В данной лекции будет рассматриваться задача численного интегрирования. Формулы численного интегрирования функций одного переменного называют *квадратурными формулами*. Задача приближенного вычисления определенного интеграла (на отрезке или по многомерной области) фактически разбивается на две самостоятельные подзадачи. Первая — это интегрирование таблично заданной функции (полученной, например, при проведении лабораторного эксперимента). В таком случае априорная информация о гладкости подынтегральной функции отсутствует, весьма ограничены возможности в выборе узлов интегрирования. Для этой задачи наиболее эффективными будут квадратурные формулы интерполяционного типа и правило Рунге оценки погрешности.

Вторая задача — подсчет значения определенного интеграла от известной функции. При этом самая ресурсоемкая операция с точки зрения вычислений — подсчет значения функции. Желательно построить численный метод, позволяющий получать как можно более высокую точность при наименьшем количестве вычислений, при этом выбор узлов квадратурных формул целиком в руках вычислителя. В этом случае наиболее эффективными окажутся квадратурные формулы типа Гаусса.

7.1. Квадратурные формулы интерполяционного типа (формулы Ньютона–Котеса)

Простейшую квадратурную формулу (формулу численного интегрирования) можно получить следующим образом. Пусть необходимо вычислить интеграл

$$I = \int_a^b f(t)dt.$$

Положим, что $f(t)$ на рассматриваемом отрезке $[a, b]$ не изменяется ($f(t) \approx const$). Тогда $I = f(\xi)(b-a), \xi \in [a, b]$. Если $\xi = \frac{a+b}{2}$, то получим формулу прямоугольников с центральной точкой

$$I \approx (b-a) \cdot f(\frac{a+b}{2}).$$

Конечно, для константы приведенная выше формула точна — говорят, что построенная квадратурная формула будет точна на полиномах степени 0. Легко можно доказать, что формула прямоугольников с центральной точкой будет давать точное значение и в случае линейной функции. Для всех других функций эту формулу будем рассматривать как приближенную.

Если предположить, что функция $f(t)$ на отрезке интегрирования $[a, b]$ достаточно близка к линейной, то можно заменить приближенное значение интеграла I площадью трапеции с высотой $(b-a)$ и основаниями $f(a)$ и $f(b)$. Тогда получается формула трапеций

$$I \approx (b-a)\frac{f(a)+f(b)}{2}.$$

В общем случае квадратурные формулы получаются при помощи интегрирования интерполяционного многочлена, аппроксимирующего подынтегральную функцию. Семейство квадратурных формул, получающихся таким образом, называется формулами интерполяционного типа (формулы Ньютона–Котеса).

Введем на отрезке интегрирования сетку, определим значения функции в узлах сетки. Узлы в дальнейшем будем именовать узлами квадратурной формулы (или квадратуры). Пусть, как и в задаче интерполяции, имеется совокупность узлов $\{t_n\}_{n=0}^{N}, t_n = a + n\tau, \tau = (b-a)/N, t \in [a, b]$. Пусть также задана таблица $f_n = \{f(t_n)\}_{n=0}^{N}$. Отрезок $[t_k, t_{k+1}]$ далее иногда будем называть элементарным отрезком.

Заменим подынтегральную функцию ее интерполяционным полиномом в форме Лагранжа. Будем полагать, что

$$\int_a^b f(t)dt \approx \int_a^b L_n(t)dt.$$

Рассмотрим некоторые частные случаи.

Формула трапеций. На отрезке $[t_k, t_{k+1}]$ проводим замену подынтегральной функции интерполяционным полиномом первой степени:

$$f(t) \approx f(t_k) + \frac{f(t_{k+1}) - f(t_k)}{t_{k+1} - t_k}(t - t_k),$$

после чего, выполнив интегрирование по элементарному отрезку, получим приближенное значение интеграла на $[t_k, t_{k+1}]$:

$$I_k \approx \frac{1}{2}(t_{k+1} - t_k)\,[f(t_{k+1}) + f(t_k)] = \frac{\tau_k}{2}\,[f(t_{k+1}) + f(t_k)]\,.$$

После суммирования интегралов по всем элементарным отрезкам $[t_k, t_{k+1}]$ получаем формулу трапеций для отрезка $[a, b]$:

$$I \approx \sum_{k=0}^{N-1} I_k = \frac{1}{2}\sum_{k=0}^{N-1} \tau_k\,[f(t_k) + f(t_{k+1})]\,,$$

$$\tau_k = t_{k+1} - t_k, \quad k = 0 \div N; \quad \sum_{k=0}^{N-1} \tau_k = b - a.$$

На равномерной сетке (сетке с равноотстоящими узлами) при $\tau_k = \tau = (b-a)/N$ полученная формула принимает вид

$$I \approx \frac{\tau}{2}\sum_{k=0}^{N-1} [f(t_k) + f(t_{k+1})] = \frac{\tau}{2}\,[f(t_0) + 2f(t_1) + \ldots + 2f(t_{N-1}) + f(t_N)]\,.$$

Формула Симпсона. Заменим подынтегральную функцию $f(t)$ на отрезке $[t_{k-1}, t_k]$ интерполяционным полиномом (в форме Лагранжа) второй степени. Для простоты положим $\tau = (b-a)/N = t_k - t_{k-1} = \text{const}$ для всех k — сетка на отрезке интегрирования равномерная. Тогда

$$F(t) = f_{k-1}\frac{(t - t_{k-1/2})(t - t_k)}{(t_{k-1} - t_{k-1/2})(t_{k-1} - t_k)} + f_{k-1/2}\frac{(t - t_{k-1})(t - t_k)}{(t_{k-1/2} - t_{k-1})(t_{k-1/2} - t_k)} +$$

$$+ f_k\frac{(t - t_{k-1})(t - t_{k-1/2})}{(t_k - t_{k-1})(t_k - t_{k-1/2})} = \frac{2}{\tau^2}\,[(t - t_{k-1/2})(t - t_k)f_{k-1} -$$

$$-2(t - t_{k-1})(t - t_{k-1/2})f_{k-1/2} + (t - t_{k-1})(t - t_{k-1/2})f_k]\,.$$

После вычисления интеграла от полинома получим приближенное значение интеграла по элементарному отрезку

$$I_k \approx \frac{\tau}{6}\,[f_{k-1} + 4f_{k-1/2} + f_k]\,.$$

Суммируя по всем элементарным отрезкам $[t_{k-1}, t_k]$, получим

$$I \approx \frac{\tau}{6}\sum_{k=1}^{N} [f_{k-1} + 4f_{k-1/2} + f_k] =$$

$$= \frac{\tau}{6}\,(f_0 + 4f_{1/2} + 2f_1 + 4f_{3/2} + \ldots + 2f_{N-1} + 4f_{N-1/2} + f_N)\,,$$

где $f_k = f(t_k)$, $f_{k+1/2} = f(\frac{t_k+t_{k+1}}{2})$. Формулу Симпсона можно также записать, не используя дробных индексов:

$$I \approx \frac{\tau}{6}(f_0 + 4f_1 + 2f_2 + 4f_3 + \ldots + 2f_{N-2} + 4f_{N-1} + f_N),$$

если локальную формулу получать путем интегрирования интерполяционного полинома второй степени по отрезку $[t_{k-1}, t_{k+1}]$:

$$I_k \approx \int_{t_{k-1}}^{t_{k+1}} F(t)dt = \frac{\tau}{3}(f_{k-1} + 4f_k + f_{k+1}),$$

где $F(t)$ — интерполяционный полином, построенный на отрезке $[t_{k-1}, t_{k+1}]$ по точкам t_{k-1}, t_k, t_{k+1}. В этом случае N — число разбиений отрезка на элементарные отрезки — должно быть четным.

Еще одна используемая на практике квадратурная формула интерполяционного типа — так называемое «правило 3/8». Она получается при замене подынтегральной функции интерполяционным полиномом третьей степени, построенным по четырем точкам. Расчетные формулы для правила 3/8 приведем без вывода:

$$\int_a^b f(t)dt \approx (b-a)\left[\frac{1}{8}f(a) + \frac{3}{8}f\left(\frac{2a+b}{3}\right) + \frac{3}{8}f\left(\frac{a+2b}{3}\right) + \frac{1}{8}f(b)\right].$$

Квадратурные формулы интерполяционного типа более высокого порядка применяются достаточно редко.

Конечно, существуют и формулы интерполяционного типа более высоких порядков. Они не применяются на практике по следующим обстоятельствам. Любая формула интерполяционного типа записывается в виде

$$I \approx \sum_{k=0}^{M} \alpha_k f_k.$$

Во всех приведенных выше формулах коэффициенты α_k были положительными. Такие квадратурные формулы называются *правильными* квадратурными формулами. При использовании полиномов более высоких степеней получаются квадратурные формулы, не являющиеся правильными.

Для степени интерполяционного полинома более 7 среди коэффициентов встречаются отрицательные. Д. Пойа показал, что

$$\lim_{n\to\infty} \sum_{k=0}^{n} |\alpha_{nk}| = \infty,$$

где α_{nk} — веса квадратурной формулы, получающейся при замене подынтегрального выражения интерполяционным полиномом степени n. Такое увеличение суммы абсолютных значений коэффициентов связано с быстрым ростом постоянной Лебега при алгебраической интерполяции на равномерной сетке.

Подробнее о свойствах квадратурных формул интерполяционного типа можно прочитать в [6].

7.2. Оценка погрешности квадратурных формул

Погрешность квадратурных формул может быть оценена, например, с использованием остаточного члена интерполяционного полинома:

$$\varepsilon_k = \left| \int_{t_k}^{t_{k+1}} R_N(t)dt \right| \leqslant \frac{\max\limits_{[t_k,t_{k+1}]} \left| f^{(N+1)}(\xi) \right|}{(N+1)!} \tau \max_{[t_k,t_{k+1}]} \left| \prod_{k=0}^{N} (t-t_k) \right|.$$

Из последней формулы следует, что квадратурная формула точна, если подынтегральная функция является многочленом степени не выше N.

Получим, например, локальную оценку погрешности для формулы трапеций, используя формулу для остаточного члена интерполяционного полинома 1-го порядка:

$$\varepsilon_k \leqslant \int_{t_k}^{t_{k+1}} \frac{\max |f''(\xi)|}{2} \max |(t-t_k)(t-t_{k+1})|\, dt = \frac{\max |f''(\xi)|}{12} \tau^3.$$

Тогда погрешность по всему отрезку $[a, b]$ будет составлять

$$\varepsilon_N \leqslant \sum_{k=1}^{N} |R_k| \leqslant \frac{\max |f''(\xi)|}{12} \tau^3 N = \frac{\max |f''(\xi)|}{12} \tau^2 (b-a).$$

Другой способ получения погрешности квадратурных формул состоит в следующем. Рассмотрим интеграл по элементарному отрезку

$$I_k = \int_{t_k}^{t_{k+1}} f(t)\, dt,$$

$$f(t) = f(z) + f'(z)(t-z) + \frac{f''}{2}(t-z)^2 + \ldots + R(t-z),$$

где $z \in [t_k, t_{k+1}]$ — некая опорная точка, тогда для приближенного значения интеграла верно

$$I_k = f(z)\tau_k + \xi\tau_k^2 + \eta\tau_k^3 + \ldots + \int_{t_k}^{t_{k+1}} R(t-z)dz, \tau_k = t_{k+1} - t_k.$$

Коэффициенты $\xi, \eta, \ldots$ зависят от производных $f'(z), f''(z), \ldots$

С другой стороны, любая из рассмотренных квадратурных формул представима в виде

$$\bar{I}_k = \tau_k(af_k + bf_{k+1/2} + cf_{k+1}).$$

Заменяя в этой формуле значения функций f в точках $f_k, f_{k+1/2}, f_{k+1}$ ее разложением по формуле Тейлора, получим

$$\bar{I}_k = f(z)\tau_k + \xi_1\tau_k^2 + \eta_1\tau_k^3 + \ldots + \bar{R},$$

где $z \in [t_k, t_{k+1}]$.

Сравнивая разложения для I_k, $\bar{I}_k$, легко заметить, что вместе с первым слагаемым совпадают и другие слагаемые до $(m-1)$-го порядка, так что $\xi = \xi_1, \eta = \eta_1, \ldots$

Разность же несовпадающих слагаемых будет, очевидно, оценкой погрешности квадратурной формулы на интервале $[t_k, t_{k+1}]$: $\varepsilon_k = |I_k - \bar{I}_k| \leqslant v \max\limits_{[t_k, t_{k+1}]} |f^{(m)}| \tau_k^{m+1}$, где v — константа.

Если просуммировать локальные погрешности по всем интервалам $[t_k, t_{k+1}]$, то получим оценку погрешности квадратурной формулы по всему отрезку $[a, b]$:

$$\varepsilon = |I - \bar{I}| \leqslant v(b-a) \max_{[a,b]} \left|f^{(m)}\right| \tau^m,$$

где $\tau = \max\limits_k \tau_k$ на неравномерной сетке, или $\tau = (b-a)/N$ на равномерной. Число m называется порядком точности квадратуры.

Получим теперь погрешность формулы прямоугольников (со средней точкой) для $\tau_k = \tau = \text{const}$:

$$\varepsilon \leqslant \frac{b-a}{24} \max_{[a,b]} |f_t''(t)| \tau^2,$$

погрешность формулы трапеций

$$\varepsilon \leqslant \frac{b-a}{12} \max_{[a,b]} |f_t''(t)| \tau^2,$$

погрешность формул Симпсона (с дробными и без дробных индексов соответственно)

$$\varepsilon \leqslant \frac{b-a}{2880} \max_{[a,b]} \left| f_t^{(4)}(t) \right| \tau^4,$$

$$\varepsilon \leqslant \frac{b-a}{180} \max_{[a,b]} \left| f_t^{(4)}(t) \right| \tau^4.$$

Заметим, что если функция $f(t)$ имеет только три непрерывных производных, то оценка погрешности формулы Симпсона ухудшается на порядок:

$$\varepsilon \leqslant \frac{b-a}{12} \max |f_t'''(t)| \tau^3$$

7.3. Кратные интегралы

Рассмотрим, для примера двукратный интеграл по прямоугольной области $\Omega(a \leqslant x \leqslant b, c \leqslant y \leqslant d)$. Аналогично одномерному случаю, в соответствии с формулой прямоугольников (со средней точкой) заменим функцию ее значением в точке пересечения диагоналей прямоугольника. В таком случае получим $I = \int_a^b \int_c^d f(x,y)dxdy \approx Sf(\frac{a+b}{2}, \frac{c+d}{2})$, где $S = \tau_x \tau_y, \tau_x = (b-a), \tau_y = (d-c)$. Если разбить прямоугольник на прямоугольные ячейки S_i, то получим аналог формулы прямоугольников со средней точкой в виде суммы по i: $I \approx \sum_i S_i f(x_i, y_i)$, где S_i — площадь i-ой прямоугольной ячейки, x_i, y_i — координаты точки пересечения ее диагоналей.

Применим теперь формулу Симпсона для вычисления двукратного интеграла путем редукции к методу вычисления одномерного интеграла, зачем представим двойной интеграл как

$$I = \int_a^b \int_c^d f(x,y)\,dx\,dy = \int_a^b dx \int_c^d f(x,y)\,dy.$$

Сначала применим формулу Симпсона для вычисления внешнего интеграла:

$$I \approx \frac{\tau_x}{6} \left[\int_c^d f(a,y)\,dy + 4 \int_c^d f(\frac{a+b}{2}, y)\,dy + \int_c^d f(b,y)\,dy \right].$$

Теперь применим формулу Симпсона для каждого из трех полученных интервалов:

$$I_1 = \int_c^d f(a,y)dy \approx \frac{\tau_y}{6}\left[f(a,c) + 4f(a,\frac{c+d}{2}) + f(a,d)\right],$$

$$I_2 = \int_c^d f(\frac{a+b}{2},y)dy \approx \frac{\tau_y}{6}\left[f(\frac{a+b}{2},c) + 4f(\frac{a+b}{2},\frac{c+d}{2}) + f(\frac{a+b}{2},d)\right],$$

$$I_3 = \int_c^d f(b,y)dy \approx \frac{\tau_y}{6}\left[f(b,c) + 4f(b,\frac{c+d}{2}) + f(b,d)\right].$$

Подставляя приближенные формулы для вычисления интегралов I_1, I_2, I_3 в формулу для вычисления I, получим

$$I \approx \frac{\tau_x\tau_y}{36}\{f(a,c) + f(a,d) + f(b,c) + f(b,d) +$$

$$+4\left[f(a,\frac{c+d}{2}) + f(b,\frac{c+d}{2}) + f(\frac{a+b}{2},c) + f(\frac{a+b}{2},d)\right] + 16f\left(\frac{a+b}{2},\frac{c+d}{2}\right)\Big\}.$$

Если область интегрирования не является прямоугольной, то ее можно сделать подобластью большей по площади прямоугольной области, которую в свою очередь разбить на прямоугольные ячейки. Ячейки в этом случае разделяются на внутренние, к ним применяются приведенные формулы численного интегрирования, и граничные, не прямоугольные, площади которых вычисляются по более сложным алгоритмам. При этом приближенное значение интеграла, например, при использовании формулы средних, можно записать как $I \approx \sum I^i_{\text{вн}} + \sum I^j_{\text{гр}}$, где $I^i_{\text{вн}}$ — приближенное значение интегралов по внутренним ячейкам, $I^j_{\text{гр}}$ — по граничным, i, j — номера ячеек с площадями S_i и S_j.

7.4. Квадратурные формулы Гаусса

Поскольку формулы Ньютона–Котеса являются интерполяционными, очевидно, что они не могут успешно использоваться для получения формул высокой точности по причине неустойчивости интерполяционного процесса для многочленов высокого порядка. Как отмечалось выше, постоянные Лебега растут с увеличением количества узлов интерполяции для равномерной сетки как 2^N. По этой причине обычно используются полиномы степени от нуля до трех (соответственно, формулы пря-

моугольников со средней точкой, трапеций, Симпсона, 3/8). Вычисление с их помощью интегралов от функций, обладающих высокой степенью гладкости, например, близким к полиномам высокой степени, представляется нерациональным. В выражение для погрешности этих формул входят первая, вторая или четвертая производные. Погрешность определяется низким порядком производной при высокой степени гладкости интегрируемой функции. Этих недостатков лишены квадратуры Гаусса.

Формулировка задачи построения квадратурных формул, поставленная Гауссом, такова.

Для заданного количества точек, а именно, для $(N+1)$ точки, найти такое расположение узлов и такие веса c_i, чтобы квадратурная формула

$$\int_a^b f(t)dt = \sum_{i=0}^{N} c_i f(t_i) + r_N(t)$$

была точной для полиномов как можно более высокой степени, т. е. чтобы $r_N(t) = 0$.

Пояснение. Для некоторых классов функций существуют квадратурные формулы с $r_N(t) = 0$, которые называются точными. Примером такого класса функций являются полиномы

$$P_N(t) = \sum_{k=0}^{N} a_k t^k$$

на отрезке $[a, b]$. Определим на этом отрезке узлы $t_i, i = 1, \ldots, N$ и веса c_i так, что

$$\int_a^b P_N(t)dt = \sum_{i=0}^{N} c_i P_N(t_i).$$

Представим $P_N(t)$ в виде интерполяционного полинома

$$P_N(t) = \sum_{i=0}^{N} P_N(t_i) \prod_{\substack{k=0 \\ k\neq i}}^{N} \frac{(t-t_k)}{(t_i-t_k)},$$

при этом остаточный член интерполяции полинома равен нулю: $P_N^{(N+1)}(t) = 0$.

Тогда из предыдущего условия следует

$$c_i = \int_a^b \prod_{\substack{k=0 \\ k\neq i}}^{N} \frac{(t-t_k)}{(t_i-t_k)}\, dt,$$

где c_i являются базисными функциями полиномов Лагранжа. Квадратурная формула

$$\int_a^b P_N(t)dt = \sum_{i=0}^{N} c_i P_N(t_i)$$

является точной для любого полинома степени N. Оказывается, эта формула может быть точной и для полиномов более высокой степени, а именно, $2N+1$, что используется при построении квадратурных формул Гаусса.

Пусть формула численного интегрирования имеет вид

$$I = \int_a^b f(t)dt = c_0 f(t_0) + c_1 f(t_1) + \ldots + c_N f(t_N) + r_N,$$

где c_i — веса, r_N — остаточный член квадратуры.

Положим, что существует многочлен $P_M(t)$ степени $M > N$, для которого квадратурная формула точна, т. е. $r_N = 0$ при $f(t) = P_M(t)$:

$$f(t) = P_M(t) = a_0 + a_1 t + a_2 t^2 + \ldots + a_M t^M,$$

где a_i — коэффициенты. В этом случае получим

$$a_0 \int dt + a_1 \int t dt + a_2 \int t^2 dt + \ldots + a_M \int t^M dt =$$

$$= c_0(a_0 + a_1 t_0 + a_2 t_0^2 + \ldots + a_M t_0^M) +$$
$$+ c_1(a_0 + a_1 t_1 + a_2 t_1^2 + \ldots + a_M t_1^M) + \ldots +$$
$$+ c_N(a_0 + a_1 t_N + a_2 t_N^2 + \ldots + a_M t_N^M).$$

Приравняем выражения в обеих частях равенства при a_j:

$$c_0 + c_1 + \ldots + c_N = \int_a^b dt = I_0,$$

$$c_0 t_0 + c_1 t_1 + \ldots + c_N t_N = \int_a^b t dt = I_1,$$

$$\ldots,$$

$$c_0 t_0^M + c_1 t_1^M + \ldots + c_N t_N^M = \int_a^b t^M dt = I_M.$$

Получается нелинейная система из $M+1$ уравнения с $2(N+1)$ неизвестными c_i, t_i. Отсюда следует, что максимальное значение M есть $2N+1$. Решение этой системы или исследование на его существование и единственность в общем случае затруднительны. Ниже будет рассмотрен пример получения квадратурной формулы Гаусса таким путем для двух узлов.

Гаусс решил эту задачу более простым (в смысле реализации, но не решения!) способом, доказав следующую теорему. Приведем ее без доказательства.

Теорема. *Если в качестве узлов $t_i, i = 0, \ldots, N$ в квадратурной формуле используются нули полиномов Лежандра $q_{N+1}(t)$, а веса c_i вычисляются по формулам*

$$c_i = \int_{-1}^{1} \prod_{\substack{k \neq i \\ k=0}}^{N} \frac{(t - t_k)}{(t_i - t_k)}\, dt,$$

то квадратурная формула

$$\int_{-1}^{1} f(t)dt = \sum_{i=0}^{N} c_i f(t_i) + r_N(t)$$

точна для полиномов степени $2N+1$.

Напомним, что *полиномы Лежандра* образуют ортогональную систему функций на отрезке [-1; 1]:

$$\int_{-1}^{1} q_i(t) q_j(t)\, dt = 0 \quad \text{при} \quad i \neq j; \int_{-1}^{1} q_i(t)\, q_j(t) dt \neq 0 \quad \text{при} \quad i = j.$$

Первые несколько полиномов Лежандра будут $q_0(t) = 1, q_1(t) = t, q_2(t) = \frac{1}{3}(3t^2 - 1), q_4(t) = \frac{1}{35}(35t^4 - 30t^2 + 3), \ldots$ рекуррентная и общая формулы имеют вид:

$$(n+1)q_{n+1}(t) = (2n+1)\, t\, q_n(t) - n q_{n-1}(t),$$

$$q_n(t) = \frac{1}{2^n (n!)} \frac{d^n}{dt^n}(t^2 - 1)^n.$$

Заметим, что рекуррентные формулы, связывающие три полинома порядка $n-1, n$ и $n+1$ уже встречались для полиномов Чебышева. Такие рекуррентные формулы существуют для всех систем *ортогональных полиномов*.

Погрешность квадратурной формулы Гаусса на отрезке будет $r_N(t) = 2^{2(N+1)+1}\alpha_N f^{(2(N+1))}(\xi)$, при этом $\xi \in [-1,1]$. Для $\xi \in [a,b]$ формула остаточного члена будет $r_N(t) = (b-a)^{2(N+1)+1}\alpha_N f^{(2(N+1))}(\xi)$, причем коэффициент α_N быстро убывает с ростом N. Здесь $\alpha_N = \frac{[(N+1)!]^4}{\{[2(N+1)]!\}^3[2(N+1)+1]}$.

Формулы Гаусса обеспечивают высокую точность уже при небольшом количестве узлов (от 4 до 10) В этом случае $\alpha_2 \approx 5 \cdot 10^{-7}$, $\alpha_3 \approx 6 \cdot 10^{-10}, \alpha_4 \approx 4 \cdot 10^{-13}$. В практических же вычислениях число узлов составляет от нескольких сотен до нескольких тысяч. Отметим также, что веса квадратур Гаусса всегда положительны, что обеспечивает устойчивость алгоритма вычисления сумм $\sum_{i=0}^{N} c_i f(t_i)$.

7.5. Вычисление интегралов от функций с особенностями

Пусть требуется вычислить несобственный интеграл

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

от функции, обращающейся в бесконечность в некоторой точке $c \in [a,b]$. В этом случае интеграл обычно разбивают на два

$$\int_a^b f(t)dt = \lim_{\substack{\delta_1 \to 0 \\ \delta_2 \to 0}} \left\{ \int_a^{c-\delta_1} f(t)dt + \int_{c+\delta_2}^b f(t)dt \right\}.$$

Числа δ_1 и δ_2 выбирают малыми величинами так, чтобы выполнялась оценка:

$$\left| \int_{c-\delta_1}^{c+\delta_2} f(t)dt \right| < \frac{1}{2}\varepsilon,$$

где ε — заданное малое положительное число (точность вычисления интеграла). После этого по квадратурным формулам вычисляют определенные интегралы $I_1 = \int_a^{c-\delta_1} f(t)dt$ и $I_2 = \int_{c+\delta_2}^b f(t)\,dt$ с точностью $\varepsilon/4$ каждый. После таких вычислений за приближенное значение интеграла с особенностью принимают $\int_a^b f(t)dt \approx I_1 + I_2$ (с точностью ε).

Другой способ вычисления интеграла от функции особенностью, называемый методом Канторовича выделения особенностей, состоит в следующем. Представим подынтегральную функцию в виде суммы:

$$\int_a^b f(t)dt = \int_a^b g(t)dt + \int_a^b [f(t) - g(t)]\,dt.$$

При этом $g(t)$ подбирают так, чтобы она была интегрируемой, а разность $[f(t) - g(t)]$ — ограниченной.

Пример. Пусть необходимо вычислить $I = \int_0^1 \frac{dt}{\sqrt{t(1+t^2)}}$.

Представим I как сумму двух интегралов $I = I_1 + I_2$, где $I_1 = \int_0^1 \frac{dt}{\sqrt{t}}$, $I_2 = \int_0^1 [\frac{1}{\sqrt{t(1+t^2)}} - \frac{1}{\sqrt{t}}]dt$.

Интеграл I_1 вычисляется аналитически, а I_2, поскольку подынтегральная функция ограничена, можно вычислить по квадратным формулам.

Аналогично можно поступить и в следующей задаче:

$$\int_0^1 \frac{e^{-t^2}}{\sqrt{t}}dt = \int_0^1 \frac{1-t^2}{\sqrt{t}}dt + \int_0^1 \frac{e^{-t^2} - 1 + t^2}{\sqrt{t}}dt.$$

Интегрирование быстро осциллирующих функций типа $I = \int_a^b f(t)e^{i\omega t}dt$ можно проводить, заменив $f(t)$ на интерполяционный полином, $I \approx \int_a^b \mathrm{P}_N(t)e^{i\omega t}dt$. Этот интеграл вычисляется явно.

7.6. Идея метода Монте-Карло

Метод Монте-Карло используется, как правило, для вычисления кратных интегралов. Рассмотрим задачу вычисления интеграла по многомерному кубу:

$$I = \int_0^1 \int_0^1 \dots \int_0^1 f(t_1, t_2, \dots, t_n)dt_1dt_2 \dots dt_n.$$

Для его вычисления можно построить кубатурные формулы, используя процедуру последовательного интегрирования, заменяя кратный

интеграл I на

$$I = \int_0^1 dt_1 \int_0^1 dt_2 \ldots \int_0^1 dt_n f(t_1, t_2, \ldots, t_n).$$

Проблема вычисления подобных интегралов заключается в том, что при росте размерности задачи объем вычисления значительно увеличивается, а задача численного интегрирования превращается из довольно простой в одну из самых сложных и трудоемких. По этой причине приведенные выше квадратурные формулы используются обычно для решения одно-, дву- и трехмерных задач.

Для вычисления интегралов по гиперкубу высокой размерности обычно используется метод Монте-Карло. Суть его состоит в том, что генерируется последовательность случайных точек единичного n-мерного куба $t_1,\ t_2, \ldots, t_n \in R^n$; очевидно, что чем больше точек участвует в вычислительном процессе, тем больше точность расчета.

Пусть теперь необходимо взять интеграл по области Ω, принадлежащей n-мерному кубу, причем, Ω выделяется неравенствами

$$g_j(t_k) \leqslant 0, \quad j = 1 \div J, \quad k = 1 \div K.$$

Далее генерируется последовательность случайных чисел, равномерно распределенная в единичном гиперкубе, и для всех точек проверяются неравенства $g_j(t_k) \leqslant 0$. Если они выполнены, т. е. $t_k \in \Omega$, то вычисляются значения $f(t_k)$, прибавляющиеся к сумме.

Пусть вычислено M точек, из которых K попали в Ω и накоплена сумма $\sum_{k=1}^{K} f(t_k)$.

Среднее по объему Ω значение функции f вычисляется по формуле $\bar{f} = \frac{I_\Omega}{S_\Omega}$, где $S_\Omega = K/M$, I_Ω — кратный интервал по Ω.

С другой стороны, это же значение можно приближенно вычислить как сумму:

$$\left[\sum f(t_k)\right] / K.$$

Приравнивая эти выражения, получим:

$$\frac{M}{K} I_\Omega \approx \frac{1}{K} \sum_{k=1}^{K} f(t_k), \text{ откуда } I_\Omega \approx \frac{1}{M} \sum_{k=1}^{K} f(t_k).$$

7.7. Задачи

1. Получить локальную и глобальную оценки погрешности для формулы трапеций.

Решение. Локальную оценку погрешности получим, используя формулы для истинного члена интерполяционного полинома:

$$\varepsilon_N = \left| \int_{t_n}^{t_{n+1}} R_N(t)dt \right| \leqslant \int_{t_n}^{t_{n+1}} \frac{\max |f^{(N+1)}(\xi)|}{(N+1)!}\tau + \max \left| \prod_{n=0}^{N} (t - t_n) \right|.$$

В случае формулы трапеции имеем

$$\varepsilon_N \leqslant \int_{t_n}^{t_{n+1}} \frac{\max |f''(\xi)|}{2} \max |(t - t_n)(t - t_{n+1})|\, dt = \frac{\max |f''(\xi)|}{12}\tau^3.$$

Погрешность для всего отрезка $[a, b]$ будет следующей ($\tau = (b - a)(N, t_m = n\tau)$:

$$\varepsilon_N \leqslant \sum_{n=1}^{N} |\varepsilon_m| \leqslant \frac{\max |f''(\xi)|}{12}\tau^3 N = \frac{\max |f''(\xi)|}{12}\tau^3(b - a).$$

2. Получить квадратурную формулу Гаусса для двух узлов на отрезке $t \in [-1, 1]$.

Решение. В случае двух узлов $N = 1$ (количество отрезков разбиения), $M = 2, N + 1 = 3$ (степень полинома). Узлы t_n и веса c_n должны удовлетворять следующей системе уравнений:

$$\sum_{n=0}^{N} c_n t_n^j = \int_{-1}^{1} t^j dt,$$

$$\sum_{n=0}^{N} c_n t_n^j = \frac{1 - (-1)^{j+1}}{j+1},\ j = 0, \ldots, M.$$

В данном случае система уравнений будет:

$$c_0 + c_1 = \int_{-1}^{1} 1\ dt = 2,$$

$$c_0 t + c_1 t_1 = \int_{-1}^{1} t\ dt = 0,$$

$$c_0 t_0^2 + c_1 t_1^2 = \int_{-1}^{1} t^2 dt = \frac{2}{3},$$

$$c_0 t_0^3 + c_1 t_1^3 = \int_{-1}^{1} t^3 dt = 0,$$

откуда получим

$$c_0 = c_1 = 1; \quad t_0 = t_1 = \frac{1}{\sqrt{3}}.$$

Формула Гаусса записывается как

$$\int_{-1}^{1} f(t)dt \approx f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right).$$

Эта формула будет точной для полиномов третьей степени.

3. Предложить способ вычисления интеграла

$$I = \int_0^1 \frac{1}{\sqrt{x}} e^{-x^2} dx.$$

Решение. Представим интеграл I как сумму двух интегралов

$$I = \int_0^1 \frac{1}{\sqrt{x}} e^{-x^2} dx = \int_0^1 \frac{1-x^2}{\sqrt{x}} dx + \int_0^1 \frac{e^{-x^2} - 1 + x^2}{\sqrt{x}} dx = I_1 + I_2.$$

Первый интеграл I_1 вычисляется аналитически: $I_1 = 1,6$. Поскольку подынтегральная функция в I_2 трижды непрерывно дифференцируема и ограничена, то I_2 можно вычислить, например, по формуле прямоугольников с центральной точкой:

$$I \approx 1,6 + h \sum_{n=1}^{N} \frac{1}{\sqrt{x_{n-1/2}}} \left(e^{-x_{n-1/2}^2} - 1 + x_{n-1/2}^2\right).$$

4. Предложить способ уточнения приближенного значения интеграла по квадратурной формуле, при вычисленных значениях этого интеграла при двух шагах интегрирования: h и $h/2$ (**правило Рунге**).

Решение. Интеграл I может быть представлен в виде

$$I = I^p(h) + Ch^p;\ I = I^p\left(\frac{h}{2}\right) + 2C_1\left(\frac{h}{2}\right)^p,$$

где $I^p(h)$ — приближенные значения интеграла, вычисленные по формуле с порядком точности p с шагом h, $I^p\left(\frac{h}{2}\right)$ — значение интеграла, вычисленное по той же формуле с шагом вдвое меньшим. При малых h константы C и C_1 близки. Этот факт тоже необходимо доказывать. Доказательство труда не представляет — пользуясь теоремой Лагранжа о среднем, легко получить, что эти величины отличаются на $O(h_p)$. Тогда получим

$$I^p(h) + Ch^p = I^p\left(\frac{h}{2}\right) + 2C_1\left(\frac{h}{2}\right) \approx I^p\left(\frac{h}{2}\right) + 2C\left(\frac{h}{2}\right)^p,$$

откуда следует

$$C \approx C_1 \approx \frac{I^p\left(\frac{h}{2}\right) - I^p(h)}{h^p - 2\left(\frac{h}{2}\right)^p}.$$

Подставив C во вторую формулу для вычисления $I(c, h/2)$, получим:

$$I \approx I^p\left(\frac{h}{2}\right) + \frac{2^{p-1}I^p\left(\frac{h}{2}\right) - I^p(h)}{2^{p-1} - 1}.$$

Во-первых, эта простая формула позволяет относительно дешевым способом уточнить вычислительное значение интеграла с шагом $h/2$. Во-вторых, получаем возможность контролировать точность численного интегрирования путем вычисления значения интеграла дважды (с шагами h и $h/2$).

Примечание. Легко получается аналог правила Рунге при вычислении интеграла для табличной функции. Необходимо лишь с использованием одних и тех же квадратурных формул вычислить интеграл с шагом таблицы h и затем повторить вычисления, выкинув половину точек, с шагом $2h$.

7.8. Задачи для самостоятельного решения

1. Предложить алгоритмы вычисления интегралов от быстро осциллирующих функций:

$$\int_0^1 \frac{\sin 100x}{1+x}dx,\ \int_1^2 \cos 100x \ln x dx$$

(формулы Филона. Подробнее о них в [3]).

2. Предложить алгоритм вычисления интегралов:

$$\int_0^{1,5} \frac{e^x}{x^2}dx, \int_0^1 \frac{\operatorname{arctg} x}{x^2}dx, \int_0^1 \frac{\sqrt{x^3+1}}{\sqrt{x}}dx,$$

$$\int_0^1 \frac{\cos x - 1}{x^2}dx, \int_0^1 \frac{\cos x}{\sqrt{x}}dx, \int_0^1 \frac{x \sin x}{\ln(1+x)}dx,$$

$$\int_1^\infty \frac{1-\cos x}{x\sqrt{x}}dx, \int_1^0 \frac{\sin x}{x}dx, \int_0^1 \frac{\ln(1+x)}{x}dx.$$

3. Доказать, что формула Симпсона точна, если подынтегральная функция есть произвольный многочлен третьей степени.

4. Определить число π по формуле

$$\pi = \int_0^1 \frac{4}{1+x^2}dx,$$

используя формулы прямоугольников с центральной точкой, трапеций, Симпсона.

5. Оценить число разбиений отрезка для вычисления интеграла

$$\int_0^1 \sin x^2 dx$$

по формуле прямоугольников с центральной точкой для достижения точности $\varepsilon = 10^{-4}$. Тот же вопрос для подсчета интеграла

$$\int_0^1 \exp(x^2)\,dx.$$

6. Оценить погрешность при вычислении интеграла

$$I = \int_0^1 \frac{dx}{1+x^2}$$

по формуле прямоугольников с центральной точкой, трапеций, Симпсона.

7. На элементарном отрезке $[x_i, x_{i-1}]$ подынтегральная функция аппроксимируется кубическим сплайном:

$$S_3^{(n)} = a_n + b_n(x - x_n) + \frac{c_n}{2}(x - x_n)^2 + \frac{d_n}{6}(x - x_n)^3.$$

Вывести соответствующую формулу сплайн-квадратуры и исследовать ее точность.

Литература

[1] *Калиткин Н.Н.* Численные методы. М.: Наука, 1978. 512 с.

[2] *Березин И.С., Жидков Н.П.* Методы вычислений. Т. 2. М.: Физматгиз, 1962. 464 с.

[3] *Косарев В.И.* 12 лекций по вычислительной математике. М.: Изд-во МФТИ, Физматкнига, 2000. 220 с.

[4] *Федоренко Р.П.* Введение в вычислительную физику. М.: Изд-во МФТИ, 1994. 528 с.

[5] *Каханер Д., Моулер К., Нэш С.* Численные методы и программное обеспечение. М.: Мир, 1998. 575 с.

[6] *Бабенко К.И.* Основы численного анализа. М.: Наука, 1986. 744 с.

Лекция 8. Численные методы решения задачи Коши для систем обыкновенных дифференциальных уравнений

Подробно рассматриваются методы типа Рунге-Кутты, менее подробно — Адамса. Формулируются и доказываются утверждения об устойчивости методов Рунге-Кутты на устойчивых и нейтральных по устойчивости траекториях.

Ключевые слова: методы Рунге-Кутты, Адамса, таблицы Бутчера, аппроксимация, устойчивость, сходимость, устойчивые траектории, нейтральные траектории.

8.1. Базовые понятия

Рассмотрим численные методы решения задачи Коши для обыкновенных дифференциальных уравнений (ОДУ) вида

$$\begin{gathered}\frac{du(t)}{dt} = f(t,u), \quad t > 0, \\ u(0) = u_0,\end{gathered} \tag{8.1}$$

а также систем ОДУ

$$\begin{gathered}\frac{d\mathbf{u}(t)}{dt} = \mathbf{f}(t,\ \mathbf{u}), \\ \mathbf{u}(0) = \mathbf{u}_0,\end{gathered}$$

где

$$\mathbf{u} = (u_1, u_2, \ldots, u_m)^T, \mathbf{f} = (f_1, f_2, \ldots, f_m)^T$$

— векторы столбцы искомых функций и правых частей соответственно.

К аналогичной форме приводится задача Коши для обыкновенного дифференциального уравнения (системы уравнений) порядка выше первого, вида

$$\frac{d^m u}{dt^m} = g\left(t,\ u,\ \frac{du}{dt}, \ldots, \frac{d^{m-1}u}{dt^{m-1}}\right), \quad t > 0,$$

$$u(0) = a_0,$$

$$\frac{du}{dt}(0) = a_1, \ldots, \frac{d^{m-1}u}{dt^{m-1}}(0) = a_{m-1},$$

если положить

$$u_1 = u, u_2 = \frac{du_1}{dt}, u_3 = \frac{du_2}{dt}, \ldots, u_m = \frac{du_{m-1}}{dt},$$

$$\frac{du_m}{dt} = g(t,\ u_1,\ u_2,\ \ldots,\ u_m),$$

$$u_i(0) = a_{i-1}, i = 1,\ 2,\ \ldots,\ m.$$

Введем в расчетной области $t \in [0,\ T]$ точки (узлы расчетной сетки) $\{t_n = n\tau, n = 0,\ 1,\ \ldots,\ N\}$, в которых вычисляется искомое решение. Совокупность узлов называется *расчетной сеткой*, (сеточной областью), τ — *шагом интегрирования*. Здесь для простоты введена равномерная сетка. В реальных расчетах применяются и неравномерные сетки.

Введем сеточную функцию u^τ, определенную в узлах сетки и представляющую собой совокупность приближенных значений искомой функции, U^τ — проекцию точного решения искомой задачи на сетку и f^τ — значения правой части в узлах сетки.

Введем вслед за [1] также операторное обозначение дифференциальной задачи

$$L(u) = F, \tag{8.2}$$

где

$$L(u) = \begin{cases} \frac{du}{dt} - f(t,\ u), & t > 0; \\ u(0), & t = 0; \end{cases} \quad F = \begin{cases} 0, & t > 0; \\ u_0, & t = 0; \end{cases}$$

и аппроксимирующей разностной задачи

$$L_\tau(u^\tau) = F_\tau, \tag{8.3}$$

где L_τ — обозначения разностного оператора, F_τ — *проекция* F на расчетную сетку. Заметим, что u и u^τ являются элементами соответственно функционального и конечномерного пространств. Определим основные понятия теории разностных схем [1, 2].

Определение. Решение задачи (8.3) u^τ сходится при $\tau \to 0$ к решению исходной задачи (8.2), если

$$\|u^\tau - U^\tau\| \to 0$$

при $\tau \to 0$.

При этом, если имеет место оценка

$$\|u^\tau - U^\tau\| \leqslant C\,\tau^p (C \neq C(\tau)),$$

то имеет место сходимость порядка p.

Определение. Говорят, что задача (8.3) аппроксимирует задачу (8.2) на ее решении, если невязка

$$\|r_\tau\| \to 0$$

при $\tau \to 0$, где $r_\tau \equiv L_\tau(U^\tau) - F_\tau$; при этом, если имеет место оценка

$$\|r_\tau\| \leqslant C_1 \tau^p (C_1 \neq C_1(\tau)),$$

то говорят, что имеет место аппроксимация порядка p.

Определение. Задача (8.3) устойчива, если из соотношений

$$L_\tau(u^\tau) - F_\tau = \xi_\tau,$$
$$L_\tau(v^\tau) - F_\tau = \eta_\tau$$

следует

$$\| u^\tau - v^\tau \| \leqslant C_2 \left(\|\xi_\tau\| + \|\eta_\tau\|\right), C_2 \neq C_2(\tau).$$

Теорема 1 (В. С. Рябенького–П. Лакса). *Решение задачи* (8.3) *сходится к решению исходной задачи* (8.2), *если задача* (8.3) *устойчива и аппроксимирует задачу* (8.2); *если аппроксимация имеет порядок p, то сходимость также имеет порядок p.*

Доказательство.

В силу аппроксимации имеем оценку: $\|r_\tau\| \leqslant C_1 \tau^p$. Тогда из определения устойчивости, положив $v^\tau = U^\tau$, получим

$$\| u_\tau - U_\tau \| \leqslant C_2 \|r_\tau\| \leqslant C_2 C_1 \tau^p = C\tau^p,$$

поскольку в данном случае

$$\|\eta_\tau\| = 0$$

и, кроме того,

$$\|r_\tau\| = \|\xi_\tau\|.$$

■

Приведем примеры простейших разностных уравнений, аппроксимирующих (8.1):

$$\frac{u_{n+1} - u_n}{\tau} = f(t_n,\ u_n), 0 \leqslant n \leqslant N-1,$$

$$\frac{u_{n+1} - u_n}{\tau} = f(t_n,\ u_{n+1}), 0 \leqslant n \leqslant N-1,$$

$$\frac{u_{n+1} - u_{n-1}}{2\tau} = f(t_n,\ u_n), 1 \leqslant n \leqslant N-1.$$

Первая из схем называется *явной* (явная схема Эйлера), вторая — *неявной* (неявная схема Эйлера). Алгоритмическая реализация первой схемы — бегущий счет (рекуррентная формула), второй — решение нелинейного алгебраического уравнения на каждом временном шаге.

Для реализации третьей схемы необходимо задание функции u_n в двух точках: t_0 и t_1. Один из возможных вариантов — решение на первом шаге нелинейного уравнения вида:

$$\frac{u_{n+1} - u_{n-1}}{2\tau} = \frac{1}{2}\left[f(t_{n-1},\ u_{n-1}) + f(t_{n+1},\ u_{n+1})\right]$$

при $n = 1$.

В данном случае проявляется несовпадение формальных порядков дифференциального и разностного уравнений (дифференциальное уравнение первого порядка, разностное — второго).

Один из первых методов приближенного решения обыкновенных дифференциальных уравнений — разложение в ряд Тейлора.

Дифференцируя по t исходное уравнение (8.1), получим

$$u'' = f'_t(t,\ u) + f'_t(t,\ u)\,u',$$

$$u''' = f''_{tt}(t,\ u) + 2f''_{tu}(t,\ u)\,u' + f''_{uu}(t,\ u)\,(u')^2 + f'_u(t,\ u)\,u'', \ldots$$

Таким образом, можно написать приближенное равенство:

$$u(t) \approx \tilde{V}_n(t) \equiv \sum_{i=0}^{I} \frac{u^{(i)}(t_n)}{i!}\,(t - t_n)^i.$$

Полагая

$$u_{n+1} = \tilde{V}_n(t_{n+1}),$$

получаем приближенное значение $u(t)$ в точке $t = t_{n+1}$. При $I = 1$ и $t_{n+1} - t_n = \tau = \text{const}$ получаем метод Эйлера:

$$u_{n+1} = u_n + \tau f_n.$$

Этот способ не получил распространения в практике решения дифференциальных уравнений из-за необходимости вычисления производных $u^{(i)}$, где $i = 1 \div I$. По затратам машинного времени он заметно уступает другим методам, о которых будет идти речь далее.

В настоящее время в практике решения *жестких систем* ОДУ применяют так называемые *многозначные методы*, основанные на разложении в ряд Тейлора и вычислении производных. О жестких системах ОДУ будет рассказано ниже.

Рассмотрим еще один способ получения простейших одношаговых расчетных схем для численного решения уравнения (8.1), для чего напишем равенство

$$u(t+\tau) = u(t) + \int_0^t u'(t+\tau)\, d\tau.$$

После аппроксимации интеграла в правой части по формуле прямоугольников и замене его на величину $\tau u'(t)$, получим

$$u(t+\tau) = u(t) + \tau u'(t) + O(\tau^2),$$

или

$$u(t+\tau) = u(t) + \tau f(t,\ u) + O(\tau^2),$$

поскольку $u'(t) = f(t,\ u)$.

Опуская член $O(\tau^2)$ и обозначая $t = t_n, t+\tau = t_{n+1}, u(t) = u_n, u(t+\tau) = u_{n+1}$, получим метод Эйлера.

Если рассматриваемый интеграл заменить формулой трапеций, получим

$$u(t+\tau) = u(t) + \frac{\tau}{2}\,[u'(t) + u'(t+\tau)] + O(\tau^3),$$

откуда имеем

$$u_{n+1} = u_n + \frac{\tau}{2}\,[f(t_n,\ u_n) + f(t_{n+1},\ u_{n+1})].$$

Этот метод называется *неявным методом трапеций*. Для того чтобы метод был явным, его делают двухэтапным:

$$\tilde{u}_{n+1} = u_n + \tau f(f_n,\ u_n),$$

$$u_{n+1} = u_n + \frac{\tau}{2}[f(t_n, u_n) + f(t_{n+1}, \tilde{u}_{n+1})],$$

где $\tilde{u}_{n+1}$ — вспомогательная величина, вычисляемая на промежуточном этапе. Если этот же интеграл приблизить формулой прямоугольников со средней точкой, то получим

$$u(t+\tau) = u(t) + \tau u'\left(t + \frac{\tau}{2}\right) + O(\tau^3).$$

Снова воспользовавшись дифференциальным уравнением (8.1), преобразуем последнее выражение к виду

$$u(t+\tau) = u(t) + \tau f\left[t + \frac{\tau}{2},\ u\left(t + \frac{\tau}{2}\right)\right] + O(\tau^3).$$

Соответствующий неявный метод имеет вид

$$u_{n+1} = u_n + \tau f\left[t_n + \frac{\tau}{2},\ u\left(t_n + \frac{\tau}{2}\right)\right];$$

для его явной реализации можно воспользоваться следующей двухэтапной формулой:

$$u_{n+1/2} = u_n + \frac{\tau}{2} f(t_n,\ u_n),$$

$$u_{n+1} = u_n + \tau f\left(t_n + \frac{\tau}{2},\ u_{n+1/2}\right).$$

8.2. Методы Рунге-Кутты

Наиболее распространенными при численном решении обыкновенных дифференциальных уравнений являются методы Рунге-Кутты. Их принято представлять в следующей форме [3].

Определение. r-шаговый *явный* метод для численного решения задачи Коши для обыкновенного дифференциального уравнения (8.1):

$$\begin{aligned}
k_1 &= f(t_n,\ u_n),\\
k_2 &= f(t_n + \alpha_2\tau,\ \ u_n + \tau\beta_{21}k_1),\\
k_3 &= f(t_n + \alpha_3\tau,\ \ u_n + \tau\left(\beta_{31}k_1 + \beta_{32}k_2\right)), \ldots,\\
k_r &= f(t_n + \alpha_r\tau,\ \ u_n + \tau\left(\beta_{r1}k_1 + \ldots + \beta_{r,r-1}k_2\right)),\\
u_{n+1} &= u_n + \tau\left(\gamma_1 k_1 + \ldots + \gamma_r k_r\right),
\end{aligned} \tag{8.4}$$

где k_i — промежуточные вспомогательные величины.

Коэффициенты, определяющие конкретный метод, могут быть представлены в виде *таблицы Бутчера* (табл. 8.1). Нулевые коэффициенты β_{ij}, как правило, в таблице Бутчера не указывают.

Таблица 8.1

0					
α_2	β_{21}				
α_3	β_{31}	β_{32}			
$\ldots$	$\ldots$	$\ldots$	$\ldots$		
α_r	β_{r1}	β_{r2}	$\ldots$	β_{rr-1}	
	γ_1	γ_2	$\ldots$	γ_{r-1}	γ_r

Обычно также используют условие, предложенное Куттой без объяснений и не являющееся обязательным [3]:

$$\alpha_n = \sum_j \beta_{nj}.$$

Получим простейшие методы Рунге-Кутты. Для этого введем погрешность

$$\xi(\tau) = u(t+\tau) - \left[u(t) + \sum_{j=0}^{r} \gamma_j k_j\right]$$

и представим ее в виде разложения в ряд Маклорена

$$\xi(\tau) = \sum_{i=0}^{p} \frac{\xi^{(i)}(0)}{i!}\tau^i + \frac{\xi^{(p+1)}(\theta\tau)}{(p+1)!}\tau^{p+1},$$

где $\frac{\xi^{(p+1)}(\theta\tau)}{(p+1)!}\tau^{p+1}$ — остаточный член ряда; $0 < \theta < 1$.

Будем полагать (что можно сделать соответствующим выбором коэффициентов)

$$\xi(0) = \xi'(0) = \ldots = \xi^{(p)}(0) = 0.$$

В таком случае разложение для $\xi(\tau)$ имеет более простой вид:

$$\xi(\tau) = \frac{\xi^{(p+1)}(\theta\tau)}{(p+1)!}\tau^{p+1},$$

где p — порядок точности метода.

1. Пусть $p = 1, r = 1$. Тогда

$$\xi(\tau) = u(t+\tau) - u(t) - \tau\gamma_1 f(t,\ u),$$

отсюда

$$\xi(0) = 0,$$
$$\xi'(0) \equiv [u'(t+\tau) - \gamma_1 f(t,\ u)]|_{t=0} = f(t,\ u)(1-\gamma_1),$$
$$\xi''(0) = u''(t+\tau).$$

Видно, что условие $\xi(0) = 0$ выполняется лишь при $\gamma_1 = 0$, что соответствует методу Эйлера, при этом

$$\frac{u(t+\tau) - u(t)}{\tau} - f(t,\ u) = \frac{\xi''(t+\theta\tau)}{2}\tau = R_\tau,$$

где R_τ — невязка, имеющая первый порядок малости по τ.

2. Рассмотрим более сложный случай: $p = 2, r = 2$. Тогда

$$\xi(\tau) = u(t+\tau) - u(t) - \tau\gamma_1 f(t,u) - \tau\gamma_2 f(t+\alpha_2\tau, u+\beta_{21}\tau f(t,u)).$$

Вводя обозначения

$$\tilde{t} = t + \alpha_2\tau, \tilde{u} = u + \beta_{21}\tau f(t,\ u),$$

получим следующие выражения для производных погрешности ξ по аргументу τ:

$$\xi'(\tau) = u'(t+\tau) - \gamma_1 f(t,u) - \gamma_2 f(\tilde{t},\tilde{u}) -$$
$$-\tau\gamma_2[\alpha_2 f'_t(\tilde{t},\tilde{u}) + \beta_{21} f'_u(\tilde{t},\tilde{u}) f(t,u)],$$
$$\xi''(\tau) = u''(t+\tau) - 2\gamma_2\,[\alpha_2 f'_t(\tilde{t},\ \tilde{u}) + \beta_{21} f'_u(\tilde{t},\ \tilde{u})\, f(t,\ u)] -$$
$$-\tau\gamma_2[\alpha_2^2 f''_{tt}(\tilde{t},\tilde{u}) + 2\alpha_2\beta_{21} f''_{tu}(\tilde{t},\tilde{u}) f(t,u) + \beta_{21}^2 f''_{uu}(\tilde{t},\tilde{u}) f^2(t,u)],$$
$$\xi'''(\tau) = u''(t+\tau) - 3\gamma_2[\alpha_2^2 f''_{tt}(\tilde{t},\tilde{u}) + 2\alpha_2\beta_{21} f''_{tu}(\tilde{t},\tilde{u}) f(t,u) +$$
$$+\beta_{21}^2 f''_{uu}(\tilde{t},\tilde{u}) f^{(}t,u)] + o(\tau).$$

Поставив в эти выражения следующие равенства:

$$u' = f, u'' = f'_t + f'_u f,$$

$$u''' = f''_{tt} + 2f''_{tu} f + f''_{uu} f^2 + f'_u u'',$$

получим

$$\xi(0) = 0, \xi'(0) = (1 - \gamma_1 - \gamma_2)\, f(t,\ u),$$
$$\xi''(0) = (1 - 2\gamma_2\alpha_2)\, f'_t(t,\ u) + (1 - 2\gamma_2\beta_{21})\, f'_u(t,\ u)\, f(t,\ u),$$
$$\xi'''(0) = (1 - 3\gamma_2\alpha_2^2) f''_{tt}(t,u) + (2 - 6\gamma_2\alpha_2\beta_{21}) f''_{tu}(t,u) f(t,u) +$$
$$+(1 - 3\gamma_2\beta_{21}^2)\, f''_{uu}(t,\ u) f^2(t,\ u) + f'_u(t,\ u)\, u''(t).$$

Второе из полученных соотношений выполняется при $\gamma_1 + \gamma_2 = 1$, третье — при $1 - 2\gamma_2\alpha_2 = 0, 1 - 2\gamma_2\beta_{21} = 0$.

Таким образом, имеется три алгебраических уравнения и четыре параметра. Эти уравнения определяют однопараметрическое семейство схем. Задавая один из параметров, можно получать различные методы Рунге-Кутты с аппроксимацией второго порядка. При формально одинаковом порядке аппроксимации они будут обладать различными свойствами (устойчивостью, реальной погрешностью).

Так, при $\gamma_1 = 1/2$, имеем $\gamma_2 = 1/2, \alpha_2 = 1, \beta_{21} = 1$; метод будет выглядеть следующим образом:

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

$$u_{n+1} = u_n + \frac{\tau}{2}\left[f(t_n,\ u_n) + f(t_{n+1},\ \tilde{u}_{n+1})\right].$$

Положив $\gamma_1 = 0$, имеем $\gamma_2 = 1, \alpha_2 = 1/2, \beta_{21} = 1/2$; соответствующий метод будет:

$$u_{n+1/2} = u_n + \frac{\tau}{2} f(t_n,\ u_n),$$

$$u_{n+1} = u_n + f\left(t_n + \frac{\tau}{2},\ u_{n+1/2}\right).$$

3. При $p = 2, r = 3$ получаем систему уравнений для коэффициентов:

$$\alpha_2 = \beta_{21}, \alpha_3 = \beta_{31} + \beta_{32},$$

$$\alpha_3(\alpha_3 - \alpha_2) - \beta_{32}\alpha_2(2 - 3\alpha_2) = 0,$$

$$\gamma_3\beta_{32}\alpha_2 = 1/6, \gamma_2\alpha_2 + \gamma_3\alpha_3 = 1/2, \gamma_1 + \gamma_2 + \gamma_3 = 1,$$

имеющую бесконечное множество решений. Расчетные формулы одного из возможных методов имеют вид

$$k_1 = \tau f(t_n,\ u_n), k_2 = \tau f\left(t_n + \frac{\tau}{2},\ u_n + \frac{k_1}{2}\right),$$

$$k_3 = \tau f(t_n + \tau,\ u_n - k_1 + 2k_2),$$

$$u_{n+1} = u_n + \frac{k_1 + 4k_2 + k_3}{6}.$$

В случае $p = 4, r = 4$ имеем двухпараметрическое семейство методов Рунге-Кутты, из которого наиболее известен следующий «классический» метод:

$$k_1 = \tau f(t_n,\ u_n), k_2 = \tau f\left(t_n + \frac{\tau}{2},\ u_n + \frac{k_1}{2}\right),$$

$$k_3 = \tau f\left(t_n + \frac{\tau}{2},\ u_n + \frac{k_2}{2}\right),$$

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

$$u_{n+1} = u_n + \frac{k_1 + 2k_2 + 2k_3 + k_4}{6}.$$

В представлении Бутчера хорошо известные методы численного решения ОДУ выглядят следующим образом. Метод Эйлера (первый порядок аппроксимации) табл. 8.2, метод Эйлера с пересчетом (второй порядок аппроксимации) — табл. 8.3. Метод Хойна третьего порядка аппроксимации — табл. 8.4. Метод Рунге-Кутты третьего порядка аппроксимации — табл. 8.5.

Таблица 8.2

0	0
	1

Таблица 8.3

0		
1/2	1/2	
	0	1

Таблица 8.4

0			
1/3	1/3		
2/3	0	2/3	
	1/4	0	3/4

Таблица 8.5

0			
1/2	1/2		
1	0	1	
	1/6	2/3	1/6

Значения коэффициентов классического метода Рунге-Кутты (четвертого порядка аппроксимации) указаны в табл. 8.6. Правило трех восьмых (четвертый порядок аппроксимации) представлено в табл. 8.7. Этот метод имеет, по-видимому, наименьшую погрешность среди явных схем Рунге-Кутты четвертого порядка аппроксимации. Метод Бутчера (шестой порядок аппроксимации) указан в табл. 8.8. Очевидна связь между методами, приведенными в таблицах (8.5) и (8.6) с формулой Симпсона численного интегрирования.

По-видимому, наивысший порядок аппроксимаций (десятый) был получен в работах Куртиса [4], позже — Хайрера [5]. Соответствующие коэффициенты не приведены ввиду их громоздкости.

В численной практике используются также так называемые вложен-

Таблица 8.6

0				
1/2	1/2			
1/2	0	1/2		
1	0	0	1	
	1/6	2/6	2/6	1/6

Таблица 8.7

0				
1/3	1/3			
2/3	-1/3	1		
1	1	-1	1	
	1/8	3/8	3/8	1/8

Таблица 8.8

0							
1/2	1/2						
2/3	2/9	4/9					
1/3	7/36	2/9	-1/12				
5/6	-35/144	-55/36	35/48	15/8			
1/6	-1/360	-11/36	-1/8	1/2	1/10		
1	-41/260	22/13	43/156	-118/39	32/195	80/39	
	13/200	0	11/40	11/40	4/25	4/25	13/200

ные формулы Рунге-Кутты [3, 6, 7], в которых наряду с соотношением

$$u_{n+1} = u_n + \tau(\gamma_1 k_1 + \ldots + \gamma_2 k_2) \tag{8.5}$$

используется соотношение того же вида, но с другим набором весовых множителей

$$\bar{u}_{n+1} = u_n + \tau(\bar{\gamma}_1 k_1 + \ldots + \bar{\gamma}_2 k_2), \tag{8.6}$$

причем порядки аппроксимации в них различаются: q — порядок аппроксимации (8.6), а p — (8.5). В этом случае порядок метода указывается, как $p(q)$, а таблица Бутчера имеет вид табл. 8.9. Последнюю строчку в этой таблице иногда называют оценщиком погрешности.

Приближения точного решения с разными остаточными членами (8.5) и (8.6) позволяют оценить погрешность численного метода, полученную в конкретном расчете, и служат для повышения точности еще на один порядок в каждой точке или (что бывает чаще) для автоматического выбора длины следующего шага интегрирования (см. 1).

Таблица 8.9

0					
α_2	β_{21}				
α_3	β_{31}	β_{32}			
...	...	...	...		
α_r	β_{r1}	β_{r2}	...	β_{rr-1}	
u_1	γ_1	γ_2	...	γ_{r-1}	γ_r
u_2	$\bar{\gamma}_1$	$\bar{\gamma}_2$	...	$\bar{\gamma}_{r-1}$	$\bar{\gamma}_r$

Приведем несколько наиболее известных вложенных методов Рунге-Кутты. Метод Фельберга 2(3) представлен в табл. 8.10, метод Ческино 2(4) — в табл. 8.11.

Таблица 8.10

0			
1	1		
1/2	1/4	1/4	
	1/2	1/2	0
	1/6	1/6	4/6

Таблица 8.11

0				
1/4	1/4			
1/2	0	1/2		
1	1	-2	2	
	1	-2	2	0
	1/6	0	4/6	1/6

Метод Кутты-Мерсона 4(5) приведен в табл. 8.12. Этот метод — наиболее простой среди вложенных методов Рунге-Кутты порядка 4(5), поскольку требует небольшого числа обращений к функциям вычисления правых частей системы уравнений (8.1) и имеет простые коэффициенты.

Таблица 8.12

0					
1/3	1/3				
1/3	1/6	1/6			
1/2	1/8	0	3/8		
1	1/2	0	-3/2	2	
	1/2	0	-3/2	2	0
	1/6	0	0	2/3	1/6

Для 5 этапного метода Кутты-Мерсона 4(5) приведем подробные расчетные формулы:

$$k_1 = \tau f(t_n,\ u_n), k_2 = \tau f\left(t_n + \frac{\tau}{3},\ u_n + \frac{k_1}{3}\right),$$

$$k_3 = \tau f\left(t_n + \frac{\tau}{3},\ u_n + \frac{k_1}{6} + \frac{k_2}{6}\right),$$

$$k_4 = \tau f\left(t_n + \frac{\tau}{2},\ u_n + \frac{k_1}{8} + \frac{3k_3}{8}\right),$$

$$k_5 = \tau f\left(t_n + \tau,\ u_n + \frac{k_1}{8} - \frac{3k_3}{8} + 2k_4\right),$$

$$u_1(t+\tau) = u(t) + \frac{k_1}{2} - \frac{3k_3}{2} + 2k_4,$$

$$u_2(t+\tau) = u(t) + \frac{k_1}{6} + \frac{2k_4}{3} + \frac{k_5}{6}.$$

Рассмотрим вкратце другие вложенные методы и соответствующие им таблицы. Метод Фельберга 4(5) представлен в табл. 8.13, метод Дормана-Принса 5(4) — в табл. 8.14 (см. также [6]).

Таблица 8.13

0						
1/4	1/4					
3/8	3/32	9/32				
$\frac{12}{13}$	$\frac{1\,932}{2\,197}$	$\frac{7\,296}{2\,197}$				
1	$\frac{439}{216}$	-8	$\frac{3\,680}{513}$	$-\frac{845}{4\,104}$		
$\frac{1}{2}$	$-\frac{8}{27}$	2	$-\frac{3\,544}{2\,565}$	$\frac{1\,859}{4\,104}$	$-\frac{14}{40}$	
	$\frac{25}{216}$	0	$\frac{1\,408}{2\,565}$	$\frac{2\,197}{4\,104}$	$-\frac{1}{5}$	0
	$\frac{16}{135}$	0	$\frac{6\,656}{12\,825}$	$\frac{28\,561}{56\,430}$	$-\frac{9}{50}$	$\frac{2}{55}$

Этот метод замечателен тем, что он не только минимизирует остаточный член в оценщике погрешностей, но и требует меньшей памяти для хранения таблицы коэффициентов метода. Действительно, последняя строка β совпадает с одной из строк γ.

С методом Фельберга более высокого порядка 7(8) можно ознакомиться в [[8]], значения его коэффициентов приведены в табл. 8.15.

Среди современных методов решения нежестких систем ОДУ наилучшие результаты дает метод Дормана-Принса 8(7). Он обладает наименьшей погрешностью среди всех схем порядка 8, коэффициенты указаны в табл. 8.16, 8.17. С вложенными формулами Рунге-Кутты, разработанными в 90-х годах, можно ознакомиться в монографии [10].

Таблица 8.14

0							
1/5	1/5						
3/10	3/40	9/40					
4/5	44/45	-56/15	32/9				
$\frac{8}{9}$	$\frac{19\,372}{6\,561}$	$-\frac{25\,360}{2\,187}$	$\frac{64\,448}{6\,561}$	$-\frac{212}{729}$			
1	$\frac{9\,017}{3\,168}$	$-\frac{355}{33}$	$\frac{46\,732}{5\,247}$	$\frac{49}{176}$	$-\frac{5\,163}{18\,656}$		
1	$\frac{35}{384}$	0	$\frac{500}{1\,113}$	$\frac{125}{192}$	$-\frac{2\,187}{6\,784}$	$\frac{11}{84}$	
	$\frac{35}{384}$	0	$\frac{500}{1\,113}$	$\frac{125}{192}$	$-\frac{2\,187}{6\,784}$	$\frac{11}{84}$	0
	$\frac{5\,179}{57\,600}$	0	$\frac{7\,571}{16\,695}$	$\frac{393}{640}$	$-\frac{92\,097}{339\,200}$	$\frac{187}{2\,100}$	$\frac{1}{40}$

Методы Рунге-Кутты — это одношаговые методы, так как позволяют найти значение искомой функции в точке t_{n+1} используя только значение функции в точке t_n. При этом не учитывается информация о функции, накопленная на предыдущих этапах расчета. Такая информация может быть полезна при решении жестких задач или при автоматическом выборе следующего шага интегрирования.

Один из способов учета предыстории заключается в использовании общих многошаговых методов. Частный случай — методы Адамса — рассмотрен в следующем параграфе. Другой способ учета описан в 1.

Таблица 8.15

0													
2/27	2/27												
1/9	1/36	1/12											
1/6	1/24	0	1/8										
5/12	5/12	0	-25/16	25/16									
1/2	1/20	0	0	1/4	1/5								
5/6	-25/108	0	0	125/108	-65/27	125/54							
1/6	31/300	0	0	0	61/225	-2/9	13/900						
2/3	2	0	0	-53/6	704/45	-107/9	67/90	3					
1/3	-91/108	0	0	23/108	-976/135	311/54	-19/60	17/6	-1/12				
1	$\frac{2383}{4100}$	0	0	$-\frac{341}{164}$	$\frac{4496}{1025}$	$-\frac{301}{82}$	$\frac{2133}{4100}$	$\frac{45}{82}$	$\frac{45}{164}$	$\frac{18}{41}$			
0	3/205	0	0	0	0	-6/41	-3/205	-3/41	3/41	6/41	0		
1	$-\frac{1777}{4100}$	0	0	$-\frac{341}{164}$	$\frac{4496}{1025}$	$-\frac{289}{82}$	$\frac{2193}{4100}$	$\frac{51}{82}$	$\frac{33}{164}$	$\frac{12}{41}$	0	1	
	41/840	0	0	0	0	34/105	9/35	9/35	9/280	9/280	41/840	0	0
	0	0	0	0	0	34/105	9/35	9/35	9/280	9/280	0	41/840	41/840

Общие условия аппроксимации методов Рунге-Кутты и барьеры Бутчера. Вернемся к общей записи метода Рунге-Кутты (8.4) и соответствующей таблице Бутчера. Рассматривается задача $\frac{d\,\mathbf{u}(t)}{d\,t} = \mathbf{f}(t,\ \mathbf{u})$,

$$\mathbf{u}(0) = \mathbf{u}_0,$$

Таблица 8.16

β_{ij}		i					
		2	3	4	5	6	7
j	1	1/18	1/48	1/32	5/16	3/80	$\frac{29\,443\,841}{614\,563\,906}$
	2		1/16	0	0	0	0
	3			3/32	-75/64	0	0
	4				75/64	3/16	$\frac{77\,736\,538}{692\,538\,347}$
	5					3/20	$-\frac{28\,693\,883}{1\,112\,000\,000}$
	6						$\frac{23\,124\,283\,393}{1\,800\,000\,000}$

β_{ij}		i					
		8	9	10	11	12	13
j	1	$\frac{16\,016\,141}{946\,692\,911}$	$\frac{39\,632\,708}{573\,591\,083}$	$\frac{246\,121\,993}{1\,340\,847\,787}$	$-\frac{1\,028\,468\,189}{846\,180\,014}$	$\frac{185\,892\,177}{718\,116\,043}$	$\frac{403\,863\,854}{491\,063\,109}$
	2,3	0	0	0	0	0	0
	4	$\frac{61\,564\,180}{158\,732\,637}$	$-\frac{433\,636\,366}{683\,701\,615}$	$-\frac{37\,695\,042\,795}{15\,268\,766\,246}$	$\frac{8\,478\,235\,783}{506\,512\,852}$	$-\frac{3\,185\,094\,517}{667\,107\,341}$	$-\frac{5\,068\,492\,393}{434\,740\,067}$
	5	$\frac{22\,789\,713}{633\,445\,777}$	$-\frac{421\,739\,975}{2\,616\,292\,301}$	$\frac{309\,121\,744}{1\,061\,227\,803}$	$\frac{1\,311\,729\,495}{1\,432\,422\,823}$	$-\frac{477\,755\,414}{1\,098\,053\,517}$	$-\frac{411\,421\,997}{543\,043\,805}$
	6	$\frac{545\,815\,736}{2\,771\,057\,229}$	$\frac{100\,302\,831}{723\,423\,059}$	$-\frac{12\,992\,083}{490\,766\,935}$	$-\frac{10\,304\,129\,995}{1\,701\,304\,382}$	$-\frac{703\,635\,378}{230\,739\,211}$	$\frac{652\,783\,627}{914\,296\,604}$
	7	$-\frac{180\,193\,667}{1\,043\,307\,555}$	$\frac{790\,204\,164}{839\,813\,087}$	$\frac{6\,005\,943\,433}{2\,108\,947\,869}$	$-\frac{48\,777\,925\,059}{3\,047\,939\,560}$	$\frac{5\,731\,566\,787}{1\,027\,545\,527}$	$\frac{11\,173\,962\,825}{925\,320\,556}$
	8		$\frac{800\,635\,310}{3\,783\,071\,287}$	$\frac{393\,006\,217}{1\,396\,673\,457}$	$\frac{15\,336\,726\,248}{1\,032\,824\,649}$	$\frac{5\,232\,866\,602}{850\,066\,563}$	$-\frac{13\,158\,990\,841}{6\,184\,727\,034}$
	9			$\frac{123\,872\,331}{1\,001\,029\,789}$	$-\frac{45\,442\,868\,181}{3\,398\,467\,696}$	$-\frac{4\,093\,664\,535}{808\,688\,257}$	$\frac{3\,936\,647\,629}{1\,978\,049\,680}$
	10				$\frac{3\,065\,993\,473}{597\,172\,653}$	$\frac{3\,962\,137\,247}{1\,805\,957\,418}$	$-\frac{160\,528\,059}{685\,178\,525}$
	11					$\frac{65\,686\,385}{487\,910\,083}$	$\frac{248\,638\,103}{1\,413\,531\,060}$

Таблица 8.17

i	1	2	3	4	5	6	7	8
α_i	1	1/18	1/12	1/8	5/16	3/8	59/400	93/200
γ_i	$\frac{14\,005\,451}{335\,480\,064}$	0	0	0	0	$-\frac{59\,238\,493}{1\,068\,277\,825}$	$\frac{181\,606\,676}{758\,867\,731}$	$\frac{561\,292\,985}{797\,845\,732}$
$\bar{\gamma}_i$	$\frac{13\,451\,932}{455\,176\,623}$	0	0	0	0	$-\frac{808\,719\,846}{976\,000\,145}$	$\frac{1\,757\,004\,468}{5\,645\,159\,321}$	$\frac{656\,045\,339}{265\,891\,186}$

i	9	10	11	12	13
α_i	$\frac{5\,490\,023\,248}{9\,719\,169\,821}$	$\frac{1\,201\,146\,811}{1\,299\,019\,798}$	3/20	1	1
γ_i	$-\frac{1\,041\,891\,430}{1\,371\,343\,529}$	$\frac{760\,417\,239}{1\,151\,165\,299}$	$\frac{118\,820\,643}{751\,138\,087}$	$-\frac{528\,747\,749}{2\,220\,607\,170}$	1/4
$\bar{\gamma}_i$	$-\frac{3\,867\,574\,721}{1\,518\,517\,206}$	$\frac{465\,885\,868}{322\,736\,535}$	$\frac{5\,301\,238}{667\,516\,719}$	2/45	0

а численный метод имеет вид

$$\mathbf{k}_1 = \mathbf{f}(t_n, \mathbf{u}_n),$$
$$\mathbf{k}_2 = \mathbf{f}(t_n + \alpha_2\tau, \mathbf{u}_n + \tau\beta_{21}\mathbf{k}_1),$$
$$\mathbf{k}_3 = \mathbf{f}(t_n + \alpha_3\tau, \mathbf{u}_n + \tau(\beta_{31}\mathbf{k}_1 + \beta_{32}\mathbf{k}_2)),$$
$$\dots$$
$$\mathbf{k}_r = \mathbf{f}(t_n + \alpha_r\tau, \mathbf{u}_n + \tau(\beta_{r1}\mathbf{k}_1 + \dots + \beta_{r,r-1}\mathbf{k}_2)),$$
$$\mathbf{u}_{n+1} = \mathbf{u}_n + \tau(\gamma_1\mathbf{k}_1 + \dots + \gamma_r\mathbf{k}_r)$$

где k_i — вспомогательные векторы.

Наряду с явными, рассмотрим также неявные методы Рунге-Кутты, определенные как

$$\mathbf{k}_1 = \mathbf{f}(t_n + \alpha_1\tau, \mathbf{u}_n + \tau\sum_{j=1}^{r}\beta_{j1}\mathbf{u}_j),$$

$$\mathbf{k}_2 = \mathbf{f}(t_n + \alpha_2\tau, \mathbf{u}_n + \tau\sum_{j=1}^{r}\beta_{j2}\mathbf{u}_j),$$

$$\dots$$

$$\mathbf{k}_r = \mathbf{f}(t_n + \alpha_r\tau, \mathbf{u}_n + \tau\sum_{j=1}^{r}\beta_{jr}\mathbf{u}_j),$$

$$\mathbf{u}_{n+1} = \mathbf{u}_n + \tau(\gamma_1\mathbf{k}_1 + \dots + \gamma_r\mathbf{k}_r);$$

таблица Бутчера для неявных методов примет вид

α_1	β_{11}	β_{12}	$\dots$	$\beta_{1,r-1}$	β_{1r}
α_2	β_{21}	β_{21}	$\dots$	$\beta_{2,r-1}$	β_{2r}
α_3	β_{31}	β_{32}	$\dots$	$\beta_{3,r-1}$	β_{3r}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
α_r	β_{r1}	β_{r2}	$\dots$	β_{rr-1}	β_{rr}
	γ_1	γ_2	$\dots$	γ_{r-1}	γ_r

Для вывода условий аппроксимации общего метода Рунге-Кутты необходимо действовать так же, как описано выше. Для этого введем погрешность

$$\xi(\tau) = u(t+\tau) - \left[u(t) + \sum_{j=0}^{r}\gamma_j k_j\right]$$

и представим ее в виде разложения в ряд Маклорена. Приравнивая члены при одинаковых степенях шага τ, получим условия аппроксимации метода. Для того чтобы метод имел порядок 3, необходимо выполнение следующих условий:

$$\sum_{i=1}^{r}\gamma_i = 1,$$

$$2\sum_{i=1}^{r}\sum_{k=1}^{r}\gamma_i\beta_{ik} = 1,$$

$$3\sum_{i=1}^{r}\sum_{k=1}^{r}\sum_{l=1}^{r}\gamma_i\beta_{ik}\beta_{il} = 1,$$

$$6\sum_{i=1}^{r}\sum_{k=1}^{r}\sum_{l=1}^{r}\gamma_i\beta_{ik}\beta_{kl} = 1,$$

причем эти выражения упрощаются, если воспользоваться необязательными условиями Кутты. При повышении порядка аппроксимации метода возникают дополнительные условия на коэффициенты, система значительно усложняется.

Для того чтобы построить аппроксимирующую схему (метод Рунге-Кутты) необходимо найти набор коэффициентов метода. Как было показано выше, в случае двух стадий метода такой набор коэффициентов — не единственный, существует континуум методов второго порядка аппроксимации. Континуум решений система уравнений порядка для явных методов Рунге-Кутты имеет и в случае явных методов с тремя или четырьмя стадиями. Но для пятистадийного метода система уравнений порядка является несовместной. Это утверждение было доказано Бутчером и носит название «первый барьер Бутчера». Его обычно формулируют в виде теоремы [3].

Теорема (первый барьер Бутчера). *Среди явных методов Рунге-Кутты с числом стадий пять не существует методов пятого порядка аппроксимации.*

Для повышения порядка до пятого приходится использовать шестистадийные методы. При увеличении числа стадий возникает второй барьер Бутчера — порядок аппроксимации метода, начиная с семи стадий, оказывается уже на 2 ниже, чем число стадий. При увеличении порядка аппроксимации метода приходится значительно увеличивать число стадий — барьеры Бутчера встречаются чаще.

Наличие такого барьера — одно из следствий быстрого роста констант Лебега при интерполяции на равномерной сетке. Дело в том, что явные методы Рунге-Кутты тесно связаны с квадратурными формулами интерполяционного типа. Достаточно очевидно, что классический метод Рунге-Кутты порядка 4 основан на применении формулы Симпсона, а правило 3/8 — на одноименной квадратурной формуле. Как было показано в лекции 7, с повышением порядка аппроксимации квадратурные формулы перестают быть правильными, а это следствие роста константы Лебега. Тогда и появляются барьеры Бутчера при построении методов решения систем ОДУ.

От этого недостатка свободны некоторые неявные методы, основанные на квадратурных формулах Гаусса.

8.3. Методы Адамса

Для решения ОДУ или систем ОДУ существуют *одностадийные методы Адамса* (линейные многошаговые методы), суть которых заключается в следующем.

Пусть известно приближенное решение в некоторых узлах расчетной сетки: $t_n, t_{n-1}, \ldots, t_{n-m}$. В окрестности этих узлов заменим $f(t,\ x(t))$ интерполяционным полиномом, записанным в форме Ньютона ([1, 2], а также лекция 6):

$$f(t) = f(t_n) + f(t_n,\ t_{n-1})\,(t - t_n) +$$

$$+ f(t_n,\ t_{n-1},\ t_{n-2})\,(t - t_n)\,(t - t_{n-1}) +$$

$$+ f(t_n,\ t_{n-1},\ t_{n-2},\ t_{n-3})\,(t - t_n)\,(t - t_{n-1})\,(t - t_{n-2}) + \ldots$$

Для того чтобы вычислить решение в точке $n+1$, запишем его в интегральном виде

$$u_{n+1} = u_n + \int\limits_{t_n}^{t_{n+1}} f(t,\ u(t))\,dt = \int\limits_{t_n}^{t_{n+1}} f(t)\,dt$$

и подставим в него интерполяционный полином с переменным шагом

$$u_{n+1} = u_n + \tau_n f(t_n) + \frac{\tau_n^2}{2} f(u_n,\ u_{n-1}) +$$

$$+ \frac{\tau_n^2}{6}\,(2\tau_n + 3\tau_{n-1})\,f(u_n,\ u_{n-1},\ u_{n-2}) + \frac{\tau_n^2}{12}\,(2\tau_n^2 + 8\tau_n\tau_{n-1} +$$

$$+ 4\tau_n\tau_{n-2} + 6\tau_{n-1}^2 + 6\tau_n\tau_{n-2}) f(u_n,\ u_{n-1},\ u_{n-2},\ u_{n-3}).$$

Здесь $\tau_n = t_{n+1} - t_n$, $f(u_n,\ u_{n-1})$, $f(u_n,\ u_{n-1},\ u_{n-2})$ и т.д. - разделенные разности.

Эта формула четвертого порядка точности. Если опустить последнее слагаемое, то получим формулу третьего порядка, если опустить еще и предпоследнее, то — второго, и т. д. Если же положить $\tau_n = \text{const}$, то формула значительно упростится:

$$u_{n+1} = u_n + \tau f_n + \frac{\tau^2}{2}\,\Delta_1 f_n + \frac{5}{12}\,\tau^3 \Delta_2 f_n + \frac{3}{8}\,\tau^4 \Delta_3 f_n,$$

где $\Delta_k f_n$ — k-я конечная разность.

Для того чтобы начать вычисления по данному варианту метода Адамса, необходимо знать решение в четырех точках. Это можно сделать, например, с помощью методов Рунге-Кутты. Кроме того, коэффициент при погрешности, например, для метода четвертого порядка точности Рунге-Кутты существенно меньше соответствующего коэффициента для метода Адамса.

Первые четыре метода Адамса, от первого до четвертого порядка точности с постоянным шагом интегрирования, представляются в виде

$$u_{n+1} = u_n + \tau f_n,$$

$$u_{n+1} = u_n + \tau \left(\frac{3}{2} f_n - \frac{1}{2} f_{n-1} \right),$$

$$u_{n+1} = u_n + \tau \left(\frac{23}{12} f_n - \frac{16}{12} f_{n-1} + \frac{5}{12} f_{n-2} \right) \tag{8.7}$$

$$u_{n+1} = u_n + \tau \left(\frac{55}{24} f_n - \frac{59}{24} f_{n-1} + \frac{37}{24} f_{n-2} - \frac{9}{24} f_{n-3} \right).$$

В общем виде методы Адамса могут быть записаны следующим образом:

$$u_{n+1} = u_n + \tau \sum_{j=0}^{k-1} \eta_j \Delta^j f_j.$$

Значительно большее значение в вычислительной практике имеют неявные методы Адамса, которые можно записать как

$$u_{n+1} = u_n - \tau \sum_{j=0}^{k} \tilde{\gamma}_j \Delta^j f_{n+1}.$$

Первые четыре неявных метода имеют вид

$$k = 0 : u_{n+1} = u_n + \tau f_{n+1},$$

$$k = 1 : u_{n+1} = u_n + \tau \left(\frac{1}{2} f_{n+1} + \frac{1}{2} f_n \right),$$

$$k = 2 : u_{n+1} = u_n + \tau \left(\frac{5}{12} f_{n+1} + \frac{8}{12} f_n - \frac{1}{12} f_{n-1} \right),$$

$$k = 3 : u_{n+1} = u_n + \tau \left(\frac{9}{24} f_{n+1} + \frac{19}{24} f_n - \frac{5}{24} f_{n-1} + \frac{1}{24} f_{n-2} \right).$$

Первая и вторая формулы — неявный метод Эйлера и неявный метод трапеций соответственно. Порядок аппроксимации приведенных методов — с первого по четвертый соответственно.

Для $k = 8$ коэффициент $\tilde{\gamma}_j$ можно представить в виде таблицы (табл. 8.19) и ввести следующие обозначения: $\Delta^{j+1} f_n = \Delta^j f_n - \Delta^j f_{n-1}$ — «разности назад», η_j — коэффициенты, которые можно представить в виде таблицы (для $k = 8$ табл. 8.20). Для $k = 1$ получается уже знакомый явный метод Эйлера.

Таблица 8.18

J	0	1	2	3	4	5	6	7	8
$\tilde{\gamma}_j$	1	$-\frac{1}{2}$	$-\frac{1}{12}$	$-\frac{1}{24}$	$-\frac{19}{720}$	$-\frac{3}{160}$	$-\frac{863}{60\,480}$	$-\frac{275}{24\,192}$	$-\frac{33\,953}{3\,628\,800}$

Таблица 8.19

j	0	1	2	3	4	5	6	7	8
η_j	1	$\frac{1}{2}$	$\frac{5}{12}$	$\frac{3}{8}$	$\frac{251}{720}$	$\frac{95}{288}$	$\frac{19\,087}{60\,480}$	$\frac{5\,257}{17\,280}$	$\frac{1\,070\,017}{3\,628\,800}$

8.4. Оценка погрешности

8.4.1. Автоматический выбор шага интегрирования

Рунге предложил простое правило оценки точности метода, основанное на проведении вычислений с разными шагами интегрирования. Основная идея правила Рунге заключается в следующем.

Погрешность численного метода порядка $p-1$ в точке $t_i = i\tau$ представляется в виде

$$u_n(t_i) - \tilde{V}_n(t_i) = C\tau^p + O(\tau^{p+1}).$$

Если выполнить аналогичные вычисления с шагом вдвое меньшим, $\tau/2$, то получим

$$u_{2n} - \tilde{V}_n(t_i) = 2C\frac{\tau^p}{2^p} + O(\tau^{p+1}).$$

После вычитания из первого соотношения второго

$$u_n - u_{2n} = C(\tau^p - \frac{\tau^p}{2^{p-1}}) + O(\tau^{p+1}),$$

откуда

$$C = \frac{2^{p-1}}{\tau^p}\frac{u_n - u_{2n}}{2^{p-1}-1} + O(\tau).$$

Подставив выражение для C во второе соотношение, получим

$$u_{2n} - \tilde{V}_n(t_i) = \frac{u_n - u_{2n}}{2^{p-1}-1} + O(\tau^{p+1}),$$

т. е. погрешность метода, с точностью $O(\tau^{p+1})$ оценивается по простой формуле

$$\varepsilon = \frac{u_n - u_{2n}}{2^{p-1}-1}.$$

Это правило используется не только для оценки погрешности вычисления, но и для автоматического выбора шага интегрирования. Для этого на каждом шаге вычисления производятся трижды: с шагом τ и с двумя шагами $\tau/2$. Полученные значения u_n и u_{2n} используются для вычисления реальной погрешности ε (точнее, оценки ее главного члена). Если величина ε превышает некую заданную (или заранее выбранную) константу ε_0, то шаг интегрирования уменьшается; если, напротив, ε существенно меньше ε_0, то τ увеличивается.

Разумеется, алгоритмы выбора шага интегрирования могут основываться и на иных принципах. Например, можно выбрать τ адаптирующимся к решению: уменьшать при увеличении абсолютной величины производной и увеличивать при ее уменьшении, т. е. вычислять τ, как функцию от $||u'_t||$.

Управление длиной шага в методах Рунге-Кутты осуществляется на основе сравнения с некоторой задаваемой величиной T, характеризующей требования к погрешности на каждом шаге численного интегрирования системы.

Пусть используется метод с порядком аппроксимации p. Тогда главный член погрешности метода ε, определяемый по правилу Рунге, или, в случае использования вложенных методов Рунге-Кутты, представляющий собой модуль разности между приближениями к решению, вычисленными по формулам (8.5) и (8.6), имеет вид

$$\varepsilon = C_2 \tau^{p+1} \leqslant T.$$

Положим теперь $T = C_2 \tau_{\text{new}}^{p+1}$. Тогда для величины максимального значения нового шага интегрирования τ_{new} получаем

$$\tau_{\text{new}} = \beta\tau \left(\frac{T}{\varepsilon}\right)^{1/(p+1)},$$

где β — так называемый гарантированный множитель. Он служит для того, чтобы в случае резкого уменьшения шага (например, при выходе на жесткий участок при решении умеренно *жестких систем*) численный метод оставался устойчивым.

Кроме того, гарантированный множитель помогает избежать слишком быстрого увеличения величины шага интегрирования в случае, когда реально полученная погрешность мала. Обычно величина гарантийного множителя принимается за 0,5; 0,8; 0,9 или $(0,25 \div 0,38)^{1/(p+1)}$.

Если при выполнении очередного шага погрешность ε не превосходит величины T, то шаг считается *принятым*, а дальнейший расчет продолжается с шагом τ_{new}; в противном случае шаг считается *отклоненным*, и проводится пересчет с новым значением шага для перехода от t_n к t_{n+1}.

На практике применяют модернизации алгоритма выбора шага. Так, если реальная ошибка ε мала, то предлагаемый алгоритм позволяет выбрать очень большой шаг по τ. В таком случае применяют алгоритм

$$\tau_{\text{new}} = \tau \cdot \min \left\{ \beta_{\max}, \ \max \left(\beta_{\min}, \ \beta \left(\frac{T}{\varepsilon} \right)^{1/(p+1)} \right) \right\}.$$

Здесь $\beta_{\max}, \beta_{\min}$ — максимальное и минимальное разрешенное изменение шага интегрирования.

Другая возможность регулировки шага численного интегрирования для методов Рунге-Кутты заключается в объединении достоинств методов Рунге-Кутты и Адамса. Первые из них допускают легко регулировать шаг интегрирования, а вторые помнят часть предыстории изменения функции.

Более тонкий алгоритм управления величиной шага получается с учетом величины ошибки на предыдущем шаге

$$\tau_{\text{new}} = \tau \left(\frac{T}{\varepsilon_n} \right)^{\alpha} \left(\frac{T}{\varepsilon_{n-1}} \right)^{-\beta},$$

т. е. фактически гарантийный множитель зависит от ошибки на предыдущем шаге. Обычно при таком выборе управления длиной шага полагается $\alpha = 1/(p+1), \beta \approx 0,08$.

Такой выбор регулировки шага повышает устойчивость численных методов Рунге-Кутты не очень высокого порядка. Для метода Дормана-Принса порядка 8(7) лучшие результаты дает $\beta \approx 0,04$.

В [10] показано, что такой метод выбора шага является решением задачи оптимального управления с учетом пропорциональной обратной связи и интегральной обратной связи. Там же показано, что оптимальный выбор показателей степени есть $\alpha \approx 0,7/(p+1), \beta \approx 0,4/(p+1)$.

8.5. Устойчивость методов Рунге-Кутты

Для исследования устойчивости методов Рунге-Кутты для численного решения задачи

$$\frac{d\mathbf{u}}{dt} = \mathbf{f}(t, \mathbf{u}), \quad \mathbf{u}(0) = \mathbf{u}_0$$

представим ее дискретный аналог в виде

$$\frac{\mathbf{u}_{n+1} - \mathbf{u}_n}{\tau} = \mathbf{F}(t_n, \mathbf{u}_n), \tag{8.8}$$

здесь $\mathbf{F}(t, \mathbf{x})$ — функция приращения метода Рунге-Кутты, которая, конечно, связана с функцией правой части системы ОДУ.

Теорема 2 (Устойчивость методов Рунге-Кутты [3, 9]). *Пусть* $\mathbf{f}(t, \mathbf{x})$ *Липшиц-непрерывна по второму аргументу, т. е.*

$$||\mathbf{f}(t, \mathbf{x}) - \mathbf{f}(t, \mathbf{y})|| \leqslant C||\mathbf{x} - \mathbf{y}||,$$

причем это условие выполняется для каждого t; $C \neq C(\tau)$ *и* $C\tau \ll 1$.

Тогда разностное уравнение (8.8) *устойчиво и имеет место оценка*

$$||\mathbf{u}_n - \mathbf{v}_n|| \leqslant e^{C_2 t}||\mathbf{u}_0 - \mathbf{v}_0|| + \frac{2\varepsilon e^{C_2 t}}{C_2}. \tag{8.9}$$

Здесь $\mathbf{u}_n, \mathbf{v}_n$ *— решения близких систем разностных уравнений:*

$$\frac{\mathbf{u}_{n+1} - \mathbf{u}_n}{\tau} = \mathbf{F}(t_n, \mathbf{u}_n) + \varepsilon_n^1, \mathbf{u}_0 = \mathbf{u}^0,$$

$$\frac{\mathbf{v}_{n+1} - \mathbf{v}_n}{\tau} = \mathbf{F}(t_n, \mathbf{v}_n) + \varepsilon_n^2, \mathbf{v}_0 = \mathbf{v}^0$$

и ε *— максимальная погрешность при вычислении правой части системы, т. е. для всех* n *(включая нуль)*

$$||\varepsilon_n^1|| \leqslant \varepsilon, ||\varepsilon_n^2|| \leqslant \varepsilon,$$

а постоянная C_2 *незначительно отличается от* C.

Сформулируем вначале следующую лемму.

Лемма 1. *Пусть* C *— постоянная Липшица для функции правых частей системы* (8.1), *тогда функция приращения* $\mathbf{F}(t, \mathbf{u})$ *для метода* (8.8) *удовлетворяет следующему неравенству:*

$$||\mathbf{F}(t, \mathbf{u}_n) - \mathbf{F}(t, \mathbf{v}_n)|| \leqslant C_2||\mathbf{u}_n - \mathbf{v}_n||,$$

где

$$C_2 = C(\sum_i |\alpha_i| + \tau C \sum_{i,j} |\alpha_i \beta_{ij}| + \tau^2 C^2 \sum_{i,j,k} |\alpha_i \beta_{ij} \beta_{jk}| + \ldots).$$

Суммирование в правой части последнего равенства ведется по каждому индексу от 1 до r — числа стадий метода. Число сумм в скобках, конечно, тоже равно r. Отметим также, что если в таблице Бутчера все коэффициенты неотрицательны (такой метод Рунге-Кутты, по аналогии с методом численного интегрирования, назовем правильным), то из условий порядка (аппроксимации) будет следовать, что в скобках стоят первые r членов разложения $e^{C\tau}$ в ряд Тейлора. Отсюда необходимое требование

(в формулировке теоремы) к малости $C\tau$. В случае наличия отрицательных коэффициентов в таблице Бутчера константа Липшица C_2 увеличится незначительно — число стадий метода конечно и невелико (в настоящее время известны методы максимум с 17 стадиями, [6]).

Доказательство данной леммы довольно простое, но громоздкое. Читатель может проделать это самостоятельно.

Доказательство теоремы.

Рассмотрим эти близкие уравнения. Вычитая из первого уравнения второе, получим

$$||\mathbf{u}_{n+1} - \mathbf{v}_{n+1}|| \leqslant ||\mathbf{u}_n - \mathbf{v}_n|| + \tau||\mathbf{F}(t, \mathbf{u}_n) - \mathbf{F}(t, \mathbf{v}_n)|| + 2\tau\varepsilon \qquad (8.10)$$

или, учитывая условие Липшица,

$$||\mathbf{u}_{n+1} - \mathbf{v}_{n+1}|| \leqslant (1 + C_2\tau)||\mathbf{u}_n - \mathbf{v}_n|| + 2\tau\varepsilon.$$

Здесь константа C_2 — постоянная Липшица функции приращения метода Рунге-Кутты, выражение для нее приведено выше в формулировке леммы 1.

Далее, применяя последовательно это неравенство, получим оценки:

$$||\mathbf{u}_1 - \mathbf{v}_1|| \leqslant (1 + C_2\tau)||\mathbf{u}_0 - \mathbf{v}_0|| + 2\tau\varepsilon,$$

$$||\mathbf{u}_2 - \mathbf{v}_2|| \leqslant (1+C_2\tau)||\mathbf{u}_1 - \mathbf{v}_1|| + 2\tau\varepsilon \leqslant (1 + C_2\tau)^2||\mathbf{u}_0 - \mathbf{v}_0|| + 2\tau\varepsilon[1 + (1+C_2\tau)],$$

$$||\mathbf{u}_3 - \mathbf{v}_3|| \leqslant (1 + C_2\tau)||\mathbf{u}_2 - \mathbf{v}_2|| + 2\tau\varepsilon \leqslant$$

$$\leqslant (1 + C_2\tau)^3||\mathbf{u}_0 - \mathbf{v}_0|| + 2\tau\varepsilon[1 + (1 + C_2\tau) + (1 + C_2\tau)^2],$$

$$||\mathbf{u}_n - \mathbf{v}_n|| \leqslant (1 + C_2\tau)^n||\mathbf{u}_0 - \mathbf{v}_0|| + 2\tau\varepsilon[1 + (1 + C_2\tau) + \ldots + (1 + C_2\tau)^{n-1}],$$

откуда, после суммирования прогрессии, имеем

$$||\mathbf{u}_n - \mathbf{v}_n|| \leqslant (1 + C_2\tau)^n||\mathbf{u}_0 - \mathbf{v}_0|| + 2\tau\varepsilon\frac{(1 + C_2\tau)^n - 1}{(1 + C_2\tau) - 1} \leqslant$$

$$\leqslant (1 + C_2\tau)^n[||\mathbf{u}_0 - \mathbf{v}_0|| + \frac{2\varepsilon}{C_2}].$$

Далее полагаем $(1 + C_2\tau)^n = (1 + C_2\tau)^{t/\tau} \approx e^{C_2 t}$ при $C_2\tau \ll 1$.

После подстановки экспоненты в последнюю оценку получим неравенство (8.9). ■

Заметим, что экспоненциальный множитель в неравенстве (8.9) при больших t велик, а оценка, в наиболее общем случае, в предположении Липшиц-непрерывности функции $\mathbf{F}(\mathbf{u})$, неулучшаема. Однако в некоторых важных частных случаях эту оценку можно улучшить, рассматривая более тонкие свойства рассматриваемой функции. Докажем следующее утверждение [9], рассматривая задачу Коши для системы ОДУ.

Утверждение. Пусть матрица

$$\mathbf{A}(\mathbf{u}) = \frac{1}{2}\left(\mathbf{f}_{\mathbf{u}}(\mathbf{u}) + \mathbf{f}_{\mathbf{u}}^{*}(\mathbf{u})\right)$$

строго отрицательна, т. е.

$$(\mathbf{A}(\mathbf{u})\xi, \xi) \leqslant -a(\xi, \xi)$$

для любых ξ, $\mathbf{u}$ и $a > 0$ (траектория, в окрестности которой выполняется это условие, называется *устойчивой*).

Тогда при интегрировании *правильным* методом Рунге-Кутты k-го порядка аппроксимации погрешность приближенного решения есть $O(\tau^k$ при любом $t > 0$ при выполнении условий $a\tau \ll 1$.

Утверждение будет доказано, если в оценке устойчивости метода не будет содержаться множитель e^{Ct}.

Доказательство.

Из полученного в теореме 2 неравенства (8.9) имеем

$$\|\mathbf{u}_{n+1} - \mathbf{v}_{n+1}\| \leqslant \|(\mathbf{u}_n - \mathbf{v}_n) + \tau\left(\mathbf{F}(\mathbf{u}_n) - \mathbf{F}(\mathbf{v}_n)\right)\| + 2\tau\varepsilon,$$

где $\mathbf{F}(\mathbf{u})$ — функция приращения метода Рунге-Кутты для систем ОДУ.

Оценим норму разности в правой части неравенства, используя более точные свойства функции $\mathbf{F}(\mathbf{u})$, чем Липшиц-непрерывность:

$$\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{v}) = \mathbf{F}(\mathbf{u} + s(\mathbf{u} - \mathbf{v}))\Big|_{s=0}^{s=1} = \int_0^1 \frac{d}{ds}\mathbf{F}(\mathbf{u} + s(\mathbf{u} - \mathbf{v}))ds =$$

$$= \int_0^1 \mathbf{F}'(\mathbf{u} + s(\mathbf{u} - \mathbf{v}))(\mathbf{u} - \mathbf{v})ds.$$

Здесь произведена замена переменных $\mathbf{u}(s) = \mathbf{v} + s(\mathbf{u} - \mathbf{v})$.

Поскольку

$$\mathbf{u} - \mathbf{v} = \int_0^1 (\mathbf{u} - \mathbf{v})\, ds,$$

то получаем, что

$$\|(\mathbf{u}-\mathbf{v})+\tau\,(\mathbf{F}(\mathbf{u})-\mathbf{F}(\mathbf{v}))\| =$$

$$\left\| \int_0^1 \left(\mathbf{E}+\tau\mathbf{F}'_{\mathbf{u}}(\mathbf{u}+s(\mathbf{u}-\mathbf{v}))\right)(\mathbf{u}-\mathbf{v})\,ds \right\| \leqslant \int_0^1 \|\mathbf{E}+\tau\mathbf{F}'_{\mathbf{u}}\|\ \|\mathbf{u}-\mathbf{v}\|\ ds. \tag{8.11}$$

Оценим норму $\|\mathbf{E}+\tau\mathbf{F}'_{\mathbf{u}}\|$. Для этого сначала оценим квадрат нормы

$$||\mathbf{E}+\tau\mathbf{F}'_{\mathbf{u}}||^2 = \sup_{||\xi\neq 0||} \frac{((\mathbf{E}+\tau\mathbf{F}'_{\mathbf{u}})\xi,(\mathbf{E}+\tau\mathbf{F}'_{\mathbf{u}})\xi)}{(\xi,\xi)} =$$

$$= \sup_{||\xi\neq 0||} \frac{(\xi,\xi)+2\tau(\mathbf{A}\xi,\xi)+\tau^2(\mathbf{F}'_{\mathbf{u}}\xi,\mathbf{F}'_{\mathbf{u}}\xi)}{(\xi,\xi)} \leqslant 1-2\tau a+O(\tau^2).$$

Здесь использовано определение нормы матрицы (лекция 2)

$$||\mathbf{B}||^2 = \sup_{||\xi\neq 0||} \frac{(\mathbf{B}\xi,\mathbf{B}\xi)}{(\xi,\xi)}$$

где $\mathbf{B}$ — матрица, ξ — вектор из рассматриваемого пространства.

При малых τ, пренебрегая членами порядка $O(\tau^2)$, из последнего неравенства получаем требуемую оценку:

$$||\mathbf{E}+\tau\mathbf{F}'_{\mathbf{u}}|| \leqslant 1-a\tau.$$

Тогда из (8.11) имеем

$$\int_0^1 ||\mathbf{E}+\tau\mathbf{F}'_{\mathbf{u}}||\,||\mathbf{u}-\mathbf{v}||ds \leqslant (1-a\tau)||\mathbf{u}-\mathbf{v}||,$$

откуда

$$||\mathbf{u}_{n+1}-\mathbf{v}_{n+1}|| \leqslant (1-a\tau)||\mathbf{u}_n-\mathbf{v}_n||+2\tau\varepsilon.$$

Рассмотрев, как и ранее, цепочку неравенств

$$||\mathbf{u}_1-\mathbf{v}_1|| \leqslant (1-a\tau)||\mathbf{u}_0-\mathbf{v}_0||+2\tau\varepsilon,$$

$$||\mathbf{u}_2-\mathbf{v}_2|| \leqslant (1-a\tau)||\mathbf{u}_1-\mathbf{v}_1||+2\tau\varepsilon,\ldots$$

получим

$$||\mathbf{u}_n-\mathbf{v}_n|| \leqslant (1-a\tau)^n||\mathbf{u}_0-\mathbf{v}_0||+2\tau\varepsilon\frac{1-(1-a\tau)^n}{1-(1-a\tau)},$$

или

$$||\mathbf{u}_n-\mathbf{v}_n|| \leqslant (1-a\tau)^n||\mathbf{u}_0-\mathbf{v}_0||+\frac{2\varepsilon}{a}. \tag{8.12}$$

Утверждение доказано. ■

Пусть теперь $(\mathbf{A}(\mathbf{u})\xi, \xi) \leqslant 0$ т.е. рассматриваются нейтральные, или «не неустойчивые» траектории исследуемой системы обыкновенных дифференциальных уравнений. В этом случае аналогичным образом показывается, что

$$||\mathbf{E} + \tau \mathbf{F}'_{\mathbf{u}}|| \leqslant 1 + C_2\tau^2.$$

Тогда, проведя аналогичные выкладки, получим

$$\|\mathbf{u}_{n+1} - \mathbf{v}_{n+1}\| \leqslant (1 + C_2\tau^2)\ \|\mathbf{u}_n - \mathbf{v}_n\| + 2\tau\varepsilon,$$

откуда

$$\|\mathbf{u}_n - \mathbf{v}_n\| \leqslant (1 + C_2\tau^2)^n\ \|\mathbf{u}_0 - \mathbf{v}_0\| + 2\tau\varepsilon\ \frac{(1 + C_2\tau^2)^n - 1}{C_2\tau^2}.$$

Для метода k-го порядка аппроксимации $\varepsilon = O(\tau^k), k \geqslant 0$. В этом случае решения близких систем на n-м шаге по времени отличаются на величину

$$||\mathbf{u}_n - \mathbf{v}_n|| \leqslant e^{C_2\tau^2 n}||\mathbf{u}_0 - \mathbf{v}_0|| + O(\tau^{k-1}). \tag{8.13}$$

Отсюда видно, что при $n \sim \tau^{-2}$ (или $t_n = n\tau \sim \tau^{-1}$) численное решение имеет точность $O(\tau^{k-1})$. Другими словами, на конечных интервалах времени $t \sim O(1)$ точность метода $O(\tau^k)$, на больших же временах $t \sim O(\tau^{-1})$ точность понижается до $t \sim O(\tau^{k-1})$.

Такие случаи возникают, когда имеется необходимость проводить численные расчеты при исследовании процессов с большим количеством колебаний, вращений и т.д. Важно отметить то, что полученные оценки погрешности численного решения получены с использованием более сильных, чем условие Липшиц- непрерывности, свойств правых частей рассматриваемых дифференциальных уравнений.

8.6. Задачи

1. **Семейство явных методов Рунге-Кутты**

   Пусть дана следующая система обыкновенных дифференциальных уравнений:

$$\dot{\mathbf{y}} = \mathbf{f}(t; \mathbf{y}),$$

$$\mathbf{y}(0) = \mathbf{y}_0.$$

   Для численного решения задачи проведем следующие вычисления:

$$\mathbf{k}_1 = \mathbf{f}(t_n, \mathbf{y}_n),$$

$$\mathbf{k}_2 = \mathbf{f}(t_n + c_2\tau, \mathbf{y}_n + \tau\, b_{21}\mathbf{k}_1),$$

$$\mathbf{k}_s = \mathbf{f}(t_n + c_s\tau, \mathbf{y}_n + \tau \sum_{\mathrm{r}=1}^{s-1} b_{s\mathrm{r}}\mathbf{k}_\mathrm{r}),$$

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \tau \sum_{p=1}^{s} a_p \mathbf{k}_p,$$

где a, b, c — действительные числа (вектор $\mathbf{y}_n$ считается известным, т. к. $\mathbf{y}_0$ — начальные условия, по этим значениям можно найти $\mathbf{y}_1$ и т. д.).

Численный метод называется *явным методом Рунге-Кутты порядка S* или *S-стадийным*.

Будем также записывать метод Рунге-Кутты в виде *таблицы Бутчера* — таблицы коэффициентов метода:

0	0	0	...	0	0
c_2	b_{21}	0	...	0	0
...	...	...	...	...	...
c_s	b_{s1}	b_{s2}	...	b_{ss-1}	0
	a_1	a_2	...	a_{s-1}	a_s

Для метода *порядка* S вывести условия, связывающие коэффициенты таблицы Бутчера, необходимые для обеспечения p-го порядка аппроксимации метода ($p \leqslant s$).

Указание 1. Разложить проекцию на сетку точного решения задачи Коши в ряд Тейлора в окрестности точки t_n. Использовать следствия исходного дифференциального уравнения, например:

$$\frac{d^2\mathbf{y}}{dt^2} = \frac{d}{dt}\mathbf{f}(t, \mathbf{y}) = \frac{\partial \mathbf{f}}{\partial t} + \frac{\partial \mathbf{f}}{\partial \mathbf{y}}\frac{\partial \mathbf{y}}{\partial t} = \frac{\partial \mathbf{f}}{\partial t} + \frac{\partial \mathbf{f}}{\partial \mathbf{y}}\mathbf{f}.$$

Указание 2. Для простоты все вычисления провести для случая одного нелинейного уравнения. Требуемые условия приведены в [3, с. 153, 162]. В качестве упражнения найти все явные методы Рунге-Кутты с $S = p = 3, S = p = 4$.

2. **Семейство неявных методов Рунге-Кутты**

Метод вида

$$\mathbf{k}_1 = \mathbf{f}(t_n + c_1\tau, \mathbf{y}_n + \tau \sum_{l=1}^{s} b_{1l}\mathbf{k}_l),$$

...

$$\mathbf{k}_s = \mathbf{f}(t_n + c_s\tau, \mathbf{y}_n + \tau\sum_{l=1}^{s} b_{sl}\mathbf{k}_l),$$

$$\mathbf{y_{n+1}} = \mathbf{y}_n + \tau\sum_{l=1}^{s} a_l\mathbf{k}_l,$$

где $\mathbf{k}_1, \ldots, \mathbf{k}_s$ определяются как решение нелинейной системы уравнений, называется *неявным методом Рунге-Кутты порядка S (S-стадийным)*. Как будет выглядеть для него таблица Бутчера?

Вывести условия аппроксимации порядка p на решении (p=1, 2, 3, 4; $S = 2$).

Следует обратить внимание на то, что для определения $\mathbf{k}_1, \mathbf{k}_2, , \mathbf{k}_s$ необходимо решать систему нелинейных уравнений. Какова ее размерность?

В чем состоит особенность методов с $b_{ij} = 0$ при $j > i$?

(Это так называемые полуявные или диагонально-неявные методы).

3. **Управление длиной шага**

(a) Система ОДУ решается с помощью метода Рунге-Кутты порядка аппроксимации $p > 1$. Описать алгоритм выбора шагов интегрирования, таких, чтобы достигнутая погрешность на каждом из них не превосходила заданную величину ε.

Решение. Пусть переход от значения y_n к y_{n+1} осуществляется с шагом h. Тогда в результате работы метода получим

$$y^{n+1} = \tilde{y}^{n+1} + Ch^{p+1} + O(h^{p+2}), \qquad (8.14)$$

где $\tilde{y}^{n+1}$ — точное решение задачи, взятое в точке $n+1$, Ch^{p+1} — главный член погрешности, постоянная C зависит как от коэффициентов конкретного метода, так и от решения задачи. (Из решения задачи 8.1 следует, что разложения y_{n+1} и $\tilde{y}^{n+1}$ в ряд Тейлора совпадают до членов с номерами $p+1$ включительно).

Осуществим теперь переход от y_n к y_{n+1} за два этапа, используя тот же метод Рунге-Кутты с шагом $h/2$. Тогда имеем

$$\tilde{\tilde{y}}^{n+1} = \tilde{y}^{n+1} + 2C(h/2)^{p+1} + O(h^{p+2}). \qquad (8.15)$$

Так как C оценивается через максимальное по абсолютной величине значение $(p+1)$-й производной на отрезке $[t_n, t_{n+1}]$,

можно считать, что в последнем выражении константа совпадает с константой в предыдущем равенстве. Используя два равенства, возможно либо явно вычислить Ch^p и исключить его из (8.15), повысив точность аппроксимации на порядок, либо поступить следующим образом. Вычитая из (8.15) равенство (8.14), имеем

$$y^{n+1} - \tilde{\tilde{y}}^{n+1} = Ch^{p+1}(1 - 2^{-p}) + O(h^{p+2}),$$

отсюда главный член погрешности (8.15) будет равен

$$Ch^{p+1} = \frac{y^{n+1} - \tilde{\tilde{y}}^{n+1}}{1 - 2^{-p}}. \tag{8.16}$$

Необходимо выбрать шаг интегрирования h_{new} таким образом, чтобы главный член погрешности не превосходил заданное значение ($Ch_{\text{new}}^{p+1} < \varepsilon$):

$$\left(\frac{h_{\text{new}}}{h}\right)^{p+1} \frac{y^{n+1} - \tilde{\tilde{y}}^{n+1}}{1 - 2^{-p}} < \varepsilon,$$

откуда имеем

$$h_{\text{new}} = \left(\frac{y^{n+1} - \tilde{\tilde{y}}^{n+1}}{1 - 2^{-p}}\right)^{1/(p+1)} h. \tag{8.17}$$

Если же на текущем шаге погрешность (8.16) превышает ε , то шаг считается *отклоненным* и расчет выполняется снова со значением, найденным по формуле (8.17).

(b) Система ОДУ решается с помощью *явных* методов p и $p+1$ порядка аппроксимации. На каждом шаге погрешность расчетов не должна превышать ε. Как выбрать длину шага интегрирования?

Рассуждаем аналогично тому, как это сделано в пункте 1. При расчете методом p порядка аппроксимации имеем

$$y_p^{n+1} = \tilde{y}^{n+1} + Ch^{p+1} + K_1 h^{p+2} + O(h^{p+3}), \tag{8.18}$$

а $p+1$ порядок дает

$$y_{p+1}^{n+1} = \tilde{y}^{n+1} + K_2 h^{p+2} + O(h^{p+3}). \tag{8.19}$$

Отсюда сразу получаем $Ch^{p+1} \sim y_p^{n+1} - y_{p+1}^{n+1}$. Эти величины сравниваются с заданной точностью.

Существует семейство методов Рунге-Кутты, таких, что члены погрешности для результата старшего $(p+1)$ порядка минимизируются, а вычисление погрешности p-го порядка используется лишь для управления длиной шага. При этом для порядка p и $p+1$ коэффициенты b, c в таблице Бутчера (2.4.1) совпадают, а различаются лишь коэффициенты a. Кроме того, $b_{si} = a_i(y_p)$. Такие методы построены Дорманом и Принсом и носят их имя.

4. Рассмотрим *уравнение Ферхюльста (логистическое уравнение)*, описывающее динамику численности популяции:

$$y' = ay(1-y), \tag{8.20}$$

$$y(0) = y_0; 0 < y_0 < 1, a > 0.$$

(a) Решить уравнение (8.20) точно, учитывая, что это уравнение в разделяющихся переменных. Получившаяся кривая — график точного решения — носит название логистической кривой.

(b) Рассмотреть для (8.20) *явный метод Эйлера*:

$$y_{n+1} = y_n + \tau a y_n (1 - y_n). \tag{8.21}$$

При каких шагах τ метод является устойчивым?

Решение. При $t \longrightarrow \infty$ точное решение (8.20) асимптотически стремится к единице. Очевидно, что последовательность $y_n \equiv 1$ является и решением (8.20). Непосредственно проверяется, что (8.21) аппроксимирует (8.20) с точностью $O(\tau)$. В случае, если метод (8.21) *устойчив*, то решение (8.21) *сходится* к решению (8.20) и должно выполняться неравенство

$$|y_{n+1} - 1| < |y_n - 1|.$$

Если рассмотреть (8.21) как запись метода простых итераций для (8.20), то он должен сходиться к корню $y = 1$ при любом начальном приближении y_0, лежащем между 0 и 1.

Перепишем (8.21) в виде

$$z_{n+1} = bz_n(1 - z_n), z_n = \frac{\tau a}{\tau a + 1} y_n; b = 1 + \tau a \tag{8.22}$$

и для (8.22) воспользуемся результатами, полученными в Лекции 5. Таким образом, условия устойчивости для (8.21) будут следующими:

$$b = 1 + \tau a < 3, \quad \text{откуда} \quad \tau < 2/a.$$

(c) Описать сценарий развития вычислительной неустойчивости для задачи (8.21), можно воспользоваться материалом лекции 5. Следует обратить внимание на то, что когда метод неустойчив, в отличие от линейных задач, модуль разности численного и точного решений остается ограниченным при любом сколь угодно большом t при $2/a \leqslant \tau < 3/a$.

8.7. Задачи для самостоятельного решения

5. Для уравнения Ферхюльста (8.20) рассмотреть *неявный метод Эйлера*

$$y_{n+1} = y_n + \tau a y_{n+1}(1 - y_{n+1}). \tag{8.23}$$

Для определения y_{n+1} можно воспользоваться (8.23), решая его как квадратное уравнение. Что можно сказать про устойчивость такого метода? Как правильно выбрать корень квадратного уравнения (8.23)?

6. Реализовать *явный* и *неявный* методы Эйлера и метод Рунге-Кутты порядка аппроксимации 4 для системы уравнений Лотки-Вольтерра «хищник–жертва», которая описывает динамику простейшей экосистемы:

$$\dot{x} = ax - xy,$$

$$\dot{y} = bxy - cy,$$

$$x(0) = x_0 > 0, y(0) = y_0 > 0,$$

Здесь x — безразмерная численность «жертв», y — численность «хищников», a, b, c — положительные константы, $b < 1$.

Построить графики численных решений на фазовой плоскости (x, y). Построить там же график точного решения. Объяснить полученные результаты. Что будет в случае, когда конечное время, до которого происходит расчет, $T \gg 1$?

Построение точного решения. в [13, с. 36]. На фазовой плоскости точному решению соответствует кривая

$$\frac{dy}{dx} = -\frac{(bx - c)y}{x(a - y)}.$$

Вводя другую параметризацию кривой $y(x)$, получим, что она может быть описана системой уравнений

$$\frac{dy}{d\tau} = \frac{y}{a-y},$$

$$\frac{dx}{d\tau} = \frac{x}{bx-c}.$$

Эта система легко интегрируется, а после исключения параметра τ находим первый интеграл исходной системы.

7. Решить численно систему из предыдущей задачи, когда скорость размножения жертв является периодической функцией времени: $a(t) = a_0(1 + \sin(\omega t))$, ω — параметр. Что происходит с решением при увеличении ω?

8. Уравнение Ван-дер-Поля [3, с. 115-119]

$$y'' + a(y^2 - 1)y' + y = 0,$$

$$y(0) = y_0 > 0; y'(0) = 0, 0 \leqslant t \leqslant 30, a > 0 (0 \div 1000)$$

описывает нелинейные колебания в различных системах.

(a) Проинтегрировать уравнение численно явными методами Рунге-Кутты различных порядков. При каких шагах τ методы становятся неустойчивыми?

Указание: представить уравнение в виде системы

$$y' = -p,$$

$$p' = ap(y^2 - 1) + y,$$

или в виде

$$y' = -a(\frac{y^3}{3} - y) + p,$$

$$p' = -y$$

(представление Льенара).

(b) Реализовать для уравнения Ван-дер-Поля следующие полуявные методы Рунге-Кутты:

$$\begin{array}{c|cc} \gamma & \gamma & 0 \\ 1-\gamma & \frac{1-2\gamma}{1/2} & \frac{\gamma}{1/2} \end{array}, \quad \gamma_1 = \frac{3+\sqrt{3}}{6} \ (\text{или } \frac{3-\sqrt{3}}{6}), \quad \gamma_2 = \frac{2+\sqrt{2}}{2} \ (\text{или } \frac{2-\sqrt{2}}{2})$$

Какой порядок аппроксимации они имеют? Устойчивы ли они?

9. **Уравнение Эйлера**

Решить численно задачу о колебаниях в системе, где и возвращающая сила, и коэффициент вязкого трения убывают со временем (уравнение Эйлера):

$$\ddot{x} + \frac{\dot{x}}{t} + 100\frac{x}{t^2} = 0,$$

где $t \in [1, 100], x(1) = 1, \dot{x}(1) = 1$.

Использовать для численного решения сетки с шагом $\tau = 0,1$

(a) явный метод Эйлера;

(b) неявный метод Эйлера;

(c) методы Рунге-Кутты порядка 2, 3, 4.

Уменьшить τ вдвое. Что происходит с решением? Какому решению верить?

Сравнить численное решение с точным.

Указание. Будем искать точное решение в виде $x = t^\alpha$, где α, вообще говоря, комплексное. Рекомендации можно найти в [14].

10. **Уравнение Эйлера**

Рассмотрим более простую систему без трения:

$$\ddot{x} + 100\frac{x}{t^2} = 0.$$

Выполнить пункты предыдущей задачи. В чем заключается разница? Точное решение уравнения можно найти в [14] или решить уравнение самостоятельно, положив $x = ct^\alpha$.

11. **Уравнение Хатчинсона**

Одной из модификаций уравнений Ферхюльста является уравнение Ферхюльста с запаздыванием (уравнение Хатчинсона):

$$\dot{y} = ay(1 - y(t - 1)).$$

Заметим, что в качестве начальных условий необходимо задавать $y(t)$ при $t \in [-1, 0]$, т. е. не в одной точке, а на отрезке.

(a) Аналитически решить уравнение так называемым *методом шагов*. Задавая конкретную функцию $y(t) = g(t)$ при $t \in [-1, 0]$ (начальные условия), интегрируем уравнение Хатчинсона на полуинтервале при $t \in]0, 1]$:

$$\frac{dy}{y} = a(1 - g(t))\,dt.$$

Зная $y(t)$ на [0,1], можно найти $y(t)$ на [1,2] и т.д.

Если положить $y(t) = 1/2$ при $t \in] - 1, 0]$, $y(t) = t, y(t) = e^{-t}$, что происходит в точках $t = 1, 2, 3, \ldots n, \ldots$?

(b) Построить численный метод интегрирования уравнение Хатчинсона. Найти решение в случае

$$y(t) = e^{-t^2}, t \in [-1, 0]$$

при $a \ll \pi/2, a = \pi/2, a \gg \pi/2$.

О численном решении уравнений с запаздыванием [3, с. 304–319].

12. **Уравнение Минорского**

$$\ddot{y} + 2r\dot{y} + \omega^2 y + 2q\dot{y}\,(t-1) = \varepsilon\dot{y}^3\,(t-1)$$

встречается в различных механических и электротехнических задачах, в которых имеется запаздывание и нелинейность (здесь r, ω, q — постоянные, $0 < \varepsilon \ll 1$). Построить численное решение, задавая начальные данные на отрезке $t \in [-1, 0]$. Рассмотреть зависимость решения от параметров r, ω, q. В качестве одного из характерных случаев значений этих параметров рассмотреть следующие: $r = -1, q = (-1)n, \omega = n\pi$.

Подробное качественное исследование уравнения Минорского представлено в [12, с. 191–211].

13. **Уравнение Тинбергена**

$$\dot{y} + by(t-1) = \varepsilon[y(t-1)]^3, \quad 0 < \varepsilon \ll 1$$

получено при исследовании циклов в судостроительной промышленности. Предполагается, что время постройки одного судна постоянно (и принято за единицу). Превышение темпа закладки новых судов над средним теоретическим темпом закладки пропорционально нелинейной функции $y - (\varepsilon/b)y^3$. Для рассматриваемого

случая $\varepsilon > 0$ уравнение показывает, что при возрастании отклонения судостроители должны реагировать относительно более консервативно из-за недостатка материалов, рабочей силы и т. п. Решить уравнение численно, задавая начальные данные на отрезке $t \in [-1, 0]$. Рассмотреть характерные режимы: 1) $0 < b < 1/e$, 2) $b \approx \pi/2$, 3) $b > \pi/2$.

Показать, что в интервале $0 < b < \pi/2$ решение уравнения ведет себя так же, как решение линейного уравнения ($\varepsilon = 0$). Исследование поведения решения уравнения Тинбергена имеется в [12, с. 180–184].

Литература

[1] *Рябенький В.С.* Введение в вычислительную математику. М.: Физматлит, 2000. 294 с.

[2] *Бахвалов Н.В., Жидков Н.П, Кобельков Г.М.* Численные методы. М: Лаборатория Базовых Знаний, 2002. 632 с.

[3] *Хайрер Э., Нерсетт С., Ваннер Г.* Решение обыкновенных дифференциальных уравнений. Нежесткие задачи. М.: Мир, 1990. 512 с.

[4] *Curtis A.R.* High-order Explicit Runge-Kutta Formula, Their Uses, and Limitations // J.Jnst.Math.Applics. 1970. V. 16. P. 35-58.

[5] *Hairer E.* A Runge-Kutta Method of Order 10 // J.Jnst.Math.Applics. 1978. V. 21. P. 47-59.

[6] *Dormand J.R., Prince P.J.* A Family of Embedded Runge-Kutta Formulae // J.Comp.Appl.Math. 1980. V. 6. P. 19-26.

[7] *Prince P.J., Dormand J.R.* High Order Embedded Runge-Kutta Formulae // J.Comp.Appl.Math. 1981. V. 7. P. 67-78.

[8] *Fehlberg E.* Classical Fifth-, Sixth-, Seventh and Eighth Order Runge-Kutta formulas with step size control. NASA Technical Report. 1968, 287. Extract published in // Comptuting. 1969. V. 4. P. 93-106.

[9] *Федоренко Р.П.* Введение в вычислительную физику. М.: Изд-во МФТИ, 1994. 528 с.

[10] *Хайрер Э., Ваннер Г.* Решение обыкновенных дифференциальных уравнений. Жесткие и дифференциально-алгебраические системы. М.: Мир, 1999. 685 с.

[11] *Лобанов А.И., Петров И.Б.* Вычислительные методы для анализа моделей сложных динамических систем. Ч. I. М.: МФТИ, 2000. 168 с.

[12] *Пинни Э.* Обыкновенные дифференциально-разностные уравнения. М.: ИЛ, 1961. 248 с.

[13] *Арнольд В.И.* Обыкновенные дифференциальные уравнения. 3-е изд. М.: Наука, 1984. 272 с.

[14] *Малинецкий Г.Г.* Задачи по курсу нелинейной динамики / В кн.: Новое в синергетике. Загадки мира неравновесных структур. М.: Наука, 1997. С. 215-262.

Лекция 9. Численные методы решения жестких систем обыкновенных дифференциальных уравнений

Дается понятие жесткой системы (ЖС ОДУ). Рассматриваются неявные методы Рунге-Кутты и Гира для решения ЖС ОДУ. Исследуется устойчивость методов.

Ключевые слова: жесткие системы ОДУ, спектр матрицы Якоби, А-устойчивость, асимптотическая устойчивость, жесткая устойчивость, неявные и диагонально-неявные методы Рунге-Кутты, методы Розенброка, методы Гира, представление Нордсика.

9.1. Явление жесткости. Предварительные сведения

Рассмотрим в качестве примера две задачи Коши для систем обыкновенных дифференциальных уравнений (ОДУ) [1, 2]:

$$\dot{u} = au + \frac{1}{\varepsilon}\, v, \dot{v} = -\frac{1}{\varepsilon}\, v,$$

с начальными данными $u(0) = u_0, v(0) = v_0$; здесь $a \sim O(1), \varepsilon \ll 1$; и линейную систему с постоянными коэффициентами

$$\dot{u} = 998u + 1998v,$$

$$\dot{v} = -999u - 1999v,$$

$$u(0) = v(0) = 1.$$

Решением первой задачи Коши являются функции

$$u(t) = u_0 e^{at} + \frac{v_0}{1 + a\varepsilon}\,(e^{at} - e^{-t/\varepsilon}),$$

$$v(t) = v_0 e^{-t/\varepsilon},$$

а второй —

$$u(t) = 4e^{-t} - 3e^{-1000t},$$

$$v(t) = -2e^{-t} + 3e^{-1000t}.$$

В обоих случаях решение состоит из двух экспонент: быстро убывающей и относительно медленно изменяющейся. Отметим, что абсолютные величины собственных значений матриц рассматриваемых линейных систем ОДУ при их представлении в виде

$$\dot{\mathbf{u}} = \mathbf{A}\mathbf{u},$$

($\mathbf{u}$ — вектор-столбец, $\mathbf{A}$ — матрица с постоянными коэффициентами) существенно различаются. Так, в первом случае $\lambda_1 \approx (4\varepsilon)^{-1}, \lambda_2 \sim O(1)$; во втором: $\lambda_1 \approx -1, \lambda_2 = 10^{-3}$. В обоих случаях имеем:

$$\frac{|\lambda_1|}{|\lambda_2|} \gg 1.$$

При моделировании физических процессов причина такой разницы в собственных числах заключена в существенно различных характерных временах процессов, описываемых системами ОДУ. Наиболее часто подобные системы встречаются при моделировании процессов в ядерных реакторах, при решении задач радиофизики, астрофизики, физики плазмы, биофизики, химической кинетики. Последние задачи часто могут быть записаны в виде [3]:

$$\frac{du_k}{dt} = \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij}^k u_i u_j, k = 1 \div N;$$

где u_k — концентрации веществ, участвующих в химических реакциях, скорости протекания которых характеризуются коэффициентами a_{ij}^k. В качестве примера приведем одну из систем химической кинетики, описывающую изменение концентрации трех веществ, участвующих в реакции для случая полного перемешивания [1].

Пример 1. Обозначим концентрации трех веществ, участвующих в реакции, через u_1, u_2 и u_3, тогда

$$\begin{aligned}
\dot{u}_1 &= -4 \cdot 10^{-2} u_1 + 10^4 u_2 u_3, \\
\dot{u}_2 &= 10^{-2} u_1 - 10^4 u_2 u_3 - 3 \cdot 10^7 u_2^2, \\
\dot{u}_3 &= 3 \cdot 10^7 u_2^2, \\
u_1(0) &= 1, u_2(0) = u_3(0) = 0.
\end{aligned}$$

Участки решения, характеризующиеся быстрым и медленным его изменением, называются *пограничным слоем* и *квазистационарным режимом*, соответственно.

Трудности численного решения подобных систем ОДУ, получивших название *жестких* (определение жесткой системы приведено ниже), связаны с выбором шага интегрирования. Дело в том, что характерные времена исследуемых процессов могут различаться более чем в 10^{12} раз. Следовательно, если при численном решении системы

$$\dot{\mathbf{u}} = \mathbf{f}(\mathbf{u})$$

выбирать шаг из условия

$$\tau \, ||\mathbf{f}'_{\mathbf{u}}(\mathbf{u})|| \ll 1,$$

то он будет соответствовать самому быстрому процессу. В данном случае затраты машинного времени для исследования самых медленных процессов будут неоправданно велики. По этой причине имеются следующие альтернативы в выборе подхода к численному решению рассматриваемых задач.

1. Численно решать систему ОДУ с шагом

$$\tau \ll ||\mathbf{f}'_{\mathbf{u}}(\mathbf{u})||^{-1},$$

т. е. с учетом характерных времен всех процессов, описываемых данной системой.

2. Решать систему ОДУ с различными шагами, соответствующими физическим процессам с существенно различными характерными временами. В этом случае необходимо задавать условия перехода к другому шагу интегрирования.

3. «Пренебречь» быстропротекающими процессами и численно рассматривать лишь медленные, проводя интегрирование с шагом, превышающим характерные времена быстрых процессов. В этом случае придется конструировать численные методы, позволяющие проводить расчеты с шагом

$$\tau \gg ||\mathbf{f}'_{\mathbf{u}}(\mathbf{u})||^{-1}.$$

Определение ([3]). Система ОДУ для задачи Коши

$$\dot{\mathbf{u}} = \mathbf{f}(\mathbf{u}), \mathbf{u}(0) = \mathbf{u}_0, t_0 \leqslant t \leqslant t_k$$

называется жесткой, если спектр матрицы Якоби

$$\mathbf{J} = \{\mathbf{f}'_{\mathbf{u}}(\mathbf{u})\}$$

разделяется на две части.

1. Жесткий спектр:

$$\operatorname{Re}\Lambda_i(u) \leqslant -\Lambda_0, |\operatorname{Im}\Lambda_k| < |\operatorname{Re}\Lambda_k|, k = 1 \div N_1$$

(Λ_i — *собственные значения матрицы Якоби*); $\Lambda_0 > 0$

2. *Мягкий спектр*:

$$|\lambda_j| \leqslant \lambda_0, j = 1 \div N_2, \lambda_0 > 0.$$

При этом $\lambda_0 \ll \Lambda_0$.

Отношение Λ_0/λ_0 называется *показателем жесткости системы*. В дальнейшем будем полагать $\lambda_0 \sim O(1)$.

Проблему численного решения жестких систем ОДУ рассмотрим на примере модельной линейной системы вида

$$\dot{\mathbf{u}} = \mathbf{B}\mathbf{u}, \mathbf{u}(0) = \mathbf{u}_0. \tag{9.1}$$

Ее точное решение задается формулой

$$\mathbf{u}(t) = \sum_{j=1}^{N_1} b_j e^{\Lambda_j t}\mathbf{\Omega}_j + \sum_{k=1}^{N_2} \bar{b}_k e^{\lambda_k t}\boldsymbol{\omega}_k, \tag{9.2}$$

где константы интегрирования $b_j, \bar{b}_k$ соответствуют жесткой и мягкой частям спектра; $\mathbf{\Omega}_j, \boldsymbol{\omega}_k$ — собственные векторы матрицы Якоби, соответствующие собственным значениям Λ_j, λ_k.

В этом решении видны две части: первая (жесткая) убывает как $e^{-\Lambda_0 t}$ на временном интервале $[t_0, O(\Lambda_0^{-1})]$ (пограничный слой), вторая заметно изменяется на интервале $[t_0, O(\Lambda_0^{-1})]$ (квазистационарный режим).

Если провести аппроксимацию линейной системы ОДУ с помощью явного метода Эйлера

$$\frac{\mathbf{u}_{n+1} - \mathbf{u}_n}{\tau} = \mathbf{B}\mathbf{u}_n,$$

или

$$\mathbf{u}_{n+1} = (\mathbf{E} + \tau\mathbf{B})\mathbf{u}_n,$$

то общее решение такой системы разностных уравнений будет иметь вид

$$\mathbf{u}_n(t) = \sum_{j=1}^{N_1} c_j(1 + \tau\Lambda_j)^n\Omega_j + \sum_{k=1}^{N_2} c_k(1 + \tau\lambda_k)^n\omega_k. \tag{9.3}$$

Второе слагаемое в этом решении аппроксимирует второе слагаемое в точном решении (9.2), а первое быстро растет и приводит к абсурдному результату.

Теперь проведем аппроксимацию линейной системы ОДУ (9.1) с помощью неявного метода Эйлера:

$$\frac{\mathbf{u}_{n+1} - \mathbf{u}_n}{\tau} = \mathbf{B}\mathbf{u}_{n+1},$$

или

$$\mathbf{u}_{n+1} = (\mathbf{E} + \tau\mathbf{B})^{-1}\mathbf{u}_n.$$

Общее решение такого разностного уравнения имеет следующий вид:

$$\mathbf{u}_n(t) = \sum_{j=1}^{N_1} c_j(1 - \tau\Lambda_j)^{-n}\mathbf{\Omega}_j + \sum_{k=1}^{N_2} \bar{c}_k(1 - \tau\lambda_k)^{-n}\boldsymbol{\omega}_k.$$

В этом случае второе слагаемое ведет себя так же, как и точное решение, а первое стремится к нулю как $(\tau\Lambda_0)^{-n}$, т. е. его поведение качественно совпадает с точным в области пограничного слоя.

В практике численных исследований жестких задач часто не нужно изучать поведение решения в пограничном слое, и можно воспользоваться неявными методами. Но в случае необходимости исследовать этот слой можно с шагом $\tau \ll \Lambda_0^{-1}$.

Устойчивость методов численного интегрирования жестких систем ОДУ обычно исследуется на примере скалярного уравнения

$$\dot{u} = \lambda u, \quad u(0) = u_0. \tag{9.4}$$

Положим, что численный метод, применяемый к решению этого уравнения, может быть записан в виде

$$u_{n+1} = R(z)u_n, z = \tau\lambda,$$

где $R(z)$ называется функцией устойчивости [1, 4]. О построении функции устойчивости речь пойдет ниже.

Определение. Численный метод для решения уравнения (9.4) является абсолютно устойчивым, если выполнено условие

$$|R(z)| \leqslant 1.$$

Из определения следует, что $|u_{n+1}| \leqslant |u_n|$.

Это требование является естественным при $\mathrm{Re}\ z \leqslant 0$, поскольку в таком случае модуль точного решения есть невозрастающая функция.

Множество всех точек z, для которых $|R(z)| \leqslant 1$, называется *областью абсолютной устойчивости.*

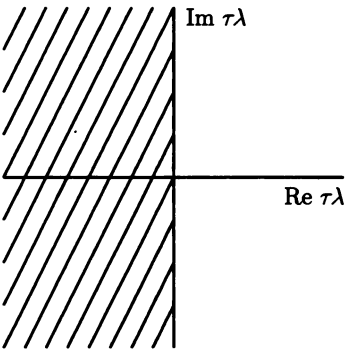


Рис. 9.1

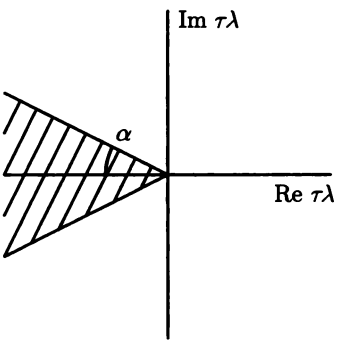


Рис. 9.2 а

Определение. Если область абсолютной устойчивости ($|R(z)| \leqslant 1$) занимает левую полуплоскость комплексной плоскости ($\mathrm{Re}\ z \leqslant 0$), то метод является A-устойчивым (заштрихованная область на рис. 9.1).

В случае, когда область абсолютной устойчивости включает в себя угол (в левой полуплоскости комплексной плоскости) с вершиной в нуле и углом полураствора α, то метод называется A(α)-устойчивым.

Области $A(\alpha)$- и $A(0)$-устойчивости изображены на рис. 9.2а и 9.2б, соответственно.

В случае, когда вся область абсолютной устойчивости включает в себя часть левой полуплоскости (граница ее лежит вне заштрихованной части на рис. 9.3), то метод называется жестко-устойчивым. Область жесткой устойчивости имеет вид, представленный, например, на рис. 9.3.

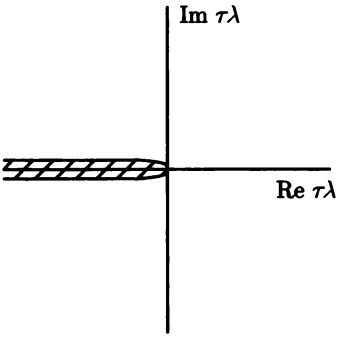


Рис. 9.2b

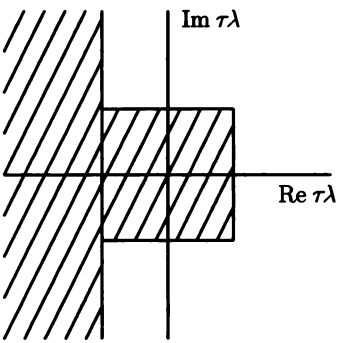


Рис. 9.3

Определение. Численный метод называется L-устойчивым, если он A-устойчив и если

$$|R(z)| \to 0, \quad \text{при } \mathrm{Re}\,\tau\lambda \to -\infty.$$

В частности, рассмотренный выше неявный метод Эйлера является L-устойчивым. Решения, полученные такими методами, будут затухающими.

9.2. Сингулярно-возмущенные задачи

Рассмотрим простейшую нелинейную жесткую систему А. Н. Тихо-

нова [5] (сингулярно-возмущенная задача Коши с малым параметром при производной):

$$\varepsilon \dot{u} = f(u,\ v),$$

$$\dot{v} = g(u,\ v),$$

$$0 < \varepsilon \ll 1,$$

или

$$\dot{u} = \frac{1}{\varepsilon}\, f(u,\ v),$$

$$\dot{v} = g(u,\ v),$$

с начальными условиями

$$u(0) = u_0, v(0) = v_0.$$

Из характеристического уравнения находим:

$$\det \begin{vmatrix} \frac{1}{\varepsilon} f'_u - \lambda & \frac{1}{\varepsilon} f'_v \\ g'_u & g'_v - \lambda \end{vmatrix} = \lambda^2 - \lambda \left(\frac{1}{\varepsilon} f'_u + g'_u \right) + \frac{1}{\varepsilon} (f'_u g'_v - g'_u f'_v),$$

$$\lambda_1 = \frac{1}{\varepsilon} f'_u + \mathrm{O}(1), \quad \lambda_2 = \mathrm{O}(1),$$

т. е. λ_1 — жесткая, а λ_2 — мягкая часть спектра.

Первое, что хотелось бы сделать, — упростить задачу, положив $\varepsilon \approx 0$. В этом случае система приобретает вид:

$$f(u,\ v) = 0,$$

$$\dot{v} = g(u,\ v),$$

$$u(0) = u_0, v(0) = v_0$$

и описывает траекторию, проходящую по кривой $f(u,\ v) = 0$. Эта кривая играет существенную роль в решении исходной системы ОДУ. Однако полностью поведение решения упрощенная (невозмущенная) система описывать не будет, поскольку начальные данные не обязательно должны лежать на этой кривой.

Для того чтобы понять эту ситуацию, рассмотрим фазовый портрет конкретной системы ОДУ вида:

$$\varepsilon \dot{u} = v - \frac{1}{3} u^3 + u,$$

$$\dot{v} = -u,$$

$$u(0) = u_0, v(0) = v_0.$$

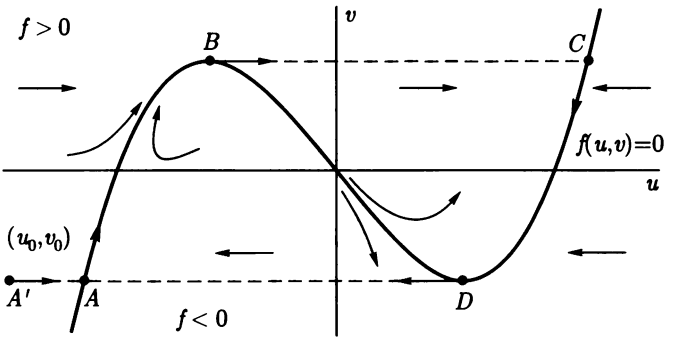


Рис. 9.4

Эта система эквивалентна нелинейному уравнению второго порядка — уравнению Ван-дер-Поля. Приведенный выше вид иногда называется представлением Льенара. Подробнее о свойствах рассматриваемой задачи можно прочитать в [6, 7, 8, 9].

Кривая $f(u,\ v) = 0$ для этого случая изображена на рис. 9.4. Она делит плоскость u, v на две части: $f > 0$ и $f < 0$; вдали от кривой, поле скоростей du, dv направлено почти горизонтально влево (вправо) в зависимости от знака f. На самой кривой выделяются две устойчивые ветви AB и CD, где $f'_u < 0$, и неустойчивая ветвь BD, на которой $f'_u > 0$. Опишем теперь качественно поведение траектории рассматриваемой системы ОДУ, состоящей из следующих участков.

1. Пограничный слой $A'A$. На этом участке за малое время $O(\varepsilon)$ траектория из точки $\{u_0,\ v_0\}$ переходит в ε-окрестность кривой $f = 0$. Здесь траектория почти горизонтальна и приближенно определяется дифференциальным уравнением

$$\dot{u} = \frac{1}{\varepsilon} f(u,\ v),$$

$$v(t) \approx v_0,$$

$$u(0) = u_0.$$

($f/\varepsilon \gg g$, так как $f, g \sim O(1), 0 < \varepsilon \ll 1$).

В окрестности кривой $f(u,\ v) = 0$ имеем $f'_u < 0$, поэтому допустима оценка

$$\frac{df}{dt} = f'_u \cdot \dot{u} = \frac{1}{\varepsilon} f \cdot f'_u,$$

откуда видно, что $f(u,\ v_0)$ стремится к нулю как экспонента с показателем $\frac{1}{\varepsilon}\ f'_u$, т. е. f становится малой величиной (≈ 0) за время $t_{A'A} = O(\varepsilon)$.

2. Квазистационарный режим. Движение точки $\{u(t),\ v(t)\}$ продолжается уже по участку кривой AB, $f(u,\ v) = 0$ и описывается системой

$$f(u,\ v) = 0,$$

$$\dot{v} = g(u,\ v).$$

На этом участке за время $t_{AB} \sim O(1)$, что видно из второго уравнения, точка подвигается от A к B, пока система ОДУ устойчива. В случае невозмущенной системы точка могла бы и далее продвигаться по участку BD, но для полной системы эта ветвь оказывается неустойчивой ($f'_u > 0$) и траектория «срывается» на устойчивую ветвь CD в точке B, в которой $f'_u = 0$.

3. Пограничный слой. На участке BC точка $\{u(t),\ v(t)\}$ «перескакивает» из B в C за малое время $O(\varepsilon)$; движение здесь, как и на ветви AB, приближенно описывается уравнениями

$$\dot{u} = \frac{1}{\varepsilon}\ f(u,\ v),$$

$$v(t) \approx \text{const}.$$

4. Квазистационарный режим. Движение по ветви CD, как и по ветви AB, описывается уравнениями

$$f(u,\ v) = 0,$$

$$\dot{v} = g(u,\ v),$$

и длится $t_{CD} \sim O(1)$.

5. Пограничный слой. На неустойчивой ветви DA также, как и на ветви BC, происходит скачок из точки D за время $t_{DA} \sim O(\varepsilon)$ в устойчивую точку A, и т. д.

Такое поведение траектории (замкнутая кривая) называется предельным циклом. Для жестких систем периодические решения называют иногда релаксационными колебаниями [6, 7].

Таким образом, это характерно для жестких систем, траектория состоит из чередующихся участков быстрого (за время $t \sim O(\varepsilon)$) и медленного (за время $t \sim O(1)$) изменения решения.

Рассмотрим проблемы, которые могут возникнуть при численном интегрировании подобных жестких систем ОДУ. Численное интегрирование в зоне пограничного слоя, если оно необходимо исследователю, проблемы не составляет. Требуется лишь выполнение условия

$$\tau \cdot \frac{1}{\varepsilon} |f'_u| \ll 1.$$

В зоне квазистационарного режима часто оказывается, что интегрирование с таким шагом слишком дорого. Можно, правда, разрешить уравнение $f(u,\ v) = 0$ $(u = \varphi(v))$ относительно u и далее интегрировать его, предварительно реализовав алгоритм перехода на другой шаг:

$$\dot{v} = g(v,\ \varphi(v)),$$

однако в случаях более сложных, построение подобных численных методов может оказаться отдельной сложной задачей.

Чаще всего в практике численных расчетов целесообразно использовать неявные схемы. В случае ЖС ОДУ неявные схемы предпочтительнее из соображений устойчивости. Так, рассматриваемую задачу можно аппроксимировать системой дискретных уравнений:

$$\frac{u_{k+1} - u_k}{\tau} = \frac{1}{\varepsilon} f(u_{k+1},\ v_{k+1}),$$

$$\frac{v_{k+1} - v_k}{\tau} = g(u_{k+1},\ v_{k+1}).$$

Эта система нелинейных уравнений может быть решена численно, например, методом Ньютона. Иногда полагают, что неявные схемы позволяют проводить численное интегрирование сквозным методом с большим шагом τ. Рассмотрим, к чему это может привести.

Первый пограничный слой будет пройден за один шаг:

$$u_1 = u_0 + \frac{\tau}{\varepsilon} f(u_1,\ v_1),$$

$$v_1 = v_0 + \tau g(u_1,\ v_1),$$

так как $\varepsilon \ll \tau, \tau g$ мало.

Далее следует процесс численного интегрирования на устойчивой ветви AB. В окрестности точки B поведение численного решения по неявной схеме осложняется. Это связано с тем, что рассматриваемая система в данной окрестности может иметь более одного решения. При этом, по крайней мере, одно из решений нелинейной алгебраической системы может лежать на неустойчивой ветви CD кривой $f(u,\ v) = 0$. Возможно, что при выборе большего шага интегрирования получится именно это нефизическое решение, к которому сойдутся итерации.

Такую опасность необходимо всегда учитывать при проведении численного интегрирования по неявным схемам с большим шагом.

9.3. Решение линейных ЖС ОДУ и вычисление матричной экспоненты

Рассмотрим один из наиболее простых численных методов решения жестких линейных систем ОДУ [10], основанный на представлении решения в явном виде

$$\mathbf{u}_{n+1} = e^{\mathbf{B}\tau}\mathbf{u}_n$$

для линейных систем жестких ОДУ вида

$$\dot{\mathbf{u}} = \mathbf{B}\mathbf{u}, \mathbf{u}(0) = \mathbf{u}_0.$$

Его численная реализация связана с вычислением матричной экспоненты. О свойствах матричной экспоненты (и в целом функций от матриц) можно прочитать в [11]. Использование для этого разложения в ряд Тейлора

$$e^{\mathbf{A}\tau} = \mathbf{E} + \sum_{k=1}^{N} \frac{\tau^k}{k!}\mathbf{B}^k$$

представляется непригодным, так как простые оценки показывают, что по вычислительным затратам этот алгоритм сопоставим с методом Эйлера с шагом $\tau\ ||\mathbf{B}|| < 1$:

$$\frac{\tau^k}{k!}\ ||\mathbf{B}^k||\ \approx \left(\frac{e\tau||\mathbf{B}||}{k}\right)^k, \text{ так как } k! \sim \left(\frac{k}{e}\right)^k.$$

Следовательно, для того, чтобы k-й член разложения ряда Тейлора был, по крайней мере, порядка $O(1)$, необходимо выполнение условия

$$\frac{e\tau||\mathbf{B}||}{k} \sim O(1), \text{ или } \tau||\mathbf{B}|| \sim O(1),$$

что соответствует условию численного интегрирования с шагом $\tau||\mathbf{B}|| < 1$. Количество членов ряда при этом $N \sim \tau||\mathbf{B}|| \gg 1$, что неприемлемо для решения.

Представим экспоненту в следующем эквивалентном виде:

$$e^{\mathbf{B}\tau} = (e^{\tau\mathbf{A}/2^p})^{2^p} \approx \left[\mathbf{E} + \frac{\tau}{2^p}\mathbf{B} + \ldots + \frac{1}{k!}\left(\frac{\tau}{2^p}\right)^k \mathbf{B}^k\right]^{2^p}.$$

При этом значение параметра p выбирают таким, чтобы

$$\frac{\tau||\mathbf{B}||}{2^p} \ll 1$$

и можно было использовать ряд Тейлора с небольшим количеством членов. Действительно, в этом случае

$$\frac{\tau||\mathbf{B}||}{k \cdot 2^p} \sim O(1) \text{ и } N \sim 2^{-p}\tau||\mathbf{B}||,$$

что вполне приемлемо при соответствующем выборе параметра p.

В этом алгоритме сначала вычисляют матрицу

$$\mathbf{D} = \mathbf{E} + \sum_{k=1}^{N} \left(\frac{\tau}{2^p}\right)^k \mathbf{B}^k,$$

затем $\mathbf{D}^{2^p}$ путем последовательных перемножений. При таком способе вычисления матрицы также имеется опасность, связанная с тем, что при некоторых p, λ_0, Λ_0 слагаемые, соответствующие мягкой части спектра, окажутся много меньше единицы (и слагаемых, соответствующих жесткой части). В этом случае предпочтение отдается представлению

$$e^{\mathbf{A}\tau} = \left[\exp\left(-\frac{\tau}{2^p}\,\mathbf{B}\right)^{-1}\right]^{2^p},$$

а последовательность вычислений имеет вид

$$\mathbf{D} = \mathbf{E} + \frac{\tau}{2^p}\,\mathbf{B}; \mathbf{C} = \mathbf{D}^{-1}; \mathbf{C}' = (\mathbf{C}^2)^p.$$

При этом на мягкой части спектра, поскольку $\tau|\lambda_j| \ll 1$, имеем

$$\left(1 - \frac{\tau}{2^p}\,|\lambda_j|\right)^{-2^p} \approx e^{\lambda\tau};$$

на жесткой части, поскольку теперь p небольшие и $\tau|\Lambda_j|/2^p \gg 1$, получим оценку

$$\left|\left(\frac{1}{1 - \tau|\Lambda_j| \cdot 2^p}\right)^{2^p}\right| \ll 1,$$

что уже приемлемо для вычислений.

9.4. Численные методы решения ЖС ОДУ. Семейства неявных методов Рунге-Кутты и Розенброка

Рассмотрим другие методы для численного решения как линейных жестких систем ОДУ вида (9.1), так и нелинейных систем общего вида

$$\dot{\mathbf{u}} = \mathbf{f}(t; \mathbf{u}), \mathbf{u}(0) = \mathbf{u}_0$$

и автономных ЖС ОДУ

$$\dot{\mathbf{u}} = \mathbf{f}(\mathbf{u}), \quad \mathbf{u}(0) = \mathbf{u}_0,$$

см. также [1, 4, 9, 13, 14, 15, 16]. Из соображений устойчивости метода предпочтение естественно отдать неявным методам. Простейшими из них являются:

1. неявный метод Эйлера (приведем его вид для случая автономной ЖС ОДУ, для неавтономной системы формулы очевидны):

$$\frac{\mathbf{u}_{n+1} - \mathbf{u}_n}{\tau} = \mathbf{f}(\mathbf{u}_{n+1}), \mathbf{u}_0 = \mathbf{a};$$

2. метод трапеций

$$\frac{\mathbf{u}_{n+1} - \mathbf{u}_n}{\tau} = \frac{\mathbf{f}(\mathbf{u}_{n+1},\ t_{n+1}) + \mathbf{f}(\mathbf{u}_n,\ t_n)}{2}, \mathbf{u}_0 = \mathbf{a};$$

3. метод прямоугольников (правило средней точки)

$$\frac{\mathbf{u}_{n+1} - \mathbf{u}_n}{\tau} = \mathbf{f}(\mathbf{u}_{n+1/2},\ t_{n+1/2}), \mathbf{u}_0 = \mathbf{a},$$

где

$$t_{n+1/2} = \frac{t_{n+1} + t_n}{2}, \mathbf{u}_{n+1/2} = \frac{\mathbf{u}_{n+1} + \mathbf{u}_n}{2}.$$

Среди одношаговых методов для решения жестких систем наиболее известны методы Рунге-Кутты. Все приведенные в данном параграфе формулы можно рассматривать как частные случаи неявных одно- и двухстадийных методов из этого семейства. Не останавливаясь на получении приводимых коэффициентов, выпишем наиболее известные из формул Рунге-Кутты, используя таблицу Бутчера. О представлении методов Рунге-Кутты в виде таблиц Бутчера — в лекции 8.

Отметим, что в отличие от рассматриваемых выше *явных* методов, при использовании *неявных* схем Рунге-Кутты матрица коэффициентов метода в таблице Бутчера — заполненная, и для определения вспомогательных векторов, входящих в функцию приращения, приходится решать систему нелинейных алгебраических уравнений.

1. Методы Гаусса соответственно 2-го, 4-го, и 6-го порядков представлены в табл. 9.1, 9.2, 9.3. Основаны эти методы на соответствующих квадратурных формулах Гаусса, рассмотренных в лекции 7. Первый метод совпадает с правилом средней точки. Второй метод (4-го порядка) носит название метода Хаммера-Холлинсворта.

Таблица 9.1

1/2	1/2
	1

Таблица 9.2

$\frac{1}{2}-\frac{\sqrt{3}}{6}$	$\frac{1}{4}$	$\frac{1}{4}-\frac{\sqrt{3}}{6}$
$\frac{1}{2}+\frac{\sqrt{3}}{6}$	$\frac{1}{4}+\frac{\sqrt{3}}{6}$	$\frac{1}{4}$
	1/2	1/2

Таблица 9.3

$\frac{1}{2}-\frac{\sqrt{15}}{10}$	$\frac{5}{36}$	$\frac{2}{9}-\frac{\sqrt{15}}{15}$	$\frac{5}{36}-\frac{\sqrt{15}}{30}$
$\frac{1}{2}$	$\frac{5}{36}+\frac{\sqrt{15}}{24}$	$\frac{2}{9}$	$\frac{5}{36}-\frac{\sqrt{15}}{24}$
$\frac{1}{2}+\frac{\sqrt{15}}{10}$	$\frac{5}{36}+\frac{\sqrt{15}}{30}$	$\frac{2}{9}+\frac{\sqrt{15}}{15}$	$\frac{5}{36}$
	$\frac{5}{18}$	$\frac{4}{9}$	$\frac{5}{18}$

2. Методы Радо IIA порядков 1, 3, 5 представлены соответственно в табл. 9.4, 9.5, 9.9. Основаны на квадратурных формулах Радо, принадлежащих к семейству формул Гаусса. Метод первого порядка является неявным методом Эйлера.

Таблица 9.4

1	1
	1

Таблица 9.5

1/3	5/12	-1/12
1	3/4	1/4
	3/4	1/4

3. Методы Лобатто IIIA 2-го, 3-го, 4-го и 6-го порядков точности см. в табл. 9.7, 9.8, 9.9 и 9.10 соответственно. Видно, что метод второго порядка точности является неявным методом трапеций.

Рассмотрим теперь, как строится функция устойчивости для методов Рунге-Кутты [4]. Уже отмечалось выше, что при построении

Таблица 9.6

$\frac{4-\sqrt{6}}{10}$	$88-7\sqrt{6}$	$\frac{296-169\sqrt{6}}{1\,800}$	$\frac{-2+3\sqrt{6}}{225}$
$\frac{4+\sqrt{6}}{10}$	$\frac{296+169\sqrt{6}}{1\,800}$	$\frac{88+7\sqrt{6}}{360}$	$\frac{-2-3\sqrt{6}}{225}$
1	$\frac{16-\sqrt{6}}{36}$	$\frac{16+\sqrt{6}}{36}$	$\frac{1}{9}$
	$\frac{16-\sqrt{6}}{36}$	$\frac{16+\sqrt{6}}{36}$	$\frac{1}{9}$

Таблица 9.7

0	0	0
1	1/2	1/2
	1/2	1/2

Таблица 9.8

0	0	0	0
1/2	5/24	1/3	−1/24
1	1/6	2/3	1/6
	1/6	2/3	1/6

Таблица 9.9

0	0	0	0	0
$5-\sqrt{5}$	$\frac{11+\sqrt{5}}{120}$	$\frac{25-\sqrt{5}}{120}$	$\frac{25-13\sqrt{5}}{120}$	$\frac{-1+\sqrt{5}}{120}$
$5+\sqrt{5}$	$\frac{11-\sqrt{5}}{120}$	$\frac{25+13\sqrt{5}}{120}$	$\frac{25+\sqrt{5}}{120}$	$\frac{-1-\sqrt{5}}{120}$
1	1/12	5/12	5/12	1/12
	1/12	5/12	5/12	1/12

функции устойчивости рассматривается модельное уравнение вида

$$\dot{v} = \lambda v, v(0) = v_0.$$

Запишем теперь метод Рунге-Кутты для решения приведенного выше уравнения в общей форме с использованием новых переменных Y:

$$v_{n+1} = v_n + \tau \sum \gamma_p f(t_n + \alpha_p \tau, Y_p) \equiv v_n + \tau\lambda \sum \gamma_p Y_p,$$

$$Y_p = v_n + \tau \sum_{k=1}^{s} \beta_{kp} f(t_n + \gamma_k \tau, Y_k) \equiv v_n + \tau\lambda \sum \beta_{kp} Y_k.$$

Приведенные выше формулы можно рассматривать как систему линейных уравнений относительно новых переменных $Y_1, \ldots, Y_k, v_{n+1}$ следующего вида:

$$\begin{pmatrix} 1-z\beta_{11} & -z\beta_{12} & \ldots & -z\beta_{1s} & 0 \\ -z\beta_{21} & 1-z\beta_{22} & \ldots & -z\beta_{2s} & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ -z\beta_{s1} & -z\beta_{s2} & \ldots & 1-z\beta_{ss} & 0 \\ -z\gamma_1 & -z\gamma_2 & \ldots & -z\gamma_s & 1 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \ldots \\ \ldots \\ Y_s \\ v_{n+1} \end{pmatrix} = \begin{pmatrix} v_n \\ v_n \\ \ldots \\ \ldots \\ v_n \\ v_n \end{pmatrix}.$$

Необходимо выразить v_{n+1} через v_n. Для этого воспользуемся правилом Крамера. Ответ можно записать в следующей форме:

$$x_{n+1} = \frac{\det(\mathbf{E} - z\mathbf{B} + z\mathbf{e}\mathbf{a}^{\mathrm{T}})}{\det(\mathbf{E} - z\mathbf{B})} x_n,$$

где $\mathbf{E}$ — единичная матрица размера $S \times S$, $\mathbf{B}$ — матрица коэффициентов β_{ij}, входящая в таблицу Бутчера, $\mathbf{e}$ — единичный вектор размерности S (столбец), $\mathbf{a}^{\mathrm{T}}$ — строка коэффициентов γ_l, входящая в таблицу Бутчера, S — число стадий метода Рунге-Кутты.

Условием устойчивости метода будет

$$\left| \frac{\det(\mathbf{E} - z\,\mathbf{B} + z\,\mathbf{e}\mathbf{a}^{\mathrm{T}})}{\det(\mathbf{E} - z\,\mathbf{B})} \right| < 1.$$

Одноитерационные методы Розенброка. Розенброком был предложен класс неявных методов, в котором не решается система нелинейных уравнений. В простейшем случае для автономной системы уравнений методы типа Розенброка могут иметь вид [3]

$$(\mathbf{E} - a\tau\mathbf{B} - b\tau^2\mathbf{B}^2)\,\frac{\mathbf{u}_{n+1} - \mathbf{u}_n}{\tau} = \mathbf{f}[\mathbf{u}_n + c\tau\mathbf{f}(\mathbf{u}_n)].$$

Здесь $\mathbf{B} = \frac{\partial \mathbf{f}}{\partial \mathbf{u}}(\mathbf{u}_n)$ — матрица, постоянная на данном шаге по времени. Параметры a, b, c подбираются таким образом, чтобы обеспечить максимально возможный порядок точности.

Например, для схемы третьего порядка точности получим

$$a = 1{,}077; \quad b = -0{,}372; \quad c = -0{,}577.$$

Такую схему иногда называют методом с одной итерацией, имея в виду, что вычисление обратной матрицы сравнимо по количеству ариф-

метических операций с одной итерацией метода Ньютона. Преимущество методов типа Розенброка перед прочими классами численных методов ЖС ОДУ заключается в том, что для определения решения на верхнем временном слое необходимо решать уже линейную систему алгебраических уравнений.

Рассмотрим связь между методами Розенброка и Рунге-Кутты.

Определение ([9]). S-стадийный метод Розенброка для решения автономной системы ЖС ОДУ имеет вид

$$\mathbf{k}_i = \tau \mathbf{f}\left(\mathbf{u}_n + \sum_{j=1}^{i-1} \beta_{i,j}\mathbf{k}_j\right) + \tau \mathbf{B} \sum_{j=1}^{i} \mu_{i,j}\mathbf{k}_j,$$

$$i = 1, \ldots, S,$$

$$\mathbf{u}_{n+1} = \mathbf{u}_n + \sum_{k=1}^{S} \gamma_k \mathbf{k}_k,$$

где β, γ, μ — управляющие коэффициенты метода. Как и выше, матрица Якоби правой части системы $\mathbf{B}$ вычисляется по данным в точке t_n. Связь последних формул с выражениями для явных методов Рунге-Кутты очевидна.

Несколько сложнее представляются методы Розенброка в случае неавтономной ЖС ОДУ [9]:

$$\mathbf{k}_i = \tau \mathbf{f}\left(t_n + \alpha_i \tau, \mathbf{u}_n + \sum_{j=1}^{i-1} \beta_{i,j}\mathbf{k}_j\right) + \tau \mathbf{B} \sum_{j=1}^{i} \mu_{i,j}\mathbf{k}_j + \mu_i \tau^2 \mathbf{B},$$

$$i = 1, \ldots, S,$$

$$\mathbf{u}_{n+1} = \mathbf{u}_n + \sum_{k=1}^{S} \gamma_k \mathbf{k}_k,$$

$$\alpha_i = \sum_j \beta_{i,j}, \mu_i = \sum_j \mu_{i,j}.$$

Вывод *условий порядка* методов Розенброка — достаточно громоздкий, и всем заинтересованным читателям можно рекомендовать разобраться в выкладках в книге [9]. Отметим, что широкое распространение получают в последнее время вложенные методы Розенброка высокого порядка аппроксимации, имеющие очень хорошие вычислительные качества. Возможно, что новые методы типа Розенброка способны будут вытеснить из вычислительной практики наиболее распространенные до этого в вычислительной практике методы Гира.

9.5. Формулы дифференцирования назад и методы Гира. Представление Нордсика

В предыдущей лекции вкратце был изложен альтернативный методам Рунге-Кутты подход к построению численных методов для решения ОДУ, основанный на расширении шаблона разностной схемы. При переходе от точки t_n к t_{n+1} использовались значения решения (или функций от него) в предыдущих точках. Полученные численные методы (первые схемы такого рода были получены Адамсом) носят название линейных многошаговых. Однако применение методов Адамса для решения ЖС ОДУ приводит к неутешительному результату. Отчасти он обоснован результатом, полученным Далквистом.

Теорема (Далквиста (или второй барьер Далквиста) [8, 9]). *Не существует A-устойчивых линейных многошаговых схем с порядком аппроксимации выше второго.*

Попробуем ввести другой класс линейных многошаговых методов — это так называемые формулы дифференцирования назад (или ФДН-методы) [8].

Общий вид ФДН-метода таков:

$$\mathbf{u}_{n+1} + \sum_{j=1}^{k} \alpha_j \mathbf{u}_{n+1-j} = \tau\beta\mathbf{f}_{n+1}, \tag{9.5}$$

где коэффициенты α_j выбираются из условий аппроксимации метода. В отличие от методов типа Рунге-Кутты, при использовании ФДН-методов нелинейная система алгебраических уравнений для определения u_{n+1} имеет меньшую размерность, следовательно, требуется меньшее число операций для нахождения решения.

Семейство $A(\alpha)$-устойчивых ФДН-методов с достаточно большим значением угла полураствора α носит общее название методов Гира. Коэффициенты методов Гира, представленных в виде (9.5), приведены в таблице 9.10 (см. также книгу [13]).

Величина k характеризует количество точек в шаблоне и порядок аппроксимации ФДН-метода. В первом столбце таблицы также приведен угол полураствора α для A(α)-устойчивых методов. ФДН-методы с порядком аппроксимации 7 и выше безусловно неустойчивы.

К очевидным недостаткам методов ФДН (как, впрочем, и других многошаговых) является необходимость разгонного участка и трудности при автоматическом выборе шага. Существует эквивалентная форма ФДН-методов, методы Нордсика, в которой эти недостатки преодолеваются сравнительно легко. В некоторых источниках они не выделяются

Таблица 9.10. Коэффициенты методов Гира

K	B	α_0	α_1	α_2	α_3	α_4	α_5
1	1	-1					
2	2/3	1/3	$-4/3$				
3(86)	6/11	$-2/11$	9/11	$-18/11$			
4(73,35)	12/25	$-3/25$	16/25	36/25	$-48/25$		
5(51,84)	$\frac{60}{137}$	$\frac{-12}{137}$	$\frac{75}{137}$	$\frac{-200}{137}$	$\frac{300}{137}$	$\frac{-300}{137}$	
6(17,84)	$\frac{600}{147}$	$\frac{10}{147}$	$\frac{-72}{147}$	$\frac{225}{147}$	$-\frac{400}{147}$	$\frac{450}{147}$	$\frac{-360}{147}$

в самостоятельный класс методов, а называются ФДН-методами в представлении Нордсика.

Рассмотрим в качестве примера построение метода Нордсика [2, 8] для модельной задачи

$$\dot{u} = f(t;\ u), u(0) = u_0 \tag{9.6}$$

и введем также в рассмотрение вектор Нордсика

$$\mathbf{z}_n = (u_n, \tau u'_n, \frac{\tau^2}{2} u''_n, \ldots, \frac{\tau^k}{k!} u_n^{(k)})^T,$$

в который, кроме искомого значения функции, включены приближенные значения первых k производных в узлах сетки. Для того чтобы осуществить переход к следующему значению z_{n+1} необходимо задать правила вычисления компонентов вектора z по данным в текущий момент времени. Вслед за [2, 8] ограничимся рассмотрением случая $k = 3$. Для этого разложим (9.6) в ряд Тейлора в окрестности t_n:

$$u_{n+1} = u_n + \tau u'_n + \frac{\tau^2}{2} u''_n + \frac{\tau^3}{3!} u_n^{(3)} + \frac{\tau^4}{4!} u^4(\xi), \tag{9.7}$$

для записи ряда использован остаточный член в форме Лагранжа. Кроме того, используем следующие следствия приведенного выше разложения:

$$\tau u'_{n+1} = \tau u'_n + 2\frac{\tau^2}{2} u''_n + 3\frac{\tau^3}{3!} u_n^{(3)} + 4\frac{\tau^4}{4!} u^{(4)}(\xi), \tag{9.8}$$

$$\frac{\tau^2}{2} u''_{n+1} = 2\frac{\tau^2}{2} u''_n + 3\frac{\tau^3}{3!} u_n^{(3)} + 6\frac{\tau^4}{4!} u^{(4)}(\xi), \tag{9.9}$$

$$\frac{\tau^3}{3!} u_{n+1}^{(3)} = 3\frac{\tau^3}{3!} u_n^{(3)} + 4\frac{\tau^4}{4!} u^{(4)}(\xi) \tag{9.10}$$

а конкретное значение ξ выберем из условий аппроксимации уравнения (9.6), т. е. $u'_{n+1} = f(t_{n+1}, u_{n+1})$.

Подставляя последнее равенство в (9.8), получим:

$$4\frac{\tau^4}{4!}u^{(4)}(\xi) = \tau(f_{n+1} - f_n^p), \tag{9.11}$$

а f_n^p выражается через компоненты вектора Нордсика:

$$f_n^p = u'_n + \tau^2 u''_n + \frac{\tau^3}{2}u_n^{(3)}.$$

Заманчивая идея подставить выражение (9.11) во все соотношения (9.7),(9.8),(9.9),(9.10) приводит к неустойчивому методу (см. [2, 8]). При этом (неустойчивый) метод имеет четвертый порядок аппроксимации. Основная идея метода Нордсика заключается в том, чтобы, несколько «испортив» порядок аппроксимации, добавляя в выражения (9.7), (9.8), (9.9), (9.10) представление (9.11), добиться максимальной устойчивости метода при приемлемом порядке аппроксимации.

Окончательный ответ для представления Нордсика можно записать в следующем кратком виде:

$$\mathbf{z}_{n+1} = \mathbf{P}\mathbf{z}_n + \mathbf{l}(\tau f_{n+1} - \mathbf{e}_1^T\mathbf{P}\mathbf{z}_n), \tag{9.12}$$

где матрица $\mathbf{P}$ — треугольная матрица Паскаля, определяемая соотношением (для произвольного числа компонентов вектора Нордсика k): $P_{ij} = C_i^j$, $(0 \leqslant i \leqslant j \leqslant k)$ здесь C_i^j — биномиальный коэффициент. В противном случае $P_{ij} = 0$. Вектор $\mathbf{e}_1$ определяется как $(0, 1, 0, 0, \ldots, 0)$, а вектор $\mathbf{l}$ — как $(l_0, 1, l_2, \ldots, l_k)$, значение 1 выбирается из условия нормировки. Относительно первого компонента вектора Норсдика u_{n+1} система уравнений нелинейна, все остальные переменные входят в систему линейным образом.

Для рассматриваемого случая $k = 3$ Норсдик получил набор коэффициентов $l = (3/8, 1, 3/4, 1/6)$ из условий минимума погрешности и обращения в нуль собственных чисел матрицы $\mathbf{M} = \mathbf{P} - \mathbf{l}\mathbf{e}_1^T\mathbf{P}$. Можно получать набор коэффициентов для метода Гира в представлении Норсдика, например, из условий жесткой устойчивости (например, ослабленного требования L-устойчивости, когда требуется лишь $|R(z)| \to 0$ при $\mathrm{Re}\ \tau\lambda \to -\infty$). В этом случае для рассмотренного выше примера $l = (6/11, 1, 6/11, 1/11)$.

Покажем, что последний из приведенных методов Нордсика эквивалентен методу ФДН Гира при $k = 3$. Для этого просто для трех последовательных шагов по времени исключаем из формул типа (9.7), (9.8), (9.9), (9.10) (точнее, из (9.12)) значения производных решения. После нетрудных преобразований получим формулу Гира. Другой рассмотренный вариант метода Нордсика приведет к одной из неявных формул Адамса. Подробнее в [8].

Отметим, что метод Гира в представлении Нордсика оказывается самостоятельно стартующим. При старте можно положить вектор Нордсика для данной задачи равным, например, $\mathbf{z}_0 = (u_0, 0, 0, \ldots, 0)$, что позволит начать вычисления, но приведет к уменьшению порядка аппроксимации. Таким образом, многозначный (по введенной в [2] терминологии) вариант метода Гира обладает переменным порядком аппроксимации: стартуя как метод первого порядка, по завершении разгонного участка метод стремится к максимально возможному для данной формулы порядку.

Отметим также, что для системы с релаксационными колебаниями лучшие результаты могут давать многозначные методы Гира не очень высокого порядка аппроксимации.

9.6. Задачи для самостоятельного решения

1. Модель Филда-Нойса **«орегонатор»**

   Простейшая математическая модель периодической химической реакции Белоусова-Жаботинского состоит из трех уравнений:

$$\dot{y}_1 = 77,27(y_2 + y_1(1 - 8,375 \cdot 10^{-6} y_1 - y_2)),$$

$$\dot{y}_2 = \frac{1}{77,27}(y_3 - (1 + y_1)y_2),$$

$$\dot{y}_3 = 0,161(y_1 - y_3).$$

   На то, что система жесткая, указывают большие различия в константах скоростей реакций — есть процессы быстрые и есть медленные.

   Так как переменные системы — концентрации ($HBrO_2$, Br^- и Ce(IV) соответственно) то начальные условия для системы следует выбирать положительными, как правило, близкими к 0. Конечное время интегрирования системы $T_k = 800$.

   О системе подробнее, например, в [8, 9, 17].

2. **Уравнение Ван-дер-Поля**

   Типичным примером жесткой задачи малой размерности является уравнение Ван-дер-Поля [8, 9, 18, 19]. Его возможно записать в виде системы

$$y_1' = y_2,$$

$$y_2' = -a(y_2(y_1^2 - 1) + y_1), \qquad (9.13)$$

или в виде

$$y_1' = -a(\frac{y_1^3}{3} - y_1) + ay_2,$$

$$y_2' = -y_1, \tag{9.14}$$

(представление Льенара). Считаем, что параметр a — большой. В расчетах рассмотреть два случая: $a = 103$ и $a = 106$. Для тестов обычно полагают $y_1 = 2$, $y_2 = 0$.

Конечное время интегрирования системы, записанной в виде (9.13), $T_k = 20$.

Периодические решения жестких систем ОДУ иногда называют релаксационными автоколебаниями [18, 19].

Дополнительный вопрос: указать преобразование, переводящее представление (9.13) в представление Льенара (9.14).

3. Система Ван-дер-Поля и **траектории-утки**

Рассмотрим неавтономную систему уравнений Ван-дер-Поля:

$$y_1' = a(-(\frac{y_1^3}{3} - y_1) + y_2),$$

$$y_2' = -y_1 + A\cos\omega t.$$

Как и в предыдущей задаче считаем, что $a = 103$ и $a = 106, y_1 = 2$, $y_2 = 0$. Рассмотреть численно случаи $0 < A < 1$ и $1 < A < \sqrt{1 + \frac{1}{64\omega^2}}$. $T_k = 200$.

О траекториях-утках в системе Ван-дер-Поля см. [19] (строгое математическое исследование) и [20](популярное изложение).

4. Суточные **колебания концентрации озона** в атмосфере

Рассмотрим простейшую математическую модель колебаний концентрации озона в атмосфере [2]. Она описывается следующей неавтономной системой ОДУ:

$$\dot{y_1} = -k_1y_1y_2 - k_2y_1y_3 + 2k_3(t)y_2 + k_4(t)y_3\ ,$$

$$\dot{y_2} = 0,$$

$$\dot{y_1} = k_1y_1y_2 - k_2y_1y_3 - k_4(t)y_3.$$

В данной модели уравнения описывают изменение концентрации атомарного кислорода, молекулярного кислорода и озона соответственно. Считается, что изменения концентрации молекулярного кислорода невелики. Начальные значения для задачи таковы:

$$y_1(0) = 10^6(\text{см}^{-3}),\ y_2(0) = 3,7 \cdot 10^{16}(\text{см}^{-3}),\ y_3(0) = 10^{12}(\text{см}^{-3}),$$

значения констант скоростей химических реакций

$$k_1 = 1,63 \cdot 10^{-16}, \quad k_2 = 4,66 \cdot 10^{-16}.$$

Две другие химические реакции зависят от локальной освещенности участка земной поверхности и приближаются следующим выражением:

$$k_i(t) = \begin{cases} \exp(-c_i/\sin\omega t), & \sin\omega t > 0, \\ 0, & \sin\omega t < 0, \end{cases}$$

где $\omega = \pi/43\,200c^{-1}, c_3 = 22,62, c_4 = 7,601$. Значения констант скоростей обращаются в нуль ночью, резко возрастают на рассвете, достигают максимума в полдень и падают до нуля на закате. Конечное время интегрирования $T_k = 172800$ с (двое суток).

Данная система является жесткой ночью и умеренно жесткой в светлое время суток.

5. **Уравнение Бонгоффера-Ван-дер-Поля**

Рассмотрим еще один пример жесткой задачи малой размерности, имеющей периодическое решение [19, 21].

$$y_1' = a(-(\frac{y_1^3}{3} - y_1) + y_2),$$

$$y_2' = -y_1 - by_2 + c.$$

Здесь $a = 103$ и $a = 106, y_1 = 2,\ y_2 = 0$.

Уравнение описывает протекание тока через клеточную мембрану. Постоянная компонента тока с в безразмерной записи системы такова, что $0 < c < 1, b > 0. T_k = 20$.

6. Сингулярно-возмущенная система — **модель** двухлампового генератора **Фрюгауфа**.

Система более высокой размерности, имеющая решение в виде релаксационного цикла, приведена в [18] (см. также [21]). Она имеет вид:

$$\varepsilon\dot{x}_1 = -\alpha(y_1 - y_2) + \varphi(x_1) - x_2,$$

$$\varepsilon \dot{x}_2 = \alpha(y_1 - y_2) + \varphi(x_2) - x_1,$$

$$\dot{y}_1 = x_1,$$

$$\dot{y}_2 = x_2.$$

Здесь $\alpha > 0$ — константа порядка единицы, функция $\varphi(u) = -\operatorname{tg}(\pi u/2), x_1(0) = x_2(0) = 0, y_1 = 2, y_2 = 0, T_k = 20, \varepsilon = 10^{-3}, 10^{-6}$.

7. Простейшая **модель гликолиза**

Простейшая модель гликолиза описывается уравнениями следующего вида [21]:

$$\dot{y}_1 = 1 - y_1 y_2,$$

$$\dot{y}_2 = \alpha y_2 \left(y_1 - \frac{1+\beta}{y_2+\beta} \right),$$

предложенными Дж. Хигинсом. В системе $\beta = 10, \alpha = 100, 200, 400, 1000$. Начальные условия для системы: $y_1(0) = 1, y_2(0) = 0,001, T_k = 50$. Решение этой системы — релаксационные автоколебания (жесткий предельный цикл).

8. Пример жесткой системы — **модель** химических реакций **Робертсона**

Один из первых и самых популярных примеров жесткой системы ОДУ принадлежит Робертсону (1966) и имеет вид, типичный для моделей химической кинетики — в правой части системы стоят полиномы второй степени от концентраций (сравните с орегонатором).

Система Робертсона имеет вид [9]

$$\dot{y}_1 = -0.04y_1 + 10^4 y_2 y_3,$$
$$\dot{y}_2 = 0.04y_1 - 10^4 y_2 y_3 - 3 \cdot 10^7 y_2^2,$$
$$\dot{y}_3 = 3 \cdot 10^7 y_2^2.$$

Начальные условия для системы таковы: $y_1(0) = 1$, $y_2(0) = 0$, $y_3(0) = 0$. Рассматриваются следующие величины отрезка интегрирования: $T_k = 40$ (в работе Робертсона рассматривался именно такой отрезок интегрирования), $T_k = 100, 1000, \ldots, 1011$. О свойствах задачи см. в [9].

9. **Модель дифференциации растительной ткани**

Данный пример из [9] — типичный случай биохимической модели «умеренной» размерности (современные модели, например, фотосинтеза включают сотни уравнений подобного типа). Хотя данная модель является *умеренно жесткой*, тем не менее, ее лучше решать с помощью методов, предназначенных для решения ЖС ОДУ.

$$\dot{y}_1 = -1.71y_1 + 0.43y_2 + 8.23y_3 + 0.0007,$$

$$\dot{y}_2 = 1.71y_1 - 8.75y_2,$$

$$\dot{y}_3 = -10.03y_3 + 0.43y_4 + 0.035y_5,$$

$$\dot{y}_4 = 8.32y_2 + 1.71y_3 - 1.12y_4,$$

$$\dot{y}_5 = -1.745y_5 + 0.43y_6 + 0.43y_7,$$

$$\dot{y}_6 = -280y_6y_8 + 0.69y_4 + 1.71y_5 - 0.43y_6 + 0.69y_7,$$

$$\dot{y}_7 = 280y_6y_8 - 1.87y_7,$$

$$\dot{y}_8 = -\dot{y}_7.$$

Начальные значения всех переменных системы равны 0, кроме $y_1(0) = 1$ и $y_8(0) = 0.0057$. Длина отрезка интегрирования $T_k = 421,8122$.

10. **Задача E5**

Еще одна модель химической реакции из [9], получившая свое название E5 в более ранних публикациях.

$$\dot{y}_1 = -Ay_1 - By_1y_3,$$

$$\dot{y}_2 = Ay_1 - MCy_2y_3,$$

$$\dot{y}_3 = Ay_1 - By_1y_3 - MCy_2y_3 + Cy_4,$$

$$\dot{y}_4 = By_1y_3 - Cy_4.$$

Начальные условия: $y_1(0) = 1,76 \cdot 10^{-3}$, а все остальные переменные равны 0. Значения коэффициентов модели следующие: $A = 7,89 \cdot 10^{-10}, B = 1,1 \cdot 10^7, C = 1,13 \cdot 10^3, M = 10^6$. Первоначально задача ставилась на отрезке $T_k = 1000$, но впоследствии было обнаружено, что она обладает нетривиальными свойствами вплоть до времени $T_k = 1013$ (подробнее см. [9]).

Обратить особое внимание, что в процессе расчетов приходится иметь дело с очень малыми концентрациями реагентов (малы значения y_2, y_3 и y_4). Как «подправить» постановку задачи E5?

11. **Уравнение Релея**

Уравнение Релея во многом похоже на уравнение Ван-дер-Поля [21]. Рассматривается задача вида

$$\ddot{x} - \mu(1 - \dot{x}^2)\dot{x} + x = 0.$$

Решить задачу, записав уравнение Релея в виде системы ОДУ. Начальные условия: $x(0) = 0, \dot{x}(0) = 0,001, \mu = 1000, T_k = 1000$.

12. **Экогенетическая модель**

Рассмотрим пример системы уравнений, которая описывает изменения численности популяций двух видов и эволюцию некого генетического признака α. Система ОДУ имеет вид

$$\dot{x} = x(1 - 0,5x - \frac{2}{7\alpha^2}y),$$

$$\dot{y} = y(2\alpha - 3,5\alpha^2 x - 0,5y),$$

$$\dot{\alpha} = \varepsilon(2 - 7\alpha x).$$

Параметры задачи таковы: $\varepsilon \leqslant 0,01, 0 \leqslant x_0 \leqslant 3, 0 \leqslant y_0 \leqslant 15, \alpha_0 = 0, T_k = 1500$. Наличие малого параметра в третьем уравнении системы показывает, что генетический признак меняется медленнее, чем численность популяций. Решение системы — релаксационные колебания.

Задача описана в статье [22].

13. **Экогенетическая модель**

Еще один пример жесткой системы описан в статье [22]. Более интересный случай — численность двух популяций зависит от взаимодействия между ними и двух медленно меняющихся генетических признаков.

$$\dot{x} = x(2\alpha_1 - 0,5x - \alpha_1^2\alpha_2^{-2}y),$$

$$\dot{y} = y(2\alpha_2 - \alpha_1^{-2}\alpha_2^2 x - 0,5y),$$

$$\dot{\alpha}_1 = \varepsilon(2 - 2\alpha_1\alpha_2^{-2}y),$$

$$\dot{\alpha}_2 = \varepsilon(2 - 2\alpha_1^{-2}\alpha_2 x).$$

Параметры задачи таковы: $\varepsilon \leqslant 0,01, 0 \leqslant x_0 \leqslant 40, 0 \leqslant y_0 \leqslant 40, \alpha_{10} = 0, \alpha_{20} = 10, T_k = 2000$.

Рассмотреть также модификацию предыдущей системы [22]:

$$\dot{x} = x(2\alpha_1 - 0,5x - \alpha_1^3\alpha_2^{-3}y),$$

$$\dot{y} = y(2\alpha_2 - \alpha_1^{-3}\alpha_2^3 x - 0,5y),$$

$$\dot{\alpha}_1 = \varepsilon(2 - 3\alpha_1^2\alpha_2^{-3}y),$$

$$\dot{\alpha}_2 = \varepsilon(2 - 3\alpha_1^{-3}\alpha_2^2 x).$$

Параметры задачи: $\varepsilon \leqslant 0,01, 0 \leqslant x_0 \leqslant 40, 0 \leqslant y_0 \leqslant 40, \alpha_{10} = 0, \alpha_{20} = 10, T_k = 2000$.

Литература

[1] *Уатт Дж. Холл, Дж.* (ред.) Современные численные методы решения обыкновенных дифференциальных уравнений. М.: Мир, 1979. 312 с.

[2] *Каханер Д., Моулер К., Нэш С.* Численные методы и программное обеспечение. М.: Мир, 1998. 575 с.

[3] *Федоренко Р.П.* Введение в вычислительную физику. М.: Изд-во МФТИ, 1994. 528 с.

[4] *Деккер К., Вервер Я.* Устойчивость методов Рунге-Кутты для жестких нелинейных дифференциальных уравнений. М.: Мир, 1988. 334 с.

[5] *Тихонов А.Н., Васильева А.Б., Свешников А.Г.* Дифференциальные уравнения. М.: Наука, 1980.

[6] *Мищенко Е.Ф., Розов Н.Х.* Дифференциальные уравнения с малым параметром и релаксационные колебания. М.: Наука, 1975. 248 с.

[7] *Мищенко Е.Ф., Колесов Ю.С., Колесов А.Ю., Розов Н.Х.* Периодические движения и бифуркационные процессы в сингулярно-возмущенных системах. М.: Наука. Физматлит, 1995. 336 с.

[8] *Хайрер Э., Нерсетт С., Ваннер Г.* Решение обыкновенных дифференциальных уравнений. Нежесткие задачи. М.: Мир, 1990. 512 с.

[9] *Хайрер Э., Ваннер Г.* Решение обыкновенных дифференциальных уравнений. Жесткие и дифференциально-алгебраические задачи. М.: Мир, 1999. 685 с.

[10] *Теоретические* основы и конструирование алгоритмов задач математической физики. / Под. ред. *К.И. Бабенко*. М.: Наука, 1979.

[11] *Гантмахер Ф.Р.* Теория матриц. 4 изд. М.: Наука, 1988. 552 с.

[12] *Оран Э., Борис Дж.* Численное моделирование реагирующих потоков. М.: Мир, 1990. 660 с.

[13] *Вычислительные* процессы и системы. Вып. 8 / Ред. Г.И. Марчук. М.: Наука, 1991. 380 с.

[14] *Ракитский Ю.В., Устинов С.М., Черноруцкий И.Г.* Численные методы решения жестких систем обыкновенных дифференциальных уравнений. М.: Наука, 1979. 160 с.

[15] *Лохов Г.М., Подзоров С.И., Щенников В.Вл.* Методы численного исследования жестких систем нелинейных обыкновенных дифференциальных уравнений. Уч. пособие. 2-е изд. М.: МФТИ, 1997. 140 с.

[16] *Лебедев В.И.* Функциональный анализ и вычислительная математика. М.: Физматлит, 2000. 296 с.

[17] *Колебания* и бегущие волны в химических системах: Пер. с англ. / Под ред. *Р. Филда, М. Бургер*. М.: Мир, 1988. 720 с.

[18] *Мищенко Е.Ф., Розов Н.Х.* Дифференциальные уравнения с малым параметром и релаксационные колебания. М.: Наука, 1975. 248 с.

[19] *Мищенко Е.Ф., Колесов Ю.С., Колесов А.Ю., Розов Н.Х.* Периодические движения и бифуркационные процессы в сингулярно-возмущенных системах. М.: Наука. Физматлит, 1995. 336 с.

[20] *Малинецкий Г.Г.* Хаос. Структуры. Вычислительный эксперимент. М.: Наука, 1998 (или Эдиториал УРСС, 2000.)

[21] *Ланда П.С.* Нелинейные колебания и волны. М.: Наука. Физматлит, 1997. 496 с.

[22] *Кондрашов А.С., Хибник А.И.* Экогенетические модели как быстро-медленные системы. / В кн.: Исследования по математической биологии. Пущино, 1996. с. 88-123.

Лекция 10. Численное решение краевых задач для систем обыкновенных дифференциальных уравнений

Рассматриваются численные методы решения краевых задач. На примере линейных краевых задач иллюстрируется применение различных вариантов метода прогонки — дифференциальной прогонки, разностной трехточечной прогонки, пятиточечной прогонки, матричной прогонки, периодической прогонки. Для нелинейных краевых задач рассмотрены методы стрельбы и квазилинеаризации. Дается представление о методах решения спектральных задач (задач на собственные значения). Обсуждается вопрос о применении метода Фурье при решении краевых задач для разностных уравнений, аппроксимирующих исходную дифференциальную задачу.

Ключевые слова: Метод стрельбы. Метод прогонки. Матричная прогонка. Дифференциальная прогонка. Жесткие краевые задачи. Метод квазилинеаризации. Задача Штурма-Лиувилля. Метод Фурье.

10.1. Краевая задача для линейной системы ОДУ первого порядка

Рассмотрим линейную систему ОДУ первого порядка

$$\frac{d\mathbf{u}}{dt} = \mathbf{A}\mathbf{u} + \mathbf{f}, \mathbf{u} \in R^n, t \in [0, L]$$

с краевыми условиями

$$\mathbf{R}\mathbf{u}(0) + \mathbf{S}\mathbf{u}(L) = q,$$

где $\mathbf{u}, \mathbf{f}, \mathbf{q}$ — n-мерные векторы, $\mathbf{A}(t), \mathbf{R}(t), \mathbf{S}(t)$ — матрицы размера $n \times n$.

Для приближенного решения задачи введем расчетную сетку $\{t_n\}_{n=0}^N$ и за приближенное решение примем сеточную функцию $\{u_n\}_{n=0}^N$. Рассмотрим методы построения приближенного решения.

Метод построения фундаментальных решений аналогичен известному по курсу дифференциальных уравнений способу построения общего решения системы линейных уравнений первого порядка. Решение представляется в виде

$$\mathbf{u}\,(t) = \bar{\mathbf{u}}(t) + \sum_{k=1}^{n} \alpha_k \mathbf{u}^k.$$

Здесь $u_k(t)$ есть полная фундаментальная система решений однородной задачи

$$\frac{d\mathbf{u}^k}{dt} = \mathbf{A}\mathbf{u}^k, k = 1, 2, \ldots, n$$

с начальными данными, например,

$$\mathbf{u}^k(0) = \{0, \ldots, 0, 1, 0, \ldots, 0\}^T,$$

где единица стоит на k месте, т. е. в качестве начальных данных используются векторы $\mathbf{u}^k(0) = \mathbf{e}_k$. Важно, чтобы решения однородной задачи составляли систему линейно независимых функций. Каждая такая функция ищется численно как решение соответствующей задачи Коши, используя методы, описанные в лекциях 8 и 9.

Пусть $\bar{\mathbf{u}}(t)$ — частное решение неоднородной системы

$$\frac{d\bar{\mathbf{u}}}{dt} = \mathbf{A}\bar{\mathbf{u}} + \mathbf{f}$$

с нулевыми начальными условиями $\bar{\mathbf{u}}(0) = 0$. Тогда неопределенные коэффициенты α_k находятся из краевых условий

$$\mathbf{R}\mathbf{u}(0) + \mathbf{S}\mathbf{u}(L) = \mathbf{q}, \text{ или}$$

$$\mathbf{R}[\sum_{k=1}^{n} \alpha_k \mathbf{u}^k(0)] + \mathbf{S}\bar{\mathbf{u}}(L) + \mathbf{S}[\sum_{k=1}^{n} \alpha_k \mathbf{u}^k(L)] = \mathbf{q}.$$

Последнее соотношение представляет собой СЛАУ относительно коэффициентов α_k размерности n.

Полную фундаментальную систему решений однородной задачи можно получить, используя, например, схему второго порядка точности (метод трапеций)

$$\frac{\mathbf{u}_{n+1} - \mathbf{u}_n}{\tau} = \mathbf{A}_{n+1/2}\frac{\mathbf{u}_n + \mathbf{u}_{n+1}}{2}, \mathbf{A}_{n+1/2} = \mathbf{A}\left(t_n + \frac{\tau}{2}\right),$$

$$\mathbf{u}_{n+1} = (\mathbf{E} - \frac{\tau}{2}\mathbf{A}_{n+1/2})^{-1}(\mathbf{E} + \frac{\tau}{2}\mathbf{A}_{n+1/2})\mathbf{u}_k,$$

$$\mathbf{u}_0 = \mathbf{e}_k.$$

Частное решение получаем аналогично:

$$\frac{\bar{\mathbf{u}}_{n+1} - \bar{\mathbf{u}}_n}{\tau} = \mathbf{A}_{n+1/2}\frac{\bar{\mathbf{u}}_{n+1} + \bar{\mathbf{u}}_n}{2} + \mathbf{f}_{n+1/2},$$

$$\mathbf{f}_{n+1/2} = \mathbf{f}\left(t_n + \frac{\tau}{2}\right),$$

$$\bar{\mathbf{u}}_0(0) = 0, \text{или}$$

$$\bar{\mathbf{u}}_{n+1} = \left(\mathbf{E} - \frac{\tau}{2}\mathbf{A}\right)^{-1}\left[\left(\mathbf{E} + \frac{\tau}{2}\mathbf{A}\right)\mathbf{u}_n + \tau \mathbf{f}_{n+1/2}\right], \bar{\mathbf{u}}_0 = 0.$$

Приведем пример, когда решение задачи методом построения фундаментальных решений не проходит. Рассмотрим систему уравнений

$$\frac{du}{dt} = av + f, \frac{dv}{dt} = bu + g,$$

$$u(0) = U_0, v(1) = 0,$$

Подобные системы ОДУ моделируют, например, процессы прохождения излучения или потоков элементарных частиц (гамма-излучение, потоки нейтронов) через разные среды (атмосфера, защита ядерных реакторов) в приближении оптически толстого слоя. Коэффициенты $a, b \sim 50$ характерны для защиты реакторов. Найдем общее решение этой задачи с помощью метода фундаментальных систем в виде линейной комбинации двух решений однородных ОДУ:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \bar{u} \\ \bar{v} \end{pmatrix} + \alpha_1 \begin{pmatrix} u^1 \\ v^1 \end{pmatrix} + \alpha_2 \begin{pmatrix} u^2 \\ v^2 \end{pmatrix},$$

где (u_1, v_1) и (u_2, v_2) — решения двух однородных систем (в дальнейшем для простоты будем полагать, что $\bar{u} = \bar{v} = 0$.) Тогда

$$\dot{u}^1 = av^1, \quad \dot{v}^1 = bu^1, \quad u^1(0) = 1, \quad v^1(0) = 0;$$

$$\dot{u}^2 = av^2, \quad \dot{v}^2 = bu^2, \quad u^2(0) = 0, \quad v^2(0) = 1,$$

а коэффициенты α_1 и α_2 находятся из краевых условий.

Общее решение такой системы, как известно, представляет собой сумму двух экспонент

$$\begin{pmatrix} u^1 \\ v^1 \end{pmatrix} = C_1 \begin{pmatrix} a_1 \\ b_1 \end{pmatrix} e^{\lambda_1 t} + C_2 \begin{pmatrix} a_2 \\ b_2 \end{pmatrix} e^{\lambda_2 t}$$

(аналогичным образом решение представляется и для u_2, v_2), где $a_i, b_i, i = 1, 2$ находятся из решения задачи, $C_i; i = 1, 2$ — произвольные постоянные; λ_1 и λ_2 — корни характеристического уравнения $\begin{vmatrix} -\lambda & a \\ b & -\lambda \end{vmatrix} = 0$. Решая характеристическое уравнение, получаем $\lambda_{1,2} = \pm\sqrt{ab} \approx \pm 50$. Полученное решение есть сумма двух экспонент: одной, быстро растущей ($\sim e^{50t}$), и второй, быстро убывающей ($\sim e^{-50t}$). Искомое же решение есть функция, близкая к $U_0 e^{-50t}$ (в случае задачи с защитой ректора U_0 —

падающий поток нейтронов; задача защиты — значительное его ослабление, приблизительно в e^{50} раз).

Слагаемые с быстрорастущими экспонентами должны взаимно уничтожиться. Получение же численного решения - весьма трудная задача, поскольку численное решение имеет большую и быстро возрастающую погрешность. Пусть $u^M = u(1+\varepsilon) \sim e^{50}t(1+10^{-12})$, т. е. начальная погрешность имеет порядок 10^{-12}. Она возрастает при вычислениях примерно в e^{50t} раз — даже при умеренных t это очень большое число.

10.2. Метод дифференциальной прогонки. Понятие о жестких краевых задачах

Рассмотрим метод дифференциальной прогонки, не приводя доказательства его устойчивости. Покажем, что с помощью этого метода можно решить рассмотренную выше задачу. Будем искать решение в виде

$$u(t) = \alpha(t)v(t) + \beta(t),$$

где α и β — пока неизвестные функции (прогоночные коэффициенты), для которых необходимо получить дифференциальные уравнения.

Продифференцируем это соотношение

$$\dot{u} = \dot{\alpha}v + \alpha\dot{v} + \dot{\beta}$$

и подставим в него уравнения системы $\dot{u} = av + f, \dot{v} = bu + g$. В результате получим, что $av + f = \dot{\alpha}v + \alpha(bu + g) + \dot{\beta}$. Подставим в полученное соотношение уравнение $u = \alpha v + \beta$. Тогда $av + f = \dot{\alpha}v + b\alpha^2 v + b\beta\alpha + \alpha g + \dot{\beta}v + \beta$.

После приведения подобных членов имеем равенство

$$v(\dot{\alpha} + b\alpha^2 - a) + (\dot{\beta} + b\beta\alpha + \alpha g - f) = 0.$$

Приравнивая к нулю коэффициенты при v и единице, получим два дифференциальных уравнения для прогоночных коэффициентов:

$$\dot{\alpha} + b\alpha^2 - a = 0,$$

$$\dot{\beta} + \alpha\beta b + \alpha g - f = 0.$$

Дополним их начальными условиями. Левое краевое условие вида $u(0) = U_0$ запишем в виде прогоночного соотношения $u(0) = \alpha(0)v(0) + \beta(0)$, полагая $\alpha(0) = 0, \beta(0) = U_0$. Таким образом, получаем начальные данные для двух задач Коши для $\alpha(t)$ и $\beta(t)$, которые могут быть решены численно.

Теперь разрешим правое краевое условие. На правой границе отрезка интегрирования имеем условие $v(1) = 0$ и прогоночное соотношение при $t = 1$: $u(1) = \alpha(1)v(1) + \beta(1)$, откуда получаем $u(1) = \beta(1)$.

Далее воспользуемся уравнением $\dot{v} = bu + g$, подставив в него прогоночное соотношение $u = \alpha v + \beta$, получим дифференциальное уравнение для v:

$$\dot{v} = \alpha b v + b\beta + g, v(1) = 0.$$

Интегрируя эту задачу справа налево, попутно определяем $u(t)$:

$$u(t) = \alpha(t)v(t) + \beta(t).$$

Метод дифференциальной прогонки оказывается весьма эффективным при решении линейных систем дифференциальных уравнений с переменными коэффициентами. Очевидно, что все приведенные выше соотношения верны без изменения и для таких систем.

Рассмотрим теперь общую постановку жесткой краевой задачи, неявно содержащую большой параметр. О жестких системах и задаче Коши для них речь шла в предыдущей лекции. Рассматривается линейная система

$$\frac{d\mathbf{u}}{dt} = \mathbf{A}\mathbf{u} + \mathbf{f}, t \in [a, b],$$

$$(\mathbf{d}_i, \mathbf{u}(a)) = \sum_{s=1}^{N} d_{is} u_s(a) = q_i, i = 1, \ldots, k, \quad k < N,$$

$$(\mathbf{d}_i, \mathbf{u}(b)) = \sum_{s=1}^{N} d_{is} u_s(b) = q_i, i = k+1, \ldots, N,$$

где $\mathbf{u}, \mathbf{q}_i, \mathbf{d}_i, \mathbf{f} \in R^n$, $\|\mathbf{d}_i\| \simeq O(1), i = 1, \ldots, N$.

Здесь $\mathbf{A}$ — постоянная матрица размером $N \times N$ (дальнейшие рассуждения будут справедливы и для систем уравнений с переменными коэффициентами). Определители систем линейных алгебраических уравнений, которыми являются краевые условия на обоих концах интервала интегрирования, полагаются отличными от нуля.

Определение. Рассматриваемая краевая задача для ОДУ является жесткой, если спектр собственных значений матрицы $\mathbf{A}$ можно разделить на три части.

1. Левый жесткий спектр, для которого $\mathrm{Re}\Lambda_i^1 \leqslant -\Lambda_0$, $|\mathrm{Im}\Lambda_i^1| < \Lambda_0, \Lambda_0 \gg 1, i = 1, \ldots, N_1$.
2. Правый жесткий спектр, для которого $\mathrm{Re}\Lambda_i^2 \geqslant \Lambda_0, |\mathrm{Im}\Lambda_i^2| < \Lambda_0, i = N_1 + 1, \ldots, N_2$.
3. Мягкий спектр $|\lambda_i| \leqslant \lambda_0, i = N_2 + 1, \ldots, N$.

Отношение $\Lambda_0/\lambda_0 \gg 1$ является параметром, характеризующим жесткость системы. В дальнейшем будем полагать $\Lambda_0(b-a) \gg 1$, $\lambda_0(b-a) \simeq O(1)$.

Общее решение такой системы имеет вид

$$\mathbf{u}(t) = \sum_{i=1}^{N_1} \gamma_i^1 e^{\Lambda_i^1 t} \mathbf{\Omega}_i^1 + \sum_{i=N_1+1}^{N_2} \gamma_i^2 e^{\Lambda_i^2 t} \mathbf{\Omega}_i^2 + \sum_{i=N_2+1}^{N} \gamma_i^3 e^{\lambda_i t} \boldsymbol{\omega}_i,$$

где $\mathbf{\Omega}_i^1, \mathbf{\Omega}_i^2, \boldsymbol{\omega}_i$ есть собственные векторы матрицы $\mathbf{A}$, соответствующие трем частям спектра. Понятна качественная структура этого решения, содержащего как левый, так и правый пограничные слои. Будем полагать, что количество собственных значений в каждой из трех частей спектра не изменяется. Особенность жестких краевых задач состоит в том, что их решениями являются ограниченные функции. Для них верно $\|\mathbf{u}\| \leqslant C(\|\mathbf{f}\| + \|\mathbf{q}\|)$, $\|\mathbf{f}\|$ и $\|\mathbf{q}\|$ — нормы правых частей в системе ОДУ и краевых условиях, соответственно. Для численного решения задачи можно использовать такую же схему второго порядка, как и ранее. Рассматриваем класс вычислительно корректных задач, для которых $C = O(1) \ll \exp(\Lambda_0(b-a))$.

В дальнейшем будем полагать величину $||\mathbf{A}||(b-a) \approx \Lambda_0(b-a)$ большой, а величину $\exp(||\mathbf{A}||(b-a)) \approx \exp(\Lambda_0(b-a))$ — очень большой. В приложениях такие задачи встречаются наиболее часто. Важно отметить и то, что не все возможные постановки задач для жесткой системы приводят к вычислительно корректным алгоритмам. Показывается, что необходимыми (и почти достаточными) условиями корректности являются следующие неравенства: $k \geqslant N_1, (N-k) \geqslant N_2 - N_1$, т. е. число краевых условий на левом конце отрезка интегрирования не должно быть меньше быстро убывающих вправо решений, на правом конце — не меньше быстро убывающих влево решений. В противном случае краевая задача оказывается вычислительно некорректной, так как $C = O(\exp(\Lambda_0(b-a)))$.

Проблемы, которые возникают при численном решении жесткой краевой задачи, были уже рассмотрены: суммирование функций порядка e^{Lt}, как известно, приводит к потере точности и накоплению вычислительных ошибок. Вторая проблема состоит в следующем. Для вычисления коэффициентов α_i, входящих в общее решение неоднородной системы (а оно состоит из суммы частного решения неоднородной системы и общего решения однородной $\mathbf{u}(t) = \bar{\mathbf{u}}_0 + \sum_{i=1}^{N} \alpha_i \mathbf{u}_i$), приходится решать плохо обусловленную СЛАУ $(\mathbf{d}_i, \bar{\mathbf{u}}(b) + \sum_i \alpha_i \mathbf{u}_i(b)) = \mathbf{q}_i, i = k+1, \ldots, N$.

10.3. Краевая разностная задача Штурма-Лиувилля для обыкновенного дифференциального уравнения второго порядка

Задача Штурма-Лиувилля для обыкновенного дифференциального уравнения второго порядка часто встречается в приложениях. Рассмотрим линейную задачу

$$\frac{d}{dt}\left[g\frac{du}{dt}\right] + h\frac{du}{dt} + su = f, t \in [a,b],$$

$$A\frac{du}{dt} + Bu = D, t = a, \tag{10.1}$$

$$A'\frac{du}{dt} + B'u = D', t = b,$$

где коэффициенты g, h, s, вообще говоря, являются функциями независимого переменного t.

Для разностной аппроксимации рассматриваемой краевой задачи введем равномерную разностную сетку $\{t_n\}_0^N$, $t_n = n\tau$, $\tau = (b-a)/N$ и определим на этой сетке сеточную функцию $\{u_n\}_0^N$.

Коэффициент g, вообще говоря, может не иметь первой производной. Такая задача может возникнуть, например, в случае расчета установившегося распределения температуры в задаче стационарной теплопроводности с контактным разрывом. Представим разностную задачу в виде

$$\frac{1}{\tau}\left(g_{n+1/2}\frac{u_{n+1}-u_n}{\tau} - g_{n-1/2}\frac{u_n - u_{n-1}}{\tau}\right) + h_n\frac{u_{n+1}-u_{n-1}}{2\tau} + s_n u_n =$$

$$= f_n, n = 1, \ldots, N-1,$$

$$g_{n+1/2} = g(t_n + \frac{\tau}{2}), h_n = h(t_n),$$

$$A\frac{u_1 - u_0}{\tau} + Bu_0 = D, \quad t = a,$$

$$A'\frac{u_N - u_{N-1}}{\tau} + B'u_N = D', \quad t = b.$$

Контактный разрыв при этом помещается в узел с номером n. В этом случае фактически аппроксимируется тепловой поток через границы ячейки разностной сетки, для уравнения теплопроводности получается консервативная разностная схема — подробнее смотри в лекциях, посвященных численному решению уравнений в частных производных.

Выписанные выше соотношения определяют простейшую *разностную схему*. Под разностной схемой здесь и ниже понимается совокупность разностных уравнений для определения значений сеточной функции внутри расчетной области, дополненная соответствующими начальными и граничными условиями для этой сеточной функции.

Для определения значений сеточной функции получается СЛАУ с трехдиагональной матрицей

$$-b_0 u_0 + c_0 u_1 = d_0,$$

$$a_n u_{n-1} - b_n u_n + c_n u_{n+1} = d_n, n = 1, \ldots, N-1,$$

$$a_N u_{N-1} - b_N u_N = d_N,$$

где $a_n = \frac{g_{n-1/2}}{\tau^2} - \frac{h_n}{2\tau}, c_n = \frac{g_{n+1/2}}{\tau^2} + \frac{h_n}{2\tau}, b_n = a_n + c_n - h_n, d_n = f_n, b_0 = \frac{A}{\tau} - \frac{B}{2}, c_0 = \frac{A}{\tau} + \frac{B}{2}, d_0 = D.$

Эта СЛАУ представима в каноническом виде $\mathbf{A}'\mathbf{u} = \mathbf{d}$, где $\mathbf{A}'$ — матрица

$$\mathbf{A}' = \begin{pmatrix} -b_0 & c_0 & 0 & \ldots & 0 \\ a_1 & -b_1 & c_1 & \ldots & 0 \\ 0 & a_2 & -b_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & -b_N \end{pmatrix},$$

$\mathbf{u}, \mathbf{d}$ есть векторы-столбцы

$$\mathbf{d} = (d_0, d_1, \ldots, d_N)^{\mathrm{T}}.$$

Трехдиагональные матрицы часто возникают при численном решении краевых задач как для обыкновенных дифференциальных уравнений, так и для уравнений в частных производных. Ранее матрица подобной структуры встречалась при построении сплайна Шонберга (лекция 6). Характерная особенность таких матриц заключается в том, что при большой размерности матрица имеет *ленточную структуру* — все элементы вне ленты (главная диагональ матрицы и по одной диагонали над и под ней) нулевые. В общем случае численное решение СЛАУ N-го порядка требует $O(N^3)$ арифметических действий и $O(N^2)$ ячеек памяти. В численных методах большую роль играют *экономичные алгоритмы*, в которых количество арифметических операций пропорционально количеству неизвестных — $O(N)$.

К таким алгоритмам относится метод *трехточечной разностной прогонки*, появление которого в России связано с именем И. М. Гельфанда,

в англоязычной литературе такая прогонка называется алгоритмом Томаса. Экономия памяти очевидна — необходимо хранить только три диагонали исходной матрицы (три одномерных массива). Рассмотрим экономичный вариант метода Гаусса, предназначенный для решения подобных систем.

Решение ищется в виде прогоночного соотношения

$$u_{n-1} = p_n u_n + q_n, n = 1, \ldots, N,$$

где p_n и q_n — прогоночные коэффициенты, подлежащие определению.

Левое краевое условие также записывается в виде прогоночного соотношения

$$u_0 = \frac{c_0}{d_0} u_1 - \frac{d_0}{b_0},$$

где $p_1 = c_0/b_0, \quad q_1 = -d_0/b_0$ (заметим, что если $A > 0$ и $B < 0$, то $b_0 > c_0$, $0 < p < 1$).

Получим рекуррентные формулы, позволяющие последовательно вычислить p_2, q_2, p_3, q_3 и т. д. вплоть до p_N, q_N.

Подставив равенство $u_{n-1} = p_n u_n + q_n$ в уравнение $a_n u_{n-1} - b_n u_n + + c_n u_{n+1} = d_n$, получим $a_n(p_n u_n + q_n) - b_n u_n + c_n u_{n+1} = d_n$, или

$$u_n = \frac{c_n}{b_n - a_n p_n} u_{n+1} + \frac{a_n q_n - d_n}{b_n - a_n p_n}.$$

Сравнивая эту запись со стандартным видом прогоночного соотношения

$$u_n = p_{n+1} u_{n+1} + q_{n+1},$$

видим, что для прогоночных коэффициентов должны выполняться равенства

$$p_{n+1} = \frac{c_n}{b_n - a_n p_n}, \quad q_{n+1} = \frac{a_n q_n - d_n}{b_n - a_n p_n}.$$

Эти формулы определяют прямой ход прогонки.

Из краевого условия на правом конце отрезка интегрирования $a_N u_{N-1} - b_N u_N = d_N$ и прогоночного соотношения $u_{N-1} = p_N u_N + q_N$ находим величину u_N.

Далее последовательно вычисляются остальные неизвестные $u_n n = = N-1, \ldots, 1 u_{n-1} = p_n u_n + q_n$. Это — обратный ход алгоритма прогонки.

Теперь, после того как описан прогоночный алгоритм, исследуем его на устойчивость.

Для этого рассмотрим вычисление прогоночного коэффициента p_n (т. е. этап прямого хода прогонки). В идеальной арифметике этот коэффициент равен $p_{n+1} = \frac{c_n}{b_n - a_n p_n}$, в конечноразрядной арифметике —

$p_{n+1}^M = p_{n+1} + \Delta_{n+1}$, где Δ_{n+1} — погрешность, связанная с округлениями на всех предшествующих этапах вычислений. Полагая Δ_{n+1} малой, исследуем изменение этой погрешности с ростом n. Для этого запишем соотношение между Δ_n и Δ_{n+1} в виде

$$p_{n+1} + \Delta_{n+1} = \frac{c_n}{b_n - a_n(p_n + \Delta_n)} + \varepsilon_n,$$

где ε_n — погрешность вычислений правой части и машинного представления коэффициентов a_n, b_n, c_n. Следовательно, Δ_{n+1} складывается из двух составляющих — локальной погрешности ε_n и наследственной Δ_{n+1}.

Полагая $\Delta_n \ll p_n$ при $\Delta_n > 0, p_n > 0$ и опуская члены порядка $O(\Delta^2)$, получим оценку

$$p_{n+1} + \Delta_{n+1} \approx \frac{c_n}{b_n - a_n p_n} + \frac{c_n a_n}{(b_n - a_n p_n)^2}\Delta_n + \varepsilon_n.$$

Отсюда, учитывая, что $p_{n+1} = \frac{c_n}{b_n - a_n p_n}$, получим требуемую оценку для эволюции погрешности

$$\Delta_{n+1} = \frac{a_n}{c_n} p_{n+1}^2 \Delta_n + \varepsilon_n, n = 0, \ldots, N-1. \tag{10.2}$$

Докажем следующую теорему.

Теорема. *Если выполнены условия диагонального преобладания* $|b_n| \geqslant |a_n| + |c_n|$ *и хотя бы для одной строки матрицы системы имеет место строгое диагональное преобладание* $(|b_n| > |a_n| + |c_n|)$. *Пусть, кроме того,* $0 < p_1 < 1$. *Тогда алгоритм прогонки устойчив.*

Доказательство.

Докажем утверждение теоремы для случая, когда во всех строках матрицы выполнено условие строгого диагонального преобладания. На случай нестрогих неравенств обобщение доказательства очевидно.

1. Пусть выполнено условие диагонального преобладания и $0 < p_1 < 1$. Для определенности положим $a_n > 0, b_n > 0, c_n > 0$.

   Тогда $p_{n+1} = \frac{c_n}{b_n - a_n p_n} > \frac{c_n}{b_n - a_n} > 0$.

   Кроме того, $p_{n+1} = \frac{c_n}{b_n - a_n p_n} < \frac{c_n}{a_n + c_n - a_n p_n} = \frac{c_n}{c_n + a_n(1 - p_n)} \leqslant 1$, откуда следует $0 < p_{n+1} < 1$.

2. Покажем, что $\frac{a_n}{c_n} \approx 1 + c\tau, \Delta_1 \leqslant (1 + c\tau)\Delta_0 + \varepsilon$.

   Для этого вспомним выражения для коэффициентов линейной системы, полученной при разностной аппроксимации исходного уравнения второго порядка.

$$a_n = \frac{g_{n-1/2}}{\tau^2} - \frac{h_n}{2\tau} \approx \frac{g_{n-1/2}}{\tau^2}, g_{n-1/2} = g(t_n - \tau/2),$$

$$c_n = \frac{g_{n+1/2}}{\tau^2} + \frac{h_n}{2\tau} \approx \frac{g_{n+1/2}}{\tau^2},$$

$$\frac{a_n}{c_n} \approx \frac{g(t-\tau/2)}{s(t+\tau/2)} = 1 + c\tau.$$

Вернемся к выражению для эволюции погрешности (10.2). С учетом полученных оценок имеем $\Delta_{n+1} \leqslant \frac{a_n}{c_n}\Delta_n + \varepsilon_n$. Так как $\frac{a_n}{c_n} \approx 1 + c\tau$, то $\Delta_{n+1} \leqslant (1 + c\tau)\Delta_n + \varepsilon_n$, где c — константа Липшица для функции $g(t)$. Считаем, что на каждом шаге ошибка округления не превосходит предельного значения, т. е. $0 \leqslant \varepsilon_n \leqslant \varepsilon$.

Из цепочки неравенств

$$\Delta_2 \leqslant (1 + c\tau)^2\Delta_0 + \varepsilon\left[1 + (1 + c\tau)\right],$$

$$\Delta_3 \leqslant (1 + c\tau)^3\Delta_0 + \varepsilon\left[1 + (1 + c\tau) + (1 + c\tau)^2\right],$$

будет следовать оценка $\Delta_n \leqslant (1 + c\tau)^n\Delta_0 + \frac{(1+c\tau)^n}{c\tau}\varepsilon$. Используя известные из математического анализа неравенства, последнюю оценку можно записать в виде

$$\Delta_n \leqslant e^{c\tau n}\Delta_0 + e^{c\tau n}\frac{\varepsilon}{c\tau}, c\tau n = ct.$$

При $ct \sim O(1)$ погрешности не накапливаются (для краевых задач это реальное условие), поскольку в расчетах $\tau \gg \varepsilon$, где ε — машинный эпсилон (подробнее в лекции 1).

Аналогичное утверждение доказывается и для второго прогоночного коэффициента $q_{n+1} = \frac{a_n q_n - d_n}{b_n - a_n p_n}$; алгоритм устойчив при выполнении тех же условий.

Покажем устойчивость обратного хода прогонки.

При обратном ходе вычисления проводятся по формулам $u_n = p_{n+1}u_{n+1} + q_{n+1}$, откуда, учитывая, что $u_n^M = u_n + \Delta_n$ и $u_{n+1}^M = u_{n+1} + \Delta_{n+1}$, получим $\Delta_n = p_{n+1}\Delta_{n+1} + \varepsilon_n$, где Δ_n — наследственная погрешность, ε_n — погрешность округления на n шаге. Очевидно, что обратный ход прогонки устойчив при выполнении условия $0 < p_n < 1$, или $0 < p_1 < 1$. ■

10.4. Пятиточечная прогонка

При численном решении краевых задач для обыкновенных дифференциальных уравнений 4-го порядка возникают СЛАУ с пятидиагональной матрицей вида:

$$c_0u_0 - d_0u_1 + e_0u_2 = f_0,$$

$$-b_1u_0 + c_1u_1 - d_1u_2 + e_1u_3 = f_1,$$

$$a_n u_{n-2} - b_n u_{n-1} + c_n u_n - d_n u_{n+1} + e_n u_{n+2} = f_n, n = 2, \ldots, N-2,$$

$$a_{N-1} u_{N-3} - b_{N-1} u_{N-2} + c_{N-1} u_{N-1} - d_{N-1} u_N = f_{N-1},$$

$$a_N u_{N-2} - b_N u_{N-1} + c_N u_N = f_N.$$

Алгоритм решения таких систем — пятиточечная прогонка, формулы которой выводятся аналогично формулам для трехточечной прогонки. Приведем их окончательный вид. В прогоночном соотношении появятся три коэффициента (p, q, r):

$$u_n = p_{n+1} u_{n+1} - q_{n+1} u_{n+2} + r_{n+1},$$

$$u_{N-1} = p_N u_N + r_N,$$

$$u_N = r_{N+1}.$$

Прогоночные коэффициенты находятся по формулам

$$p_{n+1} = [d_n + q_n(a_n p_{n-1} - b_n)] / D_n, n = 2, 3, \ldots, N-1,$$

$$p_1 = d_0/c_0, p_2 = (d_1 - q_1 b_1)/D_1,$$

$$r_{n+1} = [f_n - a_n r_{n-1} - r_n(a_n p_{n-1} - b_n)] / D_n, n = 2, 3, \ldots, N,$$

$$r_1 = f_0/c_0, r_2 = (f_1 + b_1 r_1)/D_1,$$

$$q_{n+1} = e_n/D_n, n = 1, 2, 3, \ldots, N-2, q_1 = e_0/c_0,$$

$$D_n = c_n - a_n q_{n-1} + p_n(a_n p_{n-1} - b_n), n = 2, 3, \ldots, N,$$

$$\Delta_1 = c_1 - b_1 p_1.$$

Достаточными условиями устойчивости пятиточечной прогонки являются диагональное преобразование и неравенства

$$|p_n| + |q_n| \leqslant 1, n = 1, \ldots, N-1, |p_N| \leqslant 1.$$

10.5. Матричная прогонка

Представим прогоночные соотношения для системы уравнений второго порядка

$$-\mathbf{B}_0 \mathbf{U}_0 - \mathbf{C}_0 \mathbf{U}_1 = \mathbf{D}_0,$$

$$\mathbf{A}_n \mathbf{U}_{n-1} - \mathbf{B}_n \mathbf{U}_n + \mathbf{C}_n \mathbf{U}_{n+1} = \mathbf{D}_n, n = 1, 2, 3, \ldots, N-1,$$

$$\mathbf{A}_N \mathbf{U} - \mathbf{B}_N \mathbf{U}_N = \mathbf{D}_N,$$

где $\mathbf{U}_n$ и $\mathbf{D}_n$ — векторы, $\mathbf{C}_n, \mathbf{A}_n, \mathbf{B}_n$ — матрицы.

Обратный ход прогонки выполняется в соответствии с формулами

$$\mathbf{U}_n = \mathbf{p}_{n+1}\mathbf{U}_{n+1} + \mathbf{q}_{n+1},$$

$$\mathbf{U}_N = \mathbf{q}_{N+1}.$$

Здесь $\mathbf{p}_{n+1}$ — матрица, $\mathbf{q}_{n+1}$ — вектор.

Окончательный вид формул для прямого хода матричной прогонки будет

$$\mathbf{p}_{n+1} = (\mathbf{B}_n - \mathbf{A}_n\mathbf{p}_n)^{-1}\mathbf{C}_n, n = 1, 2, 3, \ldots, N-1,$$

$$\mathbf{p}_1 = \mathbf{B}_0^{-1}\mathbf{C}_0;$$

$$\mathbf{q}_n = (\mathbf{B}_n - \mathbf{A}_n\mathbf{p}_n)^{-1}(\mathbf{A}_n\mathbf{q}_n - \mathbf{D}_n), n = 1, 2, 3, \ldots, N,$$

$$\mathbf{q}_1 = \mathbf{B}_0^{-1}\mathbf{D}_0.$$

Этот алгоритм называется методом матричной прогонки. Можно показать, что алгоритм матричной прогонки устойчив, если $\|\mathbf{p}_n\| < 1$ для $1 \leqslant n \leqslant N$; матрицы $\mathbf{B}_0$ и $(\mathbf{B}_n - \mathbf{A}_n\mathbf{p}_n)$ невырождены.

10.6. Численное решение нелинейных краевых задач

10.6.1. Метод стрельбы

Рассмотрим систему нелинейных ОДУ:

$$\begin{cases} \frac{d\mathbf{u}}{dt} = \mathbf{f}(\mathbf{u}); \quad t \in [0,\ L], \\ \mathbf{F}\,[\mathbf{u}(0),\ \mathbf{u}(L)] = 0, \quad \mathbf{u},\ \mathbf{f}, \mathbf{F} \in R^n. \end{cases}$$

Введем пока неизвестный вектор α размерности n такой, что

$$\begin{cases} \frac{d\mathbf{u}}{dt} = \mathbf{f}(\mathbf{u}), \\ \mathbf{u}(0) = \alpha \end{cases}$$

и решим соответствующую задачу Коши, например, методом Рунге–Кутты; получим решение $\mathbf{u}(t, \alpha)$. Используя краевые условия, получаем нелинейную систему алгебраических уравнений

$$\mathbf{F}(\alpha,\ \mathbf{u}\ (L, \alpha)) = 0, \quad \text{или} \quad \mathbf{F}(\alpha) = 0,$$

которая решается численно методом Ньютона или простых итераций. Отметим, что левая часть данной системы при использовании метода стрельбы задается не в виде функции, а алгоритмически — процедурой

вычисления значений функции на правом краю отрезка интегрирования как решения соответствующей нелинейной задачи Коши!

10.6.2. Метод квазилинеаризации (метод Ньютона)

Рассмотрим простейшую разностную аппроксимацию краевой задачи для ОДУ второго порядка

$$\frac{d^2u}{dt^2} = f(u), u(0) = U_1, u(L) = U_2$$

следующего вида:

$$\frac{u_{n-1} - 2u_n + u_{n+1}}{\tau^2} = f(u_n),\ n = 1, \ldots, N-1,\ u_0 = U_1,\ u_N = U_2,\ \tau = L/N.$$

Зададим некоторые начальные приближения $u_n^0 = \varphi(t_n)$ к искомой функции и построим итерационный процесс, воспользовавшись линейным приближением функции правой части

$$f(u_n^{i+1}) \approx f(u_n^i) + f'_u(u_n^i)(u_n^{i+1} - u_n^i)),$$

$$\frac{u_{n-1}^{i+1} - 2u_n^{i+1} + u_{n+1}^{i+1}}{\tau^2} = f(u_n^i) + f'(u_n^i)(u_n^{i+1} - u_n^i), u_n^0 = \varphi(t_n), i = 0, 1, \ldots$$

где i является итерационным индексом, а начальное приближение $\varphi(t_n)$ удовлетворяет граничным условиям $\varphi(0) = U_1, \varphi(L) = U_2$.

На первой итерации, т. е. при $i = 0$, получаем первое приближение к искомому решению, решая методом трехточечной прогонки разностное уравнение

$$\frac{u_{n-1}^1 - 2u_n^1 + u_{n+1}^1}{\tau^2} = f(u_n^0) + f'_u(u_n^0)(u_n^1 - u_n^0), u_n^0 = \varphi_n.$$

Полученное первое приближение вновь подставляем в линеаризованную разностную задачу и методом прогонки получаем второе приближение. Аналогично поступаем и далее, до достижения заданной точности в соответствии с условием $\left\|u^{i+1} - u^i\right\| \leqslant \varepsilon$.

10.6.3. Аппроксимация граничных условий

Для простоты выше рассматривались краевые задачи, для которых задавались граничные условия первого рода, т. е. на границе рассматриваемой области задавалось значение функции. Тогда трудностей с аппроксимацией граничных условий не возникало. Несколько сложнее обстоит дело при задании граничных условий второго и третьего рода (условий на производные и смешанные условия). Ограничимся описанием способов аппроксимации лишь для граничных условий второго рода.

Рассмотрим вначале линейную задачу с постоянными коэффициентами

$$\frac{d^2u}{dt^2} + h\frac{du}{dt} + su = f,$$
$$u'(0) = a, u'(1) = b,$$

и соответствующую ей разностную задачу

$$\frac{u_{n+1} - 2u_n + u_{n-1}}{\tau^2} + h\frac{u_{n+1} - u_{n-1}}{2\tau} + su_n = f_n,$$

заменив при этом производные в граничных условиях по формулам одностороннего дифференцирования первого порядка

$$u_1 - u_0 = \tau a, u_N - u_{N-1} = \tau b.$$

В результате получили разностную схему для аппроксимации дифференциальной задачи. Несмотря на то, что во внутренних узлах разностные формулы приближают дифференциальные уравнения со вторым порядком аппроксимации, решение по этой схеме получается только с первым порядком из-за понижения порядка аппроксимации в граничных точках. Естественно было бы повысить порядок схемы до второго во всех точках, включая граничные.

Наиболее распространены три способа.

1. Использование формул одностороннего дифференцирования более высокого порядка точности. Эти формулы разбирались в лекции 1. Подход очевиден, однако имеет недостаток — матрица системы уравнений для определения решения будет уже не трехдиагональной, следовательно, алгоритм прогонки неприменим. Можно избавиться от этого недостатка, проведя перед численным решением системы соответствующие алгебраические преобразования. Они в случае применения конкретной формулы численного дифференцирования для аппроксимации граничных условий свои, но достаточно очевидны.

2. Использование фиктивной ячейки (фиктивного узла). Рассмотрим расширение сеточной области. Введем в рассмотрение узлы с индексами 0 и $N + 1$, находящиеся формально за пределами рассматриваемой области. Пусть в них тоже определено значение сеточной функции. Тогда узел с индексом 0 выступает как внутренний узел, и в нем может быть записано соотношение

$$\frac{u_1 - 2u_0 + u_{-1}}{\tau^2} + h\frac{u_1 - u_{-1}}{2\tau} + su_0 = f_0,$$

при этом можно для граничного условия использовать формулу для вычисления производной со вторым порядком аппроксимации по формуле центральных разностей $\frac{u_1 - u_{-1}}{\tau} = a$. Из этих двух уравнений осталось только исключить значение в фиктивном узле (фиктивной ячейке).

Такой метод аппроксимации граничных условий очень хорошо зарекомендовал себя при решении нелинейных задач для уравнений в частных производных. В таком случае значения сеточной функции в фиктивных точках не исключаются из системы, а тоже находятся численно.

На правом конце отрезка формулы аналогичны. Не составляет труда обобщить этот подход и для граничных условий третьего рода.

3. Использование ряда Тейлора. Для формулы на левой границы области интегрирования $u_1 - u_0 = \tau a$ в явном виде выпишем главный член невязки:

$$\frac{u_1 - u_0}{\tau} = u'(0) + \frac{\tau}{2}u''(0) + o(\tau).$$

Используя само дифференциальное уравнение, можно выразить значение второй производной функции в окрестности границы:

$$\frac{d^2u(0)}{dt^2} = f(0) - h\frac{du(0)}{dt} - su(0),$$

откуда следует

$$\frac{u_1 - u_0}{\tau} = a + \frac{\tau}{2}(f(0) - h\frac{du(0)}{dt} - su(0)) + o(\tau).$$

Заменив в последнем соотношении производную на конечную разность, получим

$$\frac{u_1 - u_0}{\tau} = a + \frac{\tau}{2}(f_0 - h\frac{u_1 - u_0}{\tau} - su_0).$$

Осталось только привести это соотношение к виду, удобному для использования в методе прогонки.

В случае линейных уравнений с постоянными коэффициентами два последних способа приводят к одному и тому же выражению для значения сеточной функции в граничном узле. Для нелинейных уравнений эти способы будут различаться, но каждый из них приводит к повышению порядка аппроксимации граничных условий и, следовательно, разностной схемы.

10.7. Краевые задачи на собственные значения для обыкновенных дифференциальных уравнений

Краевые задачи на собственные значения достаточно часто встречаются в физических приложениях. Например, это задача определения собственных колебаний струны, сводящаяся к ОДУ вида

$$\frac{d}{dt}[k(t)\frac{du}{dt}] + \lambda r(t)u = 0.$$

В приведенном уравнении краевые условия зависят от способа закрепления струны.

Это и задачи собственных колебаний упругого стержня (ОДУ четвертого порядка), нахождения энергетических уровней атома водорода, вычисление критических нагрузок в теории стержней и оболочек и др.

В задачах на собственные значения добавляется еще один пристрелочный параметр — λ, поэтому эти задачи часто решаются методом стрельбы. Приведем простейший пример — одно дифференциальное уравнение первого порядка с двумя краевыми условиями (второе краевое условие появляется из-за присутствия неизвестного параметра λ):

$$\frac{du}{dt} + f(u,t,\lambda) = 0, u(0) = u_1,$$

Если не учитывать правое краевое условие, то получим задачу Коши. Ее численное интегрирование приводит к некому значению на правом конце, зависящему от λ и, вообще говоря, не равному u_2. Варируя параметр λ, можно добиться выполнения правого краевого условия с некоторой заданной точностью. При этом, разумеется, используются методы численного нахождения корней алгебраического уравнения, обычно метод касательных.

Второй пример — краевая задача на собственные значения для ОДУ второго порядка с нулевыми краевыми условиями:

$$\frac{d^2u}{dt^2} + b'(t)\frac{du}{dt} + [c'(t) + \lambda]u = 0,$$

$$u(0) = u(L) = 0.$$

Поскольку это уравнение второго порядка с неизвестным параметром λ (собственное значение дифференциального оператора), то для его

решения требуется третье условие. Однако в силу линейности и однородности задачи решение определяется с точностью до произвольного постоянного множителя, что и является неявным заданием третьего условия. Его можно задать, например, следующим образом:

$$\frac{du(a)}{dt} = 1.$$

Трудности при использовании метода стрельбы возникают, если соответствующая задача Коши плохо обусловлена, а также в случае жестких краевых задач. В этих случаях появляется сильная зависимость численного решения от пристрелочного параметра λ.

В качестве тестового примера для сравнения с результатами численного расчета удобно использовать модельную краевую задачу на собственные значения:

$$\frac{d^2u}{dt^2} + \lambda u = 0, u(0) = u(L) = 0,$$

имеющую точное решение

$$\lambda_k = \left(\frac{\pi k}{L}\right)^2, u_k(t) = \sin\left(\frac{\pi k t}{L}\right), k = 1, 2, \ldots$$

Непосредственной подстановкой показывается, что решениями соответствующей разностной задачи на собственные значения

$$\frac{u_{n-1} - 2u_n + u_{n+1}}{\tau^2} + \lambda u_n = 0, n = 1, 2, \ldots, N-1, \tau N = L, u_0 = u_N = 0,$$

являются собственные значения и собственные функции

$$\lambda_k = \frac{4}{\tau^2}\sin^2\frac{\pi k\tau}{2N}, \quad u_j^\tau = \sin\frac{\pi k t_n}{L}, \quad k, n = 1, 2, \ldots, N-1,$$

откуда видно, что

$$\lim_{\tau\to 0}\frac{4}{\tau^2}\sin\frac{\pi k\tau}{2L} = (\frac{\pi k}{L})^2 = \lambda_k,$$

т. е. имеет место сходимость решения разностной задачи к решению дифференциальной задачи $\lambda_k^\tau \underset{\tau\to 0}{\to} \lambda_k$.

10.8. Решение краевой задачи методом Фурье

Рассмотрим разностную задачу

$$\frac{u_{n-1} - 2u_n + u_{n+1}}{\tau^2} = -f_n, n = 1, \ldots, N-1, u_0 = u_N = 0, \tau N = L.$$

Построим ее решение в виде разложения по базису из собственных функций разностного оператора:

$$\mathbf{P}^{\tau}(u) = \frac{u_{n-1} - 2u_n + u_{n+1}}{\tau^2}.$$

Этот оператор имеет полную ортонормированную систему собственных функций $\omega_k(t_n) = \sqrt{\frac{2}{L}} \sin \frac{\pi k(\tau n)}{L}, k = 1, \ldots, N-1, \tau n = t_n$, которые соответствуют собственным значениям оператора $\mathbf{P}^{\tau}(u)$: $\lambda_k = \frac{4}{\tau^2} \sin^2 \frac{\pi k \tau}{2L}$. Будем искать решение в виде

$$u_n = u(t_n) = \sum_{k=1}^{N-1} c_k \omega_k(t_n), \quad n = 1, \ldots, N-1,$$

где c_k — пока неизвестные коэффициенты Фурье. Для нахождения этих коэффициентов представим правую часть разностного уравнения в виде суммы Фурье

$$f_n = \sum_{k=1}^{N-1} \hat{f}_k \omega_k(t_n), \quad \text{где} \quad \hat{f}_k = (f, \omega_k) = \sum_{i=1}^{N-1} f_i \omega_k(t_i)\tau.$$

Подставим выражение для u_n, f_n в виде сумм Фурье в исходное разностное уравнение

$$\mathbf{P}^{\tau}[\sum_{k=1}^{N-1} c_k \omega_k(t_n)] = -\sum_{k=1}^{N-1} \widehat{f}_k \omega_k(t_n)$$

или

$$\sum_{k=1}^{N-1} c_k \mathbf{P}^{\tau}[\omega_k(t_n)] = -\sum_{k=1}^{N-1} \hat{f}_k \omega_k(t_n),$$

откуда, с учетом соотношения

$$\mathbf{P}^{\tau}(\omega_k) = -\lambda \omega_k,$$

получим

$$-\sum_{k=1}^{N-1} c_k(\lambda_k \omega_k) = -\sum_{k=1}^{N-1} \hat{f}_k \omega_k, c_k \lambda_k = \hat{f}_k, c_k = \frac{\hat{f}_k}{\lambda_k}.$$

Таким образом, получено решение разностного уравнения в виде суммы Фурье. Несложные арифметические подсчеты показывают, что

метод прогонки, требующий $O(N)$ арифметических действий для численного решения этой же разностной задачи (из них $2(N-1)$ умножений и $N-1$ деление) оказывается более экономичным, чем метод Фурье, требующий $O(N^2)$ действий ($2(N-1)^2$ умножений и $N-1$ деление). Преимущества метода Фурье сказываются при решении двумерных разностных уравнений с постоянными коэффициентами.

10.9. Задачи

1. Пусть краевая задача имеет вид

$$y'' = f(x,y), y(0) = a, y(1) = b, \tag{10.3}$$

где нелинейная функция f не зависит явно от первой производной y'_x. В 1924 году Б. Нумеров предложил следующий метод аппроксимации задачи (10.3):

$$y_{n+1} - 2y_n + y_{n-1} = h^2[f_n + \frac{1}{12}(f_{n+1} - 2f_n + f_{n-1})], \tag{10.4}$$

где введено обозначение $f_k = f(x_k, y_k)$.

В чем заключается отличие метода Нумерова от аппроксимации вида (10.5):

$$y_{n+1} - 2y_n + y_{n-1} = h^2 f_n? \tag{10.5}$$

Описать алгоритмы численного решения нелинейных алгебраических систем (10.4) и (10.5). В случае (10.4) и $f = f(x)$ это — алгоритм прогонки.

Решение. Выпишем главный член погрешности аппроксимации разностного уравнения (10.4). Для этого подставим в разностное уравнение проекцию на сетку точного решения задачи (10.3). Следует отметить, что конкретный вид решения не важен, достаточно только, чтобы решение существовало. Предположим также, что оно четырежды непрерывно дифференцируемо.

Раскладывая проекции точного решения в правой части (10.4) в ряд Тейлора до четвертого порядка включительно, убедимся, что все нечетные производные взаимно уничтожатся, а четные дадут следующее выражение для главного члена погрешности аппроксимации:

$$y_{n+1} - 2y_n + y_{n-1} = y''(x_n) + \frac{1}{12}h^2\frac{d^4u}{dx^4}(x_n) + O(h^4).$$

Из уравнения (10.3) следует, что во всех внутренних точках области выполняется равенство

$$y^{(4)} = \frac{d^2}{dx^2} f(x, y),$$

переходя в последнем равенстве к разностной аппроксимации правой части, можно учесть явно главный член погрешности аппроксимации в (10.5). Приводя подобные слагаемые в (10.4), получаем аппроксимацию Нумерова

$$y_{n+1} - 2y_n + y_{n-1} = \frac{h^2}{12}(f_{n+1} + 10f_n + f_{n-1}),$$

которая приближает исходную задачу во внутренних точках сеточной области с четвертым порядком.

Такая идея разумного распоряжения правой частью для неоднородных и нелинейных задач приводит к компактным (возможно, название не слишком удачное — термины «компакт» и «компактный» уже давно заняты в математике под совсем другое!) разностным схемам — схемам повышенного порядка точности на нерасширенном шаблоне. Действительно, и в элементарном шаге вычислений для (10.4) и для (10.5) участвуют только три точки. Вопросам построения компактных разностных схем для нелинейных уравнений в частных производных посвящена монография [7].

В случае, когда правая часть явно зависит от первой производной, либо не получается компактной схемы, либо схема перестает быть экономичной. Действительно, чтобы не ухудшить порядок аппроксимации, необходимо вычислять значение первой производной в соответствующих узлах со вторым порядком. Если во всех точках использовать формулу с центральной разностью, то расширится шаблон схемы — в каждом шаге элементарных вычислений должно теперь участвовать пять точек. Если для точки с индексом n использовать формулу с центральной разностью, а для точек с индексами $n+1$ и $n-1$ — соответствующие формулы для односторонней производной, то для каждой точки шаблона придется вычислять заново значения функции f. При использовании классического вида аппроксимации Нумерова при каждом элементарном вычислении производится лишь однократное обращение к функции вычисления правой части — лишь для f_{n+1}, значения f_n и f_{n-1} уже получены при вычислениях для точки с индексом $n-1$. Так как время вычислений, как правило, определяется в таких задачах именно количеством обращений к правой части, то вычисления замедляются в три раза.

2. Для численного отыскания периодического с периодом «единица» решения уравнения

$$y'' + p(x)y' + q(x)y = f(x)$$

где f, q, p — заданные функции, используется разностная схема.

Предложить модификацию метода прогонки (периодическая прогонка) для решения данной задачи.

Решение. Рассмотрим следующую краевую задачу для уравнения Штурма-Лиувилля

$$y'' + p(x)y' + q(x)y = f(x)$$

с условиями периодичности

$$y^{(p)}(0) = y^{(p)}(1),$$

где p принимает значения 0 и 1. Отметим, что пока период решения считается известным. Отрезок [0, 1] возможно рассматривать без ограничения общности, так как любой отрезок можно перевести в единичный неособым линейным преобразованием, при этом вид уравнения существенно не изменится (функции p и q умножатся на постоянный множитель).

Краевая задача с условиями периодичности решается методом циклической прогонки. Запишем дискретный аналог уравнения:

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} + p_n \frac{y_{n+1} - y_{n-1}}{2h} + q_n y_n = f_n.$$

Очевидно, что оно аппроксимирует дифференциальную задачу со вторым порядком на равномерной сетке. На случай неравномерной сетки рассматриваемый метод легко обобщается.

В силу периодичности дискретное уравнение должно выполняться во всех точках сетки, включая граничные. Кроме того, в силу граничного условия при $p = 0, y_0 = y_{N+1}$, система сеточных соотношений примет вид:

$$a_0 y_N - b_0 y_0 + c_0 y_1 = \varphi_0,$$

$$\dots,$$

$$a_n y_{n-1} - b_n y_n + c_n y_{n+1} = \varphi_n,$$

$$\dots,$$

$$a_N y_{N-1} - b_N y_N + c_N y_0 = \varphi_N,$$

где $a_k = 1 - 0,5p_k h, b_k = 2 - q_k h^2, c_k = 1 + 0,5p_k h, \varphi_k = f_k h^2$. Матрица системы линейных уравнений получается «почти трехдиагональной» — от трехдиагональной ее отличают всего два элемента в углах матрицы.

Обобщение стандартных прогоночных соотношений (для трехточечной прогонки) на периодический случай будет иметь вид

$$y_{n-1} = \alpha_n y_n + \beta_n + \gamma y_N.$$

Из приведенного выше соотношения для y_0 сразу получаем, что $\alpha_1 = c_0/b_0$, $\beta_1 = -f_0/b_0$, $\gamma_1 = a_0/b_0$. Теперь несложно получить рекуррентную зависимость для прогоночных коэффициентов:

$$\alpha_{k+1} = \frac{c_k}{b_k - \alpha_k a_k}, \beta_{k+1} = \frac{a_k \beta_k - f_k}{b_k - \alpha_k a_k}, \gamma_{k+1} = \frac{a_k \gamma_k}{b_k - \alpha_k a_k}.$$

По приведенным выше формулам получаются значения коэффициентов для всех уравнений с номерами меньше, чем N. Подставим теперь прогоночные соотношения в последнее уравнение линейной системы. В итоге с учетом введенных выше обозначений получаем

$$a_N(\alpha_N y_N + \beta_N + \gamma_N y_N) - b_N y_N + c_N y_0 = \varphi_N,$$

а это соотношение, сгруппировав члены, можно переписать как:

$$y_N = \mu_N y_0 + \eta_N,$$

где введены обозначения

$$\mu_N = \frac{-c_N}{a_N(\alpha_N + \gamma_N) - b_N}, \nu_N = \frac{\varphi_N - a_N \beta_N}{a_N(\alpha_N + \gamma_N) - b_N}.$$

Теперь выражение для значения сеточной функции y_{n-1} подставляем в прогоночные соотношения. Получается выражение, связывающее y_{n-1} с y_0:

$$y_{n-1} = \alpha_n(\mu_n y_0 + \nu_n) + \beta_n + \gamma_n(\mu_N y_0 + \nu_N).$$

Отсюда получаем следующие рекуррентные соотношения:

$$\mu_{n-1} = \alpha_n \mu_n + \gamma_n \mu_N, \eta_{n-1} = \beta_n + \alpha_n \eta_n + \gamma_n \eta_N.$$

Отметим, что эти коэффициенты вычисляются в обратном порядке — аналог обратного хода прогонки. Последнее соотношение приводит к явному выражению для y_0. В результате получаем

$$y_0 = \frac{\eta_0}{1 - \mu_0}.$$

Теперь информации для определения значений искомой функции во всех точках сетки (еще один ход прогонки) достаточно.

Алгоритм периодической прогонки был предложен А. А. Абрамовым.

10.10. Задачи для самостоятельного решения

1. Рассмотреть две краевые задачи:

$$y'' = e^y, \quad y(0) = a, \quad y(1) = b, \tag{10.6}$$

$$y'' = -e^y, \quad y(0) = a, \quad y(1) = b. \tag{10.7}$$

   (a) Найти решения этих задач, положив $a = 1$ и сделав замену $y'_x = p(y)$.

   (b) Найти решение методом стрельбы при $a = 1$ и различных b. Что происходит при $0 < b < 1,499719998$? При $b > 1,499719998$? [[6] С. 110].

   (c) Решить задачу (10.6) методом Нумерова (10.4) с линеаризацией по Ньютону. Сколько узлов сетки необходимо, чтобы найти решение с точностью $\varepsilon = 10^{-4}$?

2. Рассмотреть нелинейную *сингулярно-возмущенную*[1] краевую задачу:

$$\varepsilon y'' = (y')^2, \quad y(0) = 1,\ y(0) = 0, 0 < \varepsilon \ll 1.$$

   (a) Получить точное решение задачи [[8], С. 11]. Для этого следует сделать замену $y'_x = p(y)$.

   (b) Предложить и реализовать численный метод решения задачи. Сравнить полученное решение с точным. Исследовать поведение погрешности численного метода при $\varepsilon \to 0$.

[1] Сингулярно-возмущенными задачами называются задачи с малым параметром при старшей производной.

3. Рассмотрим краевую задачу

$$\varepsilon y'' = (y - u(x))^{2q+1},$$

$$y(-1) = A,\ y(1) = B,$$

где $q \in N$ — натуральное число, $0 < \varepsilon \ll 1$.

Получить численное решение задачи в случаях:

а) $u(x) = x^2, A > 1, B > 1$,

б) $u(x) = x^2, A = 1, B > 1$ или $A > 1, B = 1$,

в) $u(x) = |x|, A > 1, B > 1$,

г) $u(x) = |x|, A = 1, B > 1$.

Что происходит с решением при увеличении q? (В численных расчетах задать $\varepsilon = 10^{-2}, 10^{-3}, 10^{-4}$).

Теоретически задача исследована в [[8], С. 170–171]. В случае $u(x) = |x|$ появляется внутренний пограничный слой — узкая область в окрестности $x = 0$, где y отличны от $|x|$.

4. Решить численно краевую задачу:

$$\varepsilon y'' = y - y^3,$$

$$y(0) = A, \quad y(1) = B,\ |A| < \sqrt{2}, \quad |B| < \sqrt{2}, \quad 0 < \varepsilon \ll 1.$$

Решением этой задачи являются так называемые пиковые, или пичковые, структуры. В [[8], с. 171–174] исследованы свойства решений и приведены графики восьми линейно независимых решений. Там же показано, что при фиксированных A и B существует по четыре линейно независимых решения, таких, что

$$\lim_{\varepsilon \to 0} y(x, \varepsilon) = 0$$

при всех x, за исключением точек $x_i = i/n, i = \overline{1, n-1}, n \in N (n \geqslant 2)$, где $\lim\limits_{\varepsilon \to 0} y(x, \varepsilon) = \sqrt{2}$.

Найти численно такие структуры для $n = 2, n = 3$, выбирая соответствующее начальное приближение при линеаризации по Ньютону. (Положить $\varepsilon = 10^{-2}, 10^{-3}, 10^{-4}$).

5. Известно, что краевая задача

$$\varepsilon y'' = y^3 - y,$$

$$y(0) = A < -1, \quad y(0) = B > 1$$

имеет решение с внутренним пограничным слоем в точке $x = 1/2$ [[8], С. 175]. Исследовать, как его толщина зависит от параметра ε.

Какое начальное приближение надо использовать при решении задачи методом линеаризации?

Известно, что при $A = 0, B = 0$ данная краевая задача имеет еще два решения, кроме тривиального ($y \equiv 0$). Найти их численно.

Всего у этой задачи счетное множество решений.

6. Исследовать следующую сингулярно-возмущенную задачу:

$$\varepsilon y'' = -y(y + a(x)),$$

$$y(0) = y_0, \quad y(1) = y_1, a'(0) = 0$$

в зависимости от вида функции $a(x)$. Рассмотреть поведение решения при $\varepsilon \to 0$. Удалось ли получить пограничный слой типа всплеска?

7. Решить численно задачу на нахождение собственных значений и собственных функций волнового уравнения [[1], С. 180–206]:

$$y'' = -k^2 y, y(0) = y(1) = 0.$$

   (a) Сравнить полученные решения с известными точными $k_n = n\pi, y_n \cong \sin(n\pi x)$ (n — положительное целое число).

   (b) Использовать этот же алгоритм для получения решения при больших k. С какими трудностями пришлось столкнуться? Как можно улучшить используемый алгоритм?

   (c) Рассмотреть данную задачу для других граничных условий, например, $y'(0) = y(1) = 0$.

8. В цилиндрических координатах задача на собственные значения имеет следующий вид:

$$\frac{d^2y}{dr^2} + \frac{1}{r}\frac{dy}{dr} = -k^2 y,$$

$$y(0) = 1, y(1) = 0.$$

Собственными функциями этой задачи являются цилиндрические функции Бесселя нулевого порядка, а собственными значениями задачи будут нули этих функций:

$$k_1 = 2,404826, \quad k_2 = 5,520078,$$

$$k_3 = 8,653728, \quad k_4 = 11,791538.$$

Показать, что подстановка $\tilde{y} = \frac{y}{\sqrt{r}}$ приводит уравнение к виду, для которого можно применять метод Нумерова. Решить численно спектральную задачу. Сравнить результаты расчета с точными собственными значениями, приведенными выше.

9. **Частица в потенциальной яме**

Найти численно *все* уровни энергии частицы в потенциальной яме с потенциалом $U(x) = -2\text{sech}^2 x$ и соответствующие им функции распределения.

Указание: уровни энергии есть собственные значения λ_k уравнения Шредингера $y'' + (\lambda - U(x))y = 0$ с условиями $y(+\infty) = y(-\infty) = 0$, а соответствующие им собственные функции и есть функции распределения.

Данный результат важен для решения уравнения Кортвега-Де Фриза, поэтому обсуждается в [9, С. 11–17].

10. **Частица в потенциальной яме**

(a) Получить аналитическое решение уравнения Шредингера для случая прямоугольной и параболической потенциальных ям и сравнить его с численными решениями, найденными обычным методом второго порядка аппроксимации с использованием алгоритма прогонки и методом Нумерова.

(b) Рассмотреть случай, когда потенциал имеет зеркальную симметрию относительно $x = 0$. При этом собственные функции системы будут четными или нечетными относительно $x = 0$, причем четность или нечетность чередуется с ростом квантового числа (энергии). Проверить этот эффект численно. Каким способом в этом случае можно в два раза сократить объем вычислений при расчете собственных значений для потенциалов?

(c) Проверить численно, что для заданного потенциала две волновые функции y_λ и $y_{\lambda'}$, соответствующие разным собственным значениям λ и λ', являются ортогональными: $\int y_\lambda(x) y_{\lambda'}(x) = 0$, как это следует из квантовой механики.

11. **Частица в поле с потенциалом Тоды**

Рассмотреть движение частицы в поле с потенциалом Тоды ([10, с. 87–89]):

$$\ddot{x} = 1 - e^x.$$

Это уравнение можно трактовать как движение частицы в поле с потенциалом $U(x) = e^x - x$.

Найти численно все периодические решения, удовлетворяющие следующим граничным условиям:

$$x(0) = x(120) = 0,$$

$$\dot{x}(0) = \dot{x}(120) = A$$

и дополнительному условию $A \geqslant 10$.

Как период колебаний зависит от A? Сколько решений получается?

Литература

[1] *Федоренко Р.П.* Введение в вычислительную физику. М.: Изд-во МФТИ, 1994. 528 с.

[2] *Калиткин Н.Н.* Численные методы. М.: Наука, 1978. 512 с.

[3] *Самарский А.А., Николаев Е.С.* Методы решения сеточных уравнений. М.: Наука, 1978. 590 с.

[4] *На Ц.* Вычислительные методы решения прикладных граничных задач. М.: Мир, 1982. 294 с.

[5] *Бахвалов Н.В., Жидков Н.П., Кобельков Г.М.* Численные методы. М: Лаборатория Базовых Знаний, 2002. 632 с.

[6] *Хайрер Э., Нерсетт С., Ваннер Г.* Решение обыкновенных дифференциальных уравнений. Нежесткие задачи. М.: Мир, 1990. 512 с.

[7] *Толстых А.И.* Компактные разностные схемы и их применение в задачах аэрогидродинамики. М.: Наука, 1990. 230 с.

[8] *Чанг К., Хауэрс Ф.* Нелинейные сингулярно-возмущенные краевые задачи. М.: Мир, 1988. 248 с.

[9] *Лэм Дж.* Введение в теорию солитонов. М.:Мир, 1981. Могилев: Бибфизмат, 1997. 294 с.

[10] *Ланда П.С.* Нелинейные колебания и волны. М.: Наука-Физматлит, 1997. 496 с.

Лекция 11. Исследование разностных схем для эволюционных уравнений на устойчивость и сходимость

В лекции рассматриваются методы исследования устойчивости разностных схем для линейных эволюционных уравнений в частных производных (гиперболического и параболического типов) Обсуждается применение спектрального признака устойчивости, энергетического признака, условия Куранта, Фридрихса и Леви для гиперболических уравнений. Формулируется и доказывается теорема (В. С. Рябенького-П. Лакса) о связи аппроксимации, устойчивости и сходимости для линейных разностных схем.

Ключевые слова: аппроксимация, устойчивость, сходимость. Теорема П. Лакса–В.С. Рябенького. Спектральный признак устойчивости. Энергетическая устойчивость.

11.1. Постановка некоторых задач для уравнений математической физики

Напомним некоторые корректные постановки задач для уравнений в частных производных, которые будут встречаться в дальнейшем.

Задача Коши для уравнения теплопроводности.

Найти функцию $u(t,x)$ в области $x \in (-\infty,\infty), t \in [0,T]$, удовлетворяющую уравнению

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

и начальным данным $u(0,x) = u_0(x)$, где $u_0(x)$ — заданная функция.

Смешанная задача для уравнения теплопроводности.

Найти функцию $u(t,x)$ в области $x \in [0,X], t \in [0,T]$, удовлетворяющую уравнению

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

начальным данным $u(0,x) = u_0(x)$ и краевым условиям, записанным в общей форме $A_1\frac{\partial}{\partial x}u(t,0) + B_1u(t,0) = \varphi_1(t), -A_2\frac{\partial}{\partial x}u(t,X) + B_2u(t,X) = \varphi_2(t)$.

Смешанная задача для уравнения переноса.

Найти функцию $u(t,x)$ в области $x \in [0,X], t \in [0,T]$, удовлетворяющую уравнению (для определенности положим $c > 0$):

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = f(t,x),$$

начальным данным Коши

$$u(0,x) = u_0(x), \quad t = 0$$

и левому краевому условию

$$u(t,0) = \varphi(t).$$

Смешанная задача для волнового уравнения.

Найти функцию $u(t,x)$ в области $x \in [0,X], t \in [0,T]$, удовлетворяющую уравнению

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

с начальными данными

$$u(0,x) = u_0, \frac{\partial u(0,x)}{\partial t} = u_1(x)$$

и краевыми условиями $u(0,t) = \varphi_1(t), u(X,t) = \varphi_2(t)$.

Эллиптическая краевая задача (уравнение Пуассона).

Найти функцию $u(x,y)$ в области $x \in [0,X], y \in [0,Y]$, удовлетворяющую уравнению

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

и краевым условиям $u(x,0) = \varphi_1(x), u(x,Y) = \varphi_2(x), u(0,Y) = \Psi_1(y)$, $u(x,Y) = \Psi_2(y)$.

Простейший способ построения численных решений для уравнений в частных производных — *метод сеток*. В дальнейшем, наряду с методом сеток, будем рассматривать и другие подходы к численному решению задач, например, вариационные, методы конечных элементов.

Рассмотрим постановку разностной задачи в методе сеток на примере одномерного уравнения теплопроводности.

Для решения одномерной смешанной задачи для уравнений в частных производных параболического типа область определения искомой функции покрывается расчетной сеткой с узлами в точках $\{t_n, x_m\}$, $n = 0,\ldots,N, m = 0,\ldots,M, t_n = n\tau, x_m = mh, \tau = T/N, h = X/M$, где τ, h — шаги сетки по времени и пространству соответственно. Приближенным

решением задачи назовем сеточную функцию $\{u_m^n\}$. Верхний индекс в такой форме записи сеточной функции традиционно указывает на номер слоя по времени, нижний (нижние) — на номер узла сетки по пространственной координате (рис. 11.1).

Рассмотрим подходы к построению численных алгоритмов для приближенного решения уравнений в частных производных.

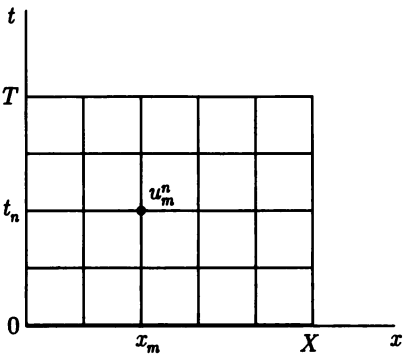


Рис. 11.1

Явная разностная схема для приближенного решения уравнения теплопроводности во внутренних узлах сетки (не принадлежащим границе сеточной области) имеет вид

$$\frac{u_m^{n+1} - u_m^n}{\tau} = a\frac{u_{m-1}^n - 2u_m^n + u_{m+1}^n}{h^2} + f_m^n, \quad f_m^n = f(t_n, x_m) \in \omega_f^\tau.$$

Под *разностной схемой* понимается совокупность разностных уравнений для определения значений сеточной функции внутри расчетной области, дополненная соответствующими начальными и граничными условиями для этой сеточной функции. *Шаблон схемы*, представляющий собой конфигурацию расчетных узлов в области интегрирования, используемых на каждом элементарном шаге вычислений, показан на рис. 11.2.

Эта схема аппроксимирует дифференциальное уравнение во внутренних точках (узлах) области интегрирования, т. е. при $n = 1, \ldots, N - 1, m = 1, \ldots, M - 1$. Проведем аппроксимацию начальных данных и краевых условий:

$$u_m^0 = u_0(x_m), m = 0 \div M, u_0 \in \omega_\tau,$$

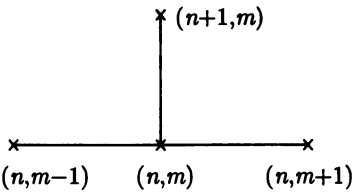


Рис. 11.2

$$-A_1\frac{u_1^n-u_0^n}{h}+B_1u_0^n=\varphi_1(t_n),n=1\div N,\varphi_1\in\omega_\varphi^\tau$$

$$A_2\frac{u_M^n-u_{M-1}^n}{h}+B_2u_M^n=\varphi_2(t_n),n=1\div N,\varphi_2\in\omega_\varphi^\tau$$

для определенности положим $A_k,\ B_k\geqslant 0,\ k=1,\ 2.$

Расчет ведется по рекуррентной формуле на каждом временном слое от $n=1$ до $n=N$ от $m=1$ до $m=M-1$ во внутренних узлах; слой $n=0(t=t_0)$ соответствует начальным данным, лучи $m=0(x=x_0)$ и $m=M(x=x_M)$ — левому и правому краевым условиям.

Запишем явную схему в виде

$$u_m^{n+1}=u_m^n+\frac{\tau}{h^2}(u_{m-1}^n-2u_m^n+u_{m+1}^n)+\tau f_m^n.$$

По этой формуле последовательно, на каждом слое вычисляется сеточная функция во внутренних узлах области интегрирования.

Для завершения расчета слоя $t=t_{n+1}$ необходимо вычислить u_{n+1}^0 и u_M^{n+1}, для чего разрешаем левое и правое краевые условия относительно этих величин:

$$u_0^{n+1}=\frac{A_1u_1^{n+1}+h\,\varphi_1^{n+1}}{A_1+h\,B_1},u_M^{n+1}=\frac{A_2u_{M-1}^{n+1}+h\,\varphi_{M-1}^{n+1}}{A_2+h\,B_2},$$

где u_1^{n+1} и u_{M-1}^{n+1} уже вычислены ранее. Реализация одного шага по времени занимает $O(M)$ арифметических операций, всех слоев — $O(NM)$ операций.

Явными схемами называются такие разностные схемы для эволюционных уравнений, когда данные на следующем слое по времени находятся непосредственно из данных на предыдущем слое без решения алгебраических систем уравнений. Если же на верхнем временном слое для определения значений сеточной функции необходимо решать систему алгебраических уравнений, то схема называется *неявной.*

Простейшая неявная разностная схема имеет вид (для простоты положим $a = 1$)

$$\frac{u_m^{n+1} - u_m^n}{\tau} = \frac{u_{m-1}^{n+1} - 2u_m^{n+1} + u_{m+1}^{n+1}}{h^2} + f_m^n,$$

ее шаблон (рис. 11.3).

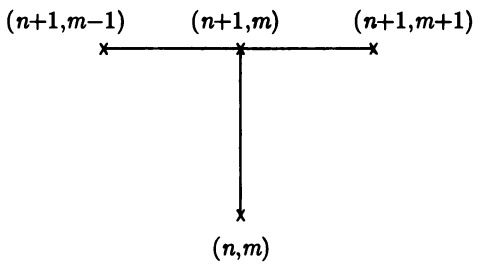


Рис. 11.3

На каждом временном слое имеем СЛАУ с трехдиагональной матрицей; алгоритм ее решения — прогонка.

Неявную схему представим в виде

$$(A_1 + hB_1)u_0^{n+1} - A_1 u_1^{n+1} = h\varphi_1^{n+1}, m = 0,$$

$$\frac{\tau}{h^2}u_{m-1}^{n+1} - (1 - 2\frac{\tau}{h^2})u_m^{n+1} + \frac{\tau}{h^2}u_{m+1}^{n+1} = u_m^n + \tau f_m^n, m = 1, \ldots, M-1,$$

$$(A_2 + hB_2)u_M^{n+1} - A_2 u_{M-1}^{n+1} = h\varphi_2^{n+1}, m = M,$$

откуда несложно получить прогоночное соотношение на каждом слое по времени.

11.2. Основные определения — сходимость, аппроксимация, устойчивость

11.2.1. Основные определения.

Дадим основные определения из теории разностных схем.

Пусть $\mathbf{L}u = F$ и $\mathbf{L}_\tau u_\tau = F_\tau$ — операторные обозначения исходной дифференциальной и аппроксимирующей ее разностной задачи (точнее,

параметрического семейства задач); $\mathbf{L}$ и $\mathbf{L}_\tau$ — соответственно, дифференциальный и разностный операторы, $u \in \Omega, u_\tau \in \Omega_\tau$ — решения дифференциального и разностного уравнений, принадлежащие соответствующим функциональным пространствам, $F \in \omega, F_\tau \in \omega_\tau$ — правая часть исходного уравнения и ее проекция на расчетную сетку. Считается известным способ получения проекции непрерывной функции на сетку. В простейшем случае используются значения функции, вычисленные в узлах сетки. Индекс τ в этой операторной записи указывает на всю совокупность сеточных параметров. Можно сказать, что для дискретной задачи имеется не один оператор, а совокупность различных операторов, зависящих от набора параметров.

Например, задачу Коши для линейного одномерного уравнения переноса

$$\frac{du}{dt} - \frac{du}{dx} = f(t,x), t \in [0,T], x \in [0,X],$$

$$u(0,x) = \varphi(x),$$

можно представить в виде

$$\mathbf{L}u = F.$$

Здесь

$$\mathbf{L}u = \begin{cases} \dfrac{du}{dt} - \dfrac{du}{dx}, t > 0 \\ u(0,x), t = 0 \end{cases}$$

$$F(t,x) = \begin{cases} f(t,x), t > 0 \\ \varphi(x), t = 0 \end{cases}$$

Одна из аппроксимирующих эту задачу разностных схем (правый уголок) имеет вид

$$\frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_{m+1}^n - u_m^n}{h} = f_m^n, n = 0, \ldots, N-1, m = 0, \ldots, M-1,$$

$$u_m^0 = \varphi_m^0,$$

или в операторной форме

$$\mathbf{L}_\tau u_\tau = F_\tau$$

где

$$\mathbf{L}_\tau u_\tau = \begin{cases} \dfrac{u_m^{n+1} - u_m^n}{\tau} - \dfrac{u_{m+1}^n - u_m^n}{h}, \\ u_m^0, \end{cases}$$

$$F_\tau = \begin{cases} f_m^n, \\ \varphi_m. \end{cases}$$

Определение 1. Говорят, что решение u_τ сходится к решению при $\tau \to 0$, если $\|u_\tau - U_\tau\| \to 0$, где U_τ — проекция точного решения на разностную сетку; причем, если имеет место оценка $\|u_\tau - U_\tau\| \leqslant c\tau^p$, $c \neq c(\tau)$, то сходимость имеет порядок p.

В качестве примера исследуем на сходимость разностную схему для задачи Коши для обыкновенного дифференциального уравнения (схема Эйлера)

$$\frac{u_{n+1} - u_n}{\tau} + \lambda u_n = 0, \quad u_0 = a,$$

аппроксимирующую простейшее ОДУ

$$\frac{du}{dt} + \lambda u = 0, \; x \in [0,1], \; u(0) = a.$$

Из разностного уравнения

$$u_n = (1 - \lambda\tau)u_{n-1}$$

найдем его общее решение:

$$u_n = a(1 - \lambda\tau)^n, \quad \text{или} \quad u_n = a(1 - \lambda\tau)^{t_n/\tau}.$$

Решение дифференциальной задачи легко находится:

$$u(t) = ae^{-\lambda t}.$$

Величина погрешности решения, входящая в определение сходимости, тогда будет $\Delta_n = a\left|(1 - \lambda\tau)^{t_n/\tau} - e^{-\lambda t_n}\right|$.

Представим $(1 - \lambda\tau)^{t_n/\tau}$ в виде

$$(1 - \lambda\tau)^{t_n/\tau} = \exp(\frac{t_n}{\tau}\ln(1 - \lambda\tau)) = \exp\frac{t_n}{\tau}\left[-\lambda\tau + \frac{\lambda^2\tau^2}{2} + O(\tau^3)\right] =$$

$$= e^{-\lambda t_n}\left[1 + \frac{\lambda^2\tau\, t_n}{2} + O(\tau^2)\right]\left[1 + O(\tau^2)\right] = e^{-\lambda t_n} + \frac{\tau}{2}\lambda^2 t_n e^{-\lambda t_n} + O(\tau^2).$$

Тогда

$$u_n = ae^{-\lambda t_n} + \tau a\frac{\lambda^2 t_n}{2}e^{-\lambda t_n} + O(\tau^2),$$

и, следовательно,

$$\Delta_n = \frac{a\tau}{2}\lambda^2 t_n e^{-\lambda t_n} + O(\tau^2) = O(\tau),$$

т. е. разностная схема имеет первый порядок сходимости. К сожалению, чтобы исследовать схему на сходимость, необходимо знать точное решение дифференциальной задачи. Обычно разностные схемы исследуются на аппроксимацию и устойчивость, откуда по теореме П. Лакса и В. С. Рябенького и следует сходимость, по крайней мере, для линейных задач.

Пример разностной задачи, аппроксимирующей рассматриваемое уравнение:

$$4\frac{u_{n+1}-u_{n-1}}{2\tau}-3\frac{u_{n+1}-u_n}{\tau}+\lambda u_n=0.$$

Можно показать, получив общее решение разностной задачи, что эта схема не является *устойчивой* и, следовательно, не имеет место сходимость решения к точному решению дифференциальной задачи.

Определение 2. Говорят, что разностная задача аппроксимирует дифференциальную на ее решении, если норма невязки, возникающей при действии разностного оператора на сеточную функцию — проекцию на сетку точного решения

$$r_\tau=\mathbf{L}_\tau U_\tau-F_\tau$$

стремится к нулю при $\tau\to 0$; если выполнена оценка $\|r_\tau\|\leqslant c_k\tau^p$, $c_k\neq c_1(\tau)$ (константа, входящая в правую часть неравенства, не зависит от сеточных параметров), то имеет место аппроксимация порядка p.

Приведем пример исследования разностной схемы на аппроксимацию, для чего напомним следующие соотношения, полученные с помощью разложения проекции точного решения на сетку в ряд Тейлора в окрестности одного из сеточных узлов:

$$U(t+\tau)=U(t)+\tau U_t'(t)+\frac{\tau^2}{2!}U_t''(t)+\frac{\tau^3}{3!}U_t^{(3)}(t)+\frac{\tau^4}{4!}U_t^{(4)}(t)+O(\tau^5)$$

$$U'(t)\approx\frac{U(t+\tau)-U(\tau)}{\tau}=U'(t)+\frac{\tau}{2}U_t''(t)+O(\tau^2),$$

$$U'(t)\approx\frac{U(t)-U(t-\tau)}{\tau}=U'(t)-\frac{\tau}{2}U_t''+O(\tau^2),$$

$$U'(t)\approx\frac{U(t+\tau)-U(t-\tau)}{2\tau}=U'(t)+\frac{\tau^2}{3}U_t^{(3)}(t)+O(\tau^3),$$

$$U(t-\tau)=U(t)-\tau U_t'(\tau)+\frac{\tau^2}{2!}U_t''(\tau)-\frac{\tau^3}{3!}U_t^{(3)}(t)+\frac{\tau^4}{4!}U_t^{(4)}+O(\tau^5);$$

$$U''(t)\approx\frac{U(t+\tau)-2U(t)+U(t-\tau)}{\tau^2}=U''(t)+\frac{\tau^2}{12}U_t^{(4)}(t)+O(\tau^4).$$

Рассмотрим задачу Коши для уравнения переноса

$$\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = f(t,x), -\infty < x < \infty, 0 \leqslant t \leqslant T,$$

$$u(0,x) = \varphi(x), -\infty < x < \infty.$$

Введем разностную сетку и положим

$$\frac{du(t,x)}{dt} \approx \frac{u(t+\tau,x) - u(t,x)}{\tau}, \quad \frac{du(t,x)}{dt} \approx \frac{u(t,x+h) - u(t,x)}{h}.$$

Получим введенную выше схему «правый уголок».

Положив, что функция $u(t,x)$ имеет ограниченные вторые производные, получим выражения для главных членов погрешности аппроксимации, т. е. тех членов, которые определяются минимальными степенями сеточных параметров.

$$\frac{U_{m+1}^n - U_m^n}{h} = \frac{\partial u(t_n,x_m)}{\partial x} + \frac{h}{2} \cdot \frac{\partial^2 u(t_n,x_m)}{\partial x^2} + O(h^2),$$

$$\frac{U_m^{n+1} - U_m^n}{\tau} = \frac{\partial u(t_n,x_m)}{\partial t} + \frac{\tau}{2} \cdot \frac{\partial^2 u(t_n,x_m)}{\partial x^2} + O(\tau^2).$$

Тогда для невязки верно равенство $\mathrm{L}_\tau U_\tau = F(t_m,x_m) + r_\tau$, или, так как в силу самого дифференциального уравнения

$$F(t_m,x_m) = \left(\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x}\right)\Big|_{t_n,x_m},$$

$$\frac{U_m^{n+1} - U_m^n}{\tau} - \frac{U_{m+1}^n - U_m^n}{h} = \left(\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x}\right)\Big|_{t_n,x_m} + r_\tau.$$

Здесь невязка r_τ определяется выражением

$$r_\tau = \frac{\tau}{2}\frac{\partial^2 u(t_n,x_m)}{\partial t^2} - \frac{h}{2}\frac{\partial^2 u(t_n,x_m)}{\partial x^2} + O(\tau^2,h^2).$$

Так как $\|r_\tau\| = O(\tau + h)$, то разностная схема имеет первый порядок аппроксимации по τ и h.

Аналогично можно получить выражение для невязки разностной схемы

$$\frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2} = 0,$$

аппроксимирующей уравнение теплопроводности

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0.$$

После соответствующих вычислений получим

$$r_\tau = \frac{\tau}{2}\frac{\partial^2 u}{\partial t^2} - \frac{h^2}{12}\frac{\partial^4 u}{\partial t^4}\bigg|_{t_n, x_m} + O(\tau^2, h^4), \|r_\tau\| = O(\tau + h^2).$$

Таким образом, схема обладает первым порядком аппроксимации по τ и вторым по h.

Определение 3. Говорят, что разностная задача является *устойчивой*, если из соотношений

$$\mathbf{L}_\tau u_\tau - F_\tau = \xi_\tau, \mathbf{L}_\tau v_\tau - F_\tau = \eta_\tau,$$

следует в смысле выбранной нормы

$\|u_\tau - v_\tau\| \leqslant c_2(\|\xi_\tau\| + \|\eta_\tau\|)$, причем эта оценка равномерная, $c_2 \neq c_2(\tau)$.

Теорема (П. Лакса–В. С. Рябенького). *Решение линейной разностной задачи* сходится *к решению дифференциальной, если разностная задача устойчива и аппроксимирует дифференциальную задачу на ее решении. При этом порядок аппроксимации совпадает с порядком сходимости.*

Дадим еще одно эквивалентное определение устойчивости разностной задачи, применимое лишь для линейных разностных операторов. Как будет видно ниже, даже для линейных дифференциальных задач возможно построение нелинейных разностных схем.

Определение 4. Линейная разностная задача устойчива, если при любой правой части F_τ она имеет единственное решение u_τ, причем $\|u_\tau\| \leqslant C\|F_\tau\|$, и данная оценка равномерна по сеточным параметрам $C \neq C(\tau)$.

Покажем, что из устойчивости задачи в смысле второго определения следует ее устойчивость в смысле первого определения.

Вычтем из разностного уравнения $\mathbf{L}_\tau u_\tau = F_\tau$ «возмущенное» разностное уравнение (положим для простоты $\eta_\tau = 0$):

$$\mathbf{L}_\tau v_\tau = F_\tau + \xi_\tau.$$

Получим в силу линейности разностного оператора

$$\mathbf{L}_\tau(u_\tau - v_\tau) = \xi_\tau.$$

В силу определения 4 справедливо неравенство

$$\|u_\tau - v_\tau\| \leqslant C\|\xi_\tau\|,$$

откуда и следует справедливость в смысле определения 3, так как ξ_τ в этом неравенстве играет роль правой части, а $u_\tau - v_\tau$ — искомой функции. Можно также показать справедливость обратного утверждения. Заметим, что в силу произвольности ξ_τ, из последнего неравенства следует единственность решения разностного уравнения.

В теории разностных схем также вводится определение корректности разностной задачи.

Определение 5. Семейство разностных уравнений

$$\mathbf{L}_\tau u_\tau = F_\tau$$

считается корректным, если:

— его решение существует и единственно при любых правых частях $F_\tau \in \omega_F$;

— существует константа C, независящая от τ, такая, что при любых F_τ выполняется оценка $\|u_\tau\| \leqslant C \|F_\tau\|$.

Первое условие эквивалентно существованию оператора $\mathbf{L}_\tau^{-1}$, второе — равномерной по τ ограниченности $\mathbf{L}_\tau^{-1}$, т. е. константа C является универсальной для всего семейства уравнений.

Заметим также, что условие $\|u_\tau\| \leqslant C \|F_\tau\|$ означает непрерывную равномерную по τ зависимость решения разностной задачи от правой части.

Это неравенство является введенным ранее определением устойчивости разностной задачи.

11.2.2. Необходимое условие сходимости разностной схемы Куранта, Фридрихса, Леви (условие КФЛ)

Рассмотрим разностное уравнение

$$u_m^{n+1} = (1-\sigma)u_m^n + \sigma\, u_{m+1}^n, \quad u_m^0 = \varphi_m^0, \quad \sigma = \tau/h,$$

аппроксимирующее задачу Коши для уравнения переноса

$$\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = 0, t \in [0,1], u(0,x) = \varphi(x).$$

Очевидно, что значение u_0^N сеточной функции в точке (1,0) выражается через значения u_0^{N-1} и u_1^{N-1} в точках $(1-\tau, 0)$, $(1-\tau,\ h)$. В свою очередь, значения u_0^{N-1} и u_1^{N-1} находятся по значениям сеточной функции u_0^{N-2}, u_1^{N-2}, u_2^{N-2} в точках $(1-2\tau, 0)$, $(1-2\tau, h), (1-2\tau,\ 2h)$,

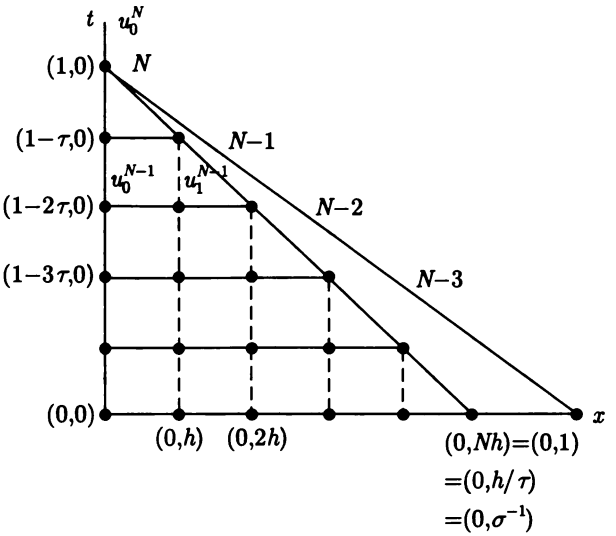


Рис. 11.4

значения сеточной функции u_0^{N-2}, u_1^{N-2}, u_2^{N-2} находятся по значениям u_0^{N-3}, u_1^{N-3}, u_2^{N-3}, u_3^{N-3} в точках $(1-3\tau, ih), i = 0, \ldots, 3$, и т.д. Значение сеточной функции u_0^N выражается через значение u_m^0 решения в точках расчетной сетки $(0, mh), m = 0, \ldots, N, N = \tau^{-1}$ Все эти точки лежат на отрезке $[0,\ h/\tau]$ или $\left[0,\ \sigma^{-1}\right]$ оси $t = 0$, на которой задано начальное условие $u(0,x) = \varphi(x)$ исходной дифференциальной задачи. Значение u_0^N не зависит от значения функции $\varphi(x)$ при x, лежащих вне отрезка $\left[0,\ \sigma^{-1}\right]$.

Из курса обыкновенных дифференциальных уравнений известно, что решением однородного уравнения переноса является функция $u(t,x) = \varphi(t+x)$, сохраняющая свое значение вдоль характеристики $t + x = \text{const}$, и, в частности, на прямой $t + x = 1$, проходящей через точки (0,1), (0,1), см. рис. 11.4.

Таким образом, при $\sigma > 1$ *область зависимости* решения дифференциальной задачи для u_0^N, являющаяся точкой (0,1), не входит в отрезок $0 \leqslant x \leqslant \sigma^{-1}$. В случае $\sigma^{-1} < 1$ и $0 \leqslant \sigma < 1$ сходимость решения разностной задачи к решению дифференциальной отсутствует. Разумеется, приведенные рассуждения не носят характера доказательства, а лишь косвенно объясняют, почему не следует ожидать сходимости при $\sigma = \tau/\sigma > 1$.

Сформулируем теперь условие Куранта-Фридрихса-Леви (условие КФЛ), необходимое для сходимости разностной задачи.

Пусть некая точка A принадлежит области определения решения $u(t,x)$ и значение функции $u(A)$ зависит от значения некоторой функции $\varphi(x)$ в точках x, принадлежащих множеству $\Omega(A)$, которое, в свою очередь, принадлежит области определения функции $\varphi(x)$.

Положим, что для приближенного вычисления решения уравнения $Lu = F$ используется разностная схема $L_\tau u_\tau = F_\tau$, причем, значение решения в точке A_x расчетной сетки, ближайшей к A, полностью определяются значениями функции φ на множестве $\Omega_x(A_x)$. Для того чтобы имела место сходимость $u_\tau \to u$ при $h \to 0$, разностная схема должна быть устроена так, чтобы при сколь угодно малых значениях пространственного шага h в произвольной окрестности любой точки области $\Omega(A)$ имелась точка множества $\Omega_x(A_x)$.

Другими словами, разностная схема должна быть устроена так, чтобы область зависимости разностного уравнения учитывала область зависимости решения исходного дифференциального уравнения. В противном случае сходимости ожидать, вообще говоря, нельзя. Если же разностная задача аппроксимирует дифференциальную, то необходимое условие сходимости КФЛ является также необходимым условием устойчивости схемы. Отметим, что условию КФЛ можно придать форму теоремы.

11.3. Элементы теории устойчивости разностных схем

Канонической формой двухслойной линейной разностной схемы называется ее запись в виде

$$\mathbf{B}\frac{\mathbf{u}_{n+1}-\mathbf{u}_n}{\tau}+\mathbf{A}\mathbf{u}_n=\mathbf{f}_n, \tag{11.1}$$

$\mathbf{B}$ и $\mathbf{A}$ — операторы, действующие в Ω_x. Рассматриваем случай, когда для этих операторов выполнено условие $(\mathbf{Au},\mathbf{u}) > \mu(\mathbf{u},\mathbf{u})$, где μ — положительное число, $\mathbf{u}$ — произвольный ненулевой элемент пространства сеточных функций. Тогда оператор $\mathbf{A}$ называется положительным, записывается $\mathbf{A} > 0$. Требуем и $\mathbf{B} > 0$. Пока не оговорено иное, рассматриваем случаи самосопряженных операторов.

Если $\mathbf{B} = \mathbf{E}$, то разностная схема называется явной. Рассмотрим для примера разностную схему с весами для одномерного уравнения теплопроводности

$$\frac{u_m^{n+1}-u_m^n}{\tau}=\xi\Lambda_{xx}u_m^{n+1}+(1-\xi)\Lambda_{xx}u_m^n, \xi\in[0,1],$$

где

$$\Lambda_{xx}u_m^{n+1}=\frac{u_{m-1}^{n+1}-2u_m^{n+1}+u_{m+1}^{n+1}}{h^2}, \Lambda_{xx}u_m^n=\frac{u_{m-1}^n-2u_m^n+u_{m+1}^n}{h^2},$$

$$n=0,\ldots,N-1, \quad m=0,\ldots,M-1.$$

В случае $\xi = 0{,}5$ эта схема называется схемой Кранка–Никольсон. Для записи схемы с весами в каноническом виде положим

$$\mathbf{A}u_n = -\Lambda_{xx}u_m^n, \mathbf{A}u_{n+1} = -\Lambda_{xx}u_m^{n+1},$$

и, обозначив $\mathbf{u}_n = (u_1^n, u_2^n, \ldots, u_{M-1}^n)^T$, получим форму записи $\mathbf{B}\frac{\mathbf{u}_{n+1}-\mathbf{u}_n}{\tau} + \mathbf{A}\mathbf{u}_n = 0$, где $\mathbf{B} = \mathbf{E} + \tau\xi\mathbf{A}$.

В случае неравномерной сетки каноническая форма записи схемы будет

$$\mathbf{B}\frac{\mathbf{u}_{n+1}-\mathbf{u}_n}{\tau_{n+1}} + \mathbf{A}\mathbf{u}_n = \mathbf{f}_n, \quad \mathbf{f}_n = (f_1^n, \ldots, f_{M-1}^n)^T. \tag{11.2}$$

Иногда схему с весами записывают в виде

$$\frac{u_m^{n+1}-u_m^n}{\tau} = \Lambda_{xx}u^{(\xi)}.$$

Заметим, что канонический вид разностной схемы аналогичен итерационному методу решения СЛАУ $\mathbf{A}u = \mathbf{f}$. Эта аналогия не является формальной — переход от СЛАУ к итерационному методу (11.1) может быть интерпретирован, как замена стационарного уравнения $\mathbf{A}u = \mathbf{f}$ нестационарным. Решение последнего при стационарных граничных условиях стремится к решению стационарного при стремлении времени к бесконечности.

Отличие состоит в том, что в последнем случае операторы $\mathbf{B}, \mathbf{A}$ и функция $\mathbf{f}$ не зависят от n и итерационный параметр τ не обязательно должен стремиться к нулю.

Введем величину $(\mathbf{A}u,\ u)$ — энергию оператора $\mathbf{A}$, а также энергетическую норму вектора:

$$||\mathbf{u}||_{\mathbf{A}} = (\mathbf{A}u,\ u)^{1/2}.$$

Говорят, что эта норма порождается оператором $\mathbf{A}$.

Рассматривается задача Коши для оператора — однородного разностного уравнения (11.1); $\mathbf{u}^0 = \varphi$.

Дадим два определения.

Определение 6. Разностная схема (11.1) устойчива по начальным данным, если для решения (11.1) выполняется оценка:

$$||\mathbf{u}^{n+1}|| \leqslant M_1||\varphi||, \forall\, t^n \in \omega^t \text{ — узлы сетки по } t, \tag{11.3}$$

причем константа M_1 не зависит от сеточных параметров.

Будем рассматривать также неоднородное уравнение, соответствующее (11.1):

$$\mathbf{B}\,\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\tau} = -\mathbf{A}\mathbf{u}^n + \mathbf{f}(x_m,\ t^n). \tag{11.4}$$

Определение 7. Говорят, что разностная схема (11.4) устойчива по правой части, если для решения (11.4) в любой момент времени выполняется условие

$$||\mathbf{u}^{n+1}|| \leqslant M_2||\mathbf{f}||, \tag{11.5}$$

причем константа M_2 не зависит от сеточных параметров.

Определение 8. Разностная схема (11.1) равномерно устойчива по начальным данным в энергетической норме, порождаемой некоторым оператором $\mathbf{R} = \mathbf{R}^* > 0$, если $\exists\, \rho > 0:\quad \forall\, t^n$ выполнено:

$$||\mathbf{u}_{n+1}||_{\mathbf{R}} \leqslant \rho||\mathbf{u}_n||_{\mathbf{R}}$$

и при этом $\rho^n \leqslant M_1$, ρ *и* M_1 не зависят от сеточных параметров.

Обычно рассматриваются случаи $\rho = M_1 = 1$ или $\rho = 1 + c\tau, M_1 = e^{cT}$.

Если разностную схему в каноничной форме представить в виде

$$\mathbf{u}_{n+1} = \mathbf{R}_\tau \mathbf{u}_n + \tau \mathbf{B}^{-1}\mathbf{f}_n, n = 0, \ldots, N-1, \mathbf{R}_\tau = \mathbf{R}_\tau(t_n),$$

то оператор $\mathbf{R}_\tau = \mathbf{E} - \tau\mathbf{B}^{-1}\mathbf{A}$ называется оператором послойного перехода разностной схемы (11.1). Нетрудно заметить, что условие равномерной устойчивости по начальным данным эквивалентно ограничению нормы оператора $\mathbf{R}_\tau$: $\|\mathbf{R}_\tau\| \leqslant \rho$, а в силу условия $\rho^n \leqslant C$ и ограниченности норм степеней оператора $\mathbf{R}$: $\|\mathbf{R}_\tau^n\| \leqslant C$.

Для оценки нормы оператора $\mathbf{R}_\tau^n$ можно воспользоваться собственными значениями этого оператора — корнями уравнения $\det\|\mathbf{R}_\tau - \lambda\mathbf{E}\| = 0$.

Если λ — собственное значение, а ω — соответствующий ему собственный вектор, то $\mathbf{R}_\tau\omega = \lambda\omega$. Поэтому $\mathbf{R}_\tau^n\omega = \lambda^n\omega$, откуда $\|\mathbf{R}_\tau^n\| \geqslant |\lambda|^n$, так как $\|\mathbf{R}_\tau^n\omega\| = |\lambda|^n\,\|\omega\| \leqslant \|\mathbf{R}_\tau^n\|\,\|\omega\|$.

Последнее неравенство должно выполняться при любом n. Оно невыполнимо, если $|\lambda|^n$ с увеличением n будет неограниченно расти, так как $\|\mathbf{R}_\tau^n\| \leqslant C$. Этого не произойдет, если на λ будет наложено условие $|\lambda| \leqslant 1 + c\tau$, константа c не зависит от сеточных параметров, $c = O(1), \tau \ll 1$. Последнее условие называется необходимым *спектральным признаком* устойчивости (признак фон Неймана).

Рассмотрим разностную задачу Коши для линейного уравнения переноса

$$\mathbf{L}_\tau u_\tau = F_\tau,$$

где

$$\mathbf{L}_\tau u_\tau = \begin{cases} \dfrac{u_m^{n+1} - u_m^n}{\tau} - \dfrac{u_{m+1}^n - u_m^n}{h}, & n = 0, \ldots, N-1, m = 0, \ldots, M-1, \\ u_m^0, & m = 0, \ldots, M-1, \end{cases}$$

$$F_\tau = \begin{cases} f_n^m,\ n = 0, \ldots, N,\ m = 0, \ldots, M-1, \\ \varphi_m,\ n = 0,\ m = 0, \ldots, M, \end{cases}$$

Условие ее устойчивости записывается в виде неравенства

$$||u_\tau|| \leqslant C||F_\tau||.$$

Для однородного уравнения переноса это условие принимает вид

$$\max_m |u_m^n| \leqslant C \max_m |u_m^0|.$$

Последнее неравенство означает устойчивость разностной схемы по начальным данным. Оно должно выполняться, в частности, в случае, если $u_m^0 = \varphi_n$ является произвольной гармоникой при представлении начальных условий в виде ряда Фурье (важно лишь знать, будет ли эта гармоника неограниченно расти по времени). Возьмем в качестве начального условия произвольную гармонику $u_m^0 = e^{i\alpha m}$, где α — вещественный параметр.

Решение однородной разностной задачи в этом случае ищется с помощью метода разделения переменных. На каждом временном слое решение разностной задачи ищется как произведение $u_m^n = \lambda^n e^{i\alpha m}$.

Спектр оператора послойного перехода $\lambda(\alpha)$ легко ищется подстановкой в разностное уравнение. Например, для однородного уравнения переноса с постоянными коэффициентами после преобразований $u_m^{n+1} = (1-\sigma)u_m^n + \sigma u_{m+1}^n$, где $\sigma = \tau/h = \text{const}$ — безразмерный параметр — число Куранта (в числитель дроби входит скорость переноса, которая в рассматриваемой задаче равна единице). Для спектра оператора перехода имеем $\lambda(\alpha) = (1-\sigma) + \sigma e^{i\alpha}$.

Для решения вида $u_m^n = \lambda^n e^{i\alpha m}$ справедливо $\|u_m^n\| = \left\|\lambda^n e^{i\alpha m}\right\| = \left\|\lambda^n \cdot u_m^0\right\| \leqslant |\lambda^n| \cdot \left\|u_m^0\right\|$, или $\max\limits_m |u_m^n| = |\lambda^n| \cdot \max\limits_m |u_m^0|$, поэтому для выполнения условия устойчивости $\max\limits_m |u_m^n| \leqslant C \cdot \max\limits_m |u_m^0|$ необходимо выполнение неравенства $|\lambda(\alpha)|^n \leqslant 1 + C\tau$.

Константа в правой части последнего неравенства не зависит от сеточных параметров.

Спектральным признаком это условие называется потому, что каждая гармоника $e^{i\alpha m}$ является собственной функцией оператора послойного перехода $u_m^{n+1} = \mathbf{R}_\tau u_m^n$. В частности, для рассмотренного выше примера с уравнением переноса $\mathbf{R}_\tau u_m^n = (1-\sigma)u_m^n + \sigma u_{m+1}^n$.

Множество точек $\lambda(\alpha)$ на комплексной плоскости состоит из собственных значений оператора перехода — спектр оператора. При этом считаем $\alpha \in [0, 2\pi]$. Сформулируем теперь спектральный признак устойчивости в этих терминах. Спектр оператора перехода с n на $n+1$ времен-

ной слой должен лежать в круге радиуса $1 + C\tau$ на комплексной плоскости.

В приведенном примере спектр оператора не зависит явно от τ поэтому условие устойчивости может быть записано в виде $|\lambda(\alpha)| \leqslant 1$.

Такое условие иногда называется условием *строгой устойчивости*. В [6] показано, что строго устойчивые схемы можно построить лишь для таких дифференциальных задач, для которых справедлив принцип максимума, т. е. максимальное и минимальное значения решение дифференциальной задачи принимает на границе расчетной области.

Вернемся к модельному уравнению переноса. В данном случае спектр представляет собой окружность с центром в точке $(1 - \sigma)$ и радиусом σ на комплексной плоскости. При $\sigma < 1$ эта окружность лежит внутри единичного круга, касаясь его в точке $\lambda = 1(\alpha = 0)$, при $\sigma = 1$ совпадает с единичной окружностью, при $\sigma > 1$ находится вне единичного круга.

Приведем еще один пример исследования на устойчивость разностной задачи:

$$\frac{u_m^{n+1} - u_m^n}{\tau} - a^2 \frac{u_{m+1}^n - 2u_m^n + u_{m+1}^n}{h^2} = 0, u_m^0 = \varphi_m,$$

$$n = 0, \ldots, N-1, m = 0, \ldots, M-1, a > 0.$$

Эта схема аппроксимирует задачу Коши для уравнения теплопроводности.

После подстановки решения в виде гармоники Фурье, умноженной на коэффициент перехода, $u_m^n = \lambda^n e^{i\alpha m}$ в разностное уравнение, получим уравнение для спектра оператора послойного перехода

$$\frac{\lambda - 1}{\tau} - a \frac{e^{-i\alpha} - 2 + e^{i\alpha}}{h^2} = 0.$$

После очевидных преобразований получим

$$\lambda(\alpha) = 1 - 4\sigma' \sin^2 \frac{\alpha}{2},$$

где $\sigma' = \tau a / h^2$ — аналог числа Куранта для параболических уравнений (иногда его называют параболическим числом Куранта)

При изменении α спектр $\lambda(\alpha)$ пробегает значения $1 - 4\sigma' \leqslant \lambda(\alpha) \leqslant 1$, а для выполнения условия устойчивости необходимо $1 - 4\sigma' \geqslant -1$, или $\sigma' \leqslant 1/2$, откуда $\tau \leqslant \frac{h^2}{2a}$.

Аппроксимация этого же дифференциального уравнения с помощью *неявной* схемы приводит к следующему выражению для спектра $\lambda(\alpha)$:

$$\lambda(\alpha) = \frac{1}{1 + 4\sigma' \sin^2 \alpha/2},$$

здесь, как и ранее, $\sigma' = \tau a / h^2$.

В этом случае условие устойчивости выполнено при любом соотношении сеточных параметров. В таких случаях говорят, что схема безусловно устойчивая.

Теорема 3. *Для задачи Коши*

$$\mathbf{B}\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\tau} = -\mathbf{A}\mathbf{u}^n; \mathbf{u}^0 = \varphi \tag{11.6}$$

условие $\mathbf{B} \geqslant \frac{\tau}{2}\mathbf{A}$ *необходимо и достаточно для устойчивости в энергетической норме, порождаемой* $\mathbf{A}$*, т. е.*

$$||\mathbf{u}^{n+1}||_{\mathbf{A}} \leqslant ||\mathbf{u}^0||_{\mathbf{A}}.$$

Неравенство в последней теореме имеет смысл операторного неравенства, т. е. для любого ненулевого вектора выполнено $(\mathbf{BU}, \mathbf{u}) \geqslant \frac{\tau}{2}(\mathbf{AU}, \mathbf{u})$.

Доказательство.

Достаточность. Умножим (11.6) скалярно на $\mathbf{y}_\tau = \frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\tau}$, тогда получим

$$(\mathbf{B}\mathbf{y}_\tau,\ \mathbf{y}_\tau) = -(\mathbf{A}\mathbf{u}^n,\ \mathbf{y}_\tau).$$

Ввиду того, что

$$\mathbf{u}^n \equiv \frac{1}{2}(\mathbf{u}^{n+1}+\mathbf{u}^n) - \frac{1}{2}\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\tau}\tau = \frac{1}{2}(\mathbf{u}^{n+1}+\mathbf{u}^n) - \frac{1}{2}\mathbf{y}_\tau$$

последнее равенство можно представить как уравнение:

$$\left(\left(\mathbf{B}-\frac{\tau}{2}\mathbf{A}\right)\mathbf{y}_\tau,\ \mathbf{y}_\tau\right) + \frac{1}{2\tau}\left(\mathbf{A}\left(\mathbf{u}^{n+1}+\mathbf{u}^n\right),\ \mathbf{u}^{n+1}-\mathbf{u}^n\right) = 0. \tag{11.7}$$

В силу самосопряженности $\mathbf{A}$, $(\mathbf{A}\mathbf{u}^n,\ \mathbf{u}^{n+1}) = (\mathbf{A}\mathbf{u}^{n+1},\ \mathbf{u}^n)$ и тогда

$$\left(\mathbf{A}\left(\mathbf{u}^{n+1}+\mathbf{u}^n\right),\ \mathbf{u}^{n+1}-\mathbf{u}^n\right) = (\mathbf{A}\mathbf{u}^{n+1},\ \mathbf{u}^{n+1}) - (\mathbf{A}\mathbf{u}^n,\ \mathbf{u}^n).$$

Из (11.7) следует

$$\left(\left(\mathbf{B}-\frac{\tau}{2}\mathbf{A}\right)\mathbf{y}_\tau,\ \mathbf{y}_\tau\right) + \frac{1}{2\tau}\left(||\mathbf{u}^{n+1}||^2_{\mathbf{A}} - ||\mathbf{u}^n||^2_{\mathbf{A}}\right) = 0. \tag{11.8}$$

В случае $\mathbf{B} \geqslant \frac{\tau}{2}\mathbf{A}$ получим, что $||\mathbf{u}^{n+1}||^2_{\mathbf{A}} \leqslant ||\mathbf{u}^n||^2_{\mathbf{A}}$. Отсюда следует устойчивость в норме $H_{\mathbf{A}}$ по начальным данным.

Необходимость. Пусть $||\mathbf{u}^{n+1}||_{\mathbf{A}} \leqslant ||\mathbf{u}^0||_{\mathbf{A}}$. Используем равенство (11.7) (оно называется *энергетическим тождеством*). В случае $n = 0$ из (11.7) следует:

$$2\tau\left(\left(\mathbf{B}-\frac{\tau}{2}\mathbf{A}\right)\mathbf{w},\ \mathbf{w}\right) + (\mathbf{A}\mathbf{u}^1,\ \mathbf{u}^1) = (\mathbf{A}\mathbf{u}^0,\ \mathbf{u}^0) = (\mathbf{A}\varphi, \varphi).$$

Это равенство может быть выполнено лишь в случае

$$\left(\left(\mathbf{B} - \frac{\tau}{2}\,\mathbf{A}\right)\mathbf{w},\ \mathbf{w}\right) > 0.$$

В силу того, что $\mathbf{B} = \mathbf{B}^* > 0$, существует $\mathbf{B}^{-1}$. Так как φ — произвольный элемент нашего пространства сеточных функций, то $\mathbf{w} = -\mathbf{B}^{-1}\mathbf{A}\varphi$ — произволен. Последнее равенство выполнено при любых φ, значит $\mathbf{B} - \frac{\tau}{2}\,\mathbf{A} \geqslant 0$.

Теорема доказана. ■

Теорема 4. *Пусть* $\mathbf{A}, \mathbf{B}$ *— постоянные самосопряженные положительные операторы. Тогда условие* $\mathbf{B} \geqslant \frac{\tau}{2}\,\mathbf{A}$ *необходимо и достаточно для устойчивости по начальным данным в энергетической норме, порождаемой оператором* $\mathbf{B}$*:*

$$\|\mathbf{u}^{n+1}\|_{\mathbf{B}} \leqslant \|\varphi\|_{\mathbf{B}}.$$

Таким образом, получим следующее правило исследования устойчивости двухслойных разностных схем.

1. Приводим схему к каноничному виду.

2. Исследуем свойства оператора $\mathbf{A}$. Если он является положительным, самосопряженным и независящим от n, проверяется условие $\mathbf{B} \geqslant 0,5\tau\,\mathbf{A}$.

Рассмотрим теперь устойчивость схемы Кранка-Николсон для уравнения теплопроводности. Эта схема введена в рассмотрение в начале данного параграфа. В операторном виде она записывается так:

$$\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\tau} = -\frac{1}{2}\,\mathbf{A}\mathbf{u}^n - \frac{1}{2}\,\mathbf{A}\mathbf{u}^{n+1}.$$

Запишем ее в каноническом виде:

$$\left(\mathbf{E} + \frac{\tau}{2}\,\mathbf{A}\right)\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\tau} = -\,\mathbf{A}\mathbf{u}^n.$$

По доказанной ранее теореме, данная схема устойчива по начальным данным в $H_{\mathbf{A}}$ в случае, если

$$\mathbf{B} = \mathbf{E} + \frac{\tau}{2}\,\mathbf{A} \geqslant \frac{\tau}{2}\,\mathbf{A} \Rightarrow \mathbf{E} > 0 \quad \text{(что всегда верно)}.$$

Тогда, в энергетической норме, порождаемой оператором $\mathbf{A}$, схема Кранка-Николсон безусловно устойчива.

Схема с весами. Можно действовать также, как и для схемы Кранка-Николсон, а можно несколько иначе. Для общей записи схемы с весами

$$\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\tau}+\mathbf{A}(\sigma\mathbf{u}^{n+1}+(1-\sigma)\mathbf{u}^n)=0$$

иногда употребляется сокращенная форма

$$\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\tau}+\mathbf{A}\mathbf{u}^{(\sigma)}=0,\quad \text{здесь}\quad \mathbf{u}^{(\sigma)}=\sigma\mathbf{u}^{n+1}+(1-\sigma)\mathbf{u}^n.$$

Умножив это разностное уравнение на $\mathbf{A}^{-1}$ слева, получаем:

$$\mathbf{A}^{-1}\,\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\tau}+\ \sigma\mathbf{u}^{n+1}+(1-\sigma)\mathbf{u}^n=0,$$

$$(\mathbf{A}^{-1}+\sigma\tau\mathbf{E})\,\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\tau}+\mathbf{u}^n=0.$$

Тогда необходимое и достаточное условие устойчивости в норме, порождаемой оператором $\mathbf{E}$ (т. е. в евклидовой норме)

$$(\mathbf{A}^{-1}+\sigma\tau\mathbf{E})\geqslant\frac{\tau}{2}\,\mathbf{E},\mathbf{A}^{-1}+\left(\sigma-\frac{1}{2}\right)\tau\mathbf{E}\geqslant 0.$$

Так как $\mathbf{A}=\mathbf{A}^*>0$, домножим обе части последнего неравенства на $\mathbf{A}$, тогда $\mathbf{E}+(\sigma-1/2)\tau\mathbf{A}\geqslant 0$ — условие устойчивости схемы с весами.

Следствие. *Неявная схема* ($\sigma=1$) *безусловно устойчива в норме* $||.||=(.,\,.)^{1/2}$ (*аналог нормы* L^2).

Докажем теперь, что из равномерной устойчивости однородной разностной схемы следует устойчивость по правой части.

Рассмотрим уравнение

$$\mathbf{B}\,\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\tau}+\ \mathbf{A}\mathbf{u}^n=\varphi^n.$$

Умножив это равенство скалярно на $2\tau\mathbf{y}_\tau=2\tau\,\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\tau}$, получим

$$2\tau\,\left(\left(\mathbf{B}-\frac{\tau}{2}\,\mathbf{A}\right)\mathbf{y}_\tau,\,\mathbf{y}_\tau\right)+(\mathbf{A}\mathbf{u}^{n+1},\,\mathbf{u}^{n+1})=\,(\mathbf{A}\mathbf{u}^n,\,\mathbf{u}^n)+2\tau\,(\varphi^n,\,\mathbf{y}_\tau)$$

(это — энергетическое множество для неоднородного уравнения). Кроме того,

$$((\mathbf{u},\,\mathbf{v})=(\mathbf{A}^{-1}\mathbf{A}\mathbf{u},\,\mathbf{v})=(\mathbf{A}\mathbf{u},\,\mathbf{A}^{-1}\mathbf{v})\leqslant\,||\mathbf{u}||_{\mathbf{A}}\,||\mathbf{v}||_{\mathbf{A}^{-1}}),$$
$$(\varphi^n,\,\mathbf{y}_\tau)\leqslant\,||\mathbf{y}_\tau||_{\mathbf{B}}\,||\varphi^n||_{\mathbf{B}^{-1}}$$

(аналог неравенства Коши–Буняковского).

Теперь используем ε-неравенство

$$0 \leqslant \left(\varepsilon a - \frac{b}{2\varepsilon}\right)^2 = \varepsilon^2 a^2 - ab + \frac{b}{4\varepsilon^2} \Rightarrow ab \leqslant \varepsilon^2 a^2 + \frac{b}{4\varepsilon^2},$$

тогда получаем

$$2\tau \left(\left(\mathbf{B} - \frac{\tau}{2}\mathbf{A}\right)\mathbf{y}_\tau,\ \mathbf{y}_\tau\right) + (\mathbf{A}\mathbf{u}^{n+1},\ \mathbf{u}^{n+1}) \leqslant$$
$$\leqslant (\mathbf{A}\mathbf{u}^n,\ \mathbf{u}^n) + \frac{\tau}{2\varepsilon_1}\,\|\boldsymbol{\varphi}^n\|^2_{\mathbf{B}^{-1}} + 2\tau\varepsilon_1\|\mathbf{y}_\tau\|^2_{\mathbf{B}},$$

$$2\tau \left(\left(\mathbf{B}\,(1-\varepsilon_1) - \frac{\tau}{2}\mathbf{A}\right)\mathbf{y}_\tau,\ \mathbf{y}_\tau\right) + (\mathbf{A}\mathbf{u}^{n+1},\ \mathbf{u}^{n+1}) \leqslant$$
$$\leqslant (\mathbf{A}\mathbf{u}^n,\ \mathbf{u}^n) + \frac{\tau}{2\varepsilon_1}\,\|\boldsymbol{\varphi}^n\|^2_{\mathbf{B}^{-1}}.$$

Выбираем ε: $\frac{1}{1-\varepsilon_1} = 1 + \varepsilon$, тогда можно написать

$$2\tau(1-\varepsilon_1)\left(\left(\mathbf{B} - \frac{1+\varepsilon}{2}\,\tau\mathbf{A}\right)\mathbf{y}_\tau,\ \mathbf{y}_\tau\right) + (\mathbf{A}\mathbf{u}^{n+1},\ \mathbf{u}^{n+1}) \leqslant$$
$$\leqslant (\mathbf{A}\mathbf{u}^n,\ \mathbf{u}^n) + \frac{\tau}{2\varepsilon_1}\,\|\boldsymbol{\varphi}^n\|^2_{\mathbf{B}^{-1}}.$$

Пусть $\mathbf{B} - \frac{1+\varepsilon}{2}\,\tau\mathbf{A} \geqslant 0, \varepsilon > 0, \varepsilon$ не зависит от сеточных параметров

$$1 = (1+\varepsilon)(1-\varepsilon_1) \Rightarrow \varepsilon - \varepsilon_1 - \varepsilon_1\varepsilon = 0, \varepsilon_1 = \frac{\varepsilon}{1+\varepsilon},$$

получаем

$$\|\mathbf{u}^{n+1}\|^2_{\mathbf{A}} \leqslant \|\mathbf{u}^n\|^2_{\mathbf{A}} + \frac{1+\varepsilon}{2\varepsilon}\,\tau\|\boldsymbol{\varphi}^n\|^2_{\mathbf{B}^{-1}}.$$

Откуда сразу следует

$$\|\mathbf{u}^{n+1}\|^2_{\mathbf{A}} \leqslant \|\mathbf{u}^0\|^2_{\mathbf{A}} + \frac{1+\varepsilon}{2\varepsilon}\,\tau\sum_{k=0}^{n}\|\boldsymbol{\varphi}^k\|^2_{\mathbf{B}^{-1}}. \tag{11.9}$$

Таким образом, для неоднородной разностной схемы доказана следующая теорема.

Теорема. *Для разностной схемы вида*

$$\mathbf{B}\,\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\tau} + \mathbf{A}\mathbf{u}^n = \boldsymbol{\varphi}^n,$$

где $\mathbf{A}$ *— постоянный (т. е. не зависящий явно от* n*) положительно определенный самосопряженный оператор, а* $\mathbf{B}$ *удовлетворяет условию*

$$\mathbf{B} \geqslant \frac{1+\varepsilon}{2}\,\tau\mathbf{A},$$

где $\varepsilon > 0$ не зависит от сеточных параметров, выполнена априорная оценка (11.9).

Вообще, в силу того, что разностная схема устойчива по начальным данным при $\mathbf{B} = \mathbf{B}^* > 0$, из равномерной устойчивости по начальным данным следует устойчивость по правой части.

Так как $\mathbf{B} = \mathbf{B}^* > 0$, то $\exists\, \mathbf{B}^{-1}$, тогда

$$\mathbf{u}^{n+1} - \mathbf{u}^n + \tau \mathbf{B}^{-1}\mathbf{A}\mathbf{u}^n = \tau \mathbf{B}^{-1}\varphi^n,$$

или

$$\mathbf{u}^{n+1} = \mathbf{T}\mathbf{u}^n + \tau \mathbf{B}^{-1}\varphi^n,$$

где $\mathbf{T} = \mathbf{E} - \tau \mathbf{B}^{-1}\mathbf{A}$ — оператор послойного перехода.

Равномерная устойчивость по начальным данным означает, что $||\mathbf{T}\mathbf{u}^n||_{\mathbf{R}} \leqslant \rho ||\mathbf{u}^n||_{\mathbf{R}}$. Тогда из предыдущего равенства с использованием неравенства треугольника получим

$$||\mathbf{u}^{n+1}||_{\mathbf{R}} \leqslant \rho ||\mathbf{u}^n||_{\mathbf{R}} + \tau ||\mathbf{B}^{-1}\varphi^n||_{\mathbf{R}}.$$

Применяя оценку n раз, получим априорную оценку для устойчивости по правой части с использованием энергетической нормы, порождаемой оператором $\mathbf{R} = \mathbf{R}^* > 0$:

$$||\mathbf{u}^{n+1}||_{\mathbf{R}} \leqslant \rho^{n+1} ||\mathbf{u}^0||_{\mathbf{R}} + \tau \sum_{k=0}^{n} \rho^{n-k} ||\mathbf{B}^{-1}\varphi^n||_{\mathbf{R}}.$$

Теперь попробуем обобщить полученные результаты на случай операторов, зависящих от времени.

Для начала рассмотрим аппроксимацию типа Кранка–Николсона

$$\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\tau} + \mathbf{A}^n \frac{\mathbf{u}^{n+1} + \mathbf{u}^n}{2} = 0, \mathbf{u}^0 = \varphi,$$

при этом $(\mathbf{A}^n\psi,\ \psi) \geqslant 0, \quad \forall\, n.$

Самый простой способ рассмотрения этого уравнения — разрешить его (в операторном виде) относительно $\mathbf{u}^{n+1}$:

$$\mathbf{u}^{n+1} = \left(\mathbf{E} + \frac{\tau}{2}\,\mathbf{A}^n\right)^{-1} \left(\mathbf{E} - \frac{\tau}{2}\,\mathbf{A}^n\right)\ \mathbf{u}^n = \mathbf{T}^n\mathbf{u}^n.$$

Оценим $||\mathbf{T}^n||$. Воспользуемся тем, что

$$\left(\mathbf{E}+\frac{\tau}{2}\mathbf{A}^n\right)^{-1} \left(\mathbf{E}-\frac{\tau}{2}\mathbf{A}^n\right) = \left(\mathbf{E}-\frac{\tau}{2}\mathbf{A}^n\right)\left(\mathbf{E}+\frac{\tau}{2}\mathbf{A}^n\right)^{-1} \text{(перестановочность)}.$$

$$\|\mathbf{T}^n\|^2 = \sup_{\|\mathbf{v}\|\neq 0} \frac{\left(\left(\mathbf{E}+\frac{\tau}{2}\mathbf{A}^n\right)^{-1}\left(\mathbf{E}-\frac{\tau}{2}\mathbf{A}^n\right)\mathbf{v}, \left(\mathbf{E}+\frac{\tau}{2}\mathbf{A}^n\right)^{-1}\left(\mathbf{E}-\frac{\tau}{2}\mathbf{A}^n\right)\mathbf{v}\right)}{(\mathbf{v},\mathbf{v})} =$$

$$\sup_{\|\psi\|\neq 0} \frac{\left(\left(\mathbf{E}-\frac{\tau}{2}\mathbf{A}^n\right)\psi, \left(\mathbf{E}-\frac{\tau}{2}\mathbf{A}^n\right)\psi\right)}{\left(\left(\mathbf{E}+\frac{\tau}{2}\mathbf{A}^n\right)\psi, \left(\mathbf{E}+\frac{\tau}{2}\mathbf{A}^n\right)\psi\right)} =$$

$$= \frac{(\psi,\ \psi) - \tau(\mathbf{A}\psi,\ \psi) + \frac{\tau^2}{4}(\mathbf{A}\psi,\ \mathbf{A}\psi)}{(\psi,\ \psi) + \tau(\mathbf{A}\psi,\ \psi) + \frac{\tau^2}{4}(\mathbf{A}\psi,\ \mathbf{A}\psi)} \leqslant 1,$$

$$\psi = \left(\mathbf{E}+\frac{\tau}{2}\mathbf{A}^n\right)^{-1}\mathbf{v}.$$

Факт, что $\|(\mathbf{E}-\sigma\mathbf{A})(\mathbf{E}+\sigma\mathbf{A})^{-1}\| \leqslant 1, \quad \forall\,\sigma>0, \mathbf{A}=\mathbf{A}^*>0$, носит название *леммы Келлога.*

Для сеточной функции используем норму $\|\mathbf{v}\| = (\mathbf{v},\ \mathbf{v})^{1/2}$. Подобный результат уже был получен при постоянном (не зависящем от времени) $\mathbf{A}$.

Везде в доказательствах существенную роль играет то, что $\mathbf{A}=\mathbf{A}^*>0$.

Может получиться так, что $\mathbf{A}>0$, но $\mathbf{A}\neq\mathbf{A}^*$:

$$\mathbf{A} = \mathbf{A}_s + \mathbf{A}_k, \mathbf{A}_s = \frac{1}{2}(\mathbf{A}+\mathbf{A}^*), \mathbf{A}_k = \frac{1}{2}(\mathbf{A}-\mathbf{A}^*).$$

В этом случае многие свойства разностных схем аналогичны доказанным выше.

Может быть так, что $\mathbf{A}_s = 0, \mathbf{A} = \mathbf{A}_k$. Тогда

$$(\mathbf{A}\mathbf{v},\ \mathbf{v}) = -(\mathbf{A}\mathbf{v},\ \mathbf{v}) = 0$$

в силу кососимметричности оператора $\mathbf{A}$. Например:

$$\frac{\mathbf{u}_m^{n+1}-\mathbf{u}_m^n}{\tau} + \frac{\mathbf{u}_{m+1}^n - \mathbf{u}_{m-1}^n}{2h} = 0.$$

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}, \quad \mathbf{A}=\mathbf{A}_k, \quad (\mathbf{A}\mathbf{v},\mathbf{v}) = 0 \quad \forall\,\mathbf{v}.$$

Теорема (об устойчивости по начальным данным для несамосопряженного оператора A (без доказательства)).

Разностная схема

$$\mathbf{B}\frac{\mathbf{u}_{n+1}-\mathbf{u}_n}{\tau} + \mathbf{A}\mathbf{u}_n = 0$$

равномерно устойчива по начальным данным, если

$$(\sigma - 0,5)\tau||\mathbf{Au}||^2 + (\mathbf{Au}, \mathbf{u}) \geqslant 0,$$

причем для ее решения справедлива оценка

$$||\mathbf{u}_{n+1}|| \leqslant ||\mathbf{u}_n||, \quad n = 0, \ldots, N-1.$$

Таким образом, рассмотрен вопрос об устойчивости разностной схемы, записанной в операторной канонической форме

$$\mathbf{B}\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\tau} + \mathbf{Au}^n = 0 \tag{11.10}$$

в случае $\mathbf{B} = \mathbf{B}^* > 0, \mathbf{A} = \mathbf{A}^* > 0$, причем $\mathbf{A}$ и $\mathbf{B}$ — постоянные матрицы. Рассмотрим более сложные случаи, в частности схему с весами

$$\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\tau} + \mathbf{Au}^{(\sigma)} = 0, \tag{11.11}$$

где $\mathbf{A}$ — кососимметрический оператор: $\mathbf{A} = -\mathbf{A}^*$, тогда $(\mathbf{Ay}, \mathbf{y}) = (\mathbf{y}, \mathbf{A}^*\mathbf{y}) = -(\mathbf{Ay}, \mathbf{y}) \forall \mathbf{y}$. Пример таких схем — простейшие разностные схемы для уравнения переноса:

$$\frac{\mathbf{u}_m^{n+1} - \mathbf{u}_m^n}{\tau} + a\frac{\mathbf{u}_{m+1}^n - \mathbf{u}_{m-1}^n}{2h} = 0.$$

Представим разностную схему (11.11) в каноническом виде (11.10)

$$\mathbf{B}\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\tau} + \mathbf{Au}^n = 0, \quad \text{где} \quad \mathbf{B} = \mathbf{E} + \sigma\tau\mathbf{A}.$$

Заметим, что $(\mathbf{Bu}, \mathbf{u}) = (\mathbf{u}, \mathbf{u}) + \sigma\tau\ (\mathbf{Au}, \mathbf{u}) = (\mathbf{u}, \mathbf{u}) = ||\mathbf{u}||^2$. Кроме того, $\forall\ \mathbf{u}: \quad ||\mathbf{u}|| \neq 0$ верно $(\mathbf{Bu}, \mathbf{u}) > 0$. Тогда $\exists\ \mathbf{B}^{-1}$, и, сделав замену, $\mathbf{u} = \mathbf{B}^{-1}\mathbf{x}$, имеем:

$$||\mathbf{u}||^2 = ||\mathbf{B}^{-1}\mathbf{x}||^2 = (\mathbf{BB}^{-1}\mathbf{x}, \mathbf{B}^{-1}\mathbf{x}) = (\mathbf{B}^{-1}\mathbf{x}, \mathbf{x}) \leqslant ||\mathbf{B}^{-1}\mathbf{x}|| \cdot ||\mathbf{x}||,$$

$$||\mathbf{B}^{-1}\mathbf{x}|| \leqslant ||\mathbf{x}|| \quad \text{и} \quad ||\mathbf{B}^{-1}|| \leqslant 1.$$

Схема (11.10) представима в виде

$$\mathbf{Bu}^{n+1} = \mathbf{Bu}^n - \tau\mathbf{Au}^n = 0; \mathbf{u}^{n+1} = \mathbf{u}^n - \tau\mathbf{B}^{-1}\mathbf{Au}^n = 0$$

(разрешенный относительно верхнего временного слоя вид).

Найдем $(\mathbf{B}\mathbf{u}^{n+1},\ \mathbf{u}^{n+1})$.

$$(\mathbf{B}\mathbf{u}^{n+1},\ \mathbf{u}^{n+1}) = (\mathbf{B}\mathbf{u}^n - \tau\mathbf{A}\mathbf{u}^n,\ \mathbf{u}^n - \tau\mathbf{B}^{-1}\mathbf{A}\mathbf{u}^n) =$$

$$= (\mathbf{B}\mathbf{u}^n,\ \mathbf{u}^n) - \tau\left[(\mathbf{A}\mathbf{u}^n,\ \mathbf{u}^n) + (\mathbf{B}\mathbf{u}^n,\ \mathbf{B}^{-1}\mathbf{A}\mathbf{u}^n)\right] + \tau^2(\mathbf{A}\mathbf{u}^n,\ \mathbf{B}^{-1}\mathbf{A}\mathbf{u}^n).$$

В этом случае идеология доказательства оказывается похожей на доказательство устойчивости методов Рунге-Кутты на нейтральных по устойчивости траекториях в лекции 8. Первое слагаемое в квадратных скобках равно нулю, оценим второе слагаемое:

$$\mathbf{B} = \mathbf{E} + \sigma\tau\mathbf{A} \equiv \mathbf{E} - \sigma\tau\mathbf{A} + 2\sigma\tau\mathbf{A} = \mathbf{B}^* + 2\sigma\tau\mathbf{A};$$

$$(\mathbf{B}\mathbf{u}^n,\ \mathbf{B}^{-1}\mathbf{A}\mathbf{u}^n) = (\mathbf{B}^*\mathbf{u}^n,\ \mathbf{B}^{-1}\mathbf{A}\mathbf{u}^n) + 2\sigma\tau\ (\mathbf{A}\mathbf{u}^n,\ \mathbf{B}^{-1}\mathbf{A}\mathbf{u}^n);$$

первое из слагаемых в левой части выражения — нуль. В результате имеем:

$$||\mathbf{u}^{n+1}||^2 = ||\mathbf{u}^n||^2 - (2\sigma - 1)\tau^2(\mathbf{B}^{-1}\mathbf{A}\mathbf{u}^n,\ \mathbf{A}\mathbf{u}^n,\ \mathbf{B}^{-1}\mathbf{A}\mathbf{u}^n). \qquad (11.12)$$

Если $\sigma \geqslant 0,5$, то $||\mathbf{u}^{n+1}|| \leqslant ||\mathbf{u}^n||$, и схема устойчива по начальным данным. Можно показать, что она также будет устойчива и по правой части.

Пусть теперь $\sigma < 0,5$. Из (11.12) следует, что

$$||\mathbf{u}^{n+1}||^2 \leqslant ||\mathbf{u}^n||^2 - (2\sigma - 1)\tau^2||\mathbf{A}||^2\ ||\mathbf{u}^n||^2 \leqslant (1 + c_0\tau)||\mathbf{u}^n||^2,$$

где $c_0 = (1 - 2\sigma)\tau||\mathbf{A}||^2$.

Если $\tau||\mathbf{A}||^2 \leqslant c_2$, то в случае $\sigma < 0,5$ разностная схема будет равномерно устойчива с $\rho = \sqrt{1 + c_0\tau} \leqslant e^{0,5c_0\tau}$. Из устойчивости по начальным данным и в этом случае следует устойчивость по правой части.

Рассмотрим пример: уравнение Шредингера движения частицы во внешнем поле ($m = 1$)

$$i\hbar\,\frac{\partial\,\psi}{\partial\,t} = -\frac{\hbar^2}{2}\,\frac{\partial^2\psi}{\partial\,x^2} + U\psi,$$

или, после обезразмеривания, получим

$$-i\,\frac{\partial\,\psi}{\partial\,t} - \frac{1}{2}\,\frac{\partial^2\psi}{\partial\,x^2} + \tilde{U}\psi = 0.$$

Если внешнего поля нет, можно потенциальную энергию частицы $\tilde{U}$ положить равной нулю. На отрезке [0; 1] можно записать разностную схему «с комплексными коэффициентами», аппроксимирующую уравнение

Шредингера:

$$-i\,\frac{\psi^{n+1}-\psi^{n+1}}{\tau}-\frac{1}{2}\,\Lambda\psi^{(\sigma)}=0,\qquad \Lambda\psi^{n}\equiv-\frac{\psi^{n}_{m+1}-2\psi^{n}_{m}+\psi^{n}_{m-1}}{h^2}. \tag{11.13}$$

Можно формально умножить правую и левую части (11.13) на i и воспользоваться приведенными выше результатами. Получим, что в случае $\sigma \geqslant 0,5$ схема устойчива, а при $\sigma < 0,5$ возникает условие устойчивости вида

$$\frac{\tau}{4}\,||\Lambda||^2 \leqslant 1.$$

Так как $||\Lambda|| = \frac{4}{h^2}$, то окончательно получаем условие

$$\tau \leqslant \frac{h^4}{4}\,c_2,$$

где c_2 — константа, определяющая ρ-устойчивость схемы. При этом погрешности все равно экспоненциально возрастают! В энергетической норме $H_{\mathbf{A}}$ условие устойчивости будет

$$Re\,(i\mathbf{E}) \geqslant \frac{1}{2}\,\tau\mathbf{A},$$

которое не выполняется при любом τ. Следовательно, в энергетической норме устойчивых разностных схем для уравнения Шредингера при $\sigma < 0{,}5$ нет.

Рассмотрим разностную схему

$$\frac{\mathbf{u}^{n+1}_{m}-\mathbf{u}^{n}_{m}}{\tau}+a\left(\frac{\mathbf{u}^{n}_{m+1}-\mathbf{u}^{n}_{m-1}}{2h}\right)^{(\sigma)}=0.$$

Здесь $\mathbf{B}=\mathbf{E}$; $\mathbf{A}=\operatorname{diag}\left(-\frac{a}{2h}\quad 0\quad \frac{a}{2h}\right)$ — трехдиагональная матрица; $\frac{1}{2}\,(\mathbf{A}+\mathbf{A}^{*})=0$ — оператор $\mathbf{A}$ кососимметрический.

В случае $\sigma \geqslant 0,5$ схема устойчива. При $\sigma < 0,5$ имеем условие устойчивости

$$\tau\,\frac{a^2}{4h^2} < c_2 \Rightarrow \frac{a\tau}{h} < \frac{4c_2 h}{a},$$

но ошибки все же экспоненциально возрастают как $e^{0,5(1-2\sigma)k^2 t}$, где k — число Куранта. По спектральному признаку при $\sigma = 0$ схема неустойчива.

Очевидно, что для уравнения Шредингера нет устойчивых схем в случае $\sigma < 0,5$ и при использовании метода конечных элементов. Действительно, будем рассматривать дискретную схему

$$i\hat{\mathbf{B}}\,\frac{\psi^{n+1}-\psi^{n+1}}{\tau}-\frac{1}{2}\,\Lambda\psi^{(\sigma)}=0,$$

где $\hat{\mathbf{B}} = \hat{\mathbf{B}}^*$. Условие устойчивости в энергетической норме приведет к операторному неравенству вида

$$\frac{i\hat{\mathbf{B}} - i\hat{\mathbf{B}}^*}{2} \geqslant \frac{(-\tau\Lambda)}{4}(-\Lambda > 0).$$

В силу положительности $-\Lambda$, приходим к противоречию $0 \geqslant \frac{\tau}{4}(-\Lambda) > 0$.

В завершение экскурса в теорию устойчивости разностных схем рассмотрим разностную схему более общего вида:

$$\mathbf{B}(t)\,\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\tau} + \mathbf{A}(t)\mathbf{u}^n = \varphi.$$

Пусть, как и ранее, выполнено условие самосопряженности

$$\mathbf{A}(t) : \mathbf{A}(t) = \mathbf{A}^*(t) > 0, t \in [0;\ T], \quad \mathbf{B}(t) > 0;$$

$\mathbf{A}(t)$ — Липшиц-непрерывен по t.

$$|([\mathbf{A}(t) - \mathbf{A}(t-\tau)]\,\mathbf{x},\ \mathbf{x})| \leqslant \tau c(\mathbf{A}(t-\tau)\,\mathbf{x},\ \mathbf{x}).$$

Введем энергетические нормы, зависящие от времени:

$$H_{\mathbf{A}}: \quad (\mathbf{A}(t)\,\mathbf{u},\ \mathbf{u}) = ||\mathbf{u}||^2_{\mathbf{A}(t)}.$$

Воспользуемся введенным ранее энергетическим тождеством:

$$2\tau\left(\left[\mathbf{B}(t) - \frac{\tau\mathbf{A}(t)}{2}\right]\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\tau},\ \frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\tau}\right) + (\mathbf{A}(t)\mathbf{y}^{n+1},\ \mathbf{y}^{n+1}) =$$

$$= (\mathbf{A}(t)\mathbf{u}^{n+1},\ \mathbf{u}^{n+1}) + 2\tau\left(\varphi,\ \frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\tau}\right).$$

Учтем, что

$$(\mathbf{A}(t)\mathbf{y}(t), \mathbf{y}(t)) \equiv (\mathbf{A}(t-\tau)\mathbf{y}(t), \mathbf{y}(t)) + ([\mathbf{A}(t) - \mathbf{A}(t-\tau)]\mathbf{y}(t), \mathbf{y}(t)) \leqslant$$

$$\leqslant (1 + \tau c)(\mathbf{A}(t-\tau)\mathbf{y}(t), \mathbf{y}(t))$$

в силу Липшиц-непрерывности. Тогда получаем

$$2\tau\left(\left[\mathbf{B}(t) - \frac{\tau\mathbf{A}(t)}{2}\right]\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\tau},\ \frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\tau}\right) + (\mathbf{A}(t)\mathbf{u}(t+\tau),\ \mathbf{u}(t+\tau)) \leqslant$$

$$\leqslant (1 + \tau c) \times (\mathbf{A}(t-\tau)\mathbf{u}^n,\ \mathbf{u}^n) + 2\tau\left(\varphi(t),\ \frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\tau}\right).$$

Теперь если $\mathbf{B}(t) - \frac{\tau \mathbf{A}(t)}{2} \geqslant 0$, то при $\varphi(t) \equiv 0$ сразу имеем

$$(\mathbf{A}(t)\mathbf{u}^{n+1},\ \mathbf{u}^{n+1}) = \|\mathbf{u}^{n+1}\|^2_{\mathbf{A}(t)} \leqslant (1+\tau c)(\mathbf{A}(t-\tau)\mathbf{u}^n,\ \mathbf{u}^n) =$$
$$= (1+\tau c)\,\|\mathbf{u}^n\|^2_{\mathbf{A}(t)} \leqslant e^{ct}(\mathbf{A}(0)\mathbf{u}^1,\ \mathbf{u}^1) \leqslant e^{ct}\|\mathbf{u}^0\|^2_{\mathbf{A}(t)}.$$

Отсюда и следует устойчивость: из энергетического тождества при $n = 1$ получим, что схема устойчива по начальным данным. Устойчивость по правой части будет следовать из устойчивости по начальным данным.

Таким образом, случай операторов, зависящих от времени, принципиально не отличается от случая постоянных операторов, лишь нормы устроены несколько по-другому.

Аналогичные теории строятся и для трехслойных разностных схем, условия устойчивости также получаются на основе энергетических тождеств в виде операторных неравенств.

11.4. Задачи

1. Построить разностную схему, аппроксимирующую задачу Коши для уравнения переноса

$$\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = f(t,x), u(0,x) = \varphi(x), -\infty < x < \infty, t > 0$$

с помощью аппроксимации первых производных со вторым порядком точности.

Решение. Аппроксимация уравнения переноса со вторым порядком точности по схеме с центральными разностями

$$\frac{u_m^{n+1} - u_m^{n-1}}{2\tau} - \frac{u_{m+1}^n - u_{m-1}^n}{2h} = f_m^n, n = 0, \ldots, N-1,\ m = 1, \ldots, M-1.$$

Для аппроксимации начальных условий необходимо задать не только $u_m^0 = \varphi_m$, но и значение u_m^1, которое можно вычислить с помощью, например, формулы Тейлора:

$$u_m^1 = u_m^0 + \tau \left(\frac{\partial u}{\partial t}\right)_m^0 + O(\tau^2) \approx u_m^0 + \tau \left(\frac{\partial u}{\partial t}\right)_m^0.$$

Поскольку $\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = f(t,x)$ и $u(0,x) = \varphi(x)$, то

$$\left(\frac{\partial u}{\partial t}\right)_m^0 = \left(\frac{\partial u}{\partial x}\right)_m^0 + f(0,x) = [\varphi_x'(x)]_m^0 + f_m^0.$$

В таком случае аппроксимация задачи будет

$$\frac{u_m^{n+1}-u_m^{n-1}}{2\tau}-\frac{u_{m+1}^n-u_{m-1}^n}{2h}=f_m^n, n=0,\ldots,N-1,\ m=1,\ldots,M-1.$$

$$u_m^0=\varphi_m, m=0,\ldots,M,$$

$$u_m^1=\varphi_m+\tau\left[(\varphi_x')_m^0+f_m^0\right]$$

В этом примере пришлось конструировать дополнительное второе начальное условие, поскольку исходное дифференциальное уравнение имеет первый порядок, а разностное — второй.

2. Исследовать устойчивость разностной схемы

$$\frac{u_m^{n+1}-u_m^n}{\tau}-\frac{u_{m+1}^n-u_{m-1}^n}{2h}-\frac{\tau}{2h^2}\left(u_{m-1}^n-2u_m^n+u_{m+1}^n\right)=f_m^n,$$

$$n=0,\ldots,N-1,\ m=1,\ldots,M-1,$$

аппроксимирующей задачу Коши для уравнения переноса.

Решение. Воспользуемся спектральным признаком. Подставляем в разностную схему решение в виде: $u_m^n=\lambda^n e^{im\alpha}$ и рассматриваем однородное уравнение.

Получим

$$\frac{\lambda-1}{\tau}-\frac{e^{im\alpha}-e^{-im\alpha}}{2h}-\frac{\tau}{2h^2}\left(e^{-im\alpha}-2+e^{im\alpha}\right)=0,$$

откуда

$$\lambda(\alpha)=1+i\sigma\sin\alpha-2\sigma^2\sin^2\frac{\alpha}{2}, \sigma=\frac{\tau}{h}\text{ — число Куранта.}$$

Вычислим границы спектра. Для этого вычислим

$$|\lambda(\alpha)|^2=\left(1-2\sigma^2\sin^2\frac{\alpha}{2}\right)^2+\sigma^2\sin^2\alpha,$$

а затем расстояние $1-|\lambda|^2$. Проведем необходимые вычисления:

$$\left(1-2\sigma^2\sin^2\frac{\alpha}{2}\right)^2+\sigma^2\sin^2\alpha=1-4\sigma^2\sin^2\frac{\alpha}{2}+4\sigma^2\sin^2\alpha=$$

$$=1+4\sigma^4\sin^4\frac{\alpha}{2}+\left(-4\sigma^4\sin^4\frac{\alpha}{2}+\sigma^2\sin^2\alpha\right)=$$

$$= 1 + 4\sigma^4 \sin^2 \frac{\alpha}{2} + \left(-4\sigma^2 \sin^2 \frac{\alpha}{2} + \sigma^2 4 \sin^2 \frac{\alpha}{2} \cos^2 \frac{\alpha}{2}\right) =$$

$$= 1 + 4\sigma^4 \sin^2 \frac{\alpha}{2} + \left[-4\sigma^2 \sin^2 \frac{\alpha}{2} + \sigma^2 4 \sin^2 \frac{\alpha}{2} \left(1 - \sin^2 \frac{\alpha}{2}\right)\right] =$$

$$= 1 + 4\sigma^4 \sin^2 \frac{\alpha}{2} - 4\sigma^2 \sin^2 \frac{\alpha}{2}.$$

Условие устойчивости выполнено, если правая часть неотрицательна, что достигается при $\sigma \leqslant 1$.

3. Исследовать устойчивость явной разностной схемы

$$\frac{u_{ml}^{n+1} - u_{ml}^n}{\tau} - \frac{u_{m-1,l}^n - 2u_{ml}^n + u_{m+1,l}^n}{h^2} - \frac{u_{m,l-1}^n - 2u_{m,l}^n + u_{m,l+1}^n}{h^2} = 0,$$

$$n = 0, \ldots, N-1, m = 1, \ldots, M-1,\ l = 1, \ldots, L-1,$$

аппроксимирующую задачу Коши для уравнения в частных производных

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = 0.$$

Решение. Воспользуемся спектральным признаком устойчивости фон Неймана, представив разностное решение в виде

$$u_{ml}^n = \lambda^n e^{i\alpha m + i\beta n}.$$

После подстановки в разностное уравнение, получим

$$\lambda(\alpha, \beta) = 1 - 4\sigma \sin^2 \frac{\alpha}{2} - 4\sigma \sin^2 \frac{\beta}{2}, \quad \sigma = \frac{\tau}{h^2},$$

откуда следует $1 - 8\sigma \leqslant \lambda(\alpha, \beta) \leqslant 1$. Окончательно условие устойчивости разностной схемы будет $\sigma \leqslant 1/4$.

4. Исследовать на устойчивость разностную схему

$$\frac{\mathbf{u}_m^{n+1} - \mathbf{u}_m^n}{\tau} - \mathbf{A}\frac{\mathbf{u}_{m+1}^n - \mathbf{u}_m^n}{h} = 0, n = 0, \ldots, N-1, m = 0, \ldots, M-1,$$

(где $\mathbf{u} = (u_1, u_2)^{\mathrm{T}}$, — вектор, $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, — матрица 2×2), аппроксимирующую систему уравнений в частных производных

$$\frac{\partial u_1}{\partial t} - \frac{\partial u_2}{\partial x} = 0, \ \frac{\partial u_2}{\partial t} - \frac{\partial u_1}{\partial x} = 0; \ \frac{\partial \mathbf{u}}{\partial t} - \mathbf{A}\frac{\partial \mathbf{u}}{\partial x} = 0, t > 0, -\infty < x < \infty.$$

Решение. Подставим в разностную схему решение в виде

$$\mathbf{u}_m^n = \lambda^n \begin{pmatrix} u_1^0 \\ u_2^0 \end{pmatrix} e^{i\alpha m} = \lambda^n \mathbf{u}^0 e^{i\alpha m}.$$

После подстановки в разностное уравнение, получим

$$(\lambda - 1)\mathbf{u}^0 - \sigma(e^{i\alpha} - 1)\mathbf{A}\mathbf{u}^0 = 0, \sigma = \tau/h \text{ — число Куранта.}$$

Это уравнение можно рассматривать как векторную запись СЛАУ относительно $\mathbf{u}^0$

$$\begin{pmatrix} \lambda - 1 & -\sigma(e^{i\alpha} - 1) \\ -\sigma(e^{i\alpha} - 1) & \lambda - 1 \end{pmatrix} \begin{pmatrix} u_1^0 \\ u_2^0 \end{pmatrix} = 0.$$

Система имеет нетривиальное решение только тогда, когда определитель матрицы обращается в нуль, т. е.

$$(\lambda - 1)^2 = \sigma^2(e^{i\alpha} - 1)^2,$$

откуда

$$\lambda_1 = 1 - \sigma + \sigma e^{i\alpha}, \quad \lambda_2 = 1 + \sigma - \sigma e^{i\alpha}.$$

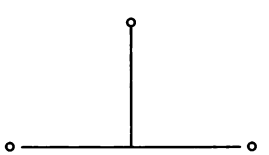
Эти два корня пробегают окружности радиуса σ с центрами в точках $1 - \sigma$ и $1 + \sigma$. Условие устойчивости не выполнено ни при каком значении числа Куранта.

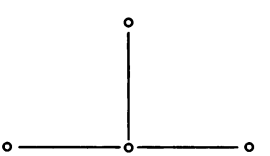
11.5. Задачи для самостоятельного решения

1. Для задачи Коши для линейного уравнения переноса

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0,$$

построить разностные схемы и исследовать их на сходимость, используя шаблоны:



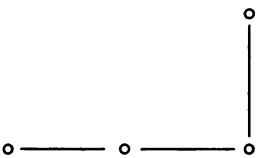


Разностная схема, построенная на симметричном трехточечном шаблоне, называется схемой П. Лакса.

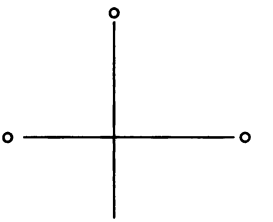
2. Для линейного уравнения теплопроводности

$$\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial x^2}, t > 0, -\infty < x < \infty,$$

построить разностные схемы и исследовать их на сходимость, используя шаблоны:



(несимметричная схема)

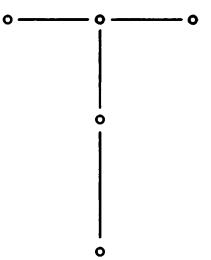


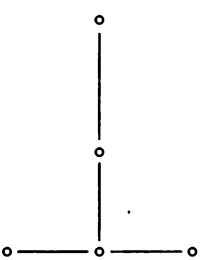
(схема Ричардсона)

3. Построить разностные схемы для линейного волнового уравнения

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2},$$

и исследовать их на сходимость, используя шаблоны





4. Построить разностную схему, сходящуюся к решению акустической системы

$$\frac{\partial u_1}{\partial t} - \frac{\partial u_2}{\partial x} = f_1(t,x), \frac{\partial u_2}{\partial t} - \frac{\partial u_1}{\partial x} = f_2(t,x).$$

5. Построить разностную схему П. Лакса, аппроксимирующую задачу Коши для линейного двумерного уравнения переноса

$$\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} - \frac{\partial u}{\partial y} = 0,$$

и исследовать ее на сходимость.

Литература

[1] *Годунов С.К., Рябенький В.С.* Разностные схемы, введение в теорию. М.: Наука, 1977. 400 с.

[2] *Федоренко Р.П.* Введение в вычислительную физику. М.: Изд-во МФТИ, 1994. 526 с.

[3] *Самарский А.А.* Теория разностных схем. М.: Наука, 1983. 656 с.

[4] *Самарский А.А., Вабищевич П.Н., Матус Г.П.* Разностные схемы с операторными множителями. Минск, 1998. 441 с.

[5] *Самарский А.А., Гулин А.В.* Численные методы математической физики. М.: Научный мир, 2003. 316 с.

[6] *Жуков А.И.* Метод Фурье в вычислительной математике. М.: Наука, 1992. 128 с.

Лекция 12. Численное решение дифференциальных уравнений в частных производных параболического типа на примере уравнения теплопроводности

В лекции рассматриваются разностные схемы для решения линейного уравнения теплопроводности, нелинейного уравнения теплопроводности. Приводится пример интегро- интерполяционного метода для построения разностных схем. Отдельно рассматриваются экономичные схемы решения многомерных задач для уравнения теплопроводности — переменных направлений, дробных шагов, Дугласа-Ганна.

Ключевые слова: явная разностная схема. Неявная разностная схема. Интегро-интерполяционный метод. Квазилинейное уравнение теплопроводности. Консервативная схема. Методы переменных направлений. Схема Дугласа-Ганна.

12.1. Постановки задач для уравнений параболического типа

Рассмотрим численные методы решения уравнений параболического типа.

Одномерное линейное уравнение теплопроводности (диффузии). Напомним постановку соответствующей смешанной задачи:

$$\frac{\partial u}{\partial t} = a\frac{\partial^2 u}{\partial x^2} + f(t,x), t \in [0,T], x \in [0,X]. \tag{12.1}$$

Здесь $a = a(x,t) > 0$. Для того чтобы задача была поставлена корректно, необходимо задать начальное условие

$$u(0,x) = u_0(x), t = 0,$$

и граничные условия

$$-A_1\frac{\partial u}{\partial x} + B_1 u = \varphi_1(t), x = 0,$$

$$A_2\frac{\partial u}{\partial x} - B_2 u = \varphi_2(t), x = X.$$

О свойствах решений линейного уравнения теплопроводности подробнее в [1, 3].

Одномерное квазилинейное уравнение теплопроводности (диффузии):

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left[a(u)\frac{\partial u}{\partial t}\right] + f(u), \quad t \in [0, \mathrm{T}], x \in [0, X]. \tag{12.2}$$

Уравнения такого вида встречаются в теории горения, астрофизике, физике плазмы, теории сверхпроводимости Гинзбурга-Ландау, динамике популяций и других приложениях. Здесь $a(u) > 0$ при любых значениях u, кроме того, $\int\limits_0^u a(z)dz < +\infty$. Для глобальной ограниченности решения также требуется выполнение условия $\int\limits_1^\infty \frac{dz}{f(z)} = \infty$. Для корректной постановки задачи необходимо задать одно начальное и два граничных условия. Подробнее о квазилинейных уравнениях теплопроводности в книгах [4, 5, 6, 7, 8].

Двухмерное линейное уравнение теплопроводности (диффузии):

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, t \in [0, \mathrm{T}], x \in [0, X], y \in [0, Y]. \tag{12.3}$$

Для численного решения уравнения (12.1), по-видимому, наиболее известной является параметрическая двухслойная шеститочечная разностная схема вида

$$\frac{u_m^{n+1} - u_m^n}{\tau} = a\left[\xi\frac{u_{m-1}^{n+1} - 2u_m^{n+1} + u_{m+1}^{n+1}}{h^2} + (1-\xi)\frac{u_{m-1}^n - 2u_m^n + u_{m+1}^n}{h^2}\right], \tag{12.4}$$

где $\xi \in [0, 1]$.

При $\xi = 0$ имеем явную схему $\frac{u_m^{n+1}-u_m^n}{\tau} = a\frac{u_{m-1}^n - 2u_m^n + u_{m+1}^n}{h^2}$, устойчивую при $\sigma = \frac{a\tau}{h^2} \leqslant \frac{1}{2}$, с порядком аппроксимации $O(\tau, h^2)$.

При $\xi = 1$ имеем неявную схему $\frac{u_m^{n+1}-u_m^n}{\tau} = a\frac{u_{m-1}^{n+1} - 2u_m^{n+1} + u_{m+1}^{n+1}}{h^2}$, устойчивую при любых τ, h, с порядком аппроксимации $O(\tau, h^2)$.

При $\xi = 1/2$ разностный метод называется схемой Кранка–Никольсон: $\frac{u_m^{n+1}-u_m^n}{\tau} = \frac{a}{2}\left(\frac{u_{m-1}^{n+1} - 2u_m^{n+1} + u_{m+1}^{n+1}}{h^2} + \frac{u_{m-1}^n - 2u_m^n + u_{m+1}^n}{h^2}\right)$. Схема устойчива при любых шагах τ, h и имеет порядок аппроксимации $O(\tau^2, h^2)$. Эта схема, в отличие от двух предыдущих, не является *монотонной*, т. е. она может давать осцилляции разностного происхождения на решениях, имеющих большие градиенты.

Схема, имеющая второй порядок точности по τ и четвертый по h, получается на расширенном шаблоне с учетом разностной аппроксимации главного члена невязки. При исследовании аппроксимации явной

двухслойной схемы получим

$$\Lambda_\tau U_\tau = \Lambda u + \frac{\tau}{2} u''_{tt} - \frac{ah^2}{12} u^{(4)}_x + O(\tau^2, h^4).$$

Произведя аппроксимацию первого и второго слагаемого в правой части (в невязке) рассматриваемого равенства и учитывая следствия исходного уравнения теплопроводности

$$u'_t = a u''_{xx}, u''_{tt} = a(u'_t)''_{xx},$$

$$u''_{tt} = a^2 u^{(4)}_x,$$

получим новую схему повышенного порядка точности:

$$\frac{u^n_m - u^n_m}{\tau} - a\frac{u^n_{m-1} - 2u^n_m + u^n_{m+1}}{h^2} -$$

$$-\frac{a^2\tau}{2}\left(1 - \frac{1}{6}\frac{h^2}{a\tau}\right)\frac{u^n_{m-2} - 4u^n_{m-1} + 6u^n_m - 4u^n_{m-1} + u^n_{m+2}}{h^4} = 0.$$

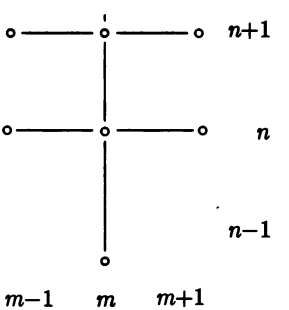
Трехслойная параметрическая схема для численного решения одномерного линейного уравнения теплопроводности имеет вид

$$\frac{(1-\eta)(u^{n+1}_m - u^n_m)}{\tau} - \eta\frac{u^n_m - u^{n-1}_m}{\tau} -$$

$$-a\left[(1-\xi)\frac{u^n_{m-1} - 2u^n_m + u^n_{m+1}}{h^2} + \xi\frac{u^{n+1}_{m-1} - 2u^{n+1}_m + u^{n+1}_{m+1}}{h^2}\right] = 0;$$

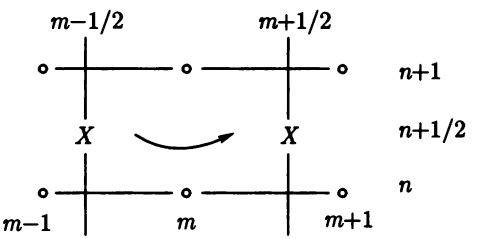
при $\eta = 0,5, \quad \xi = 1$ ее порядок аппроксимации равен $O(\tau^2, h^2)$. Недостатком схемы является трехслойность и, следовательно, необходимость ставить дополнительное условие на $u'_t(0, x)$.

Соответствующий шаблон имеет вид



В случае если коэффициент теплопроводности a зависит от времени и координат, *консервативную* схему можно получить, используя *интегро-интерполяционный* метод (положим, для простоты $f(t,x) = 0$). Напомним, что разностная схема называется консервативной, если выполняются следующие условия. В дифференциальной задаче выполняется некий закон сохранения. Соответствующий закон сохранения выполняется и на сеточном уровне. Если же в дифференциальной задаче имеется несколько законов сохранения, а при переходе к сеточному описанию все они получаются как следствие нашей разностной схемы в результате алгебраических преобразований, то схема называется *полностью консервативной.*

Как правило, при записи уравнений в частных производных законам сохранения соответствует дивергентная форма записи. Для уравнения теплопроводности роль такого закона сохранения играет непрерывность теплового потока.



Для этого запишем уравнение в *дивергентной форме*:

$$\frac{\partial u}{\partial t} + \frac{\partial W}{\partial x} = 0, \quad \text{где} \quad W = -a(t,x)\frac{\partial u}{\partial x} \text{ — тепловой поток,}$$

$$\text{или} \int\limits_S (\frac{\partial u}{\partial t} + \frac{\partial W}{\partial t})dtdx = \oint\limits_\Gamma u dx - W dt = 0.$$

Произведем аппроксимацию последнего интеграла по прямоугольному контуру с узловыми точками $(n, m-1/2), (n, m+1/2), (n+1, m+1/2), (n+1, m-1/2)$:

$$u_m^n h - W_{m+1/2}^{n+1/2}\tau - u_m^{n+1} \cdot h + W_{m-1/2}^{n+1/2}\tau = 0,$$

$$\frac{u_m^{n+1} - u_m^n}{\tau} + \frac{W_{m+1/2}^{n+1/2} - W_{m-1/2}^{n+1/2}}{h} = 0.$$

Отсюда, учитывая вид $W_{m\pm1/2}^{n+1/2}$:

$$W_{m+1/2}^{n+1/2} = \frac{1}{2}\left(a_{m+1/2}^n \frac{u_{m+1}^n - u_m^n}{h} + a_{m+1/2}^{n+1} \frac{u_{m+1}^{n+1} - u_m^{n+1}}{h}\right),$$

$$W_{m-1/2}^{n+1/2} = \frac{1}{2}\left(a_{m-1/2}^{n}\frac{u_m^n - u_{m-1}^n}{h} + a_{m-1/2}^{n+1}\frac{u_m^{n+1} - u_{m-1}^{n+1}}{h}\right),$$

$$\frac{u_m^{n+1} - u_m^n}{\tau} + \frac{1}{2h}\left[\left(a_{m+1/2}^{n}\frac{u_{m+1}^n - u_m^n}{h} - a_{m-1/2}^{n}\frac{u_m^n - u_{m-1}^n}{h}\right) + \right.$$

$$\left. + \left(a_{m+1/2}^{n+1}\frac{u_{m+1}^{n+1} - u_m^{n+1}}{h} - a_{m-1/2}^{n+1}\frac{u_m^{n+1} - u_{m-1}^{n+1}}{h}\right)\right] = 0$$

12.2. Разностные схемы для численного решения нелинейного уравнения теплопроводности

12.2.1. Неявная схема с нелинейностью на нижнем слое

$$\frac{u_m^{n+1} - u_m^n}{\tau} = \frac{1}{h}\left[a_{n+1/2}\frac{u_{m+1}^{n+1} - u_m^{n+1}}{h} - a_{n-1/2}\frac{u_m^{n+1} - u_{m-1}^{n+1}}{h}\right] + f_m^n,$$

где $a_{m+1/2}$ вычисляется следующим образом:

1. $a_{m+1/2} = \frac{1}{2}(a(u_m^n) + a(u_{m+1}^n))$,
2. $a_{m+1/2} = a\left(\frac{u_m^n + u_{m+1}^n}{2}\right)$,
3. $a_{m+1/2} = a\left(\frac{2u_m^n u_{m+1}^n}{u_m^n + u_{m+1}^n}\right)$,
4. $a_{m+1/2} = \frac{2a(u_m^n)a(u_{m+1}^n)}{a(u_m^n)+a(u_{m+1}^n)}$.

На верхнем слое по времени решение находится с помощью метода прогонки. Недостаток схемы заключается в необходимости выполнения условия, ограничивающего шаг по времени: $\tau \|f_u'\| \ll 1$.

12.2.2. Схема с нелинейностью на верхнем слое

(необходимость ее реализации появляется, когда условие $\tau \|f_u'\| \ll 1$ оказывается трудновыполнимым):

$$\frac{u_m^{n+1} - u_m^n}{\tau} = \frac{1}{h}\left(a_{m+1/2}^{n+1}\frac{u_{m+1}^{n+1} - u_m^{n+1}}{h} - a_{m-1/2}^{n+1}\frac{u_m^{n+1} - u_{m-1}^{n+1}}{h}\right) + f_m^{n+1}.$$

Для реализации алгоритма прогонки проведем линеаризацию функции в правой части, то есть используем итерационный метод Ньютона

в функциональных пространствах (метод квазилинеаризации). Пусть u_m^i есть i приближение к u_m^{n+1}; необходимо вычислить u_m^{n+1}.

Тогда

$$f(u_m^{i+1}) = f[u_m^i + (u_m^{i+1} - u_m^i)] \approx f(u_m^i) + f'_u(u_m^i)(u_m^{i+1} - u_m^i),$$

$$a(u_m^{i+1}) = a[u_m^i + (u_m^{i+1} - u_m^i)] \approx a(u_m^i) + a'_u(u_m^i)(u_m^{i+1} - u_m^i).$$

Для вычисления значений сеточной функции на следующем временном слое имеем СЛАУ с трехдиагональной матрицей:

$$\frac{u_m^{i+1} - u_m^n}{\tau} = \frac{1}{h}\left(a^i_{m+1/2}\frac{u_{m+1}^{i+1} - u_m^{i+1}}{h} - a^i_{m-1/2}\frac{u_m^{i+1} - u_{m-1}^{i+1}}{h}\right) +$$

$$+ f(u_m^i) + f'_u(u_m^i)(u_m^{i+1} - u_m^i).$$

Итерации продолжаются до выполнения условия

$$\left\|u_m^{i+1} - u_m^i\right\| \leqslant \varepsilon.$$

При реализации шеститочечной схемы с переменным коэффициентом теплопроводности и нелинейной правой частью $f(u)$ итерационный процесс может иметь следующий вид:

$$\frac{u_m^{i+1} - u_m^n}{\tau} - \frac{1}{2h}\left[\left(a^n_{i+1/2}\frac{u_{m+1}^{i+1} - u_m^{i+1}}{h} - a^n_{i-1/2}\frac{u_m^{i+1} - u_{m-1}^{i+1}}{h}\right) +\right.$$

$$\left. + \left(a^n_{i+1/2}\frac{u_{m+1}^n - u_m^n}{h} - a^n_{n-1/2}\frac{u_m^n - u_{m-1}^n}{h}\right)\right] = f(u_m^i) + f'_u(u_m^i)(u_m^{i+1} - u_m^i).$$

Здесь также была применена линеаризация функции $f(u)$.

Отдельно рассмотрим важный частный случай $a = u^k$. В довольно грубом приближении уравнение описывает тепловые волны, образующиеся в высокотемпературной плазме и при образовании сверхновых звезд.

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}(u^k \frac{\partial u}{\partial x}).$$

Начальное условие для этой задачи $u(x, 0) = 0$, граничные условия —

$$u(0, t) = Ct^{\frac{1}{k}}, \lim_{x \to +\infty} u(x,t) = 0.$$

Будем искать *автомодельное* решение задачи, т. е. решение, зависящее не от двух переменных t и x, а от одной, являющейся их комбинацией:

$$\eta = x - Dt, \quad D = \text{const}.$$

Часто под автомодельными понимают решения, зависящие от безразмерных комбинаций независимых переменных. Введенную переменную иногда называют переменной бегущей волны. В этом случае

$$\frac{\partial u}{\partial t} = \frac{du}{d\eta} \cdot \frac{\partial \eta}{\partial t} = -D\frac{du}{d\eta}, \frac{\partial u}{\partial x} = \frac{\partial u}{\partial \eta},$$

тогда рассматриваемое уравнение приобретает вид

$$-Du' = (u^k u')'.$$

Штрихом здесь обозначено дифференцирование по новой переменной η. Скорость распространения волны, обозначенная здесь через D, будет определена ниже. После интегрирования обыкновенного дифференциального уравнения получим

$$-Du = u^k \cdot u'_\eta.$$

Постоянную интегрирования полагаем равной нулю, так как должно выполняться условие непрерывности теплового потока на фронте тепловой волны. Далее

$$-D = u^{k-1} \cdot u',$$

или $(u^k)'_\eta = -kD$, откуда $u^k(\eta) = -kD\eta$ или $u(\eta) = \sqrt[k]{kD} \cdot (-\eta)^{1/k}$.

Интересуют только положительные решения при $\eta \geqslant 0$, т.е. при $x \leqslant Dt$. При $\eta > 0$ положим $u(\eta) = 0$. В таком случае *обобщенное* решение (так как в точке $\eta = 0$ получается разрыв первой производной) рассматриваемой задачи будет иметь вид

$$u(\eta) = \begin{cases} \sqrt[k]{kD(-\eta)}, & \eta < 0 \\ 0 \quad , & \eta > 0. \end{cases}$$

Скорость фронта тепловой волны легко определяется из граничного условия:

$$D = \sqrt{C^k/k}.$$

Квазилинейное уравнение вида

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}(a_0 u^k \frac{\partial u}{\partial x}) + q_0 u^l$$

имеет качественно различные решения при разных параметрах k и l. Но в окрестности теплового фронта при распространении тепла по нулевому фону все эти решения имеют одинаковую асимптотику, решения в окрестности фронта устроены так же, как и у рассмотренной выше задачи о распространении тепловой волны.

12.3. Разностные схемы для численного решения многомерного уравнения теплопроводности

Численное решение даже простейших уравнений параболического типа сильно усложняется, если в задаче имеется в наличии более одного пространственного измерения. Условие устойчивости для многомерных схем накладывает столь жесткие ограничения на шаги по времени, что расчет по ним практически невозможен. Необходимо применять неявные схемы. Представим разностную схему для численного решения двумерного уравнения теплопроводности

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}$$

в виде

$$\frac{u_{ml}^{n+1} - u_{ml}^{n}}{\tau} = \Lambda_1 u_{ml}^{n+1} + \Lambda_2 u_{ml}^{n+1},$$

здесь

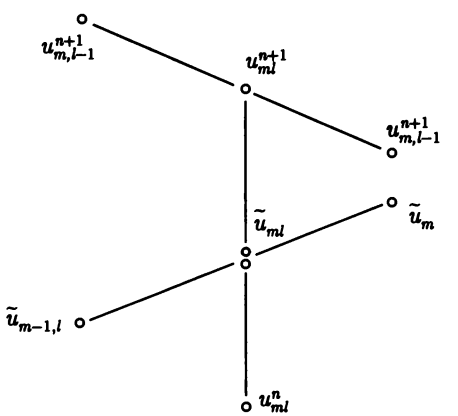
$$\Lambda_1 u_{ml}^{n+1} = \frac{u_{m-1,l}^{n+1} - 2u_{m,l}^{n+1} + u_{m+1,l}^{n+1}}{h_x^2}, \Lambda_2 u_{ml}^{n+1} = \frac{u_{m,l-1}^{n+1} - 2u_{ml}^{n+1} + u_{m,l+1}^{n+1}}{h_y^2}.$$

Получена линейная система с разреженной (блочной) матрицей. Однако вид этой матрицы таков, что алгоритм пятиточечной прогонки в данном случае не применим.

Можно предложить *схему расщепления по направлениям*, или *локально-одномерную* схему (*метод дробных шагов*, Н. Н. Яненко [7]):

$$\frac{\tilde{u}_{ml} - u_{ml}^{n}}{\tau} = \Lambda_1 \tilde{u}_{ml}, \frac{u_{ml}^{n+1} - \tilde{u}_{ml}}{\tau} = \Lambda_2 u_{ml}^{n+1}.$$

Соответствующий пространственный шаблон схемы будет



Для аналогичной трехмерной задачи

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}$$

можно предложить локально-одномерную схему дробных шагов

$$\frac{u_{mlp}^{n+1/3} - u_{mlp}^{n}}{\tau} = \Lambda_1 u_{mlp}^{n+1/3}, \quad \frac{u_{mlp}^{n+2/3} - u_{mlp}^{n+1/3}}{\tau} = \Lambda_2 u_{mlp}^{n+2/3},$$

$$\frac{u_{mlp}^{n+1} - u_{mlp}^{n+2/3}}{\tau} = \Lambda_3 u_{mlp}^{n+1}.$$

Порядок аппроксимации этих схем: $O(\tau, h_x^2, h_y^2)$ в двумерном случае и $O(\tau, h_x^2, h_y^2, h_z^2)$ — в трехмерном.

Порядок аппроксимации этой схемы по времени можно увеличить до второго, если провести усреднение операторов $\Lambda_i u_{mlp}^n (i = 1 \div 3)$, аппроксимирующих вторые производные по координатам $x_i (i = 1 \div 3)$:

$$\frac{u_{mlp}^{n+1/3} - u_{mlp}^{n}}{\tau} = \Lambda_1 [\xi u_{mlp}^{n+1/3} + (1 - \xi) u_{mlp}^{n}],$$

$$\frac{u_{mlp}^{n+2/3} - u_{mlp}^{n+1/3}}{\tau} = \Lambda_2 [\xi u_{mlp}^{n+2/3} + (1 - \xi) u_{mlp}^{n+1/3}],$$

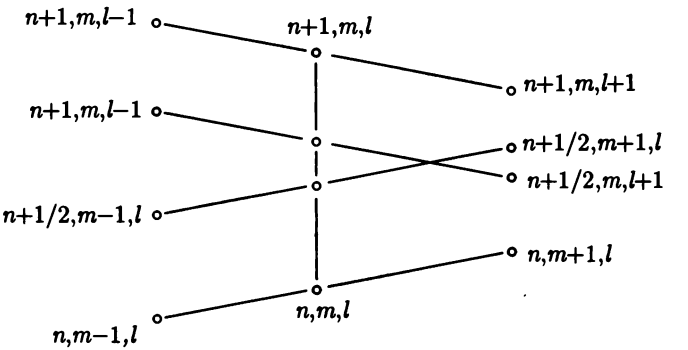
$$\frac{u_{mlp}^{n+1} - u_{mlp}^{n+2/3}}{\tau} = \Lambda_3 [\xi u_{mlp}^{n+1} + (1 - \xi) u_{mlp}^{n+2/3}],$$

где $0 \leqslant \xi \leqslant 1$, причем при $\xi = 1/2$ порядок аппроксимации схемы будет $O(\tau, h_{x^2}, h_{y^2}, h_z^2)$ Эта схема Кранка–Никольсон устойчива при любых τ, h_x, h_y, h_z; ее шаблон для двумерной задачи

$$\frac{u_{ml}^{n+1/2} - u_{ml}^{n}}{\tau} = \Lambda_1 \left[\xi u_{ml}^{n+1/2} + (1 - \xi) u_{ml}^{n}\right],$$

$$\frac{u_{ml}^{n+1} - u_{ml}^{n+1/2}}{\tau} = \Lambda_2 \left[\xi u_{ml}^{n+1} + (1 - \xi) u_{ml}^{n+1/2}\right]$$

представлен на рисунке.



Приведем еще одну схему, имеющую второй порядок аппроксимации по τ и h:

$$\frac{\tilde{u}_{ml} - u^n_{ml}}{\tau} = \frac{1}{2}(\Lambda_1 \tilde{u}_{ml} + \Lambda_2 u^n_{ml}), \quad \frac{u^{n+1}_{ml} - \tilde{u}_{ml}}{\tau} = \frac{1}{2}(\Lambda_1 \tilde{u}_{ml} + \Lambda_2 u^{n+1}_{ml});$$

ее пространственный шаблон:

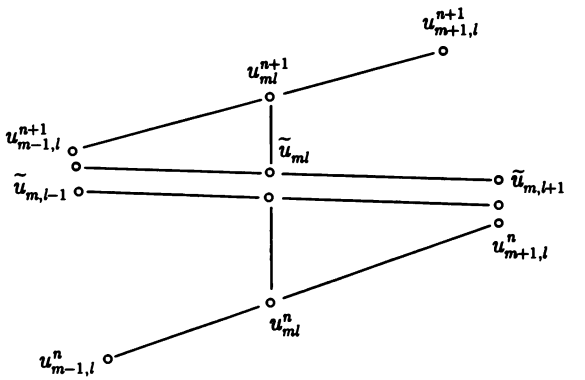


Схема Дугласа-Ганна — это общий метод построения неявных разностных схем переменных направлений для трехмерного уравнения теплопроводности, имеющих второй порядок точности и безусловно устойчивых [13]:

$$u^* - u^n = \frac{r_x}{2}\Lambda_1 (u^* + u^n) + r_y \Lambda_2 (u^n) + r_z \Lambda_3 (u^n),$$

$$u^{**} - u^n = \frac{r_x}{2}\Lambda_1 (u^* + u^n) + \frac{r_y}{2}\Lambda_2 (u^{**} + u^n) + r_z \Lambda_3 (u^n),$$

$$u^{n+1} - u^n = \frac{r_x}{2}\Lambda_1 (u^* + u^n) + \frac{r_y}{2}\Lambda_2 (u^{**} + u^n) + \frac{r_z}{2}\Lambda_3 (u^{n+1} + u^n).$$

Верхние индексы * и ** обозначают промежуточные значения, координатные индексы i, j, k опущены во всех членах уравнений, Λ_i — компоненты разностного оператора теплопроводности. На каждом шаге метода возникает система линейных уравнений с трехдиагональной матрицей, решаемая методом прогонки. Схема Дугласа-Ганна безусловно устойчивая. В настоящее время это наиболее удачная из схем расщепления в трехмерном случае.

12.4. Исследование сходимости разностных схем для многомерного уравнения теплопроводности

Простейшая явная разностная схема для численного решения двумерного уравнения теплопроводности

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}$$

получается путем замены производных конечными разностями

$$\frac{u_{ml}^{n+1} - u_{ml}^n}{\tau} = \frac{u_{m-1,l}^n - 2u_{ml}^n + u_{m+1,l}^n}{h_x^2} + \frac{u_{m,l-1}^n - 2u_{ml}^n + u_{m,l+1}^n}{h_y^2}$$

или, в операторной форме,

$$\frac{u_{ml}^{n+1} - u_{ml}^n}{\tau} = \Lambda_1 u_{ml}^n + \Lambda_2 u_{ml}^n.$$

Исследование спектральной устойчивости этот схемы ($u_{me}^n = \lambda^n e^{i\alpha m + i\beta l}$, $\alpha, \beta \in [0, 2\pi]$) приводит к следующему результату для спектра оператора послойного перехода:

$$\lambda(\alpha, \beta, \tau, h_x, h_y) = 1 - 4\frac{\tau}{h_x^2}\sin^2\frac{\alpha}{2} - 4\frac{\tau}{h_y^2}\sin^2\frac{\beta}{2},$$

откуда получаем условие устойчивости

$$\tau \leqslant \frac{1}{2(h_x^{-2} + h_y^{-2})}.$$

Для неявной схемы

$$\frac{u_{ml}^{n+1} - u_{ml}^n}{\tau} = \Lambda_1 u_{ml}^{n+1} + \Lambda_2 u_{ml}^{n+1}$$

исследование на устойчивость по спектральному признаку дает

$$\lambda = \frac{1}{1 + 4\frac{\tau}{h_x^2}\sin^2\frac{\alpha}{2} + 4\frac{\tau}{h_y^2}\sin^2\frac{\beta}{2}},$$

т. е. схема устойчива при любых α, β.

Приведем исследование спектральной устойчивости для схемы переменных направлений.

Рассмотрим переход с нижнего на верхний временный слой. В таком случае можно положить $u_{ml}^n = e^{i\alpha m + i\beta l}$; сомножитель λ^n опускаем, так как рассматривается один переход с n-го на $(n+1)$-й слой в предположении, что известно решение на n-ом слое (можно было бы написать, $u_{ml}^n = Ce^{i\alpha m + i\beta l}$, где $C = \lambda^n$, но в этом нет смысла, так как C в дальнейших выкладках сократится).

Тогда на первом этапе получим $\tilde{u}_{ml} = \lambda_1 u_{ml}^n$, а на втором — $u_{ml}^{n+1} = \lambda_2 \tilde{u}_{ml} = (\lambda_1\lambda_2)u_{ml}^n$.

Вычисление λ_1 и λ_2 дает

$$\lambda_1 = \frac{1 - 4\frac{\tau}{h_x^2}\sin^2\frac{\alpha}{2}}{1 + 4\frac{\tau}{h_y^2}\sin^2\frac{\alpha}{2}}, \lambda_2 = \frac{1 - 4\frac{\tau}{h_x^2}\sin^2\frac{\beta}{2}}{1 + 4\frac{\tau}{h_y^2}\sin^2\frac{\beta}{2}}.$$

Окончательно спектр оператора послойного перехода представим в виде произведения спектров на каждом промежуточном этапе.

$$\lambda(\alpha, \beta, \tau, h_x, h_y) = \lambda_1\lambda_2 = \frac{(1 - 4\frac{\tau}{h_x^2}\sin^2\frac{\alpha}{2})(1 - 4\frac{\tau}{h_x^2}\sin^2\frac{\beta}{2})}{(1 + 4\frac{\tau}{h_y^2}\sin^2\frac{\alpha}{2})(1 + 4\frac{\tau}{h_y^2}\sin^2\frac{\beta}{2})}.$$

Схема безусловно устойчива.

Исследование схемы расщепления на аппроксимацию проведем на примере локально-одномерной схемы для двумерного уравнения теплопроводности

$$\frac{\tilde{u}_{ml} - u_{ml}^n}{\tau} = \Lambda_1\tilde{u}_{ml}, \quad \frac{u_{ml}^{n+1} - \tilde{u}_{ml}}{\tau} = \Lambda_2 u_{ml}^{n+1}.$$

Запишем эти уравнения в операторной форме

$$(\mathbf{E} - \tau\Lambda_1)\tilde{u}_{ml} = u_{ml}^n, (\mathbf{E} - \tau\Lambda_2)u^{n+1} = \tilde{u}_{ml}.$$

Подействуем на обе части второго уравнения оператором $(\mathbf{E} - \tau\Lambda_1)$:

$$(\mathbf{E} - \tau\Lambda_1)(\mathbf{E} - \tau\Lambda_2)u_{ml}^{n+1} = (\mathbf{E} - \tau\Lambda_1)\tilde{u}_{ml}.$$

Так как

$$(\mathbf{E}-\tau\Lambda_1)\tilde{u}_{ml}=u^n_{ml},\quad \text{то}\quad (\mathbf{E}-\tau\Lambda_1)(\mathbf{E}-\tau\Lambda_2)u^{n+1}_{ml}=u^n_{ml}.$$

Это уравнение приводится к виду

$$\frac{u^{n+1}_{ml}-u^n_{ml}}{\tau}=\Lambda_1 u^{n+1}_{ml}+\Lambda_2 u^{n+1}_{ml}-\tau^2\Lambda_1\Lambda_2 u^{n+1}_{ml}.$$

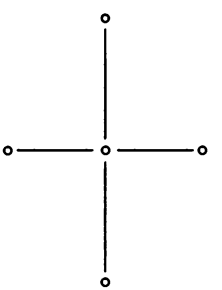
Последнее слагаемое в правой части является величиной порядка $O(\tau)$ и определяет погрешность аппроксимации.

Если правая часть $f(t,x,y)$ не нулевая, то схему можно переписать, например

$$\frac{\tilde{u}_{ml}-u^n_{ml}}{\tau}=\Lambda_1\tilde{u}_{ml}+\frac{1}{2}f^n_{ml},\ \frac{u^{n+1}_{ml}-u^n_{ml}}{\tau}=\Lambda_2 u^{n+1}_{ml}+\frac{1}{2}\tilde{f}_{ml}.$$

12.5. Задачи

1. Исследовать на аппроксимацию и устойчивость схему Ричардсона $\frac{u^{n+1}_m-u^{n-1}_m}{2\tau}=\frac{a}{h^2}\Lambda_{xx}u^n$, с шаблоном



аппроксимирующую уравнение теплопроводности $\frac{\partial u}{\partial t}=a\frac{\partial^2 u}{\partial x^2}$.

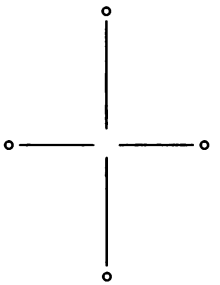
Решение. В силу симметрии шаблона схемы несложно заметить, что ее порядок аппроксимации $O(\tau^2,h^2)$.

Подстановка решения в виде $u^n_m=\lambda^n e^{i\alpha m}$ дает квадратное уравнение для определения спектра оператора послойного перехода λ^2+ $+\frac{8a\tau}{h^2}\lambda\sin^2\frac{\alpha h}{2}-1=0$, один из корней которого при любом значении параметра α по модулю больше единицы, т. е. рассматриваемая схема *безусловно неустойчива.*

Выходом из данной ситуации оказывается замена в выражении $\Lambda_{xx}u_m^n = \frac{u_{m-1}^n - 2u_m^n + u_{m+1}^n}{h^2}$ величины u_m^n на $\frac{u^{n-1}+u^{n+1}}{2}$ (схема Франкела-Дюффорта):

$$\frac{u_m^{n+1} - u_m^{n-1}}{2\tau} = \frac{a}{h^2}\left[u_{m-1}^n - \left(u_m^{n+1} + u_m^{n-1}\right) + u_{m+1}^n\right]$$

со следующим шаблоном:



которая разрешается явно относительно u_m^{n+1} и *безусловно устойчива.*

Однако она обладает лишь условной аппроксимацией: $O(\tau^2, h^2, \tau^2/h^2$ и таким образом, сходимость возможна лишь при $\tau/h \to 0$.

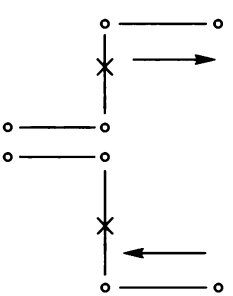
2. Исследовать на аппроксимацию и устойчивость схему Саульева бегущего счета

$$\frac{u_m^n - u_m^{n-1}}{\tau} = a\frac{u_{m-1}^n - (u_m^n + u_m^{n-1}) - u_{m+1}^{n-1}}{h^2}$$

(четные слои, счет справа налево)

$$\frac{u_m^{n+1} - u_m^n}{\tau} = a\frac{u_{m-1}^n - (u_m^n + u_m^{n+1}) + u_{m+1}^{n+1}}{h^2}$$

(нечетные слои, счет слева направо) со следующим шаблоном:



Решение. Путем подстановки решения в виде $u_m^n = \lambda^n e^{i\alpha m}$, несложно проверить, что схема безусловно устойчива.

Невязка каждого из двух уравнений, вычисленная относительно точек, отмеченных крестиками, имеет порядок $O(\tau^2, h^2, \tau, h, \tau/h)$.

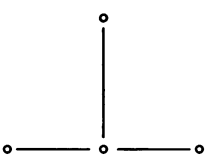
Сложение же невязок из двух рассматриваемых уравнений дает погрешность $O(\tau^2, h^2, \tau^2/h^2)$, аналогичную схеме Франкела-Дюфорта.

Схема Саульева допускает значительное улучшение. Достаточно только вычислить значения функции в рамках перехода с данного слоя на следующий два раза — бегущим счетом слева направо и бегущим счетом справа налево — и усреднить результаты. Свойства такого метода расчета предлагается исследовать самостоятельно.

3. Исследовать на аппроксимацию и устойчивость схему Алена-Чена

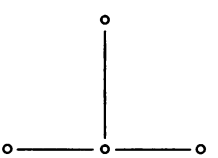
$$\frac{u_m^{n+1} - u_m^n}{\tau} = a\frac{u_{m-1}^n - 2u_m^{n+1} + u_{m+1}^n}{h^2}$$

с шаблоном



Ответ. Во-первых, несмотря на то, что в правую часть входит значение функции u_m^{n+1}, вычисляемое на верхнем слое, разностное уравнение разрешается относительно u_m^{n+1}. Схема безусловно устойчива, что является ее достоинством при реализации, однако она имеет порядок аппроксимации $O(\tau, h^2, \tau/h^2)$, т. е. схема является условно аппроксимирующей.

4. Выяснить, является ли явная четырехточечная схема $\frac{u_m^{n+1} - u_m^n}{\tau} = a\Lambda_{xx}u^n$ с шаблоном



монотонной. Монотонные разностные схемы (по Фридрихсу) — это схемы, которые при записи в виде, разрешенном относительно u_m^{n+1} при значениях сеточной функции во всех остальных точках шаблона, имеют неотрицательные коэффициенты. Монотонные разностные схемы не дают паразитные осцилляции в решении. Придумать

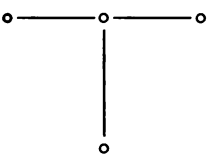
доказательство утверждения, что из монотонности разностной схемы следует ее устойчивость, предлагается самостоятельно.

Решение. Представим рассматриваемую схему в виде

$$u_m^{n+1} = \sum_{k=-1}^{1} \alpha_m u_{m+k}$$

где $\alpha = 1 - \frac{2a\tau}{h^2}, \alpha_1 = \alpha_{-1} = \frac{a\tau}{h^2}$, откуда видно, что схема является монотонной при выполнении условия $\alpha_k \geqslant 0, k = -1, 0, 1$ то есть при $\tau \leqslant \frac{h^2}{2a}$. Эта схема условно монотонная.

Заметим, что неявная четырехточечная схема является монотонной, а схема Кранка-Никольсон — условно монотонная.



Исследование монотонности параметрической двуслойной схемы в виде $u_m^{n+1} = \sum\limits_{k=0}^{\infty} \alpha_m u_{m+k}$, дает (выкладки опускаем ввиду их громоздкости):

$$u_m^{n+1} = \alpha_0 u_m + \sum_{k=1}^{\infty} \alpha_k (u_{m-k} + u_{m+k}),$$

$$\alpha_0 = 1 - \frac{4a\tau \left(h + \sqrt{h^2 + 4\xi a\tau}\right)^{-1}}{\sqrt{h^2 + 4\xi a\tau}},$$

$$\alpha_1 = \frac{4a\tau}{\sqrt{h^2 + 4\xi a\tau}} \left(h + \sqrt{h^2 + 4\xi a\tau}\right)^{-2},$$

$$\alpha_k = \alpha_{k-1} \cdot 4\xi a\tau \left(h + \sqrt{h^2 + 4\xi a\tau}\right)^{-2}, k \geqslant 2.$$

Коэффициенты α_x при $x \geqslant 1$ неотрицательны, коэффициент α_0 неотрицателен при выполнении условия:

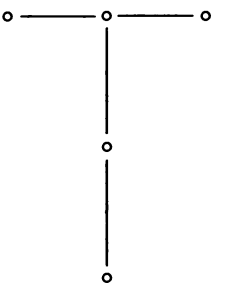
$$\tau \leqslant \frac{(2 - \xi) h^2}{4a (1 - \xi)^2}.$$

Отсюда видно, что, за исключением неявной четырехточечной схемы с $\xi = 1$, все неявные схемы являются монотонными лишь при условии $\tau \sim h^2$.

5. Исследовать на устойчивость и аппроксимацию трехслойную схему

$$\frac{1,5(u_m^{n+1}-u_m^n)}{\tau}+\frac{0,5(u_m^n-u_m^{n-1})}{h}=\Lambda_{xx}u_m^{n+1}$$

с шаблоном



Ответ. Простое исследование на аппроксимацию данной схемы путем разложения функций в ряд Тейлора дает $O(\tau^2, h^2)$. Схема безусловно устойчива. Кроме того, схема монотонна.

6. Показать, что решение задачи

$$\frac{\partial u}{\partial t}=\frac{\partial}{\partial x}au^{\alpha}\frac{\partial u}{\partial x},(\alpha>0), \tag{12.5}$$

$$u(+\infty,t)=0,\quad u(0,t)=ct^{1/\alpha},\quad u(x,0)=0$$

представляет собой бегущую волну, распространяющуюся с конечной скоростью, причем при $\alpha \geqslant 0$ на фронте волны решение терпит разрыв первой производной (т. е. является обобщенным решением).

Предложить какие-нибудь разностные схемы для численного решения данной задачи, например, применив интегро-интерполяционный метод [3]. Почему в этом случае нельзя переписать уравнение в виде

$$\frac{\partial u}{\partial t}=au^{\alpha}\frac{\partial^2 u}{\partial x^2}+a\alpha u^{\alpha-1}\left(\frac{\partial u}{\partial x}\right)^2,$$

продифференцировав правую часть по x, а затем заменить производные их разностными аппроксимациями?

Решение. Построим решение задачи в переменных бегущей волны:

$$u=u(\xi)=u(x-vt),(x-vt\geqslant 0),u\equiv 0. \tag{12.6}$$

Подставив (12.6) в исходное уравнение, получим следующее обыкновенное дифференциальное уравнение:

$$-vu'_{\xi}=a(u^{\alpha}u'_{\xi})'_{\xi}. \tag{12.7}$$

Это уравнение интегрируется по ξ :

$$k - vu = au^{\alpha}u'_{\xi}, \tag{12.8}$$

где k — константа интегрирования.

Выражение, стоящее в правой части (12.8), есть поток величины u, а уравнение (12.8) является некоторым законом сохранения. В точке фронта $u = 0$, в силу непрерывности потока справа и слева от фронта (12.8) должно выполняться, отсюда следует $k = 0$.

Так как интересуют только нетривиальные решения (12.5), то $u \neq 0$, и можно разделить правую и левую части на u. Тогда получим

$$-v = au^{\alpha-1}u'_{\xi} = \frac{a}{\alpha}(u^{\alpha})'_{\xi}, \tag{12.9}$$

откуда $u = \left[\frac{\alpha v}{a}(x - vt)\right]^{1/\alpha}$.

Из краевых условий при $x = 0$ имеем $v = ac^{\alpha}/\alpha$. Наличие разрывов производной легко проверяется. Уравнение (12.5) записано в *дивергентном* виде, которому будет соответствовать *консервативная* схема.

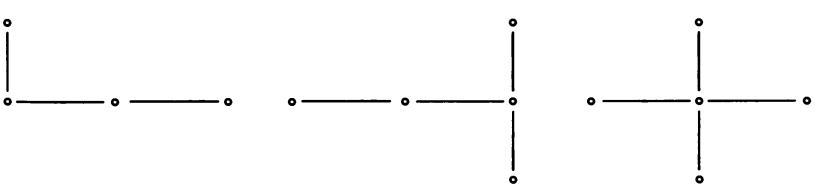
В [9] использован несколько отличный подход к решению задачи. В [4] рассмотрен общий случай, когда коэффициент теплопроводности есть функция $K(u)$:

$$\Phi(u) = \int_{o}^{u} \frac{K(\eta)}{\eta} d\eta; \Phi(0) = 0; \Phi(1) < \infty.$$

Подробнее про разностные схемы для аппроксимации уравнения (5.3.1) в [4, с. 441–463].

12.6. Задачи для самостоятельного решения

1. Используя условие устойчивости Куранта-Фридрихса-Леви, определить, какие из разностных схем, шаблоны которых приведены ниже, не будут устойчивыми:



2. При каком соотношении шагов τ и h явная разностная схема

$$\frac{u_m^{n+1}-u_m^n}{\tau}=\frac{u_{m-1}^n-2u_m^n+u_{m+1}^n}{h^2}$$

для уравнения теплопроводности имеет порядок аппроксимации $O(\tau^2,h^4)$?

3. Показать, что параметрическая разностная схема

$$\frac{u_m^{n+1}-u_m^n}{\tau}=\xi\frac{u_{m-1}^{n+1}-2u_m^{n+1}+u_{m+1}^{n+1}}{h^2}+(1-\xi)\frac{u_{m-1}^n-2u_m^n+u_{m+1}^n}{h^2},$$

при весе $\xi=\frac{1}{2}-\frac{h^2}{12\tau}$ имеет порядок аппроксимации $O(\tau^2,h^4)$.

4. Для аппроксимации уравнения (12.2) использована схема

$$\frac{u_{\mathrm{m}}^{n+1}-u_{\mathrm{m}}^n}{\tau}=\frac{1}{h}\left(k_{\mathrm{m}+1/2}\frac{u_{\mathrm{m}+1}^{n+1}-u_{\mathrm{m}}^{n+1}}{h}-k_{\mathrm{m}-1/2}\frac{u_{\mathrm{m}}^{n+1}-u_{\mathrm{m}-1}^{n+1}}{h}\right),$$

где $k_{\mathrm{m}+1/2}$ вычисляется следующим образом:

(a) $k_{\mathrm{m}+1/2}=\frac{a}{2}((u_{\mathrm{m}}^n)^\alpha+(u_{m+1}^n)^\alpha)$;

(b) $k_{\mathrm{m}+1/2}=a\left(\frac{u_{\mathrm{m}}^n+u_{\mathrm{m}+1}^n}{2}\right)^\alpha$;

(c) $k_{\mathrm{m}+1/2}=a\left(\frac{2u_{\mathrm{m}}^n u_{\mathrm{m}+1}^n}{u_{\mathrm{m}}^n+u_{\mathrm{m}+1}^n}\right)^\alpha$;

(d) $k_{\mathrm{m}+1/2}=a\frac{2(u_{\mathrm{m}}^n)^\alpha(u_{\mathrm{m}+1}^n)^\alpha}{(u_{\mathrm{m}}^n)^\alpha+(u_{\mathrm{m}+1}^n)^\alpha}$.

Выражения a)–d) суть некоторые аппроксимации au^α, взятые на предыдущем слое по времени между узлами $u_{\mathrm{m}},u_{\mathrm{m}+1}$. Какой из вариантов предпочтительнее? Почему не работают в окрестности фронта средние гармонические — c) и d)?

Реализовать схемы a) и b) на ЭВМ, сравнить численное решение с точным (см. выше). Почему при больших числах Куранта наблюдается отставание фронта волны в численном решении от точного значения [9]?

5. **Режимы с обострением**

Рассмотрим уравнение

$$\frac{\partial u}{\partial t}=\frac{\partial}{\partial x}\left(u^\alpha\frac{\partial u}{\partial x}\right)+au^\beta, a>0,\alpha>0,\beta>0. \qquad (12.10)$$

Построить численно решение (12.10) в случаях а) $\beta = \alpha + 1$, б) $\beta < \alpha + 1$, в)$\beta > \alpha + 1$,

$$u(-\infty,\ t) = 0, u(+\infty,\ t) = 0, u(x,\ 0) = \begin{cases} u_0,\ |x| \leqslant 1, \\ 0, |x| > 1. \end{cases}$$

В чем качественное различие случаев а) (так называемый S-режим с обострением), б) (*HS*-режим), в) (*LS*-режим)? (См. [5, глава 2], [4, С. 172–206]). В случае а) численно проверить справедливость формулы полуширины локализации тепла: $l = \pi\sqrt{\frac{\alpha+1}{a\alpha^2}}$. Как решение зависит от амплитуды начального возмущения u_0?

6. **Автомодельные решения и автомодельные переменные**

В случаях 1–3 будем искать решение уравнения (12.10) в виде [5, С. 32–62], [4]:

$$u(x,\ t) = g(t)f(\xi),$$

где $\xi = x/\varphi(t)$.

(a) Найти функции $g(t), \varphi(t)$.

(Ответ: $g(t) = (1 - t/t_f)^{-1/2}, \varphi(t) = (1 - t/t_f)^{\frac{\beta-(\alpha+1)}{2(\beta-1)}}$, t_f — положительный параметр).

(b) Какому дифференциальному уравнению удовлетворяет при этом функция $f(\xi)$? Решить численно получившееся уравнение для $f(\xi)$ с условиями

$$f'_\xi(\xi = 0) = 0; f(\xi = l) = 0, l < +\infty$$

с дополнительным требованием $f^\alpha f'_\xi|_{\xi=l} = 0$.

(c) Задавая $f(\xi)$ при $t = 0$ в качестве начальных условий для (12.10), сравните поведение численного решения с автомодельным. (Так как известны $g(t), \phi(t)$ и $f(\xi)$, то тем самым найдено $u(x,t) \forall t 0 < t < t_f$). Что происходит при $t \to t_f$?

7. **Тепловой кристалл**

Рассмотрим уравнение

$$\frac{\partial u}{\partial t} = k_1 \frac{\partial}{\partial x} u^\alpha \frac{\partial u}{\partial x} + k_2 \frac{\partial}{\partial y} u^\alpha \frac{\partial u}{\partial y} + a u^\beta.$$

Попытайтесь качественно исследовать свойства решений этого уравнения.

В случае $\alpha = \beta - 1 > 0, k_1 \neq k_2$ рассмотреть задачу со следующими граничными условиями:

$$u(t, x, 0) = A_0(1-t)^n(1-\lambda_1 x)^{2/\alpha}$$

при $x \leqslant 1/\lambda_1$,

$$u(t, 0, y) = A_1(1-t)^n(1-\lambda_2 y)^{2/\alpha}$$

при $x \leqslant 1/\lambda_2$, иначе 0. Здесь $n < 0$ — действительное число.

Рассмотреть случаи

$$n = -1/\alpha, \quad n < -1/\alpha, \quad -1/\alpha < n < 0.$$

Решить задачу численно и сравнить полученное решение с аналитическим. Решение тепловой кристалл описано в [4, с. 148-155].

8. **Остановка тепловой волны**

Модифицируем задачу 6:

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(u^\alpha \frac{\partial u}{\partial x}\right) + au^\beta - \varepsilon u, \tag{12.11}$$

положив $\varepsilon = 0,1$.

(a) Выполнить пункты задачи 6. Как влияет малый линейный сток на поведение решения?

(b) В случае $\alpha = \beta - 1$ найти автомодельное решение, аналогичное рассматриваемому в задаче (7).

Указание. Рассмотреть последовательность замен

$$v = ue^{\varepsilon t}, \frac{\partial v}{\partial t} = e^{\varepsilon t}\left(\frac{\partial u}{\partial t} + \varepsilon u\right)$$

и $d\tau = e^{\alpha \varepsilon t} dt$.

Каким станет уравнение (12.11) в переменных v, x, t?

(c) Рассмотреть задачу о формировании теплового кристалла для уравнения (12.11).

9. **Неустойчивость Тьюринга**

Рассмотрим систему

$$\frac{\partial u}{\partial t} = D_1 \frac{\partial^2 u}{\partial x^2} + au + bv,$$

$$\frac{\partial v}{\partial t} = D_2 \frac{\partial^2 v}{\partial x^2} + cu + dv$$

с условиями

$$\left.\frac{\partial u}{\partial x}\right|_{x=0} = \left.\frac{\partial u}{\partial x}\right|_{x=1} = 0, \left.\frac{\partial v}{\partial x}\right|_{x=0} = \left.\frac{\partial v}{\partial x}\right|_{x=1} = 0,$$

причем $a + d < 0, ad - bc > 0$. При этих условиях особая точка $(0, 0)$ системы $\dot{u} = au + bv, \dot{v} = cu + dv$ устойчива. Пусть, без ограничения общности, $a > 0$.

(a) Найти условие, когда внесение в систему диффузии приводит к потере устойчивости однородного стационарного решения.

Указание. Рассмотреть преобразование Фурье по пространственной переменной. Исследовать на устойчивость особые точки получившейся системы ОДУ.

(b) Подобрав коэффициенты a, b, c, d и D_1, D_2, удовлетворяющие условиям, найденным в пункте 1, получить при численном счете так называемые *структуры Тьюринга*. Под структурой Тьюринга здесь понимается пространственно-неоднородное решение с волновым числом k, таким, что $\mathrm{Re}\lambda(k^2) = 0; \frac{\partial(\mathrm{Re}\lambda)}{\partial(k^2)} = 0$ при записи решения в виде $\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} e^{\lambda t - ikx}$. В случае *линейной* задачи эти структуры будут возрастать по амплитуде при $t \to \infty$, в случае нелинейного уравнения (следующая задача) бесконечный рост становится невозможным, структура стабилизируется за счет нелинейности.

10. **Распределенный брюсселятор**

Рассмотрим систему уравнений типа реакция-диффузия, где для описания химических реакций использована модельная система «брюсселятор». В литературе эта постановка имеет название «распределенный брюсселятор»:

$$\frac{\partial u}{\partial t} = A - (B + 1)u + u^2 v + D_1 \frac{\partial^2 u}{\partial x^2},$$

$$\frac{\partial v}{\partial t} = Bu - u^2 v + D_2 \frac{\partial^2 v}{\partial x^2}$$

с граничными условиями

$$\frac{\partial u}{\partial x}(0, t) = \frac{\partial u}{\partial x}(l, t) = \frac{\partial v}{\partial x}(0, t) = \frac{\partial v}{\partial x}(l, t) = 0$$

и начальным условием $u(x,0) = A(1+\varepsilon\cos\omega x), v(x,0) = B/A$. Это — система уравнений «брюсселятор» с учетом диффузии компонентов.

(a) Пусть $B > (1+A\sqrt{D_1/D_2})^2, D_1 < D_2$. Рассмотреть образование структур Тьюринга в случаях $\mu = A\sqrt{D_1/D_2}, \mu < 0.207, 0.207 < \mu < 2.418, \mu > 2.418$.

Когда можно пользоваться явной разностной схемой? Когда необходима неявная схема?

(b) Пусть теперь $D_1 > D_2, A > 1$. Что происходит в системе? Почему для расчетов необходимо применять неявные схемы? Какую схему расщепления по физическим процессам можно предложить для решения задачи? Примеры расчетов приведены в книге [15], качественное исследование — в [14, С. 403-407].

11. **Схема «Классики»**

Для двумерного уравнения теплопроводности используется схема «классики». Как и в схеме Саульева, расчет осуществляется в два этапа:

$$\frac{u_{lm}^{n+1}-u_{lm}^{n}}{\tau} = \frac{u_{l-1m}^{n}-2u_{lm}^{n+1}+u_{l+1m}^{n}}{h_x^2} + \frac{u_{lm-1}^{n}-2u_{lm}^{n+1}+u_{lm+1}^{n}}{h_y^2}$$

в случае, если $l+m+n$ — четное,

$$\frac{u_{lm}^{n+1}-u_{lm}^{n}}{\tau} = \frac{u_{l-1m}^{n+1}-2u_{lm}^{n+1}+u_{l+1m}^{n+1}}{h_x^2} + \frac{u_{lm-1}^{n+1}-2u_{lm}^{n+1}+u_{lm+1}^{n+1}}{h_y^2}$$

в случае, если $l+m+n$ — нечетное.

Явная или неявная эта схема? Исследовать ее на аппроксимацию и устойчивость. Зачем нужно «перепрыгивание» — смена порядка обхода узлов при переходе со слоя на слой по времени?

Применить эту схему к расчету предыдущей задачи.

Литература

[1] *Тихонов А.Н., Самарский А.А.* Уравнения математической физики. М., Изд-во МГУ, 2002.

[2] *Владимиров В.С.* Уравнения математической физики. М.: Наука, 1984.

[3] *Соболев С.Л.* Уравнения математической физики. М.: Наука, 1992.

[4] *Самарский А.А., Галактионов В.А., Курдюмов С.П., Михайлов А.П.* Режимы с обострением в задачах для квазилинейных параболических уравнений. М.: Наука, 1987. 480 с.

[5] *Ахромеева Т.С., Курдюмов С.П., Малинецкий Г.Г., Самарский А.А.* Нестационарные структуры и диффузионный хаос. М.: Наука, 1992. 544 с.

[6] *Курдюмов С.П., Куркина Е.С.* Тепловые структуры в среде с нелинейной теплопроводностью. / В кн.: Будущее прикладной математики. Лекции для молодых исследователей. Под ред. *Г.Г. Малинецкого.* М.: Едиториал УРСС, 2005. 512 с.

[7] *Яненко Н.Н.* Метод дробных шагов решения многомерных задач математической физики. Новосибирск: Наука, 1967. 196 с.

[8] *Годунов С.К., Рябенький В.С.* Разностные схемы, введение в теорию. М.: Наука, 1977. 400 с.

[9] *Федоренко Р.П.* Введение в вычислительную физику. М.: Изд-во МФТИ, 1994. 526 с.

[10] *Самарский А.А.* Теория разностных схем. М.: Наука, 1983. 656 с.

[11] *Самарский А.А., Вабищевич П.Н., Матус Г.П.* Разностные схемы с операторными множителями. Минск, 1998. 441 с.

[12] *Марчук Г.И.* Методы расщепления. М: Наука, 1988. 263 с.

[13] *Андерсон Д., Таннехилл Дж., Плетчер Р.* Вычислительная гидромеханика и теплообмен: в 2-х т., Т.1: Пер. с англ. М.: Мир, 1990. 384 с.

[14] *Ланда П.С.* Нелинейные колебания и волны. М.: Наука-Физматлит, 1997. 496 с.

[15] *Хайрер Э., Нернсет С., Ваннер Г.* Решение обыкновенных дифференциальных уравнений. Нежесткие задачи. М.: Мир, 1990. 512 с.

Лекция 13. Численные методы решения уравнений в частных производных гиперболического типа (на примере уравнения переноса)

В лекции дается понятие о простейших разностных схемах для решения линейного уравнения переноса. Приводится вид некоторых часто употребляемых схем. Обсуждаются способы конструирования гибридных разностных схем. Обсуждаются вопросы обобщения на квазилинейный случай. Дается первоначальное представление о способах регуляризации решений с большими градиентами. Вводится понятие схем с уменьшением полной вариации (TVD). Рассматриваются основные идеи метода конструирования разностных схем в пространстве неопределенных коэффициентов.

Ключевые слова: линейное уравнение переноса, характеристика, квазилинейное уравнение переноса, дивергентная форма, консервативная разностная схема, регуляризация, сглаживание, схемная вязкость, антидиффузия, гибридные схемы, TVD-схемы, пространство неопределенных коэффициентов, первое дифференциальное приближение.

13.1. Простейшее линейное уравнение переноса

Рассмотрим простейший пример уравнений в частных производных. Пусть в некотором объеме движущейся жидкости находится пассивная примесь, т. е. такая, наличие которой не меняет принципиально характер движения. Например, это может быть маркер — краска, чернила или мелкие частицы, которые специально добавлены в жидкость для визуализации течений. Тогда изменение концентрации примеси в любом сколь угодно малом объеме равно потоку примеси через границы объема в единицу времени (закон сохранения массы), и можно записать, устремляя рассматриваемый объем к нулю

$$\frac{\partial u}{\partial t} = -\mathrm{div}(u\mathbf{v}),$$

где u — концентрация пассивной примеси, $\mathbf{v}$ — скорость течения жидкости. Пусть жидкость несжимаема и

$$\mathrm{div}\mathbf{v} = 0.$$

Тогда из двух предыдущих соотношений сразу следует уравнение

$$\frac{\partial u}{\partial t} + \mathbf{v}\,\mathrm{grad}\,u = 0. \tag{13.1}$$

Равенство (13.1) будем в дальнейшем называть уравнением переноса пассивной примеси или линейным уравнением переноса. Кроме (13.1), будем рассматривать и одномерное уравнение переноса

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0, \tag{13.2}$$

а также неоднородное уравнение переноса

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = f, \tag{13.3}$$

где f — заданная функция, играющая роль источника (стока).

Неоднородным уравнением переноса описываются системы, в которых пассивная примесь может вступать, например, в химические реакции. Если уравнения переноса описывают распространение планктона, то в правой части будет стоять функция, описывающая размножение планктона и его пассивную утечку (например, за счет его поедания рыбами). О неоднородных линейных моделях речь пойдет ниже.

Уравнения вида (13.1, 13.2, 13.3) традиционно рассматриваются в курсах обыкновенных дифференциальных уравнений [1, 2].

Причина этого кроется в следующем обстоятельстве. Рассмотрим систему ОДУ

$$\dot{\mathbf{x}} = \mathbf{v}, \tag{13.4}$$

соответствующую (13.1). Это — уравнение характеристик для линейного уравнения переноса. Если функция u является *первым интегралом* (13.4), то она является решением (13.1). Иными словами, вдоль характеристики решение однородного уравнения переноса сохраняет постоянное значение.

Упражнение. Найти в явном виде уравнение характеристики для (13.2). В какое уравнение перейдет неоднородное уравнение переноса (13.3) вдоль характеристики?

Для корректной постановки задач для линейного уравнения переноса начальные и граничные условия необходимо ставить на некоторой гиперповерхности. Так как решение уравнений переноса распространяется вдоль характеристик, то начальная гиперповерхность должна быть трансверсальной ко всем характеристикам (не иметь точек касания с характеристиками). Кроме того, если для однородного уравнения переноса

какая-либо характеристика имеет с начальной гиперповерхностью более одной общей точки, то значения начальной функции во всех этих точках должны быть равны между собой. Все эти условия достаточно очевидны, если вспомнить физический смысл уравнения переноса.

Отметим, что наличие характеристик можно считать условием того, что система имеет гиперболический тип. Так, если система уравнений произвольного порядка n имеет n действительных характеристик, будем называть ее гиперболической. Таким образом, *линейное уравнение переноса имеет гиперболический тип.*

13.2. Квазилинейные уравнения гиперболического типа. Характеристики квазилинейных уравнений

Рассмотрим теперь более сложные примеры уравнений первого порядка в частных производных. Уравнение вида

$$\frac{\partial u}{\partial t} + \sum c_i \frac{\partial u}{\partial x_i} = f, \tag{13.5}$$

где $f = f(t, x_i, u), c_k = c_k(t, x_i, u)$, будем называть *квазилинейным уравнением* (переноса) [1, 2]. Естественно предположить, что все функции, входящие в запись (13.5), достаточно гладкие (подробнее в [1]). Для квазилинейных уравнений также существует понятие характеристики. В соответствии с [1, 2], уравнения характеристик для (13.5) будут иметь вид

$$\dot{\mathbf{x}} = \mathbf{v},$$

$$\dot{\mathbf{u}} = \mathbf{f}.$$

Последние соотношения для характеристик легко переписываются в координатах:

$$\frac{dx_i}{dt} = c_i(t, \mathbf{x}, u),$$

$$\frac{du}{dt} = f(t, \mathbf{x}, u).$$

Рассмотрим пример нахождения характеристик для квазилинейного уравнения.

Уравнения газовой динамики моделируются уравнением Хопфа:

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0.$$

Для вывода уравнения Хопфа рассмотрим одномерную среду, состоящую из частиц, которые движутся по инерции. Считаем, что движение происходит вдоль прямой. Запишем для частиц второй закон Ньютона:

$$0 = \ddot{x} = \frac{d}{dt}u = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}\frac{dx}{dt} = \frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x}.$$

Найдем характеристики уравнения Хопфа: $\frac{dx}{dt} = u, \frac{du}{dt} = 0$. Таким образом, характеристики уравнения Хопфа — семейство линий $x = a + bt, u = b$.

Отметим, что линейное уравнение можно рассматривать как частный случай квазилинейного уравнения. При этом можно ввести определение характеристики для линейного уравнения аналогично характеристике для квазилинейного уравнения. Заметим, что характеристика квазилинейного уравнения, вообще говоря, определяется в пространстве более высокой размерности, чем для линейного уравнения. Характеристики линейного уравнения, определенные выше, в этом случае совпадают с проекциями на гиперплоскость $(x_1, \ldots, x_n)$ характеристик квазилинейного уравнения. В частном случае для однородных квазилинейных уравнений можно рассматривать характеристики на гиперплоскости $(x_1, \ldots, x_n)$. Тогда характеристики уравнения Хопфа можно считать прямыми линиями $x(t) = x(0) + u(x(0), 0)t$. На практике при построении численных методов часто именно эти линии и считаются характеристиками уравнения Хопфа.

Используя метод характеристик, можно решать уравнения Хопфа. Рассмотрим в качестве примера уравнение Хопфа с начальным условием $u(x, 0) = ch^{-2}(x)$. Построим характеристики для данной задачи, пользуясь начальными данными. Вдоль каждой характеристики значение функции остается постоянным. Можно построить решение вплоть до момента времени, когда характеристики начинают пересекаться. После этого решение уравнения Хопфа в классическом смысле перестает существовать. Далее существует лишь разрывное *обобщенное решение* типа ударной волны. На разрыве (фронте ударной волны), вообще говоря, необходимо ставить дополнительные условия. Пересечение характеристик и образование ударной волны в решении также называют *градиентной катастрофой*. Более подробная информация о свойствах уравнения Хопфа в [3]. Конечно, возможно найти численное решение уравнения Хопфа в виде ударной волны. Речь об этом пойдет ниже.

13.3. Численные методы решения уравнений в частных производных гиперболического типа на примере линейного уравнения переноса

Двухслойные разностные схемы. Линейное одномерное уравнение переноса, как отмечалось выше, имеет вид (13.3)

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0, \quad c = \text{const};$$

Наряду с линейным уравнением переноса, в следующем параграфе будем рассматривать также *квазилинейные* уравнения. Простейшее квазилинейное уравнение гиперболического типа — уравнение Хопфа (13.5). Оно интересно тем, что моделирует уравнения газовой динамики. Запишем его в двух формах:

$$\frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} = 0, f = \frac{u^2}{2}, \tag{13.6}$$

(*дивергентная* форма). Отметим, что дивергентная форма записи отражает наличие некоторого закона сохранения. В случае уравнения Хопфа это закон сохранения импульса. В другой форме,

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0, \quad 0 \leqslant t \leqslant T, \quad 0 \leqslant x < \infty. \tag{13.7}$$

Эту форму записи иногда в дальнейшем будем называть *характеристической*. Отметим, что для линейных уравнений с постоянными коэффициентами нет принципиальной разницы между характеристической и дивергентной формами. Для уравнений с переменными коэффициентами эта разница возникает даже в линейном случае. При использовании различных форм записи для построения разностных схем возникают численные методы с разными свойствами.

Для корректной постановки начальной и краевой задач необходимо корректно задать начальные и граничные условия. Они для рассматриваемых ниже задач имеют вид

$$u(0,\ x) = \varphi_1(x), \quad 0 \leqslant x < \infty;$$

$$u(t,\ 0) = \varphi_2(t), \quad 0 \leqslant t \leqslant T.$$

Для определенности пока положим $c > 0$.

Решением задачи Коши для уравнения (13.3) является «бегущая волна»:

$$u(t, x) = \Psi(x - ct),$$

где c — скорость переноса, а функция $\Psi(x-ct)$ определяется из начальных или граничных условий. Характеристики уравнения имеют вид $x - ct = \text{const}$ и при постоянной скорости переноса являются прямыми линиями. Решение однородного уравнения (13.3), как отмечено выше, остается постоянным вдоль характеристики, поэтому начальные и граничные условия переносятся вдоль этих линий. В случае неоднородного уравнения вдоль характеристики оно превращается в обыкновенное дифференциальное уравнение.

Приведем вид двухпараметрического семейства двухслойных разностных схем первого и второго порядков точности [4]:

$$\frac{u_m^{n+1} - u_m^n}{\tau} + c\frac{\Delta^+ + \Delta^-}{2h}u_m^{(\gamma)} = \frac{h^2 q}{2\tau}\frac{\Delta^+\Delta^-}{h^2}u_m^{(\gamma)}, \qquad (13.8)$$

где применены обозначения

$$u_m^{(\gamma)} = \gamma u_m^{n+1} + (1-\gamma)u_m^n,$$

$$\Delta^+ u_m = u_{m+1} - u_m, \Delta^- u_m = u_m - u_{m-1},$$

$$(\Delta^+ + \Delta^-)u_m = \Delta^+ u_m + \Delta^- u_m = u_{m+1} - u_{m-1},$$

$$(\Delta^+\Delta^-)u_m = \Delta^+(\Delta^- u_m) = \Delta^+(u_m - u_{m-1}) =$$

$$= \Delta^+ u_m - \Delta^+ u_{m-1} = u_{m+1} - 2u_m + u_{m-1}.$$

Рассмотрим теперь конкретные разностные схемы решения модельного линейного уравнения переноса.

Схема П. Лакса (трехточечная схема) получается при $\gamma = 0, q = 1$. Ее порядок аппроксимации $O(\tau + h^2 + h^2/\tau)$. Здесь и далее порядок аппроксимации приводится для модельного уравнения переноса с постоянными коэффициентами. В случае переменных коэффициентов в схему надо внести необходимые изменения. При этом естественно линейное уравнение рассматривать как частный случай квазилинейного.

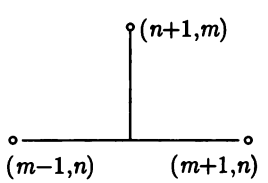


Схема является *условно устойчивой*, т. е. при выполнении условия Куранта $\sigma = c\tau/h \leqslant 1$. Отметим здесь, что величина σ играет определяющую роль при исследовании разностных схем на аппроксимацию и

устойчивость. Она называется *числом Куранта*. Исследование разностной схемы на устойчивость для линейного эволюционного уравнения с постоянными коэффициентами можно провести с использованием *спектрального признака* (фон Неймана).

Приведем разностные уравнения для схемы Лакса во внутренних точках расчетной области:

$$\frac{u_m^{n+1} - 0,5(u_{m+1}^n + u_{m-1}^n)}{\tau} + c\frac{u_{m+1}^n - u_{m-1}^n}{2h} = 0.$$

На рисунке приведен шаблон для схемы Лакса. Напомним, что шаблоном разностной схемы называется конфигурация узлов, значения сеточной функции в которых определяют вид разностных уравнений во внутренних (не приграничных) точках сетки. Как правило, на рисунках с изображениями шаблонов точки, участвующие в вычислении производных, соединяются линиями.

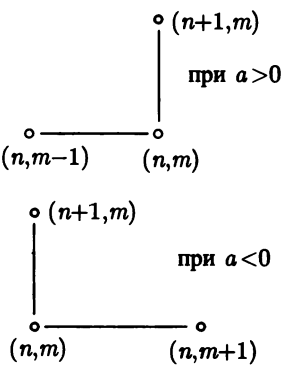


Схема Куранта-Изаксона-Риса (КИР), которую иногда также связывают с именем С.К. Годунова, получается при $\gamma = 0$, $q = \sigma \operatorname{sign} c$. Ее порядок аппроксимации $O(\tau + h)$. Схема КИР условно устойчива, т. е. при выполнении условия Куранта $\sigma = c\tau/h \leqslant 1$. Приведем разностные уравнения для схемы Куранта-Изаксона-Риса во внутренних точках расчетной области:

$$\frac{u_m^{n+1} - u_m^n}{\tau} + c\frac{u_m^n - u_{m-1}^n}{h} = 0, \quad c > 0,$$

$$\frac{u_m^{n+1} - u_m^n}{\tau} + c\frac{u_{m+1}^n - u_m^n}{h} = 0, \quad c < 0.$$

Эти схемы, имеющие также название схемы с разностями против потока (в англоязычной литературе — upwind) могут быть записаны в ви-

де

$$u_m^{n+1} = u_m^n - \sigma \begin{cases} u_{m+1}^n - u_m^n, & a < 0, \\ u_m^n - u_{m-1}^n, & a \geqslant 0. \end{cases}$$

Их преимущество состоит в более точном учете области зависимости решения. Если ввести обозначения

$$a^+ = \frac{1}{2}(a + |a|) = \begin{cases} a, & a \geqslant 0, \\ 0, & a < 0, \end{cases} \quad a^- = \frac{1}{2}(a - |a|) = \begin{cases} 0, & a \geqslant 0, \\ a, & a < 0, \end{cases}$$

то обе схемы можно записать в следующих формах:

$$u_m^{n+1} = u_m^n - \frac{\tau}{h}\left[a^+(u_m^n - u_{m-1}^n) + a^-(u_{m+1}^n - u_m^n)\right];$$

$$u_m^{n+1} = u_m^n - \frac{\tau}{h}(f_{m+1/2}^n - f_{m-1/2}^n), \quad f_{m+1/2}^n =$$

$$= \frac{1}{2}\left[a(u_{m+1}^n + u_m^n) - |a|\,(u_{m+1}^n - u_m^n)\right],$$

$$f_{m-1/2}^n = \frac{1}{2}\left[a(u_m^n + u_{m-1}^n) - |a|\,(u_m^n - u_{m-1}^n)\right]$$

(потоковая форма разностного уравнения);

$$u_m^{n+1} = u_m^n - \frac{1}{2}\sigma(u_{m+1}^n - u_{m-1}^n) + \frac{|\sigma|}{2}(u_{m+1}^n - 2u_m^n + u_{m-1}^n)$$

(здесь явно выделен член со второй разностью, придающий устойчивость схеме);

$$\Delta_\tau u_m^n = \frac{|\sigma| - \sigma}{2}\Delta^+ u_m^n - \frac{|\sigma| + \sigma}{2}\Delta^- u_m^n$$

(уравнение в конечных приращениях).

Рассмотрим также метод неопределенных коэффициентов для построения разностной схемы правый уголок первого порядка точности для уравнения переноса

$$\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = \varphi(t, x), \text{ т. е. при } a = -1.$$

Схему можно представить в виде

$$L_\tau u^\tau = b_1 u_m^{n+1} + b_2 u_m^n + b_3 u_{m+1}^n = \varphi_m^n,$$

$$b_1 = \tau^{-1}, b_2 = h^{-1} - \tau^{-1}, b_3 = -h^{-1}.$$

Схема Куранта-Изаксона-Риса тесно связана с численными методами характеристик. Дадим краткое описание идеи таких методов.

Две последние полученные схемы (при разных знаках скорости переноса) можно интерпретировать следующим образом. Построим характеристику, проходящую через узел (t_{n+1}, x_m), значение в котором необходимо определить, и пересекающую слой t_n в точке $x' = x_m - c\tau$. Для определенности считаем, что скорость переноса c положительна.

Проведя линейную интерполяцию между узлами x_{m-1} и x_m на нижнем слое по времени, получим

$$u_m(x') = u_{m-1}\frac{x_m - x'}{h} + u_m\frac{x' - x_{m-1}}{h} =$$

$$= \frac{c\tau}{h}u_{m-1} + (1 - \frac{c\tau}{h})u_m = \sigma u_{m-1} - (1-\sigma)u_m.$$

Далее перенесем вдоль характеристики значение $u_n(x')$ без изменения на верхний слой t_{n+1}, т. е. положим $u_m^{n+1} = u_n(x')$. Последнее значение естественно считать приближенным решением однородного уравнения переноса. В таком случае

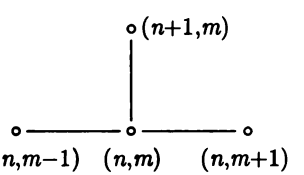
$$u_m^{n+1} = (1-\sigma)u_m + \sigma u_{m-1},$$

или, переходя от числа Куранта снова к сеточным параметрам,

$$\frac{u_m^{n+1} - u_m^n}{\tau} + a\frac{u_m^n - u_{m-1}^n}{h} = 0,$$

т. е. другим способом пришли к уже известной схеме «левый уголок», устойчивой при $\sigma = c\tau/h \leqslant 1, \quad c > 0$. При $\sigma > 1$ точка пересечения характеристики, выходящей из узла (t_{n+1}, x_m), с n-м слоем по времени расположена левее узла (t_n, x_{m-1}). Таким образом, для отыскания решения u_m^{n+1} используется уже не интерполяция, а экстраполяция, которая оказывается неустойчивой.

Неустойчивость схемы «правый уголок» при $c > 0$ также очевидна. Для доказательства этого можно использовать либо спектральный признак, либо условие Куранта, Фридрихса и Леви. Аналогичные рассуждения можно провести и для случая $c < 0$ и схемы «правый уголок».



Неустойчивая **четырехточечная схема** получается при $\gamma = 0, q = 1$, ее порядок аппроксимации $O(\tau + h^2)$. Сеточные уравнения для разностной

схемы будут иметь следующий вид:

$$\frac{u_m^{n+1}-u_m^n}{\tau}+c\frac{u_{m+1}^n-u_{m-1}^n}{2h}=0.$$

Схема Лакса-Вендроффа возникает при $\gamma=0, q=\sigma$. Порядок аппроксимации схемы Лакса-Вендроффа есть $O(\tau^2+h^2)$. Схема устойчива при выполнении условия Куранта $\sigma=c\tau/h\leqslant 1$.

Эту схему можно получить либо методом неопределенных коэффициентов, либо путем более точного учета главного члена погрешности аппроксимации. Рассмотрим процесс вывода схемы Лакса-Вендроффа подробнее. Проводя исследование предыдущей четырехточечной схемы на аппроксимацию (а исследование это довольно элементарно и сводится к разложению функции проекции на сетку точного решения дифференциальной задачи в ряд Тейлора), получим для главного члена погрешности

$$\frac{u_m^{n+1}-u_m^n}{\tau}+c\frac{u_{m+1}^n-u_{m-1}^n}{2h}=\frac{\partial u}{\partial t}+c\frac{\partial u}{\partial x}\bigg|_{t_n,x_m}+$$

$$+c^2\frac{\tau}{2}\frac{\partial^2 u}{\partial x^2}+O(\tau^2+h^2).$$

При выводе выражения для главного члена погрешности аппроксимации использовано следствие исходного дифференциального уравнения переноса $\frac{\partial^2 u}{\partial t^2}=c^2\frac{\partial^2 u}{\partial x^2}$, которое получается путем дифференцирования исходного уравнения (13.3) сначала по времени t, затем по координате x и вычитанием одно из другого получившихся соотношений.

Далее, заменяя вторую производную во втором слагаемом в правой части с точностью до $O(h^2)$, получим новую разностную схему, аппроксимирующую исходное дифференциальное уравнение с точностью $O(\tau^2+h^2)$. Сеточные уравнения для схемы Лакса–Вендроффа во внутренних узлах расчетных сеток есть

$$\frac{u_m^{n+1}-u_m^n}{\tau}+c\frac{u_{m+1}^n-u_{m-1}^n}{2h}-c^2\frac{\tau}{2}\frac{u_{m-1}^n-2u_m^n+u_{m+1}^n}{h^2}=0.$$

Неявная шеститочечная схема возникает при $q=0$; при $\gamma=0,5$ ее порядок аппроксимации $O(\tau^2+h^2)$, при $\gamma=1-O(\tau+h^2)$.

Построить шаблоны схемы при $\gamma=0,5$ и при $\gamma=1$.

Неявная нецентральная схема. Рассмотрим случай $q=\sigma\operatorname{sign} c$. При $\gamma=0,5$ порядок аппроксимации — $O(\tau^2+h)$. При $\gamma=1-O(\tau+h)$. Упражнение. Нарисовать шаблон схемы при $\gamma=0,5$ и при $\gamma=1$.

Последние две разностные схемы носят названия схем Ландау-Меймана-Халатникова и Карлсона, соответственно.

Явная схема Бима-Уорминга

Бим и Уорминг предложили изменить метод Мак-Кормака, используя на обоих этапах односторонние разности одинаковой направленности; для линейного уравнения переноса эта схема будет

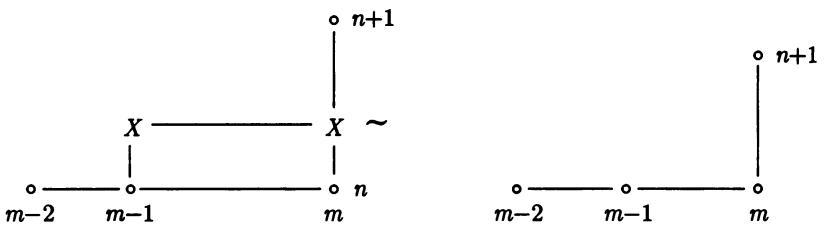
$$\frac{\tilde{u}_m - u_m}{\tau} + a\frac{u_m^n - u_{m-1}^n}{h} = 0,$$

$$\frac{u_m^{n+1} - \frac{1}{2}(u_m^n + \tilde{u}_m)}{\tau} + a\frac{\tilde{u}_m - \tilde{u}_{m-1}}{h} + a\frac{u_m^n - 2u_{m-1}^n + u_{m-2}^n}{h^2} = 0.$$

При подстановке первого уравнения во второе, получим

$$u_m^{n+1} = u_m^n - \sigma(u_m^n - u_{m-1}^n) + \frac{\sigma}{2}(1-\sigma)(u_m^n - 2u_{m-1}^n + u_{m-2}^n) = 0.$$

Шаблоны двух- и одноэтапной схем имеют вид



Эта же схема может быть записана в приращениях

$$\Delta_\tau u_m^n = \frac{\sigma(\sigma-1)}{2}\Delta^- u_m^n - \frac{\sigma(1+\sigma)}{2}\Delta^- u_{m-1}^n.$$

13.4. Численные методы решения уравнений в частных производных гиперболического типа для квазилинейного уравнения переноса

Рассмотрим теперь простейшие разностные схемы для уравнения Хопфа.

Обобщение на случай уравнения Хопфа схемы П. Лакса имеет вид

$$\frac{u_m^{n+1} - 0,5(u_{m+1}^n - u_{m-1}^n)}{\tau} + \frac{f_{m+1}^n - f_{m-1}^n}{h} = 0.$$

Здесь, очевидно, используется дивергентный вид уравнения (13.6).

Упражнения. Рассмотрим схему Лакса-Вендроффа для уравнения Хопфа. Пусть начальные условия для задачи Коши поставлены следующим образом: $u(x,0) = ch^{-2}(x)$. Тогда уравнение Хопфа имеет первый интеграл: $\int\limits_{-\infty}^{+\infty} u(x,t)dx = \text{const}$. Проверить, что приведенная выше схема является *консервативной*, т. е. в ней на сеточном уровне автоматически выполняется тот же закон сохранения.

Построить аналогичную схему с использованием *характеристической формы* записи уравнения Хопфа (13.9). Будет ли она консервативной?

Схема условно устойчива при выполнении условия Куранта (точнее, обобщения условия Куранта) $\frac{\tau}{h}\max\limits_m |u_m^n| \leqslant 1$. Здесь и ниже, как и ранее в (13.7), $f = 0{,}5\,u^2$. При этом предполагается, что течение достаточно гладкое, момент градиентной катастрофы еще не наступил, в решении нет ударных волн и других разрывов.

Схема Куранта-Изаксона-Риса. Обобщение схем КИР на квазилинейный случай (при использовании дивергентной формы записи уравнений) очевидно.

$$\frac{u_m^{n+1} - u_m^n}{\tau} + \frac{f_{m+1}^n - f_m^n}{h} = 0, \quad u_m^n < 0,$$

$$\frac{u_m^{n+1} - u_m^n}{\tau} + \frac{f_m^n - f_{m-1}^n}{h} = 0, \quad u_m^n > 0.$$

Схема устойчива при выполнении условия Куранта $\frac{\tau}{h}\max\limits_m |u_m^n| \leqslant 1$.

Обобщение **схемы Лакса-Вендроффа** (схема предиктор-корректор). Для квазилинейных уравнений (а также линейных уравнений с переменными коэффициентами, неоднородных уравнений и т. п.) схема Лакса-Вендроффа становится более сложной. Для ее построения необходимо ввести так называемые полуцелые точки (точки с дробными индексами). На первом этапе (предиктора) значения в полуцелых точках вычисляются по приведенной выше схеме — обобщению на квазилинейный случай схемы Лакса:

$$\frac{u_{m+1/2}^{n+1/2} - 0{,}5(u_{m+1}^n + u_m^n)}{\tau/2} + \frac{f_{m+1}^n - f_m^n}{h} = 0,$$

$$\frac{u_{m-1/2}^{n+1/2} - 0{,}5(u_m^n + u_{m-1}^n)}{\tau/2} + \frac{f_m^n - f_{m-1}^n}{h} = 0,$$

на втором этапе (корректор) используется схема «чехарда» (трехслойная схема на крестообразном шаблоне, которая не входит в семейство (13.8)):

$$\frac{u_m^{n+1} - u_m^n}{\tau} + \frac{f_{m+1/2}^{n+1/2} - f_{m-1/2}^{n+1/2}}{h} = 0.$$

Схема Лакса-Вендроффа принадлежит к так называемым *центральным* схемам. Ее шаблон симметричен. На первом этапе рассчитываются значения сеточной функции в полуцелых точках шаблона на промежуточном слое $(t_{m-1/2},\ x_{m-1/2}), (t_{n+1/2},\ x_{m+1/2})$, на втором этапе вычисляется решение на верхнем слое в точке $(t_{n+1},\ x_m)$. Схема устойчива при выполнении условия Куранта.

Аналогично строятся схемы Лакса-Вендроффа для линейных неоднородных уравнений.

Нецентральная **схема Мак-Кормака** (предиктор-корректор).

Как и приведенная выше схема Лакса-Вендроффа, схема МакКормака состоит из двух этапов. Рассмотрим построение схемы МакКормака для однородного уравнения (13.7). Первый этап (предиктор) имеет вид

$$\frac{\tilde{u}_m - u_m^n}{\tau} + \frac{f_{m+1}^n - f_m^n}{h} = 0,$$

$$\frac{\tilde{u}_{m-1} - u_m^n}{\tau} + \frac{f_m^n - f_{m-1}^n}{h} = 0,$$

т. е. используется схема «явный правый уголок». Второй этап — корректор:

$$\frac{u_m^{n+1} - 0,5(u_m^n + \tilde{u}_m)}{\tau} + \frac{\tilde{f}_m - \tilde{f}_{m-1}}{2h} = 0.$$

Таким образом, расчет на первом этапе по схеме «правый уголок», на втором — «левый уголок».

Другая схема Мак-Кормака имеет вид

$$\frac{\tilde{u}_m - u_m^n}{\tau} + \frac{f_m^n - f_{m-1}^n}{h} = 0,$$

$$\frac{\tilde{u}_{m+1} - u_{m+1}^n}{\tau} + \frac{f_{m+1}^n - f_m^n}{h} = 0,$$

$$\frac{u_m^{n+1} - 0,5(u_m + \tilde{u}_m)}{\tau} + \frac{\tilde{f}_{m+1} - \tilde{f}_m}{2h} = 0.$$

Такие разностные схемы называют *нецентральными*. К их преимуществам относят отсутствие полуцелых индексов, более простую постановку граничных условий. В линейном случае схемы Мак-Кормака совпадают

со схемой Лакса-Вендроффа. Схемы имеют второй порядок аппроксимации по обеим переменным, схемы устойчивы при выполнении условия Куранта.

Схема Русанова (центральная схема третьего порядка точности).

Для построения схемы Русанова вводятся не только полуцелые точки, но и два слоя промежуточных точек с дробными индексами. Первый этап схемы Русанова (переход к слою 1/3) имеет вид

$$\frac{u_{m+1/2}^{n+1/3} - 0,5(u_m^n + u_{m+1}^n)}{\tau/3} + \frac{f_{m+1}^n - f_m^n}{h} = 0,$$

$$\frac{u_{m-1/2}^{n+1/3} - 0,5(u_m^n + u_{m-1}^n)}{\tau/3} + \frac{f_m^n + f_{m-1}^n}{h} = 0,$$

ее второй этап есть схема «чехарда»

$$\frac{u_m^{n+2/3} - u_m^n}{2\tau/3} + \frac{f_{m+1/2}^{n+1/3} - f_{m-1/2}^{n+1/3}}{h} = 0,$$

а третий этап

$$\frac{u_m^{n+1} - u_m^n}{\tau} + \frac{3}{8}\frac{f_{m+1}^{n+2/3} - f_{m-1}^{n+2/3}}{h} +$$

$$+ \frac{-2f_{m+2}^n + 7f_{m+1}^n - 7f_{m-1}^n + 2f_{m-2}^n}{24h} +$$

$$+ \frac{\omega}{24}(u_{m+2}^n - 4u_{m+1}^n + 6u_m^n - 4u_{m-1}^n + u_{m-2}^n) = 0.$$

На первом этапе производится расчет по схеме Лакса, на втором — по схеме «крест» («чехарда»). Последнее слагаемое третьего этапа вводится для обеспечения устойчивости схемы (член, пропорциональный разностной аппроксимации 4-й производной).

Схема является условно устойчивой при выполнении условия Куранта и условия $4\sigma^2 - \sigma^4 \leqslant \omega \leqslant 3$.

Нецентральная **схема Уорминга-Кутлера-Ломакса** 3-го порядка точности.

Первый этап:

$$\frac{u_m^{n+1/3} - u_m^n}{2\tau/3} + \frac{f_{m+1}^n - f_m^n}{h} = 0,$$

$$\frac{u_{m-1}^{n+1/3} - u_{m-1}^n}{2\tau/3/3} + \frac{f_m^n - f_{m-1}^n}{h} = 0,$$

Второй этап:

$$\frac{u_m^{n+2/3} - 0,5(u_m^n + u_m^{n+1/3})}{2\tau/3} + \frac{f_m^{n+1/3} - f_{m-1}^{n+1/3}}{2h} = 0,$$

$$\frac{u_{m+1}^{n+2/3} - 0,5(u_{m+1}^n + u_{m+1}^{n+1/3})}{2\tau/3} + \frac{f_{m+1}^{n+1/3} - f_m^{n+1/3}}{2h} = 0,$$

Третий этап:

$$\frac{u_m^{n+1} - u_m^n}{\tau} + \frac{3}{8} \cdot \frac{f_{m+1}^{n+1/3} - f_{m-1}^{n+1/3}}{h} +$$

$$+ \frac{-2f_{m+2}^n + 7f_{m+1}^n - 7f_{m-1}^n + 2f_{m-2}^n}{24h} +$$

$$+ \frac{\omega}{24}(u_{m+2}^n - 4u_{m+1}^n + 6u_m^n - 4u_{m-1}^n + u_{m-2}^n) = 0.$$

Последний член добавляется для устойчивости схемы, которая является условно устойчивой при выполнении условий Куранта.

Неявная схема Бима-Уорминга.

Схема Бима-Уорминга имеет вид:

$$\frac{u_m^{n+1} - u_m^n}{\tau} + \frac{1}{2}\left[\frac{\partial f}{\partial x}(t^n, x_m) + \frac{\partial f}{\partial x}(t^{n+1}, x_m)\right] = 0.$$

Рассмотрим алгоритм расчетов по схеме Бима-Уорминга. Линеаризуем функцию f_m^{n+1}:

$$f_m^{n+1} \approx f_m^n + \frac{\partial f(t^n, x_m)}{\partial u}(u_m^{n+1} - u_m^n) = f_m^n + F_m^n(u_m^{n+1} - u_m^n).$$

Здесь мы воспользовались тем, что функция f не зависит явно от времени, иначе формулы будут выглядеть несколько иначе. Тогда будем иметь равенство

$$\frac{u_m^{n+1} - u_m^n}{\tau} + \frac{1}{2}\left\{2\frac{\partial f(t^n, x_m)}{\partial u} + \frac{\partial}{\partial x}[F_m^n(u_m^{n+1} - u_m^n)]\right\} = 0.$$

Проведя аппроксимацию производной $(\partial f/\partial x)_m^n$ центральными разностями, получим искомую неявную разностную схему:

$$-\frac{\tau}{4h}A_{m-1}^n \cdot u_{m-1}^{n+1} + u_m^{n+1} + \frac{\tau}{4h}A_{m+1}^n \cdot u_{m+1}^{n+1} =$$

$$= -\frac{\tau}{h} \cdot \frac{f_{m+1}^n - f_{m-1}^n}{2} - \frac{\tau}{4h}A_{m-1}^n \cdot u_{m-1}^n + u_m^n + \frac{\tau}{4h}A_{m+1}^n \cdot u_{m+1}^n.$$

В результате для аппроксимации уравнения Хопфа получается система линейных алгебраических уравнений с трехдиагональной матрицей, алгоритм численного решения которой — один из вариантов трехточечной прогонки.

13.5. Методы регуляризации численных решений с большими градиентами

Кроме основных понятий теории разностных схем — аппроксимации, устойчивости, сходимости — на практике существенную роль играют дополнительные свойства разностных схем. Среди таких свойств упомянем *монотонность*. Не существует общепринятого определения монотонности разностной схемы. Пока, до рассмотрения разностных схем в пространстве неопределенных коэффициентов, воспользуемся определением, данным Борисом и Буком в [5, 6].

Под монотонными далее будем понимать такие разностные схемы, в которых не увеличивается число локальных экстремумов (минимумов и максимумов численного решения) по сравнению с числом локальных экстремумов в решении точной задачи. Кроме того, монотонные схемы не должны увеличивать по абсолютному значению уже имеющиеся экстремумы.

Свойство монотонности разностной схемы очень полезно при расчете разрывных решений. К сожалению, доказанная для линейного уравнения переноса теорема С. К. Годунова гласит, что среди линейных разностных схем (термин «линейный» уточним чуть ниже) не существует монотонных с порядком аппроксимации выше первого.

Зачастую в практических задачах схемы первого порядка аппроксимации не могут обеспечить требуемую точность численного решения.

При использовании *немонотонных* схем для получения численных решений с большими градиентами, появляются осцилляции разностного происхождения. По этой причине в численных методах часто используется *регуляризация* численных решений. Рассмотрим наиболее распространенные методы регуляризации.

Сглаживание численного решения на верхнем слое (метод Л. А. Чудова).

Вычисленное с помощью немонотонной разностной схемы значение функции в точке x_n корректируется следующим образом:

$$\tilde{u}_m = (1 - 2\alpha)u_m + \alpha u_{m-1} + \alpha u_{m+1},$$

причем при $\alpha = 0,5$ значение $\tilde{u}_m$ является средним арифметическим

$$\tilde{u}_n = 0,5(u_{m-1} + u_{m+1}).$$

Для того чтобы понять действие сглаживающего оператора, представим явную схему для численного решения линейного одномерного уравнения теплопроводности

$$\frac{\partial u}{\partial t} = a\frac{\partial^2 u}{\partial x^2}:$$

$$u_m^{n+1} = (1 - 2\frac{a\tau}{h^2})u_m^n + \frac{a\tau}{h^2}u_{m-1}^n + \frac{a\tau}{h^2}u_{m+1}^n;$$

т. е. при $\alpha = a\tau/h^2$ сглаживание аппроксимирует диссипативный член, пропорциональный второй производной по координате. Понятно, что скорректированное решение будет устойчивым при $\alpha \leqslant 0,5$. Для снятия этого ограничения можно ввести алгоритм неявного сглаживания. Он имеет вид

$$\tilde{u}_m = (1 - 2\alpha)\tilde{u}_m + \alpha\tilde{u}_{m-1} + \alpha\tilde{u}_{m+1}.$$

Подробнее о сглаживании по Л. А. Чудову можно прочитать, например, в [7].

Аппроксимационная вязкость.

Рассмотрим схему первого порядка аппроксимации, для численного решения модельного однородного уравнения переноса (13.3)

$$\mathbf{L}_\tau u^\tau = \frac{u_m^{n+1} - u_m^n}{\tau} + c\frac{u_m^n - u_{m-1}^n}{h} = 0,$$

исследование которой на аппроксимацию дает следующее выражение для главных членов ошибки аппроксимации (главных членов невязки):

$$\mathbf{L}_\tau u^\tau = \mathbf{L}u + \frac{\tau}{2}\cdot\frac{\partial^2 u}{\partial t^2} - \frac{ch}{2}\cdot\frac{\partial^2 u}{\partial x^2} + O(\tau^2 + h^2) =$$

$$= \mathbf{L}u - \frac{ch}{h}(1 - \frac{c\tau}{2})\frac{\partial^2 u}{\partial x^2} + O(\tau^2 + h^2)$$

Таким образом, с точностью до членов второго порядка, аппроксимируется уравнение

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = \gamma\frac{\partial^2 u}{\partial x^2},$$

где аналог коэффициента теплопроводности $\gamma = \frac{ch}{2}(1 - \frac{c\tau}{h})$ — так называемый коэффициент аппроксимационной вязкости, а член

$$\eta = \frac{ch}{2}(1 - \frac{c\tau}{h})\frac{\partial^2 u}{\partial x^2}$$

именуется аппроксимационной вязкостью. Его действие — сглаживание численного решения. Слагаемые в правой части уравнения переноса, пропорциональные второй производной, моделируют диссипативный

эффект. Однако эта вязкость является свойством выбранной разностной схемы первого порядка аппроксимации. Здесь в уравнение не вводятся никакие дополнительные члены и не используются сглаживающие операторы. Рассматриваемое уравнение носит название *первого дифференциального приближения* разностной схемы. Все схемы первого порядка аппроксимации будут обладать схемной вязкостью.

Упражнение. Некоторая разностная схема для решения уравнения переноса (13.3) первого порядка аппроксимации обладает отрицательным коэффициентом схемной вязкости. Что можно сказать о других свойствах этой схемы (устойчивость, сходимость, монотонность)?

Искусственная вязкость.

Для регуляризации решения, полученного с помощью немонотонных разностных схем с порядком аппроксимации выше первого можно вводить так называемую *искусственную вязкость*. Этот способ регуляризации решения был предложен фон Нейманом и Рихтмайером для численного решения системы уравнений газовой динамики. В рамках данной лекции ограничимся рассмотрением модельного линейного уравнения переноса.

Идею искусственной вязкости можно проиллюстрировать на примере, вообще говоря, неустойчивой схемы

$$\mathrm{L}_\tau u^\tau = \frac{u_m^{n+1} - u_m^n}{\tau} + c\frac{u_{m+1}^n - u_{m-1}^n}{2h} = 0$$

для решения линейного одномерного уравнения переноса (13.3).

Для того чтобы сделать эту схему устойчивой, введем в правую часть член порядка погрешности аппроксимации, моделирующий диссипативный эффект, так называемую искусственную вязкость. Получим следующую задачу:

$$\mathrm{L}_\tau u^\tau = \xi\tau \cdot \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2},$$

где $\frac{\xi\tau}{h^2}(u_{m+1}^n - 2u_m^n + u_{m-1}^n)$ — искусственная вязкость, величина порядка главного члена погрешности аппроксимации (невязки) ξ — коэффициент искусственной вязкости. Исследование этой схемы на устойчивость с помощью *спектрального признака* дает $r \leqslant 0,5$, $\frac{\sigma^2}{2r} \leqslant 1$, где $r = \frac{\xi\tau}{h^2}$, $\sigma = \frac{c\tau}{h}$. При $\xi = 0,5$ для рассматриваемого линейного уравнения получим схему Лакса-Вендроффа.

Упражнение. Построить четырехточечную схему с искусственной вязкостью для квазилинейного уравнения (уравнения Хопфа в форме (13.5) и в форме (13.6)).

Методы коррекции потоков Бориса–Бука.

Рассмотрим идею введения схемной антидиффузии, предложенную Борисом и Буком в [5, 6]. Пусть $\mathrm{L}_\tau u^\tau = 0$ — схема первого порядка

точности, аппроксимирующая линейное одномерное уравнение переноса (13.3) и обладающая аппроксимационной вязкостью.

Уменьшим влияние последней на численное решение, введя так называемые потоки антидиффузии (терминология авторов метода):

$$u_m^{n+1} = \tilde{u}_m - \frac{1}{h}(\Phi_{m+1/2} - \Phi_{m-1/2}),$$

где $\tilde{u}_m$ — решение, полученное по упомянутой схеме, $\Phi_{m\pm1/2}$ — потоки, имеющие вид

$$\Phi_{m+1/2} = \frac{\xi_{m+1/2} \cdot \tau}{h}(\tilde{u}_{m+1} - \tilde{u}_m),$$

$$\Phi_{m-1/2} = \frac{\xi_{m-1/2} \cdot \tau}{h}(\tilde{u}_m - \tilde{u}_{m-1}).$$

При $\xi_{m-1/2} = \xi_{m+1/2}$ разностное уравнение имеет вид

$$u_m^{n+1} = \tilde{u}_m - \frac{\xi\tau}{h}(\tilde{u}_{m+1} - 2\tilde{u}_m + \tilde{u}_{m-1}).$$

Таким образом, идея метода коррекции потоков состоит во введении сглаживающего оператора определенного вида. Метод коррекции потоков описан, например, в [5].

13.6. Гибридные схемы (метод Р. П. Федоренко)

Идею построения гибридных схем, изложенную в [11], рассмотрим на традиционном примере схемы "уголок"

$$L_\tau u^\tau = \frac{u_m^{n+1} - u_m^n}{\tau} + \frac{u_m^n - u_{m-1}^n}{h} = 0$$

для численного решения модельного уравнения переноса

$$Lu = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0.$$

Разложения сеточных функций проекции на сетку точного решения дифференциальной задачи в ряд Тейлора дает

$$L_\tau u^\tau = Lu + \frac{\tau}{2} \cdot \frac{\partial^2 u}{\partial t^2} - \frac{h}{2} \cdot \frac{\partial^2 u}{\partial x^2} + O(\tau^2 + h^2).$$

Поскольку $\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}$, что показывается дифференцированием рассматриваемого уравнения переноса по t и по x, полученное выражение может быть представлено в виде

$$\frac{u_m^{n+1} - u_m^n}{\tau} + \frac{u_m^n - u_{m-1}^n}{h} + \frac{1}{2\tau}\left(\frac{\tau}{h} - \frac{\tau^2}{h^2}\right)(u_{m-1}^n - 2u_m^n + u_{m+1}^n) = 0.$$

Аналогичным образом можно получить и схему третьего порядка точности:

$$\frac{u_m^{n+1}-u_m^n}{\tau}+\frac{u_m^n-u_{m-1}^n}{h}+\frac{1}{2\tau}\left(\frac{\tau}{h}-\frac{\tau^2}{h^2}\right)\left(u_{m-1}^n-2u_m^n+u_{m+1}^n\right)+$$

$$+\frac{1}{6}\left(\frac{\tau}{h}-\frac{\tau^3}{h^3}\right)\left(u_{m+1}^n-3u_{m-1}^n+3u_{m-1}^n-u_{m-2}^n\right)=0.$$

Введем разностный анализатор гладкости численного решения, сравнивая конечные разности первого и второго порядков:

$$\gamma=\begin{cases}1, & \left|u_{m-1}^n-2u_m^n+u_{m+1}^n\right|<\lambda\left|u_m^n-u_{m-1}^n\right|\\ 0, & \left|u_{m-1}^n-2u_m^n+u_{m+1}^n\right|>\lambda\left|u_m^n-u_{m-1}^n\right|\end{cases}$$

и представим полученную схему в виде:

$$\frac{u_m^{n+1}-u_m^n}{\tau}+\frac{u_m^n-u_{m-1}^n}{h}+\frac{\gamma}{2\tau}\left(\frac{\tau}{h}-\frac{\tau^2}{h^2}\right)\left(u_{m+1}^n-2u_m^n+u_{m-1}^n\right)=0.$$

Таким образом, в областях с большим градиентом численного решения $\gamma=0$ и расчет ведется по схеме первого порядка точности, в области же гладкого решения $\gamma=1$ и расчет ведется по схеме второго порядка, (при $\lambda=0$ имеем схему первого, при $\lambda=\infty$ – второго порядка точности). Аналогичный анализатор можно ввести и для схемы третьего порядка аппроксимации.

Заметим, что гибридные схемы, построенные выше для аппроксимации линейного уравнения переноса, уже нелинейные — коэффициенты переключения зависят от локальных свойств решения. Таким образом, в соответствие линейному дифференциальному оператору ставится нелинейный. Для таких схем не обязана выполняться теорема С. К. Годунова, и можно ожидать, что на пути введения нелинейности для гиперболических систем и уравнений можно построить монотонные или близкие к монотонным схемы высокого порядка аппроксимации.

13.7. Схемы с уменьшением полной вариации (Total Variation Diminishing, схемы Хартена)

Схемы с уменьшением полной вариации (сокращенно их называют TVD-схемами) описаны, например, в [8, 12]. Для построения схемы рассмотрим полную вариацию численного решения. Она определяется следующим образом:

$$\mathrm{Var}\left(u^n\right)=\sum_{j=-\infty}^{\infty}\left|u_{j+1}^n-u_j^n\right|. \tag{13.9}$$

Схема будет TVD, если

$$\mathrm{Var}\,(u^{n+1}) \leqslant \mathrm{Var}\,(u^n). \tag{13.10}$$

Суть построения TVD-схем можно понять, представив схему Лакса-Вендроффа для численного решения модельного уравнения переноса в виде

$$u_m^{n+1} = u_m^n - (\sigma)(u_m^n - u_{m-1}^n) - (f_{m+1/2}^n - f_{m-1/2}^n), \tag{13.11}$$

где $\sigma = c\tau/h$, $f_{m+1/2} = 0,5\sigma(1-\sigma)(u_{m+1} - u_m)$ — антидиффузионные потоки. Схема похожа на метод коррекции потоков, но одношаговый. Эта схема не монотонна. Сделаем ее монотонной, ограничив антидиффузионные потоки введением функций $\varphi(r_m)$:

$$f'_{m+1/2} = 0,5\varphi(r_m)\sigma(1-\sigma)(u_{m+1} - u_m),$$

аналогично для $f'_{m-1/2}$. Здесь φ — ограничитель,

$$r_m = \frac{u_m - u_{m-1}}{u_{m+1} - u_m},$$

отношение прилежащих градиентов $\varphi(r_m)$ выбирается так, чтобы (13.11) была TVD-схемой, причем в расчетах обычно полагают $r_m = \frac{u_m - u_{m-1} + \varepsilon}{u_{m+1} - u_m + \varepsilon}$, где ε — малое число.

Можно показать, что условием устойчивости схемы является неравенство

$$0 < \varphi(r_m) \leqslant \min(2r_m,\, 2) \text{ при } r_m > 0 \text{ и } \varphi(r_m) = 0 \text{ при } r_m \leqslant 0.$$

Простейшим ограничителем является выбор конечных разностей в соответствии с принципом минимальных производных Колгана:

$$\varphi(r_m) = 1, \quad f'_{m+1/2} = \frac{\sigma}{2}(1-\sigma) \cdot \begin{cases} \Delta u_m^n, & |\Delta u_m^n| \leqslant \Delta^- u_m^n, \\ \Delta^- u_m^n, & |\Delta u_m^n| > \Delta^- u_m^n. \end{cases}$$

Приведем пример другого ограничителя:

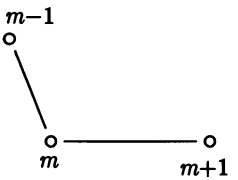
$$\varphi(r) = \begin{cases} \min(2, r_m), & r_m > 1, \\ \min(2r_m, 1), & 0 < r_m \leqslant 1, \\ 0, & r_m \leqslant 0. \end{cases} \quad \text{или}$$

$$\varphi(r_m) = \begin{cases} \min \mathrm{mod}(2, r_m), & r_m > 1 \\ \min \mathrm{mod}(1, 2r_m), & 0 < r_m < 1, \end{cases}$$

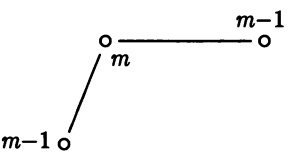
где $\min \mathrm{mod}(x,y) = \frac{\mathrm{sign}x + \mathrm{sign}y}{2} \min(|x|, |y|)$.

Пояснение.

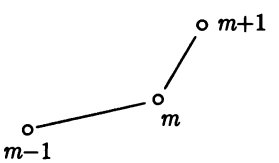
1. $r_m > 1 : u_m - u_{m-1} > u_{m+1} - u_m$(если и числитель, и знаменатель в выражении для r_m положительны) или $u_{m-1} - u_m > u_m - u_{m+1}$(если и числитель, и знаменатель отрицательны); численное решение сглаживается, так как его градиент убывает.



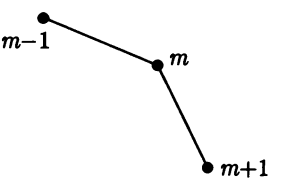
или



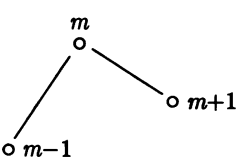
2. $0 < r_m \leqslant 1; u_m - u_{m-1} \leqslant u_{m+1} - u_m$ (если и числитель, и знаменатель в выражении для r_m положительны) или $u_{m-1} - u_m < u_m - u_{m+1}$ (если и числитель, и знаменатель отрицательны); градиент численного решения растет (или не убывает).



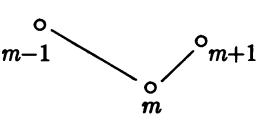
или



3. $r_m \leqslant 0 : u_m - u_{m-1} > 0$ и $u_{m+1} - u_m < 0$, или $u_m - u_{m-1} < 0$ и $u_{m+1} - u_m > 0$, — численное решение осциллирует.



или



Для обеспечения второго порядка точности необходимо $\varphi(1) = 1$. Различные TVD-алгоритмы соответствуют различному выбору $\varphi(r_m)$.

Дивергентный вариант TVD-схемы:

$$u_m^{n+1} = u_m^n - \frac{\tau}{h}(f_{m+1/2} - f_{m-1/2}), \text{ где}$$

$$f_{m+1/2} = \frac{f_m^n + f_{m+1}^n}{2} - \frac{q}{2}(f_{m+1}^n - f_m^n) + \frac{\varphi(r_m)}{2}(q - \sigma)(f_{m+1}^n - f_m^n),$$

$$q = \operatorname{sign}\sigma_{m+1/2}, \text{ число Куранта } \sigma = \frac{a\tau}{h}.$$

13.8. Идеи построения сеточно-характеристических методов и анализ разностных схем в пространстве неопределенных коэффициентов

Идея построения метода.

Продемонстрируем основные идеи построения сеточно-характеристических методов на примере решения простейшего линейного уравнения переноса:

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = f(x,t), \quad c = \text{const} > 0 \tag{13.12}$$

с соответствующими начальными и граничными условиями.

Известно, что (13.12) имеет характеристику, определяемую *уравнением характеристики*:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = a$$

и вдоль нее превращается в обыкновенное дифференциальное уравнение (ОДУ):

$$\frac{du}{d\xi} = f(\xi).$$

В частности, для однородного уравнения (1) значение функции u вдоль характеристики не меняется. На сведении уравнения в частных производных к ОДУ построены так называемые *методы характеристик*.

Методы сеток основаны на введении в области интегрирования уравнения разностной сетки, сеточного шаблона, т. е. совокупности узлов сетки, используемых для замены дифференциального уравнения (в малой области) разностным аналогом (аппроксимацией). При анализе аппроксимации обычно используются разложения проекции на сетку точного решения дифференциальной задачи в ряд Тейлора.

Рассмотрим класс методов, в которых одновременно использован и сеточный, и характеристический подход.

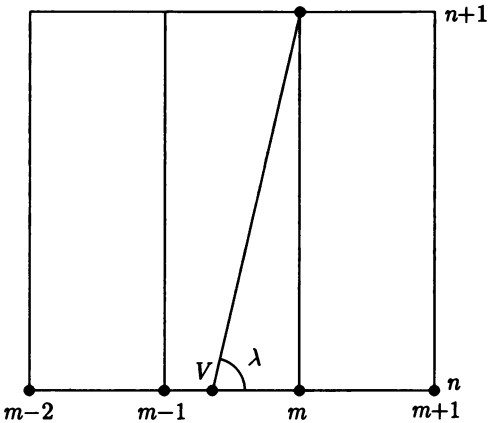


Рис. 13.1

В области интегрирования введем *разностную сетку*:

$$t^n = n \cdot \tau, x_m = m \cdot h; n = 0, 1, 2, 3, \ldots; m = 0, \pm 1, \pm 2, \pm 3, \ldots$$

Значение решения (1) в узлах сетки обозначим через u_m^n. Для перехода на следующий временной шаг сетки будем использовать явный пятиточечный шаблон (рис. 13.1). Необходимо вычислить значение u_m^{n+1}, используя значения u на n слое по времени. Проведем из точки $n+1, m$ характеристику с наклоном λ до пересечения со слоем t^n в точке V. Тогда вдоль характеристики значение u_m^{n+1} выражается через u_v^n, например, как

$$u_m^{n+1} = u_v^n + \frac{\tau}{2}(f_m^{n+1} + f_v^n) + O(\tau^2). \qquad (13.13)$$

Выражение (13.13) получено, как аппроксимация обыкновенного дифференциального уравнения вдоль характеристики методом трапеций. Проводя ту или иную интерполяцию для нахождения значения u_v^n по точкам слоя t^n, получим различные разностные схемы для определения u_m^{n+1} на данном шаблоне.

О физическом смысле условия Куранта.

Пусть $\sigma = \tau \cdot \lambda / h$ — число Куранта. В случае $\lambda = \text{const}$ очевидно, что выбором τ и h можно добиться $\sigma = 1$ или $\sigma = 2$. В этом случае характеристика, проведенная через точку u_m^{n+1}, попадет в узел разностной сетки, где значение u_{m-1}^n (или u_{m-2}^n соответственно) известно точно. Для однородного уравнения происходит перенос значения в точку u_m^{n+1} вдоль

характеристики. Если $2 \geqslant \sigma > 0$, то для определения значения u_v^n приходится решать задачу интерполяции, а для $\sigma > 2$ — задачу экстраполяции по узлам сетки. Известно, что вторая задача, как правило, неустойчива.

Точку V назовем *областью влияния* дифференциального уравнения (1), а точки шаблона, по которым проводится интерполяция — областью влияния разностного уравнения. Для устойчивости разностной схемы необходимо, чтобы область влияния дифференциальной задачи лежала внутри области влияния разностной — это проявление условия КФЛ.

Пространство неопределенных коэффициентов.

Запишем разностную схему в виде

$$u_m^{n+1} = \sum_{k=-2}^{1} \alpha_k \cdot u_{m+k}^n, \tag{13.14}$$

где коэффициенты α_k будем искать из условий аппроксимации, раскладывая u_m^{n+1}, u_{m+k}^n в ряды Тейлора в окрестности точки (t^n, x_m). Оставляя свободными два коэффициента (например, α_{-2} , α_0), получаем условия аппроксимации порядка $O(\tau + h)$:

$$\alpha_{-1} = \frac{1}{2}(1 + \sigma - 3 \cdot \alpha_{-2} - \alpha_0),$$

$$\alpha_1 = \frac{1}{2}(1 - \sigma + \alpha_{-2} - \alpha_0).$$

Примем свободные коэффициенты за координатные оси линейного пространства с евклидовой метрикой. Каждая точка этого пространства будет соответствовать разностной схеме первого порядка аппроксимации. Кроме того, можно выделить множество схем порядка $O(\tau + h^2)$

$$\alpha_0 = 1 - \sigma + 3 \cdot \alpha_{-2} \tag{13.15}$$

и единственную на данном шаблоне схему третьего порядка аппроксимации с порядком $O(\tau + h^3)$:

$$\alpha_{-2} = \frac{\sigma \cdot (\sigma - 1)}{6}. \tag{13.16}$$

Остальные коэффициенты находятся с использованием (13.15) и (13.16).

Введем пространство коэффициентов (α_{-2}, α_0). Тогда любая точка в этом пространстве есть разностная схема с порядком аппроксимации $O(\tau+h)$. Прямая (13.15) отделяет в пространстве множество схем с порядком $O(\tau + h^2)$ (рис. 13.2), на ней лежит единственная точка — аппроксимация $O(\tau + h^3)$. Должна быть также и точка с порядком аппроксимации $O(\tau^2 + h^2)$.

Зафиксируем какое-либо число Куранта, например, $\sigma = 0,5$. Применим к разностной схеме (13.14) с неопределенными коэффициентами спектральный признак устойчивости (фон Неймана). Получится кривая, которая определяет границу устойчивости разностных схем в пространстве неопределенных коэффициентов.

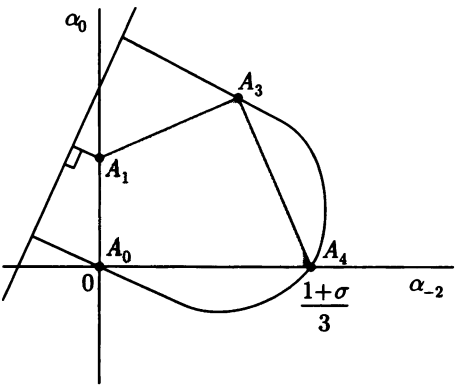


Рис. 13.2

Для схем первого порядка выпишем первое дифференциальное приближение:

$$u'_t + au'_x = \frac{h^2}{\tau}\left(1 - \sigma^2 - \alpha_0 + 3\alpha_{-1}\right) u''_{xx}.$$

Можно также выделить множество таких схем, что $\alpha_\mu \geqslant 0$. Это — монотонные схемы (заштрихованный многоугольник на рис. 13.2). Среди монотонных схем можно найти схему с наименьшей ошибкой аппроксимации. Это точка многоугольника, которая при данном σ лежит ближе всего к прямой со схемами второго порядка аппроксимации.

Закончим рассмотрение примера с модельным линейным уравнением переноса:

$$u'_t + cu'_x = 0.$$

На выбранном шаблоне любая разностная схема, как указывалось ранее, представляется в виде

$$u_m^{n+1} = \sum_{\mu \in III} \alpha_\mu u_{m+\mu}^n.$$

В случае монотонной схемы можно оценить норму погрешности. Заметим, что погрешность v определяется тем же разностным уравнени-

ем (13.14), тогда с использованием первой нормы (максимум абсолютной величины)

$$||v^{n+1}|| = ||v^n|| \sum_{\mu \in III} \alpha_\mu$$

в силу аппроксимации

$$\sum_{\mu \in III} \alpha_\mu \leqslant 1.$$

Отсюда следует, что монотонные разностные схемы всегда устойчивы. В общем случае можно рассматривать многослойные шаблоны для уравнения переноса (рис. 13.3)

$$u_m^{n+1} = \sum_{\mu,\nu \in III} \alpha_\mu^\nu u_{m+\mu}^{n+\nu}$$

и записывать условия порядка для аппроксимации соответствующего порядка:

$$\delta_0 = -1 + \sum_{\mu,\nu \in III} \alpha_\mu^\nu = 0,$$

$$\delta_1 = \sigma + \sum_{\mu,\nu \in III} (\mu - \nu\sigma)\alpha_\mu^\nu$$

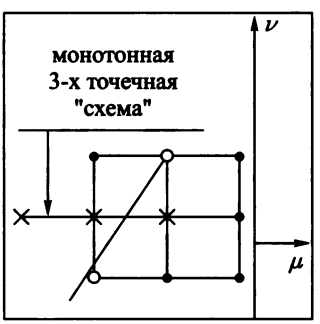


Рис. 13.3

Исключая два коэффициента из условий порядка, можно от пространства неопределенных коэффициентов $\{\alpha_\mu^\nu\}$ перейти к пространству $\{\tilde{\alpha}_\mu^\nu\}$, размерность которого на 2 меньше, например:

$$\alpha_{-1}^0 = \frac{1}{2}\left(1 + \sigma - \sum (\mu - \nu\sigma)\alpha_\mu^\nu\right),$$

$$\alpha_1^0 = \frac{1}{2}\left(1 - \sigma - \sum (\mu - \nu\sigma)\alpha_\mu^\nu\right),$$

где, конечно, точки $(0; -1)$ и $(0; 1)$ не включаются в суммирование. Условие устойчивости в пространстве неопределенных коэффициентов имеет вид

$$|q(\alpha_\mu^\nu,\ \varphi)| \leqslant 1,$$

где q есть спектр оператора послойного перехода. Эта величина определяется из условия

$$q = \sum_{\mu,\nu} q^\nu e^{i\mu\varphi}$$

и дополнительного требования

$$\frac{\partial}{\partial\varphi}\,(|q| - 1) = 0.$$

Основная гипотеза: *Разностным схемам, которым в пространстве* $\{\tilde{\alpha}_\mu^\nu\}$, *соответствуют близкие друг к другу точки (в смысле* $\rho = \sqrt{(\tilde{\alpha}_1,\ \tilde{\alpha}_1) - (\tilde{\alpha}_2,\ \tilde{\alpha}_2)}$*) по своим свойствам также близки.*

Расширяя шаблон, как и в случае пятиточечного шаблона, можно строить области схем высокого порядка аппроксимации, монотонные схемы ($\alpha_\mu^\nu \geqslant 0$) и т. д.

Монотонные разностные схемы в пространстве неопределенных коэффициентов занимают некий выпуклый многоугольник, вершины которого определяются довольно просто: это все возможные при данном числе Куранта σ трехточечные разностные схемы, причем для характеристики, проходящей через u_m^{n+1}, одна точка схемы лежит выше (левее) характеристики, а другая ниже (правее), см. рис. 13.3.

Метод построения разностных схем в пространстве неопределенных коэффициентов для квазилинейных систем уравнений гиперболического типа (к ним относятся системы уравнений механики сплошных сред, в частности, газовой динамики, механики деформируемого твердого тела (МДТТ) и т.п.) допускает обобщение и на многомерные случаи. Подробное описание можно найти в монографии [9]. Здесь же многомерные обобщения рассматриваться не будут. Они приводят к эффективным численным методам для нестационарных многомерных задач.

Исследовав схемы (13.14) на устойчивость по спектральному признаку, получаем множество устойчивых схем, а потребовав выполнения условия $\alpha_k \geqslant 0$ для всех точек шаблона, получаем множество *схем с положительной аппроксимацией* (монотонных по Фридрихсу схем). На рассматриваемом шаблоне устойчивые схемы существуют при $2 \geqslant \sigma > 0$ (для $a > 0$).

Множество схем с положительной аппроксимацией не пересекается с множеством схем с порядком аппроксимации выше первого, как это следует из теоремы С.К. Годунова.

Первое дифференциальное приближение. Дисперсионная и диссипативная ошибки

Поскольку решения дифференциальной задачи и разностного уравнения принадлежат разным функциональным пространствам, что порождает определенные трудности при теоретическом анализе свойств разностных схем, для такого исследования возможно рассматривать разностные операторы в том же пространстве. Будем считать, что разностные схемы удовлетворяются функциями непрерывного аргумента в каждой точке рассматриваемой области.

Обычно ограничиваются рассмотрением уравнений, в которых оставлены члены в разложении в ряд Тейлора проекции точного решения на сетку по τ и h, порядок которых совпадает с порядком погрешности аппроксимации схемы. Получающиеся при этом уравнения называют первым дифференциальным приближением (ПДП).

Для схемы первого порядка (5) при выполнении условия (6) первым дифференциальным приближением будет

$$u'_t + a \cdot u'_x = \delta_1 \cdot h \cdot u''_{xx}, \text{ где } \delta_1 = \text{const}. \qquad (13.17)$$

Из уравнения (13.17) исключены члены со второй производной u'_{tt}, с использованием так называемой продолженной системы:

$$(u'_t + a \cdot u'_x)'_x = 0,$$

$$(u'_t + a \cdot u'_x)'_t = 0.$$

Иногда уравнение (13.17) называют П-*формой* (параболической формой) первого дифференциального приближения. Если производные по времени не исключаются из ПДП, то имеем Г-форму (гиперболическую форму) ПДП, которая, как правило, не применяется в исследованиях как малоинформативная.

При $\delta_1 > 0$ уравнение (13.17) можно трактовать как присутствие в схеме некоторой диссипации (схемной вязкости). Ее наличие проявляется в расчетах в виде размазывания точного решения, причем его интенсивность увеличивается при ухудшении аппроксимации (увеличении шага h). В этом случае говорят, что ошибка схемы носит диссипативный характер. Если схемная вязкость получается отрицательной, то приходим к обратной задаче теплопроводности. Как известно из курсов математической физики, такая задача поставлена некорректно. А соответствующая разностная схема при исследовании по спектральному признаку оказывается неустойчивой — по ПДП можно сделать вывод об устойчивости схемы.

Для более высокого (второго) порядка ПДП имеет вид

$$u'_t + a \cdot u'_x = \delta_2 \cdot h^2 \cdot u'''_{xxx}. \qquad (13.18)$$

Уравнение (13.18) обладает *дисперсией*, т. е. разные пространственные гармоники разложения начального возмущения в ряд Фурье распространяются по сетке с разными скоростями. Говорят, что ошибка носит дисперсионный характер. Сеточная дисперсия легко получается, если искать частное решение последнего уравнения в виде комплексной экспоненты: $u = \lambda(t)\exp(ikx)$. Подробнее о ПДП в [4].

Интересна связь ПДП и исследования свойств схем в пространствах неопределенных коэффициентов. Так, для схем первого порядка расстояние от точки в пространстве неопределенных коэффициентов до прямой схем высокого порядка (13.15) по абсолютной величине равен коэффициенту δ_1 в уравнении (13.17), а знак определяется положением внутри области устойчивости.

Понятие о гибридных схемах

Численные расчеты по разностным схемам высокого порядка показывают, что осцилляции нефизического характера появляются в окрестности разрывов решения или его первой производной. В связи с этим возникает идея построения численного метода, имеющего высокий порядок аппроксимации на участках гладкого решения, в то время как в окрестности разрывов функций или их производных применяется монотонная схема (с положительной аппроксимацией) первого порядка.

Можно формализовать этот подход:

$$u_m^{n+1} = \sum_{p=-2}^{1} (\gamma \cdot \alpha(1)_p + (1-\gamma) \cdot \alpha(2)_p) \cdot u_{m+p}^n$$

$$\gamma = \frac{1 + (2 \cdot b - 1)}{2 \cdot |2 \cdot b - 1|^{(k-1)/k}}$$

$$b = \frac{\left| |u_{m+1}^n - u_m^n| - |u_m^n - u_{m-1}^n| \right|}{|u_{m+1}^n - u_m^n| + |u_m^n - u_{m-1}^n|},$$

где $\alpha(1)_p$ — коэффициенты первой схемы (высокого порядка аппроксимации), применяемой в области гладкого решения, $\alpha(2)_p$ — коэффициенты второй (монотонной) схемы, γ — весовой коэффициент, вспомогательный параметр b характеризует гладкость решения (очевидно, что $b = 0$, при $u \equiv \text{const}$), k — коэффициент гибридности — целое число из диапазона $2 \leqslant k \leqslant 10$.

При этом реализовано достаточно гладкое переключение со схем высокого порядка аппроксимации на монотонные схемы.

13.9. Задачи

1. Для линейного уравнения переноса

$$\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = 0$$

предложить схему $N+1$ порядка аппроксимации.

Указание. Рассмотреть невязку, образующуюся при замене дифференциального уравнения разностным

$$\frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_{m+1}^n - u_m^n}{h} = 0.$$

Решение. Проведем исследование разностного уравнения на аппроксимацию. Для этого представим сеточные функции u_m^{n+1} и u_{m+1}^n в виде разложения проекции на сетку точного решения дифференциальной задачи в ряд Тейлора:

$$u_{m+1}^n = u_m^n + h\,(u_x')_m^n + \frac{h^2}{2!}\,(u_x'')_m^n + \frac{h^3}{3!}\,(u_x''')_m^n + \frac{h^4}{4!}\,(u_x^{IV})_m^n + \ldots +$$

$$+\frac{h^N}{N!}\left(u_x^{(N)}\right)_m^n + O(h^{N+1}),$$

$$u_m^{n+1} = u_m^n + \tau\,(u_t')_m^n + \frac{\tau^2}{2!}\,(u_t'')_m^n + \frac{h^3}{3!}\,(u_t''')_m^n + \frac{h^4}{4!}\,(u_t^{IV})_m^n + \ldots +$$

$$+\frac{\tau^N}{N!}\left(u_t^{(N)}\right)_m^n + O(\tau^{N+1}).$$

Отсюда получим с учетом того, что $u_t^{(k)} = u_x^{(k)}$:

$$\frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_{m+1}^n - u_m^n}{h} = (u_t' - u_x')_m^n + r_\tau,$$

где r_τ — невязка, вычисляемая в точке (t^n, x_m) по формуле

$$r_\tau = \frac{1}{2!}(\tau - h)u_x'' + \frac{1}{3!}(\tau^2 - h^2)u_x''' + \frac{1}{4!}(\tau^3 - h^3)u_x^{IV} + \ldots +$$

$$+\frac{1}{(N+1)!}(\tau^N - h^N)u_x^{(N+1)} + O(\tau^{N+1}, h^{N+1}) =$$

$$= \sum_{k=1}^{N} \frac{h^k}{(k+1)!}(\sigma^k - 1)u_x^{(k+1)} + O(\tau^{N+1}, h^{N+1}),$$

$\sigma = \tau/h$ — число Куранта.

Аппроксимация частных производных высоких порядков, входящих в выражение для невязки с помощью конечных разностей, приводит к соотношению

$$r_\tau = \frac{1}{h}\sum_{k=1}^{N}\frac{\sigma^k - 1}{(k+1)!}\Delta^{k+1}u_m^n + O(\tau^{N+1}, h^{N+1}).$$

В таком случае, после переноса суммы по k в выражении для невязки в левую часть, получим разностную схему искомого порядка аппроксимации:

$$\frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_{m+1}^n - u_m^n}{h} - \frac{1}{h}\sum_{k=1}^{N}\frac{\sigma^k - 1}{(k+1)!}\Delta^{k+1}u_m^n = 0,$$

или

$$u_m^{n+1} = u_m^n + \sigma\Delta^+ u_m^n - \sigma\sum_{k=1}^{N}\frac{(1-\sigma^k)}{(k+1)!}\Delta^{k+1}u_m^n = 0.$$

Эта схема легко распространяется на уравнение

$$\frac{\partial u}{\partial t} - a\frac{\partial u}{\partial x} = 0,$$

если учесть, что

$$\frac{\partial^{k+l}u}{\partial t^k \partial x^l} = (-1)^k a^k \frac{\partial^{k+l}u}{\partial x^{k+l}}.$$

2. Доказать устойчивость разностной схемы

$$\frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_{m+1}^n - u_m^n}{h} = f_m^n,$$
$$n = 0, \ldots, N-1,\ m = 0, \pm 1, \ldots, \pm(M-1),$$

$$u_m^0 = \varphi_m^n, m = 0, \pm 1, \ldots, \pm(M-1),$$

аппроксимирующую задачу Коши для уравнения переноса

$$\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = f(t,x) t \in [0,T],$$
$$x \in (-\infty, \infty), \quad u(0,x) = \varphi(x), \quad x \in (-\infty, \infty),$$

используя второе определение устойчивости:

$$\|u_\tau\| \leqslant c\|f\| \text{ при} \sigma = \frac{\tau}{h} < 1.$$

Решение. Перепишем разностную схему в виде, разрешенном относительно следующего слоя по времени

$$u_m^{n+1} = (1-\sigma)\, u_m^n + \sigma u_{m+1}^n + \tau f_m^n,$$

$$u_m^0 = \varphi_m.$$

Определим норму в пространстве сеточных функций как

$$\|u_\tau\| = \sup_{n,m} |u_m^n| = \max_n \sup_m |u_m^n|.$$

Если $1-\sigma \geqslant 0$, то справедлива оценка

$$\left\|(1-\sigma)u_m^n + \sigma u_{m+1}^n\right\| \leqslant |(1-\sigma)+\sigma| \max_n \left(|u_m^n|, |u_{m+1}^n|\right) =$$

$$= \max_n \left(|u_m^n|, |u_{m+1}^n|\right) \leqslant \max_m |u_m^n|.$$

В таком случае

$$|u_m^{n+1}| \leqslant \max_m |u_m^n| + \tau |f_m^n| \leqslant \max_m |u_m^n| + \tau \max_{m,n} |f_m^n|.$$

Отсюда видно, что при $f_m^n = 0$ норма решения $|u_m^n|$ не возрастает при возрастании n — выполняется принцип максимума. Поскольку правая часть в полученном неравенстве не зависит от m, то $\max\limits_m |u_m^{n+1}| \leqslant \max\limits_m |u_m^n| + \tau \max\limits_{m,n} |f_m^n|$.

Аналогично

$$\max_m |u_m^n| \leqslant \max_m |u_m^{n-1}| + \tau \max_{m,n} |f_m^n|,$$

$$\max_m |u_m^{n-1}| \leqslant \max_m |u_m^{n-2}| + \tau \max_{m,n} |f_m^n|, \ldots,$$

$$\max_m |u_m^1| \leqslant \max_m |u_m^0| + \tau \max_{m,n} |f_m^n|.$$

Сложение этих неравенств дает

$$\max_m |u_m^{n+1}| \leqslant \max_m |u_m^0| + (n+1)\tau \max_{m,n} |f_m^n|,$$

откуда получаем, с учетом того, что $t_n = n\tau$:

$$\max_m |u_m^{n+1}| \leqslant \max_m |\varphi_m| + t_{n+1} \max_{m,n} |f_m^n| \leqslant$$

$$\leqslant \|f_\tau\| + t_{n+1} \|f_\tau\| = (1+t_{n+1}) \|f_\tau\|.$$

При этом учтено, что

$$\|\varphi_\tau\| = \|f_\tau\| = \max_{n,m} |f_m^n| + \max_m |\varphi_m^n|.$$

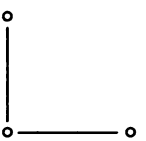
Таким образом, получим

$$\|u_\tau\| \leqslant c\,\|f_\tau\|, \text{ где } c = 1 + t_{n+1}.$$

3. Построить явную разностную схему первого порядка точности для аппроксимации линейного уравнения переноса $\frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = f(t,x)$, используя ее запись в общем виде через неопределенные коэффициенты

$$\mathbf{L}_\tau u_\tau = c_1 u_m^{n+1} + c_2 u_m^n + c_3 u_{m+1}^n = f_m^n,$$

на шаблоне «явный правый уголок».



Решение. Будем подбирать коэффициенты c_i таким образом, чтобы выполнялось условие аппроксимации первого порядка $\mathbf{L}_\tau U_\tau = \mathbf{L}u|_{t_n,x_m} + O(\tau + h)$.

Разложение сеточных функций u_m^{n+1} и u_{m+1}^n в ряды Тейлора в окрестности точки x_m, t_n приводит к равенствам

$$u_m^{n+1} = u_m^n + \tau\,(u_t')_m^n + O(\tau^2),$$

$$u_{m+1}^n = u_m^n + h\,(u_x')_m^n + O(h^2).$$

Подстановка этих разложений в разностную схему с неопределенными коэффициентами дает

$$\mathbf{L}_\tau U_\tau = c_1 u_m^{n+1} + c_2 u_m^n + c_3 u_{m+1}^n =$$

$$= (c_1 + c_2 + c_3) u_m^n + c_1 \tau (u_t')_m^n + c_3 h (u_x')_m^n + O(\tau^2 + h^2) =$$

$$= (c_1 + c_2 + c_3) u_m^n + c_1 \sigma h (\mathbf{L}u)_m^n + (c_1\sigma + c_3) h (u_x')_m^n + O(h^2),$$

поскольку можно считать независимыми переменными шаг по пространству и число Куранта, а для шага по времени получаем очевидное выражение $\tau = \sigma h, \sigma = \text{const}$.

Здесь использованы очевидные равенства для дифференциального оператора

$$\frac{\partial u}{\partial t} = \mathbf{L}u + \frac{\partial u}{\partial x}, \frac{\partial u}{\partial x} = \frac{\partial u}{\partial t} - \mathbf{L}u.$$

Для выполнения условия аппроксимации первого порядка

$$\mathbf{L}_\tau \mathrm{U}_\tau |_{t_n, x_m} = \mathbf{L}u|_{t_n, x_m} + O(h)$$

необходимо выполнение *условий порядка*

$$c_1 \sigma h = 1 + O(h),$$

$$c_1 + c_2 + c_3 = 0 + O(h),$$

$$(c_1 + c_3)h = 0 + O(h).$$

Если положить $O(h) = 0$, поскольку это некое число, стремящееся к нулю при измельчении шага, то условия порядка примут вид

$$c_1 \sigma h = 1,$$

$$c_1 + c_2 + c_3 = 0,$$

$$c_1 + c_3 = 0.$$

Эта схема линейных алгебраических уравнений имеет единственное решение:

$$c_1 = \frac{1}{\sigma h} = \tau^{-1}, c_2 = \frac{\sigma - 1}{\sigma h} = h^{-1} - \tau^{-1}, c_3 = -h^{-1}.$$

Получили коэффициенты уже известной разностной схемы первого порядка аппроксимации на заданном шаблоне:

$$\frac{u_m^{n+1} - u_m^n}{\tau} - \frac{u_{m+1}^n - u_m^n}{h} = f_m^n.$$

Несложно проверить, что учет $O(h)$ привел бы к незначительной коррекции результата.

Конечно, выбранная разностная схема легко получается и из других соображений. Но с помощью метода неопределенных коэффициентов удается построить схемы с хорошими свойствами для более сложных уравнений и систем.

4. Исследовать на спектральную устойчивость разностную схему на шаблоне квадрат

$$\frac{1}{2\tau}\left[(u_{m+1}^{n+1}+u_m^{n+1})-(u_{m+1}^n-u_m^n)\right]+$$

$$+\frac{1}{h}\left[(u_{m+1}^{n+1}+u_{m+1}^n)-(u_m^{n+1}-u_m^n)\right]=f_{m+1/2}^{n+1/2},$$

аппроксимирующую линейное уравнение переноса

$$\frac{\partial u}{\partial t}+\frac{\partial u}{\partial x}=0.$$

Решение. После подстановки в схему гармоник Фурье

$$u_m^n=\lambda^n e^{i\alpha m},$$

получим выражение для спектра оператора послойного перехода

$$\lambda(\alpha)=\frac{1-i\sigma\operatorname{tg}\alpha/2}{1+i\sigma\operatorname{tg}\alpha/2},\sigma=\tau/h\text{ — число Куранта.}$$

Тогда $|\lambda(\alpha)|=1$ и рассматриваемая схема безусловно устойчива. Этого следовало ожидать, так как аппроксимация производной по времени производится как на нижнем, так и на верхнем временном слое. Однако решение на верхнем слое выписывается в явном виде, поскольку правое условие ставится на левой границе. Поэтому схему иногда относят к явным схемам бегущего счета. Расчетная формула будет

$$u_{m+1}^{n+1}=u_m^n+\frac{\sigma-1}{\sigma+1}(u_m^{n+1}-u_{m+1}^n)+\tau f_{m+1/2}^{n+1/2}.$$

Схема обладает вторым порядком аппроксимации по времени и пространственной координате.

5. Получить дисперсионное соотношение для разностных схем первого и второго порядка сходимости для аппроксимации нелинейного уравнения

$$\frac{\partial v}{\partial t}+u\frac{\partial v}{\partial x}=0,\quad\text{где}\quad u\neq\text{const},$$

и, вообще говоря, зависит от решения.

Указание. Использовать дисперсионное соотношение для данного дифференциального уравнения в виде $e^{\lambda t}e^{ikx}$, где k — волновое число, $\lambda=-uki$.

Решение. Будем искать дисперсионное соотношение для разностного уравнения в виде

$$v_m^n = e^{\lambda(k)t_n} \cdot e^{ikx_m} = e^{\lambda n\tau + ikmh}.$$

После его подстановки в схему правый уголок первого порядка аппроксимации, получим

$$\frac{1}{\tau}\left(e^{\lambda\tau} - 1\right) + \frac{u}{h}\left(1 - e^{-ikh}\right) = 0,$$

откуда следует

$$\lambda(\tau, h, k) = \tau^{-1}\ln\left\{1 - \frac{u\tau}{h} + \frac{u\tau}{h}e^{-ikh}\right\}.$$

Сравним дисперсионное соотношение для дифференциального и соответствующих разностных уравнений.

Пусть $\frac{u\tau}{h} = 1$. Имеет место совпадение $\lambda(k)$ и $\lambda(\tau,h,k)$, однако этот случай мало интересен для вычислительной практики, поскольку u может меняться в ходе решения задачи. Для проведения дисперсионного анализа схем положим $kh \ll 1$, имея в виду то, что h — малый параметр, а аппроксимация тем лучше, чем меньше волновое число, т. е. чем более гладким является исследуемое частное решение. Заметим, что для расчетной сетки с шагом h обычно реализуются волновые числа $k \approx 2\pi h^{-1}$.

В таком случае дисперсионное соотношение будет

$$\lambda(\tau, h, k) \approx -iuk - \frac{uhk^2}{2}\left(1 - \frac{u\tau}{h}\right) = \lambda(k) - \frac{k^2}{2}uh(1-\sigma),$$

$$\sigma = u\tau/h - \text{число Куранта},$$

а частное решение примет вид

$$\exp(ik(mh - un\tau)) \cdot \exp\left(-\frac{1}{2}uk^2(h - u\tau)n\tau\right).$$

Первый сомножитель совпадает с соответствующим частным решением дифференциального уравнения, второй является особенностью разностного решения; его поведение при различных τ, h, k представляет особенный интерес. При $u < 0, h - u\tau > 0$, как видно из этого решения, рассматриваемую схему нельзя использовать для проведения расчетов. Второй сомножитель имеет порядок

$\exp(k^2 |u| ht_n)$, он быстро растет при $k \sim h^{-1}, h \to 0, \tau \to 0$. Аналогичная ситуация получается и при $u > 0, h - u\tau < 0$. При $u < 0, h - u\tau > 0$ второй сомножитель затухает с ростом t_n тем быстрее, чем больше k, или чем меньше λ — длина волны частного решения вида $\exp(ikx)$. Таким образом, численное решение, полученное по схеме уголок, отличается от точного, в котором все гармоники сохраняют амплитуду с ростом t_n, наличием затухающего множителя для гармоник с большими волновыми числами (или малыми длинами волн). Действие этого сомножителя приводит к сглаживанию решений, имеющих разрыв в начальных данных.

Для разностной схемы Лакса-Вендроффа второго порядка

$$(v_{m+1}^{n+1} + v_m^n) - \sigma(v_m^n - v_{m-1}^n) + \frac{\sigma}{2}(v_{m-1}^n - 2v_m^n + v_{m+1}^n) = 0$$

дисперсионное соотношение будет

$$\lambda = \tau^{-1} \ln \left\{ 1 + \sigma(e^{-ikh} - 1) + 2uh(1 - \sigma)\sin^2 \frac{kh}{2} \right\},$$

для длинноволновых гармоник с $kh \ll 1$

$$\lambda(\tau, h, k) \approx -iku + i\frac{k^3 u}{6}(h^2 - u^2\tau^2) = \lambda(k) + ik^3 \frac{uh^2}{6}(1 - \sigma^2).$$

Из полученных соотношений можно сделать следующий вывод. Решения исходного уравнения переноса имеют вид волн с волновым числом k, которые движутся вправо со скоростью u, так как

$$v(t, x) = \exp(ikx + \lambda(k)t) = \exp(ik(x - ut)).$$

Решения разностного уравнения имеют вид

$$v(t_n, x_m) = \exp(ikx_m + \lambda(\tau, h, k)t_n) =$$

$$= \exp\left(ik \left\{ x_m - u \left[1 - \frac{k^2}{6} \left(h^2 - u^2\tau^2 \right) \right] t_n \right\} \right) =$$

$$= \exp\left(ik \left\{ x_m - u \left[1 + \mathrm{U}_k(k) \right] t_n \right\} \right),$$

где введено обозначение

$$\mathrm{U}_k = -\frac{k^2}{6} \left(h^2 - u^2\tau^2 \right).$$

Последнее выражение означает, что разностные решения уже не имеют вид волн, движущихся с одной и той же скоростью. Теперь каждая волна со своей частотой движется с собственной скоростью $u = U_k[1 + u(k)]$. Разумеется, при малых k и скорости u_k мало отличаются от u, но высокочастотные волны движутся со скоростями, заметно отличающимися от скорости переноса. Кроме того, в схеме Лакса-Вендроффа гармоники со временем не затухают, численное решение не сглаживается со временем. Начальный волновой пакет u и определенный им начальный профиль решения, изменяются с течением времени из-за рассогласования фаз. Это приводит к потере монотонности профиля $v(x)$, если он был вначале монотонным, и появлению осцилляций разностного происхождения. Появление *сеточной дисперсии* — одно из проявлений эффекта Гиббса, подробнее об эффекте Гиббса в [16].

6. Акустическая система описывает распространение плоских звуковых волн. Поставим для нее задачу Коши:

$$\frac{\partial u}{\partial t} + \frac{1}{\rho}\frac{\partial p}{\partial x} = 0, \frac{\partial p}{\partial t} + \rho c^2 \frac{\partial u}{\partial x} = 0,$$
$$\rho = \text{const}, \quad c = \text{const},$$
$$u(x,0) = u_0(x), \quad p(x,0) = p_0(x),$$
$$t > 0, -\infty < x < \infty,$$

где u — скорость движения среды, p — давление, ρ — плотность среды, c — скорость звука в среде.

a) Представить акустическую систему в интегральной форме.

b) Преобразовать акустическую систему к виду системы с разделенными переменными.

c) Предложить разностную схему первого порядка точности для ее решения.

Решение.

a) Проинтегрировав акустическую систему по произвольной области с границей G, получим:

$$\oint_\Gamma \rho u dx - p dt = 0,$$

$$\oint_\Gamma \frac{p}{c^2} dx - \rho u dt = 0.$$

b) Умножим второе уравнение на $(\rho c)^{-1}$, затем сложим с первым и вычтем из него. Получим систему двух уравнений

$$\frac{\partial}{\partial t}\left(u+\frac{p}{\rho c}\right)+c\frac{\partial}{\partial x}\left(u+\frac{p}{\rho c}\right)=0$$
$$\frac{\partial}{\partial t}\left(u-\frac{p}{\rho c}\right)-c\frac{\partial}{\partial x}\left(u-\frac{p}{\rho c}\right)=0,$$

или, после введения обозначений (инвариантов Римана) $R = u+\frac{p}{\rho c}, S = u-\frac{p}{\rho c}$ система запишется в виде двух уравнений переноса:

$$\frac{\partial R}{\partial t}+c\frac{\partial R}{\partial x}=0 \quad \frac{\partial S}{\partial t}-c\frac{\partial S}{\partial x}=0.$$

Эта система позволяет выписать общее решение:

$$R=f(x-ct), \quad S=g(x+ct).$$

Здесь f, g — произвольные непрерывно дифференцируемые функции, определяемые из начальных условий. В терминах инвариантов Римана можно определить и значения естественных переменных u, p:

$$u=\frac{1}{2}(f(x-ct)+g(x+ct)),$$
$$p=\frac{\rho c}{2}(f(x-ct)-g(x+ct)).$$

Замечательное свойство инвариантов Римана R, S заключается в их постоянстве вдоль прямых $x-ct=\text{const}$, и $x+ct=\text{const}$ или $\frac{dx}{dt}=\pm c$ — характеристик акустической системы.

Если задавать начальные данные $u(x,0)=\varphi(x) p(x,0)=\psi(x)$, $-\infty<x<\infty$, то связь начальных данных и инвариантов Римана будет

$$\varphi(x)=\frac{1}{2}(f(x)+g(x)), \psi(x)=\frac{\rho c}{2}[f(x)-g(x)],$$

откуда получим

$$f(x)=\varphi(x)+\frac{\psi(x)}{\rho c}, g(x)=\varphi(x)-\frac{\psi(x)}{\rho c}.$$

Следовательно, решение системы в этом случае

$$u(t,x)=\frac{1}{2}(\varphi(x-ct)+\varphi(x+ct))+\frac{1}{2\rho c}(\psi(x-ct)+\psi(x+ct)),$$
$$p(t,x)=\frac{1}{2}(\varphi(x-ct)+\varphi(x+ct))-\frac{\rho c}{2}(\psi(x-ct)+\psi(x+ct)).$$

Для численного решения системы, записанной в инвариантах Римана, можно применить схему Куранта-Изаксона-Риса:

$$\frac{R_m^{n+1}-R_m^n}{\tau}+c\frac{R_m^{n+1}-R_{m-1}^n}{h}=0,$$

$$\frac{S_m^{n+1}-S_m^n}{\tau}-c\frac{S_{m+1}^n-S_m^n}{h}=0,$$

или

$$R_m^{n+1}=(1-\sigma)\,R_m^n+\sigma R_{m-1}^n,$$

$$S_m^{n+1}=(1+\sigma)\,S_m^n-\sigma S_{m+1}^n.$$

Известно, что данная схема имеет первый порядок аппроксимации и является устойчивой при выполнении условия КФЛ: $\frac{c\tau}{h}\leqslant 1$.

7. Заменить смешанную задачу для волнового уравнения

$$\frac{\partial^2 u}{\partial t}=c^2\frac{\partial^2 u}{\partial x^2}+f(t,x),\quad 0<t\leqslant T,\quad 0<x<X, u(0,x)=\varphi_1(x),$$

$$\frac{\partial u(0,x)}{\partial t}=\varphi_2(x), 0<x<X, u(t,0)=\varphi_3(t), u(t,X)=\varphi_4(t), 0\leqslant t\leqslant T$$

эквивалентной ему парой уравнений в частных производных первого порядка.

Решение. Введем новые переменные

$$v(t,x)=\int\limits_0^x\frac{\partial u(t,\eta)}{\partial t}d\eta, F(t,x)=\int\limits_0^x f(t,\eta)d\eta.$$

Функции $u(t,x), v(t,x)$, что несложно проверить, удовлетворяют акустической системе уравнений

$$\frac{\partial u}{\partial t}-\frac{\partial v}{\partial x}=0, \frac{\partial v}{\partial t}-c^2\frac{\partial u}{\partial x}=F(t,x),$$

$$u(0,x)=\varphi_1(x), v(0,x)=\int\limits_0^x\varphi_2(\eta)d\eta, u(t,0)=\varphi_3(t), u(t,X)=\varphi_4(t).$$

Введем в рассмотрение параметрическую схему для численного решения этой системы:

$$\frac{u_m^{n+1}-u_m^n}{\tau}=\frac{c^2}{h}\left\{\xi_1(v_{m+1}^{n+1}-v_m^{n+1})+(1-\xi_1)(v_{m+1}^n-v_m^n)\right\}+F_{m+1/2}^{n+1/2},$$

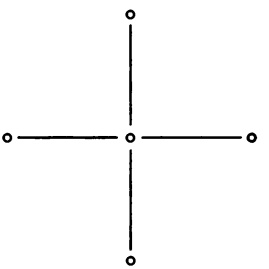
$$\frac{v_m^{n+1}-v_m^n}{\tau}=\frac{1}{h}\left\{\xi_2(u_m^{n+1}-u_{m-1}^{n+1})+(1-\xi_2)(u_m^n-u_{m-1}^n)\right\},$$

$$0\leqslant\xi_1,\xi_2\leqslant 1.$$

Исследование полученной двухпараметрической разностной схемы на спектральную устойчивость дает условия устойчивости: $\xi_1 + \xi_2 \geqslant 1, \sigma^2(2\xi_1 - 1)(2\xi_2 - 1) \geqslant -1, \sigma = \frac{c\tau}{h}$. Отсюда видно, что если $\xi_1 \geqslant 0,5$ и $\xi_2 \geqslant 0,5$, то схема устойчива при любых числах Куранта и сеточных параметрах, если $\xi_1 + \xi_2 \geqslant 1$, но один из параметров меньше 0,5, то схема условно устойчива при $\tau \leqslant \frac{h}{c}(2\xi_1 - 1)^{-1}(2\xi_2 - 1)^{-1}$.

Пусть $\xi_1 = \xi_2 = 0,5$. Тогда полученная схема является безусловно устойчивой и обладает вторым порядком аппроксимации по обоим переменным.

8. Исследовать на сходимость схему «крест» с шаблоном



для численного решения волнового уравнения $\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0$.

Решение. Запишем разностную схему:

$$\mathrm{L}_\tau u_\tau = \frac{u_m^{n+1} - 2u_m^n + u_m^{n-1}}{\tau^2} - \frac{u_{m-1}^n - 2u_m^n + u_{m+1}^n}{h^2} = 0.$$

Исследование на аппроксимацию дает выражение для главного члена невязки

$$r_\tau = \frac{\tau^2}{12}u_t^{(4)} - \frac{h^2}{12}u_x^{(4)} + O(\tau^2, h^2).$$

Схема имеет второй порядок аппроксимации. Исследуем схему на устойчивость по спектральному признаку. Для спектра оператора послойного перехода получается

$$\lambda^2 - 2\left(1 - 2\sigma^2 \sin^2\frac{\alpha}{2}\right)\lambda + 1 = 0, \sigma = \tau/h \text{ — число Куранта.}$$

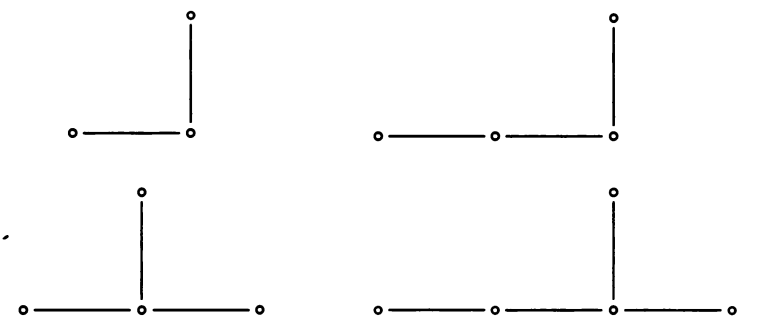
По теореме Виета, произведение корней $\lambda_1\lambda_2 = 1$, и условие устойчивости выполняется, если $|\lambda_1| = |\lambda_2| = 1$. Для полученного квадратного уравнения с действительными коэффициентами это означает, что его корни являются комплексно сопряженными. Это возможно лишь в случае, если дискриминант

$$D(\alpha) = 4\sigma^2 \sin^2\alpha\left(\sigma^2 \sin^2\frac{\alpha}{2} - 1\right)$$

будет отрицательным. Условие устойчивости будет выполнено для всех гармоник, если $\sigma < 1$.

13.10. Задачи для самостоятельного решения

1. Для линейного однородного уравнения переноса $\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0, a = \text{const} > 0$ построить разностные схемы первого, второго и третьего порядков аппроксимации, используя шаблоны, приведенные на рисунках.



2. Для линейного однородного уравнения переноса $\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0$, построить разностную схему, имеющую второй порядок по времени и четвертый по координате, используя шаблон с точками $(t^n, x_m)(t^{n+1}, x_m)(t^n, x_{m\pm1})(t^{n+1}, x_{m\pm1})(t^n, x_{m\pm2})(t^{n+1}, x_{m\pm2})$.

3. Исследовать устойчивость параметрической схемы

$$\frac{u_m^{n+1}-u_m^n}{\tau}+\frac{1}{2}\left[\alpha\frac{u_{m+1}^{n+1}-u_{m-1}^{n+1}}{2h}+(1-\alpha)\frac{u_{m+1}^{n}-u_{m-1}^{n}}{2h}\right]=0, \alpha\in[0,1]=0$$

для численного решения линейного уравнения переноса.

4. Определить, при каких значениях α схема

$$\frac{u_m^{n+1}-u_m^n}{\tau}+\alpha\frac{u_{m+1}^n-u_m^n}{h}+(1-\alpha)\frac{u_m^n-u_{m-1}^n}{h}$$

аппроксимирующая линейное однородное уравнение переноса, является устойчивой.

5. Для численного решения линейного уравнения переноса

$$\frac{\partial u}{\partial t} = \frac{\partial u}{\partial x} + \frac{\partial u}{\partial y}$$

предложить разностные схемы первого и второго порядков точности; исследовать их на сходимость.

6. Исследовать на сходимость схему предиктор-корректор:

предиктор

$$\frac{u^{n+1/2}_{m+1/2} - u^n_{m+1/2}}{\tau/2} + \frac{u^n_{m+1} - u^n_m}{h} = 0,$$

$$\frac{u^{n+1/2}_{m-1/2} - u^n_{m-1/2}}{\tau/2} + \frac{u^n_m - u^n_{m-1}}{h} = 0,$$

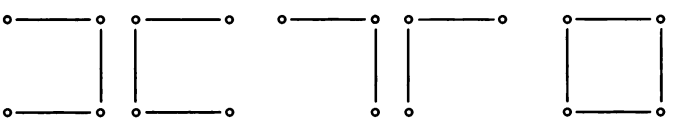
корректор

$$\frac{u^{n+1}_m - u^n_m}{\tau/2} + \frac{u^{n+1/2}_{m+1/2} - u^{n+1/2}_{m-1/2}}{h} = 0,$$

аппроксимирующую уравнение переноса

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0.$$

7. Исследовать на сходимость схемы, имеющие шаблоны



для численного решения уравнения переноса

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0, \quad a = \text{const} > 0.$$

8. Предложить устойчивые схемы первого и второго порядка аппроксимации для численного решения акустической системы

$$\frac{\partial u_1}{\partial t} - \frac{\partial u_2}{\partial x} = 0, \quad \frac{\partial u_2}{\partial t} - \frac{\partial u_1}{\partial x} = 0,$$

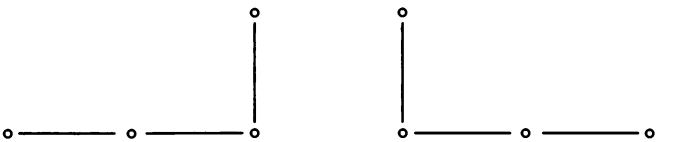
которую можно представить в матричной форме:

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A}\frac{\partial \mathbf{u}}{\partial x} = 0, u = (u_1, u_2)^{\mathrm{T}}, \mathbf{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

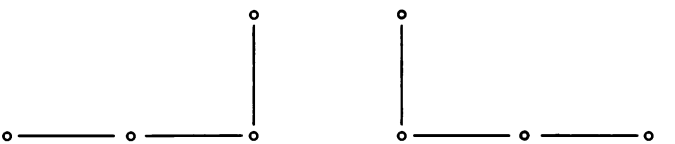
9. Используя характеристические свойства уравнения переноса

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0, \quad a = \text{const} > 0$$

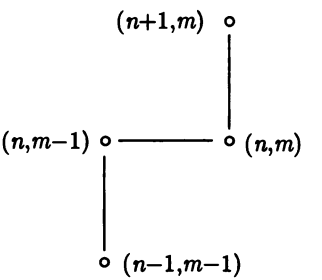
построить аппроксимирующую разностную схему на шаблонах, приведенных на рисунках



10. Предложить гибридную разностную схему для численного решения линейного однородного уравнения переноса, использующую два опорных шаблона, изображенных ниже:



11. Исследовать на сходимость схему с шаблоном



для численного решения линейного однородного уравнения переноса.

12. Исследовать на устойчивость схему с искусственной вязкостью

$$\frac{u_m^{n+1} - u_m^n}{\tau} + a\frac{u_{m+1}^n - u_{m-1}^n}{2h} = \xi\tau\frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2}$$

для решения уравнения переноса

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0, \quad a = \text{const} > 0.$$

13. При каких значениях ξ эта схема совпадает со схемами Лакса, Куранта-Изаксона-Риса, Лакса-Вендроффа?

14. Показать, что разностные схемы Лакса и Куранта-Изаксона-Риса удовлетворяют условию TVD.

15. **Акустика**

Рассмотрим нестационарное течение жидкости, охватывающее ограниченную область. Требуется предсказать звуковое поле, излучаемое на большом расстоянии от области течения. Акустическая аналогия этой задачи состоит в решении неоднородного волнового уравнения:

$$\left(\frac{1}{a_0^2}\frac{\partial^2}{\partial t^2} - \Delta\right)\varphi = q,$$

здесь a_0 — скорость распространения звука в покоящейся жидкости, φ — искомый потенциал скорости, q — интенсивность источника.

Если жидкость несжимаемая, то $a_0 \to \infty$ и волновое уравнение переходит в уравнение Лапласа.

Все волны обладают некоторыми общими свойствами. Одним из важных свойств большинства волн является сохранение энергии. Другое свойство состоит в том, что полное смещение среды в волне обычно весьма мало. Поскольку волновое уравнение имеет второй порядок как по пространственным переменным, так и по времени, оно обратимо, если только в уравнении не присутствуют члены, описывающие диссипативные процессы. Следовательно, в алгоритмах для волнового уравнения нежелательны диссипативные ошибки, приводящие к уменьшению амплитуды решения. Выполнение условия обратимости решения в алгоритме гарантирует, что моды не затухают, когда алгоритм устойчив. Поскольку волны часто распространяются на весьма большие расстояния, алгоритм, применяемый для расчета их поведения, должен давать минимальное численное затухание. Наилучшие из подобных алгоритмов являются почти неустойчивыми или могут быть легко переведены в неустойчивое состояние за счет малых ошибок аппроксимации, дополнительных нецентрированных слагаемых в разностном уравнении или нелинейных эффектов, которые не появляются при линейном анализе устойчивости.

(a) Реализовать разностный алгоритм для численного решения и рассмотреть поведение решений волнового уравнения, представленных в виде суммы двух волн $\varphi(x,t) = f_1(\xi) + f_2(\eta)$, где

$\xi = x - a_0 t, \eta = x + a_0 t$. Для этого задать в начальный момент времени (t = 0) потенциал возмущенного движения $\varphi(x)$ в виде функции, например, $f_1(\xi)$, отличной от нуля только на участке $0 \leqslant x_0 \leqslant \xi_0$.

(b) Исследовать распространение волн, если область их распространения с одной стороны ограничена. Как учесть взаимодействие волны со стенкой в численном расчете?

(c) Рассмотрим систему дифференциальных уравнений:

$$\frac{\partial \rho}{\partial t} = -a_0 \frac{\partial v}{\partial x},$$
$$\frac{\partial v}{\partial t} = -a_0 \frac{\partial \rho}{\partial x}.$$

Уравнения имеют периодические решения вида

$$\rho(x,t) = \rho(k) e^{ikx} e^{-i\omega t},$$
$$v(x,t) = v(k) e^{ikx} e^{-i\omega t},$$

где $\rho(k)$ и $v(k)$ — комплексные величины, k — волновое число, отвечающее направлению x. Это действительная величина, которая с длиной волны моды связана соотношением $\lambda = 2\pi/k$. Реализовать следующую конечно-разностную аппроксимацию системы:

$$\frac{\rho_j^n - \rho_j^{n-1}}{\Delta t} = -a_0 \frac{v_{j+1/2}^{n-1/2} - v_{j-1/2}^{n-1/2}}{\Delta x},$$
$$\frac{v_{j-1/2}^{n+1/2} - v_{j-1/2}^{n-1/2}}{\Delta t} = -a_0 \frac{\rho_j^n - \rho_{j-1}^n}{\Delta x}.$$

Здесь переменная ρ определена в центрах ячеек в старый и новый момент времени. Скорость v определена на границах ячеек и в моменты, промежуточные по отношению к тем, в которые определена плотность. Подобные сетки называются *разнесенными*. Разностная схема называется *явным алгоритмом «чехарда» на разнесенной сетке*. Вывести условие устойчивости для этой схемы.

(d) Для акустической системы реализовать неявный конечно-разностный алгоритм:

$$\rho_j^n - \rho_j^{n-1} = -\frac{\Delta t a_0}{2}\left[\frac{(v_{j+1/2}^n - v_{j-1/2}^n)}{\Delta x} + \frac{(v_{j+1/2}^{n-1} - v_{j-1/2}^{n-1})}{\Delta x}\right],$$
$$v_{j+1/2}^n - v_{j+1/2}^{n-1} = -\frac{\Delta t a_0}{2}\left[\frac{(\rho_{j+1}^n - \rho_j^n)}{\Delta x} + \frac{(\rho_{j+1}^{n-1} - \rho_j^{n-1})}{\Delta x}\right].$$

Расчетная сетка здесь разнесена по пространству, но не по времени. Новые значения ρ и v теперь определяются одновременно на одном и том же шаге по времени. Исследовать устойчивость этой схемы.

Одной из хороших проверок алгоритма является следующая процедура. Провести расчет на большое число шагов по времени, затем остановить его, изменить знак δt и затем выполнить обратный расчет задачи до начального времени $t = 0$.

Более подробно качественные свойства рассмотренных здесь алгоритмов обсуждаются в [18, С. 145–154].

Литература

[1] *Тихонов А.Н., Васильева А.Б., Свешников А.Г.* Дифференциальные уравнения. 3-е изд. М.: Наука. Физматлит, 1998. 232 с.

[2] *Арнольд В.И.* Дополнительные главы теории обыкновенных дифференциальных уравнений. М.: Наука, 1982. 302 с. Современное переиздание этой книги: Арнольд В.И. Геометрические методы в теории обыкновенных дифференциальных уравнений. Ижевск: Ижевская республиканская типография, 2000. 400 с.

[3] *Самарский А.А., Попов Ю.П.* Разностные методы решения задач газовой динамики. М.: Едиториал УРСС, 2002. 480 с.

[4] *Шокин Ю.И., Яненко Н.Н.* Метод дифференциального приближения. Применение к газовой динамике. Новосибирск, Наука, 1985. 364 с.

[5] *Борис Дж.П., Бук Д.Л.* Решение уравнения непрерывности методом коррекции потоков. / В кн. Вычислительные методы в физике. Управляемый термоядерный синтез. М.: Мир, 1980. с. 92-141.

[6] *Boris J.P., Book D.L.*// J. Comput. Phys., 1973, Vol. 11, pp. 38-69.

[7] *Пасконов В.М., Полежаев В.И., Чудов Л.А.* Численное моделирование процессов тепло- и массообмена. М.: Наука, 1984. 288 с.

[8] *Куликовский А.Г., Погорелов Н.В., Семенов А.Ю.* Математические вопросы численного решения гиперболических систем уравнений. М.: Физматлит, 2001. 608 с.

[9] *Магомедов М.-К.М., Холодов А.С.* Сеточно-характеристические численные методы. М.: Наука, 1988. 288 с.

[10] *Годунов С.К., Рябенький В.С.* Разностные схемы, введение в теорию. М.: Наука, 1977. 400 с.

[11] *Федоренко Р.П.* Введение в вычислительную физику. М.: Изд-во МФТИ, 1994. 526 с.

[12] *Флетчер К.* Вычислительные методы в динамике жидкостей. М.: Мир, 1991. 240 с.

[13] *Белоцерковский О.М.* Численное моделирование в механике сплошных сред. М.: Физматлит, 1994. 442 с.

[14] *Лобанов А.И., Петров И.Б., Старожилова Т.К.* Вычислительные методы для анализа моделей сложных динамических систем. ч. II. Учебное пособие. М.: МФТИ, 2002. 154 с.

[15] *Рождественский Б.Л., Яненко Н.Н.* Системы квазилинейных уравнений и их приложения к газовой динамике. М.: Наука, 1978. 687 с.

[16] *Жуков А.И.* Метод Фурье в вычислительной математике. М.: Наука, 1992. 128 с.

[17] *Галанин М.А.* Численное решение уравнения переноса. / В кн.: Будущее при-кладной математики. Лекции для молодых исследователей. Под ред. Г.Г. Малинецкого. М.: Едиториал УРСС, 2005. 512 с.

[18] *Оран Э., Борис Дж.* Численное моделирование реагирующих потоков. М.: Мир, 1990. 661 с.

Лекция 14. Введение в методы численного решения уравнений газовой динамики

Лекция не обязательна при первом прочтении книги. В лекции приводятся некоторые часто употребляемые численные методы решения уравнений газовой динамики. Особое внимание уделено идее конструирования разностных схем из семейства сеточно- характеристических.

Ключевые слова: уравнения газовой динамики. Характеристическая форма. Дивергентная форма. Пространство неопределенных коэффициентов. Сеточно-характеристические разностные схемы. Искусственная вязкость. Метод частиц в ячейках.

Рассмотренные в предыдущих лекциях разностные методы решения уравнений в частных производных демонстрировались на примере либо линейных задач, либо достаточно простых нелинейных уравнений с хорошо изученными свойствами. Такие задачи в реальной вычислительной практике обычно служат тестами для отработки методов решения более сложных нелинейных систем. Традиционным объектом приложения численных методов служат уравнения механики сплошной среды (МСС). Основных причин тому три. Первая — все математические модели МСС известны достаточно давно, хорошо исследованы и являются частью научной классики. Вторая — математические модели МСС нелинейны. Как правило, у линеаризованных уравнений очень узкая область применения. Третья — в результатах решения задач МСС на протяжении всего XX века была практическая заинтересованность, вызванная бурным развитием авиации, осуществлением наукоемких ядерных и космических программ в разных странах.

В данной книге ограничимся самой простой моделью МСС — уравнениями газовой динамики. Основная идея лекции — демонстрация идей и методов вычислительной математики, рассмотренных выше, в приложении к реальным задачам.

14.1. Формы записи одномерных уравнений газовой динамики

В основе построения математических моделей, описывающих поведение жидкостей и газов, лежит понятие о сплошной среде. Из молекулярной физики известно, что среда состоит из отдельных частиц (молекул, ионов, электронов, атомов), расстояние между которым существен-

но больше их собственных размеров. Длина свободного пробега частицы l (расстояние, пройденное частицей между двумя столкновениями) тем меньше, чем больше частиц заключено в единице объема, чем больше плотность среды. В механике жидкостей и газов рассматриваются среды, содержащие в единице объема большое количество частиц (много больше, чем число Авогадро — число частиц в одной грамм-молекуле вещества, $N_A \approx 6 \cdot 10^{23}$ (г $\cdot$ моль)$^{-1}$).

В таких средах можно рассматривать лишь некоторые усредненные характеристики, не занимаясь изучением поведения каждой частицы в отдельности. В этом предположении заключается идея модели сплошной среды, непрерывно заполняющей пространство.

Количественным критерием применимости приближения сплошной среды может служить неравенство $l/L \ll 1$, где L — характерный пространственный размер задачи (например, размер тела при внешнем обтекании потоком газа). В газах при нормальных условиях $l \approx (10^{-5} \div 10^{-6})$ см, поэтому приведенное условие для тел с размером более 1 см выполняется с достаточной точностью. Предложение о сплошности среды, по-видимому, берет свое начало от Эйлера, впервые рассмотревшего газ как непрерывную деформируемую субстанцию. Не вдаваясь в особенности получения уравнений газовой динамики (этой теме посвящено большое количество литературы, [1, 2, 3]), приведем их в окончательный вид.

Эйлерова (недивергентная) форма одномерной системы уравнений газодинамики имеет вид

$$\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} + \rho\frac{\partial u}{\partial x} = 0 \quad \text{(уравнение неразрывности)}, \tag{14.1}$$

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = -\frac{1}{\rho}\cdot\frac{\partial p}{\partial x} \quad \text{(уравнение движения)},$$

$$\frac{\partial e}{\partial t} + u\frac{\partial e}{\partial x} = -\frac{1}{\rho}\frac{\partial (pu)}{\partial x} \quad \text{(уравнение энергии)},$$

где e — удельная энергия, равная $e = \varepsilon + \frac{u^2}{2}$, ε — удельная внутренняя энергия, u — скорость газа, ρ — плотность среды, p — давление, t — время, x — декартова координата. Эту же систему уравнений в частных производных можно представить в матричной (характеристической) форме

$$\frac{\partial \mathbf{U}}{\partial t} + \mathbf{A}\frac{\partial \mathbf{U}}{\partial x} = 0, \tag{14.2}$$

где $\mathbf{U}$ — вектор-столбец, $\mathbf{A}$ — квадратная матрица 3×3.

Система замыкается уравнением состояния

$$f(\rho,\ \varepsilon,\ p) = 0,$$

которое, например, для идеального газа имеет вид

$$\frac{p}{\rho(\gamma-1)} - \varepsilon = 0.$$

где γ — безразмерная постоянная, равная отношению теплоемкости газа при постоянном давлении и теплоемкости при постоянном объеме — постоянная адиабаты.

В системе, записанной в матричной форме (14.2), учтено, что давление есть функция температуры (или удельной внутренней энергии) и плотности $p = p(\rho, \varepsilon)$ следовательно,

$$\frac{\partial p}{\partial x} = \frac{\partial p}{\partial \rho}\frac{\partial \rho}{\partial x} + \frac{\partial p}{\partial \varepsilon}\frac{\partial \varepsilon}{\partial x}.$$

Не занимаясь выводом формул (это делается простыми алгебраическими преобразованиями), представим другие виды записи уравнения энергии, справедливые для приведенного выше уравнения состояния:

$$\frac{\partial p}{\partial t} + u\frac{\partial p}{\partial x} + \gamma p\frac{\partial u}{\partial x} = 0,$$

$$\frac{\partial p}{\partial t} + u\frac{\partial p}{\partial x} + c^2\rho\frac{\partial u}{\partial x} = 0,$$

где $c = \sqrt{\gamma p/\rho}$ — адиабатическая скорость звука,

$$\frac{\partial S}{\partial t} + u\frac{\partial S}{\partial x} = 0,$$

$$\text{или} \quad \frac{dS}{dt}|_{tr} = 0, \quad \text{где} \quad \frac{d}{dt}|_{tr} = \frac{\partial}{\partial t} + u\frac{\partial}{\partial x},$$

$S = m(\varepsilon\rho)^{1-\gamma}$ — энтропия. Она, как следует из последней формулы, сохраняется вдоль траектории частицы идеального газа, т. е. на траектории уравнения

$$\frac{dX}{dt} = u(t,X), X(0) = X_0.$$

Дивергентная форма уравнений газовой динамики получается при записи соответствующих законов сохранения в интегральной форме. Если затем совершить предельный переход, то дифференциальная запись уравнений будет

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) = 0,$$

$$\frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(\rho u^2 + p) = 0, \qquad (14.3)$$

$$\frac{\partial}{\partial t}\left[\rho\left(\varepsilon+\frac{u^2}{2}\right)\right]+\frac{\partial}{\partial x}\left[\rho u\left(\varepsilon+\frac{u^2}{2}+\frac{p}{\rho}\right)\right]=0,$$

или в матричной форме

$$\frac{\partial \mathbf{R}}{\partial t}+\frac{\partial \mathbf{Q}}{\partial x}=0,$$

где $\mathbf{R}=\{\rho,\rho u,\rho(\varepsilon+\frac{u^2}{2})\}^{\mathrm{T}}$, $\mathbf{Q}=\{\rho u,\rho u^2+p,\rho u(\varepsilon+\frac{u^2}{2}+\frac{p}{\rho})\}^{\mathrm{T}}$.

Интегральная форма этих уравнений получается при использовании теоремы Гаусса-Остроградского

$$\int\limits_{\Omega}\int(\frac{\partial \mathbf{R}}{\partial t}+\frac{\partial \mathbf{Q}}{\partial x})\,dtdx=\oint\limits_{\Gamma}(\mathbf{R}dx-\mathbf{Q}dt)=0,$$

где Γ — граница замкнутой области интегрирования Ω в плоскости t, x. Здесь $u(t,x),\rho(t,x),p(t,x)$ — скорость, плотность и давление газа соответственно. В механике сплошных сред вводится эйлерово и лагранжево описание поведения среды. В первом случае наблюдатель полагается неподвижным, например, стоящим на берегу реки. Соответственно расчетная сетка будет неподвижной (фиксированная эйлерова сетка). Во втором случае полагаем, что наблюдатель движется вместе со средой, например, находится на лодке, плывущей по течению реки. В этом случае лагранжева расчетная сетка будет двигаться вместе с частицами среды.

Лагранжева форма одномерных уравнений газовой динамики имеет вид

$$\frac{du}{\partial t}+\frac{1}{\rho}\frac{\partial p}{\partial \xi}I=0,$$

$$\frac{d\rho}{\partial t}+\rho\frac{\partial u}{\partial \xi}I=0,$$

$$\frac{\partial e}{dt}+\frac{1}{\rho}\frac{\partial(pu)}{\partial \xi}I=0,$$

$$\frac{dx}{dt}=u(t,\xi),$$

где $I=\left(\frac{\partial x}{\partial \xi}\right)^{-1}$, x и ξ соответственно эйлерова и лагранжева координаты, связь между которыми дается последним уравнением. Эта система в одномерном случае может быть записана в другом виде. Если ввести лагранжеву массовую координату $\eta(\xi)$, связанную с лагранжевой координатой ξ дифференциальным уравнением $\frac{d\eta}{d\xi}=\rho(0,\xi)$, в массовых координатах последняя система запишется как

$$\frac{du}{dt}+\frac{\partial p}{\partial \eta}=0,$$

$$\frac{d\mathrm{v}}{dt} - \frac{\partial u}{\partial \eta} = 0,$$

$$\frac{de}{dt} + \frac{\partial (pu)}{\partial \eta} = 0. \qquad (14.4)$$

Эта система дополняется уравнениями, связывающими лагранжевы и эйлеровы координаты

$$\frac{dx}{dt} = u(t, \eta), \quad \frac{dx(0, \eta)}{d\eta} = \mathrm{v}(0, \eta),$$

где $\mathrm{v} = \rho^{-1}$ — удельный объем.

14.2. Методы Лакса-Вендроффа и Мак-Кормака

Если система уравнений газодинамики записана в дивергентной форме

$$\frac{\partial \mathbf{R}}{\partial t} + \frac{\partial \mathbf{Q}}{\partial x} = 0,$$

то запись разностных схем, соответствующих методам Лакса-Вендроффа и Мак-Кормака, аналогична их записи для численного решения уравнения переноса (лекция 13).

Так, схема Лакса-Вендроффа может быть представлена в следующем виде:

$$\frac{\tilde{\mathbf{R}}_{m+1/2} - 0.5(\mathbf{R}^n_{m+1} + \mathbf{R}^n_m)}{\tau/2} + \frac{\mathbf{Q}^n_{m+1} - \mathbf{Q}^n_m}{h} = 0,$$

$$\frac{\tilde{\mathbf{R}}_{m-1/2} - 0.5(\mathbf{R}^n_m + \mathbf{R}^n_{m-1})}{\tau/2} + \frac{\mathbf{Q}^n_m - \mathbf{Q}^n_{m-1}}{h} = 0.$$

(первый этап);

$$\frac{\mathbf{R}^{n+1}_m - \mathbf{R}^n_m}{\tau} + \frac{\tilde{\mathbf{Q}}_{m+1/2} - \tilde{\mathbf{Q}}_{m-1/2}}{h} = 0$$

(второй этап).

Схему МакКормака представим следующим образом:

$$\frac{\tilde{\mathbf{R}}_m - \mathbf{R}^n_m}{\tau} + \frac{\mathbf{Q}_{m+1} - \mathbf{Q}_m}{h} = 0,$$

$$\frac{\tilde{\mathbf{R}}_{m-1} - \mathbf{R}^n_{m-1}}{\tau} + \frac{\mathbf{Q}_m - \mathbf{Q}_{m-1}}{h} = 0$$

(первый этап);

$$\frac{\mathbf{R}_m^{n+1} - 0.5(\mathbf{R}_m^n + \tilde{\mathbf{R}}_m)}{\tau} + \frac{\tilde{\mathbf{Q}}_m - \tilde{\mathbf{Q}}_{m-1}}{2h} = 0$$

(второй этап).

Так как использована дивергентная форма записи исходных уравнений, то можно ожидать, что полученные таким образом схемы окажутся *консервативными*. О консервативных схемах газовой динамики подробнее в [4, 5].

14.3. Сеточно-характеристический метод для численного решения уравнений газовой динамики (М.-К. М. Магомедова–А. С. Холодова)

Этот метод (точнее, семейство методов) наиболее эффективен при численном решении задач, для которых существенным являются учет волновых процессов в сплошной среде. Подробнее о методе в книге [6], детальное описание — в монографии [7].

Матрица $\mathbf{A}$ системы уравнений газовой динамики, записанной в форме

$$\frac{\partial \mathbf{U}}{\partial t} + \mathbf{A}\frac{\partial \mathbf{U}}{\partial x} + \mathbf{f} = 0,$$

$$\mathbf{U} = (\rho, u, \varepsilon)^T, \mathbf{A} = \begin{pmatrix} u & \rho & 0 \\ \frac{1}{\rho}\frac{\partial p}{\partial \rho} & u & \frac{1}{\rho}\frac{\partial p}{\partial \varepsilon} \\ 0 & p/\rho & u \end{pmatrix},$$

имеет вещественные собственные числа $\lambda_1 = u + c, \lambda_2 = u, \lambda_3 = u - c$ и соответствующие им собственные векторы, ω_i (например, левые, которые находятся при решении системы линейных алгебраических уравнений $\omega_i \mathbf{A} = \lambda_i \omega_i$). Таким образом, одномерная система уравнений газовой динамики является квазилинейной системой гиперболического типа.

Матрица из левых собственных векторов, записанных в строку, является матрицей перехода в базис из собственных векторов матрицы $\mathbf{A}$:

$$\Omega = \begin{pmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{pmatrix} = \begin{pmatrix} \frac{\partial p}{\partial \rho} & \rho c & \frac{\partial p}{\partial \varepsilon} \\ p & 0 & -\rho^2 \\ p & -\rho c & \frac{\partial p}{\partial \rho} \end{pmatrix}.$$

Матрица $\mathbf{A}$ представима в виде $\mathbf{A} = \Omega^{-1}\Lambda\Omega$, где Λ — диагональная матрица, состоящая из собственных чисел матрицы $\mathbf{A}$:

$$\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \lambda_3).$$

Для получения разностной схемы введем обозначения

$$\mathbf{\Lambda}^+ = \frac{1}{2}(\mathbf{\Lambda} + |\mathbf{\Lambda}|), \mathbf{\Lambda}^- = \frac{1}{2}(\mathbf{\Lambda} - |\mathbf{\Lambda}|),$$

$$\mathbf{A}^+ = \frac{1}{2}(\mathbf{A} + |\mathbf{A}|), \boldsymbol{A}^- = \frac{1}{2}(\mathbf{A} - |\mathbf{A}|), \sigma = \tau/h.$$

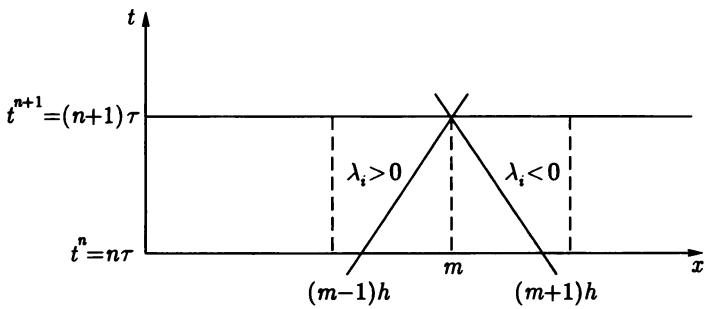


Рис. 14.1

Умножим каждое из уравнений исходной системы газодинамики, записанных в матричной форме на ω_i; в результате получим

$$\omega_i \frac{\partial \mathbf{U}}{\partial t} + \omega_i A \frac{\partial \mathbf{U}}{\partial x} + \omega_i \mathbf{f} = 0,$$

или, учитывая, что ω_i — левый собственный вектор,

$$\omega_i \frac{\partial \mathbf{U}}{\partial t} + (\lambda_i \omega_i) \frac{\partial \mathbf{U}}{\partial x} + \omega_i \mathbf{f} = 0.$$

Построим разностную аппроксимацию полученной системы уравнений в частных производных с учетом знака собственных чисел (или направления характеристик). Для облегчения восприятия ограничимся простейшим методом первого порядка аппроксимации

$$\omega_i \frac{\mathbf{U}_m^{n+1} - \mathbf{U}_m^n}{\tau} \mp \lambda_i \omega_i \frac{\mathbf{U}_{m\mp1}^n - \mathbf{U}_m^n}{h} + \omega_i \mathbf{f} = 0.$$

Приведенное выше разностное уравнение может быть записано в матричной форме

$$\mathbf{\Omega}(\mathbf{U}_m^{n+1} - \mathbf{U}_m^n) - \sigma \mathbf{\Lambda}^+ \mathbf{\Omega}(\mathbf{U}_{m-1}^n - \mathbf{U}_m^n) +$$
$$+ \sigma \mathbf{\Lambda}^- \mathbf{\Omega}(\mathbf{U}_{m+1}^n - \mathbf{U}_m^n) + \tau \mathbf{\Omega} \mathbf{f}_m^n = 0,$$

или в виде, разрешенном относительно верхнего слоя:

$$\mathbf{U}_m^{n+1} = \mathbf{U}_m^n - \sigma\left[(\mathbf{\Omega}^{-1}\mathbf{\Lambda}^+\mathbf{\Omega})_m^n(\mathbf{U}_{m-1}^n - \mathbf{U}_m^n) - \right.$$

$$\left. -(\mathbf{\Omega}^{-1}\mathbf{\Lambda}^-\mathbf{\Omega})_m^n(\mathbf{U}_{m+1}^n - \mathbf{U}_m^n)\right] + \tau\mathbf{f}_m^n,$$

или в компактной форме:

$$\mathbf{U}_m^{n+1} = \mathbf{U}_m^n + \sigma(-\mathbf{B}^+\Delta^-\mathbf{U} + \mathbf{B}^-\Delta^+\mathbf{U})_m^n + \tau\mathbf{f}_m^n,$$

где

$$\mathbf{B}^+ = \mathbf{\Omega}^{-1}\mathbf{\Lambda}^+\mathbf{\Omega},\ \mathbf{B}^- = \mathbf{\Omega}^{-1}\mathbf{\Lambda}^-\mathbf{\Omega}.$$

Учет направления характеристик позволяет получать устойчивые разностные схемы для системы уравнений газовой динамики. Сеточно-характеристические схемы позволяют гибко менять форму шаблона в зависимости от локальных свойств решения задачи. Существует обобщение методов на случаи двух и трех пространственных измерений. В сочетании с методом неопределенных коэффициентов сеточно-характеристические схемы дали очень хорошие результаты не только в традиционной газовой динамике, но и в механике деформируемого твердого тела, магнитной гидродинамике [6, 7].

14.4. Разностная схема И.М. Гельфанда для численного решения одномерной системы уравнений газовой динамики

Система уравнений газодинамики решается в области $t \geqslant 0$, $\eta \in [0, X]$, отрезок интегрирования разбивается на интервалы узлами $\{\eta_i\}_0^N$. Все интервалы заполнены газом, что соответствует приближению механики сплошной среды. Величины на интервалах считаются кусочно-постоянными. При численном решении определяются шесть функций: $u, p, \varepsilon, T, v, x$ — скорость, давление, удельная внутренняя энергия, температура, удельный объем, эйлерова координата.

Для решения задачи используется система одномерных нестационарных уравнений в частных производных, описывающих поведение газа в лагранжевых переменных (η — лагранжева координата):

$$\frac{dv}{dt} - \frac{du}{\partial\eta} = 0,$$

$$\frac{du}{dt} + \frac{\partial(p+Q)}{\partial\eta} = 0,$$

$$\frac{d}{dt}\left(\varepsilon+\frac{u^2}{2}\right)+\frac{\partial[(p+Q)u]}{\partial\eta}=\frac{\partial}{\partial\eta}\left[a(T,v)\frac{\partial T}{\partial\eta}\right],$$

$$\frac{dx}{dt}=u. \tag{14.5}$$

Здесь $a(T,v)$ — заданный коэффициент теплопроводности,

$$Q=\frac{\mu}{v}\left(\frac{\partial u}{\partial\eta}-\left|\frac{\partial u}{\partial\eta}\right|\right)\frac{\partial u}{\partial\eta}$$

— искусственная вязкость Рихтмайера-Неймана, μ — коэффициент искусственной вязкости. Очевидно, что $Q\neq 0$ если $\frac{\partial u}{\partial\eta}<0$, при этом $\frac{\partial v}{\partial t}<0$ или $\frac{\partial\rho}{\partial t}>0$. Так как плотность со временем увеличивается, происходит сжатие газа. В зонах разрежения, где $\frac{\partial u}{\partial x}>0$ и $Q=0$, искусственная вязкость действует только в зонах сжатия.

Коэффициент теплопроводности зависит от температуры и задается как

$$a(T,v)=T^{\gamma}a(T,v),\quad \gamma>1,\quad a_1\leqslant a\leqslant a_2,\quad a_1>0.$$

Начальные данные для рассматриваемой задачи будут

$$u(0,\eta)=u_0,\quad x(0,\eta)=x_0,\quad v(0,\eta)=v_0,\quad T(0,\eta)=T_0.$$

Условия на границах выбираем следующие:

$$u(t,0)=u_1(t),\quad a\frac{\partial T}{\partial\eta}(t,0)=P_1(t),\quad T(t,X)=T_2(t),$$

При численном интегрировании полагаем, что значения функций $u,T,,v,x$ известны на n слое и задача состоит в вычислении этих же функций на $n+1$ слое. Сеточные функции $u_{m^{n+1}},T^{n+1}_{m+1/2},v^{n+1}_{m+1/2},x_{m^{n+1}}$ вычисляются с помощью неявной разностной схемы.

В результате разностная схема записывается как

$$\frac{v^{n+1}_{m+1/2}-v^{n}_{m+1/2}}{\tau}-\frac{u^{n+1}_{m+1}-u^{n+1}_{m}}{2}=0, m=1,\ldots,M,$$

$$\frac{u^{n+1}_{m}-u^{n}_{m}}{\tau}+\frac{(p+Q)^{n+1}_{m+1/2}-(p+Q)^{n+1}_{m-1/2}}{h_m}=0$$

$$h_m=\eta_{m+1/2}-\eta_{m-1/2}, m=1,\ldots,M-1, \tag{14.6}$$

$$\frac{1}{\tau}\left[\varepsilon^{n+1}_{m+1/2}+\frac{(u^{n+1}_{m})^2+(u^{n+1}_{m+1})^2}{4}-\varepsilon^{n}_{m+1/2}-\frac{(u^{n}_{m})^2+(u^{n}_{m+1})^2}{4}\right]+$$

$$+\frac{1}{h_{m+1/2}}\left[(p+Q)_{m+1}^{n+1}u_{m+1}^{n+1}-(p+Q)_{m}^{n+1}u_{m}^{n+1}\right]=$$

$$=\frac{1}{h_{m+1/2}}\left[a_{m+1}\frac{T_{m+3/2}^{n+1}-T_{m+1/2}^{n+1}}{h_{m+1}}-a_m\frac{T_{m+1/2}^{n+1}-T_{m-1/2}^{n+1}}{h_m}\right],$$

$$h_{m+1/2}=\eta_{m+1}-\eta_m, m=1,\ldots,M-1,$$

$$\frac{x_m^{n+1}-x_m^n}{\tau}-\frac{u_m^{n+1}-u_m^n}{2}=0,\quad m=0,\ldots,M.$$

К этим уравнениям системы добавим разностное выражение для вычисления искусственной вязкости

$$Q_{m+1/2}=\frac{\mu}{(hv)_{m+1/2}}\left[(u_{m+1}-u_m)-|u_{m+1}-u_m|\right](u_{m+1}-u_m)$$

с фиктивным краевым условием $Q_{M+1/2}=0$ и интерполяционное выражение для p_m

$$p_m=\frac{h_{m-1/2}\cdot p_{m+1/2}+h_{m+1/2}\cdot p_{m-1/2}}{h_{m-1/2}+h_{m+1/2}}.$$

Схема имеет второй порядок аппроксимации по координате, а при весовом коэффициенте, равном 0,5, и второй порядок по времени, причем все точки спектра лежат на единичной окружности. Подробнее о данной схеме в [10].

14.5. Метод частиц в ячейках Харлоу (PIC method:Particle-In-Cell)

Метод PIC разработан Харлоу в Лос-Аламосской лаборатории (США) в 60-х годах прошлого века для расчета процессов с большими деформациями исходной области интегрирования (расплескивание, разрушение).

Область интегрирования покрывается фиксированной в пространстве расчетной сеткой, шаг которой h постоянен по обеим координатам x, y, ячейки занумерованы двумя индексами k, l.

В центре ячейки вычисляются величины u_{1kl}^n, u_{2kl}^n (компоненты скорости газа), $\varepsilon_{ikl}^n, m_{ikl}^n$ где i — номер вещества. ε_{ikl}^n — удельная внутренняя энергия газа с номером i, m_{ikl}^n — масса этого вещества. Если этого вещества в ячейке нет, то в ней и энергия, и масса полагаются равными нулю.

Предположим, что в каждой ячейке содержится несколько частиц (5–10), каждая из которых характеризуется координатами X_j^n, Y_j^n массой μ_j, i_j — номер вещества, из которого состоит частица с номером j.

Шаг численного интегрирования состоит в расчете величин $\{u_1, u_2, \varepsilon_i, m_i\}_{kl}^{n+1}$ и $\{X,\ Y\}_j^{n+1}$ на верхнем временном слое t_{n+1} по вычисленным величинам $\{u_1, u_2, \varepsilon_i, m_i\}_{kl}^{n}$, $\{X,\ Y\}_j^{n}$ на нижнем слое t_n.

На первом этапе расчета учитываются изменения искомых функций только за счет сил давления. При этом предположении разностные соотношения аппроксимируют уравнения

$$\frac{\partial \rho}{\partial t} = 0,$$

$$\frac{\partial(\rho u_1)}{\partial t} = -\frac{\partial p}{\partial x},$$

$$\frac{\partial(\rho u_2)}{\partial t} = -\frac{\partial p}{\partial y},$$

$$\frac{\partial E}{\partial t} + \frac{\partial(p u_1)}{\partial x} + \frac{\partial(p u_2)}{\partial y} = 0,$$

где

$$E = \rho e = \rho\left[\varepsilon + \frac{1}{2}(u_1^2 + u_2^2)\right].$$

В расчетах участвуют также уравнения состояния для каждого газа

$$p_i = F_i(\varepsilon_i, \rho_i).$$

На втором этапе аппроксимируются конвективные члены

$$\frac{\partial \rho}{\partial t} + \frac{\partial(\rho u_1)}{\partial x} + \frac{\partial(\rho u_2)}{\partial y} = 0,$$

$$\frac{\partial \rho}{\partial t} + \frac{\partial(\rho u_1)}{\partial x} + \frac{\partial(\rho u_2)}{\partial y} = 0,$$

$$\frac{\partial(\rho u_1)}{\partial t} + \frac{\partial(\rho u_1^2)}{\partial x} + \frac{\partial(\rho u_1 u_2)}{\partial y} = 0,$$

$$\frac{\partial(\rho u_2)}{\partial t} + \frac{\partial(\rho u_1 u_2)}{\partial x} + \frac{\partial(\rho u_2)}{\partial y} = 0,$$

$$\frac{\partial E}{\partial t} + \frac{\partial(E u_1)}{\partial x} + \frac{\partial(E u_2)}{\partial y} = 0.$$

Опишем вычислительную процедуру на первом этапе. Известны u_1^n, u_2^n, m^n, ε^n, X^n, Y^n (остальные индексы для простоты изложения

опускаются). Сначала рассчитывается давление p_{ij}^n, исходя из предложения равенства давлений на границе двух сред $p_1 = p_2 = \ldots$, или

$$F_1(\varepsilon_1^n, \gamma_1^{-1} m_1^n) = F_2(\varepsilon_2^n, \gamma_2^{-1} m_2^n) = \ldots$$

К этим уравнениям добавляется условие $\sum \gamma_i = h^2$, поскольку γ_i — часть объема h^2 ячейки, занимаемого газом с номером i. По известной массе i газа находим его плотность: $\rho_i = m_i^n/\gamma_i$, а по известной удельной энергии ε_i^n — давление $p_i = F_i(\varepsilon_i^n, \gamma_i^{-1} m_i^n)$.

Затем по закону Дальтона находится давление, p_{ikl}^n, которое приписывается к центру ячейки k, l. Система нелинейных алгебраических уравнений решается, вообще говоря, итерационным методом. В случае $p = \rho f_i(\varepsilon)$ выписывается ее явное решение. Далее находим предварительные значения расчитываемых величин, которые обозначим как $\bar{u}_1$, $\bar{u}_2$, $\bar{m}$, $\bar{\varepsilon}$. Первое из уравнений $\frac{\partial \rho}{\partial t} = 0$ — закон сохранения массы — на сетке приобретает вид $\bar{m}_{ikl} = m_{ikl}^n$. Поскольку на первом этапе $\frac{\partial \rho}{\partial t} = 0$, то два следующих разностных уравнения — уравнения движения — записываются как

$$\rho_{kl}^n \frac{\bar{u}_{1kl} - u_{1kl}^n}{\tau} + \frac{p_{k+1/2,l}^n - p_{k-1/2,l}^n}{h} = 0,$$

$$\rho_{kl}^n \frac{\bar{u}_{1kl} - u_{2kl}^n}{\tau} + \frac{p_{k,l+1/2}^n - p_{k,l-1/2}^n}{h} = 0.$$

Здесь

$$p_{k+1/2,l}^n = \frac{1}{2}(p_{kl}^n + p_{k+1,l}^n),$$

$$m_{kl}^n = \sum_i m_{ikl}^n, \rho_{kl}^n = m_{kl}^n/h^2.$$

Последняя из рассчитываемых величин — энергия. Дискретный аналог уравнения энергии в методе частиц в ячейках будет

$$\frac{\bar{E}_{kl} - E_{kl}^n}{\tau} + \frac{p_{k+1/2,l}^n \cdot u_{1,k+1/2,l}^{n+1/2} - p_{k-1/2,l}^n \cdot u_{1,k-1/2,l}^{n+1/2}}{h} +$$

$$+ \frac{p_{k,l+1/2}^n \cdot u_{2,k,l+1/2}^{n+1/2} - p_{k,l-1/2}^n \cdot u_{2k,l-1/2}^{n+1/2}}{h} = 0.$$

Здесь

$$u_{1,k+1/2,l}^{n+1/2} = \frac{\bar{u}_{1,kl} + u_{1,kl}^n + \bar{u}_{1,k+1,l} + u_{1,k+1,l}^n}{4},$$

$$u_{2,k,l+1/2}^{n+1/2} = \frac{\bar{u}_{2kl} + u_{2kl}^n + \bar{u}_{2,k,l+1} + u_{2k,l-1}}{4}.$$

Вычислим величину e^n_{kl} — энергию. Напомним, что $e = \rho(\varepsilon + \frac{u_1^2+u_2^2}{2})$. Тогда eh^2 есть энергия в ячейке $h \times h$:

$$\mathrm{E}^n_{kl} \cdot h^2 = \left[(h^2 \cdot \rho) \cdot \varepsilon + (h^2\rho)\frac{u_1^2 + u_2^2}{2}\right]^n_{kl},$$

где ρh^2 — масса ячейки, равная $m^n_{kl} = \sum\limits_i m^n_{ikl}$. В таком случае, с учетом закона сохранения массы,

$$\left[(h^2\rho)\frac{u_1^2 + u_2^2}{2}\right]^n_{kl} = \frac{1}{2} m^n_{kl}\left[(u^n_{1,kl})^2 + (u^n_{2,kl})^2\right].$$

Вычислим величину $\left[(h^2\rho)\varepsilon\right]^n_{kl}$, имеющую смысл полной внутренней энергии в ячейке, зная массу m^n_{ikl} и удельную внутреннюю энергию ε^n_{ikl} вещества.

$$\left[(h^2\rho)\varepsilon\right]^n_{kl} = \sum_i m^n_{ikl} E^n_{ikl},$$

теперь имеется алгоритм вычисления E^n_{kl} и, следовательно, $\tilde{E}^n_{kl}$.

Из соотношения

$$\bar{e}_{kl} \cdot h^2 = (h^2\rho_{kl})\bar{\varepsilon}_{kl} + (h^2\rho_{kl}) \cdot \frac{(\bar{u}_{1,kl})^2 + (\bar{u}_{2,kl})^2}{2}$$

находим величину полной удельной внутренней энергии $\bar{\varepsilon}_{kl}$. Однако искомыми являются значения удельной внутренней энергии для каждого вещества. Пусть $\Delta\varepsilon_i$ — изменение удельной внутренней энергии i вещества за первый этап шага по времени по i веществу. Зная m_i (масса i вещества), запишем полное приращение полной удельной внутренней энергии в ячейке (Δ_ε) и приравняем его к уже полученному полному приращению

$$\sum_i m^n_{ikl}\,\Delta\varepsilon_{ikl} = m^n_{kl}(\bar{\varepsilon}_{kl} - \varepsilon^n_{kl}).$$

Для определения изменения количества каждого газа нужно сделать некое правдоподобное предположение, например, считать, что все $\Delta\varepsilon_{ikl}$ одинаковы. Тогда $N\,\Delta\varepsilon_{ikl} = \bar{\varepsilon}^n_{kl} - \varepsilon^n_{kl}$ и, соответственно, $\bar{\varepsilon}_{ikl} = \varepsilon^n_{ikl} + \Delta\varepsilon_{ikl}$. На этом первый этап расчета (предиктор) закончен.

Рассмотрим второй этап расчета.

Движение частиц описывается обыкновенными дифференциальными уравнениями

$$\dot{X}_p = u_1(t, X_p, Y_p), \dot{Y}_p = u_2(t, X_p, Y_p),$$

которые могут быть приближены, например, с использованием явного метода Эйлера. Тогда дифференциальные уравнения заменяются разностными уравнениями

$$X_p^{n+1} = X_p + \tau \tilde{u}_{1p}, Y_p^{n+1} = Y_p^n + \tau \tilde{u}_{2p},$$

где скорости частиц $\tilde{u}_k, \tilde{v}_k$ определяются интерполяцией величин $\bar{u}_1, \bar{u}_2$ в ячейках, окружающих p частицу.

После этого рассчитывается перенос массы и вычисляется новая масса каждой ячейки. Для этого выделяются три группы частиц:

- частицы, оставшиеся при переходе на $n+1$ слой в пределах ячейки, которые, очевидно, не вносят изменений в массу, импульс, энергию новой ячейки, т. е. $(X_p^n, Y_p^n) \in \omega_{ij}, (X_p^{n+1}, Y_p^{n+1}) \in \omega_{ij}$, где ω_{ij} — обозначение старой ячейки,

- частицы, покинувшие ячейку ω_{ij}: $(X_p^n, Y_p^n) \in \omega_{ij}$, $(X_p^{n+1}, Y_p^{n+1}) \notin \omega_{ij}$,

- частицы, перешедшие из соседних ячеек: $(X_p^n, Y_p^n) \notin \omega_{ij}$, $(X_p^{n+1}, Y_p^{n+1}) \in \omega_{ij}$,

На шаг по времени накладывается ограничение

$$\tau < \frac{h}{\sqrt{u_1^2 + u_2^2}},$$

что означает запрет на перемещение частицы за один шаг больше, чем на одну ячейку. Перемещение в данном методе возможно только в соседнюю ячейку. Предположим, что каждая p частица, перешедшая на $n+1$ шаге по времени в соседнюю ячейку, переносит в нее массу m_p. Это означает, что значение массы m_{ikl} считается путем сложения масс всех частиц типа i, для которых $(X_p^{n+1}, Y_p^{n+1}) \in \omega_{ij}$

Процедура вычисления импульса выполняется следующим образом. Компоненты полного импульса частиц в ячейке (k, l) могут быть вычислены, как $m_{kl}^n \bar{u}_{1,kl}, m_{kl}^n \bar{u}_{2,kl}$, при этом p частица, покинувшая ячейку ω_{ij}, уносит импульс $m_p \bar{u}_{1,kl}, m_p u_{2,kl}$. Изменение импульса в S_{kl} за один шаг по времени будет

$$m_k^{n+1} u_{1,kl}^{n+1} = m_{kl}^n \bar{u}_{1,kl} - (\sum \mu_p \bar{u}_{1,kl})_1 + (\sum \mu_p \bar{u}_{1,kl})_2,$$

$$m_{kl}^{n+1} u_{2kl}^{n+1} = M_{kl}^n \bar{u}_{2kl} - (\sum \mu_p \vec{u}_{2kl})_1 + (\sum \mu_p \bar{u}_{2k'l'})_2,$$

здесь символы суммирования означают суммирование по частицам (p), покинувшим данную ячейку и пришедшим в нее, соответственно. После

вычисления компонентов импульса каждой ячейки вычисляются компоненты скорости u_{1kl}^{n+1}, u_{2kl}^{n+1}.

Частица типа i, переходящая из S_{ke} в другую ячейку, переносит полную энергию

$$\Delta E_p = m_p \left[\bar{\varepsilon}_{ikl} + \frac{(\bar{u}_{1kl})^2 + (\bar{u}_{2kl})^2}{2} \right].$$

Тогда можно вычислить энергию i вещества в ячейке ω_{kl} на промежуточном шаге:

$$E_{ikl} = \sum_{i_p \in i} m_p \left[\bar{\varepsilon}_{ikl} + \frac{(\bar{u}_{1kl})^2 + (\bar{u}_{2kl})^2}{2} \right].$$

При $t = t_{n+1}$ полная удельная энергия изменится на величину

$$E_{ikl}^{n+1} = \bar{E}_{ikl} - (\sum_{i_p \in i} \cdot \Delta E_p)_1 + (\sum_{i_p \in i} \cdot \Delta E_p)_2,$$

где знаки суммирования снова означают суммы по всем частицам, покинувшим ячейку ω_{kl} и пришедшим в нее, соответственно. Далее получим

$$\varepsilon_{ikl}^{n+1} = \frac{h^2}{m_{ikl}^{n+1}} E_{ikl}^{n+1} - \frac{1}{2} \left[(u_{1kl}^{n+1})^2 + (u_{2kl}^{n+1})^2 \right].$$

Отметим недостатки этого метода. Во-первых, это дискретность плотности, что приводит при небольшом количестве частиц в ячейке к скачкообразным изменениям плотности. Во-вторых, аппроксимация исходных уравнений достигается при количестве частиц, стремящемся к бесконечности. Кроме того, этот метод требует значительно большего количества памяти, чем конечно-разностные методы, так как наряду с физическими характеристиками узлов необходимо хранить и свойства частиц в ячейках.

Подробное описание метода частиц в ячейках в [11].

14.6. Задачи для самостоятельного решения

1. **Волны Римана**

Выпишем нелинейную систему уравнений одномерных движений идеальной сжимаемой жидкости в случае баротропных процессов. Она состоит из уравнения Эйлера

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + \frac{1}{\rho} \frac{\partial p}{\partial x} = 0,$$

уравнения неразрывности

$$\frac{\partial \rho}{\partial t} + \rho \frac{\partial u}{\partial x} + u \frac{\partial \rho}{\partial x} = 0$$

и условия баротропности

$$p = f(\rho).$$

Уравнения позволяют определить плотность ρ и скорость u в зависимости от координаты x и времени t. Система не имеет решений, зависящих только от $x \pm a_0 t$, но оказывается возможным найти решение этой системы, представляющее собой плоскую волну и являющееся обобщением решений вида $f(x \pm a_0 t)$. Будем искать такие решения системы, для которых скорость u является функцией только плотности ρ. Частные решения системы уравнений носят названия решений Римана; соответствующие этим решениям движения называются волнами Римана $f(x \pm a_0 t)$.

(a) Доказать, что в рассматриваемом течении скорость можно определить по формуле

$$u = \pm \int \sqrt{\frac{dp}{d\rho}} \frac{d\rho}{\rho}.$$

Обозначим $\frac{dp}{d\rho} = a^2(\rho)$ и введем величину $c = u + a$. Какой физический смысл имеет величина c?

(b) Задав начальный профиль возмущения плотности, численно решить уравнение для $\rho(x,t)$:

$$\frac{\partial \rho}{\partial \mathrm{t}} + c(\rho) \frac{\partial \rho}{\partial x} = 0,$$

а) для случая адиабатических движений совершенного газа ($\gamma = 1,4$):

$$c(\rho) = \sqrt{A\gamma} \left[1 + \frac{2}{\gamma - 1}\right] \rho^{\frac{1}{2}(\gamma - 1)},$$

б) задав самостоятельно некоторую зависимость давления от плотности, $p = f(\rho)$.

(c) Описать качественное поведение решения $\rho(x,\ t)$. Указать, какие требования к численному методу предъявляет возникновение в потоке скачков уплотнения. Вывести зависимость $p(\rho)$, при которой не возникает эффекта опрокидывания волны сжатия Римана. Дать физическую трактовку полученного соотношения. Провести численный расчет течения с полученной зависимостью $p(\rho)$.

(d) Доказать, что рассмотренные решения Римана можно определить как такие решения, для которых имеется семейство прямолинейных характеристик.

(e) Поставить условия существования центрированных волн Римана, когда

$$u = u_0\, f(x/t), \rho = \rho_0\, \varphi(x/t).$$

Течения подобного типа — частный случай автомодельных течений, когда решение зависит от некоторой комбинации независимых переменных. Проиллюстрировать численными расчетами особенности распространения центрированных волн Римана.

Литература

[1] *Седов Л.И.* Механика сплошной среды, т. 1, 2. М.: Наука, 1976.

[2] *Лойцянский Л.Г.* Механика жидкости и газа. М.: Дрофа, 2003. 840 с.

[3] *Овсянников Л.В.* Лекции по основам газовой динамики. Москва-Ижевск: Институт компьютерных исследований, 2003. 336 с.

[4] *Самарский А.А., Попов Ю.П.* Разностные методы решения задач газовой динамики. М.: Едиториал УРСС, 2002. 480 с.

[5] *Попов Ю.П.* О консервативности разностных схем. / В кн.: Будущее прикладной математики. Лекции для молодых исследователей. Под ред. Г.Г. Малинецкого. М.: Едиториал УРСС, 2005. 512 с.

[6] *Белоцерковский О.М.* Численное моделирование в механике сплошных сред. М.: Физматлит, 1994. 442 с.

[7] *Магомедов М.-К.М., Холодов А.С.* Сеточно-характеристические численные методы. М.: Наука, 1988. 288 с.

[8] *Годунов С.К., Забродин А.В. и др.* Численное решение многомерных задач газовой динамики. М.: Наука, 1976. 400 с.

[9] *Рождественский Б.Л., Яненко Н.Н.* Системы квазилинейных уравнений. М.: Наука, 1978. 687 с.

[10] *Федоренко Р.П.* Введение в вычислительную физику. М.: Изд-во МФТИ, 1994. 526 с.

[11] *Харлоу Ф.Х.* Численный метод частиц в ячейках для задач газовой динамики. Вычислительные методы в гидродинамике. М.: Мир. 1967. 460 с.

[12] *Флетчер К.* Вычислительные методы в динамике жидкостей. М.: Мир, 1991. 240 с

[13] *Андерсен Д., Таннехилл Дж., Плетчер Р.* Вычислительная гидромеханика и теплообмен. М.: Мир, 1990. т. 1, 2.

[14] *Оран Э., Борис Дж.* Численное моделирование реагирующих потоков. М.: Мир, 1990. 661 с.

Лекция 15. Численное решение уравнений в частных производных гиперболического типа с большими градиентами решений

Лекция продолжает тему предыдущей лекции и также является необязательной. В ней рассматриваются некоторые идеи, нашедшие свое применение для построения разностных схем решения задач механики сплошной среды. Рассматриваются способы построения гибридных схем для задач с большими градиентами решения, описываются идеи TVD- и ENO-схем. Вкратце описываются разностные схемы, построенные на основе решения задачи о распаде произвольного газодинамического разрыва (схемы С.К. Годунова).

Ключевые слова: потоковая форма представления разностных схем, регуляризация, сглаживание, антидиффузия, гибридные схемы, пространство неопределенных коэффициентов, TVD-схемы, ENO-схемы, метод С. К. Годунова.

15.1. Потоковая форма представления разностных схем

Рассмотрим линейное одномерное уравнение переноса

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0, a = \text{const}.$$

Приблизим его при помощи схем «левый уголок» и «правый уголок» в зависимости от направления переноса (схемы Куранта–Изаксона–Риса [1] или С. К. Годунова [2]):

$$\frac{u_m^{n+1} - u_m^n}{\tau} + a\frac{u_m^n - u_{m-1}^n}{h} = 0,\ a > 0, \frac{u_m^{n+1} - u_m^n}{\tau} + a\frac{u_{m+1}^n - u_m^n}{h} = 0,\ a < 0.$$

Можно объединить форму записи приведенных выше выражений и записать схему как

$$u_m^{n+1} = u_m^n - \sigma \begin{cases} u_{m+1}^n - u_m^n, & a < 0 \\ u_m^n - u_{m-1}^n, & a \geqslant 0 \end{cases} \quad \sigma = a\frac{\tau}{h} - \text{число Куранта}.$$

Введя обозначения $a^+ = \frac{1}{2}(a + |a|), a^- = \frac{1}{2}(a - |a|)$, приведенную выше объединенную запись можно представить еще в двух формах:

$$u_m^{n+1} = u_m^n - \frac{\tau}{h}(a^-(u_{m+1}^n - u_m^n) + a^+(u_m^n - u_{m-1}^n)),$$

или

$$u_m^{n+1} = u_m^n - \frac{\sigma}{2}(u_{m+1}^n - u_{m-1}^n) + \frac{|\sigma|}{2}(u_{m+1}^n - 2u_m^n + u_{m-1}^n).$$

Представим разностную схему в потоковом виде, для чего введем функции $f_{m+1/2}^{n+1/2}, f_{m-1/2}^{n+1/2}$ так, что данные на следующем слое по времени можно представить в виде

$$u_m^{n+1} = u_m^n - \sigma(f_{m+1/2}^n - f_{m-1/2}^n).$$

Тогда простое сравнение выражений даст возможность записать потоки, для них легко получаются выражения

$$f_{m+1/2}^n = \frac{1}{2}((u_{m+1}^n + u_m^n) - \operatorname{sign} a(u_{m+1}^n - u_m^n)),$$
$$f_{m-1/2}^n = \frac{1}{2}((u_m^n + u_{m-1}^n) - \operatorname{sign} a(u_m^n - u_{m-1}^n)).$$

В [3] приведен общий вид неявной схемы, записанной в потоковом виде

$$u_m^{n+1} = u_m^n - \sigma(\bar{f}_{m+1/2} - \bar{f}_{m-1/2}),$$
$$\text{где } \bar{f}_{m\pm1/2} = (1-\theta)f_{m\pm1/2}^n + \theta f_{m\pm1/2}^{n+1}, 0 \leqslant \theta \leqslant 1,$$

выражение для $f_{m\pm1/2}^{n+1}$ аналогично выражению для $f_{m\pm1/2}^n$.

Введем также двухпараметрическое семейство разностных схем [4]

$$u_m^{n+1} = u_m^n + a\frac{\tau}{h}(\bar{f}_{m+1/2}^n - \bar{f}_{m-1/2}^n),$$

в которых потоки будут определяться в зависимости от направления переноса a. Они также будут зависеть от значений параметров:

$$\bar{f}_{m+1/2} = \begin{cases} \alpha u_{m-1}^n + (1-\alpha-\beta)\,u_m^n + \beta u_{m+1}^n, & a \geqslant 0, \\ \alpha u_{m+2}^n + (1-\alpha-\beta)\,u_{m+1}^n + \beta u_m^n, & a < 0, \end{cases}$$

$$\bar{f}_{m-1/2} = \begin{cases} \alpha u_{m-2}^n + (1-\alpha-\beta)\,u_{m-1}^n + \beta u_m^n, & a \geqslant 0, \\ \alpha u_{m+1}^n + (1-\alpha-\beta)\,u_m^n + \beta u_{m-1}^n, & a < 0. \end{cases}$$

15.2. Гибридные схемы

Для построения первой гибридной разностной схемы в работе [5] использовалось следующее представление ($a > 0$):

$$u_m^{n+1} = u_m^n - \frac{\tau a}{h}(u_m^n - u_{m-1}^n) + \frac{\gamma}{2}\frac{a\tau}{h}\left(1 - \frac{a\tau}{h}\right)(u_{m+1}^n - 2u_m^n + u_{m-1}^n) = 0,$$

где γ — параметр гибридности. От него зависит порядок разностной схемы, по которой будет производиться расчет в областях с большими локальными градиентами решения. В цитируемой работе [5]

$$\gamma = \begin{cases} 1, & \left|u^n_{m+1} - 2u^n_m - u^n_{m-1}\right| < \lambda \left|u^n_m - u^n_{m-1}\right|, \\ 0, & \left|u^n_{m+1} - 2u^n_m - u^n_{m-1}\right| \geqslant \lambda \left|u^n_m - u^n_{m-1}\right|. \end{cases} \tag{15.1}$$

При $\lambda = 0$ схема имеет первый порядок точности, при сколь угодно больших λ — второй. Можно повысить порядок аппроксимации этой схемы до третьего:

$$u^{n+1}_m = u^n_m - \frac{\tau a}{h}(u^n_m - u^n_{m-1}) +$$

$$+\frac{\gamma}{2}\frac{a\tau}{h}\left(1 - \frac{a\tau}{h}\right)\left\{(u^n_{m+1} - 2u^n_m + u^n_{m-1}) + (u^n_{m+1} - 3u^n_m + 3u^n_{m-1} - u^n_{m-2})\right\}.$$

В [21] схему, записанную в потоковой форме представления, предложено сделать гибридной, вводя потоки:

$$f^n_{m+1/2} = \frac{a}{2}\left\{(u^n_{m+1} - u^n_m) - \bar{\sigma}(u^n_{m+1} - u^n_m)\right\},$$

$$f^n_{m-1/2} = \frac{a}{2}\left\{(u^n_m - u^n_{m-1}) - \bar{\sigma}(u^n_m - u^n_{m-1})\right\},$$

при такой записи $\bar{\sigma} = a\frac{\tau}{h}$ в области гладкого решения и $\bar{\sigma} = \operatorname{sign} a$ в области с большими градиентами решения. Переключатель между «гладким» и «негладким» решениями может быть построен аналогично (15.1). Другие способы построения переключателей описаны в [6, 7]. В [6] построены *сеточно-характеристические* гибридные схемы.

15.3. Гибридные схемы и пространство неопределенных коэффициентов

Повышать точность метода также можно, если использовать разложение сеточной функции $u^{n+1}_{m+\nu}$ в ряд Тейлора в окрестности точки (t^n, x_m). В общем случае любая разностная схема представляется в виде суммы по точкам шаблона с неопределенными весовыми множителями [8]:

$$u^{n+1}_m = \sum_{\mu,\nu} \alpha^\nu_\mu u^{n+\nu}_{m+\mu}, \tag{15.2}$$

где $\nu = 1, 0, -1$ — номера слоев по времени, входящих в шаблон (шаблоны с более чем 3 слоями по времени рассматривать не будем), $\mu = 0, \pm 1, \pm 2, \ldots$ — пространственные узлы точек сеточного шаблона

$(t^{n+\nu}, x_{m+\mu}), \alpha_\mu^\nu$ — неопределенные коэффициенты. Если ν не принимает положительных значений, то схема явная, в противном случае — неявная. Если все α_μ^ν неотрицательные, то *схема монотонная* по Фридрихсу или схема *с положительной аппроксимацией.*

Учитывая продолжения исходного дифференциального уравнения

$$\frac{\partial^{k+l} u}{\partial t^k \partial x^l} = (-1)^k a^k \frac{\partial^{k+l} u}{\partial x^{k+l}},$$

полученные дифференцированием исходного однородного уравнения переноса по независимым переменным $k + l - 1$ раз, получим после подстановки разложения проекции точного решения уравнения переноса на сетку $u_{m+\mu}^{n+\nu}$ в разностную схему (15.2):

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = \frac{\delta_0}{\tau}u + \frac{h\delta_1}{\tau}\frac{\partial u}{\partial x} + \frac{h^2\delta_2}{2\tau}\frac{\partial^2 u}{\partial x^2} + \frac{h^2\delta_3}{3!\tau}\frac{\partial^3 u}{\partial x^3} + \frac{h^2\delta_4}{4!\tau}\frac{\partial^4 u}{\partial x^4} + \ldots, \tag{15.3}$$

откуда получим выражения для условий порядка, соблюдение которых необходимо для того, чтобы разностная схема аппроксимировала дифференциальную задачу

$$\delta_k = -(\sigma)^k + \sum_{\mu,\nu} (\mu - \nu\sigma)^k \alpha_\mu^\nu,$$

$\sigma = a\tau/h$ — число Куранта, $k = 0, 1, 2, \ldots$ — порядок аппроксимации, который может быть достигнут.

Из условий аппроксимации видно, что для получения первого порядка точности $O(\tau, h)$ необходимо и достаточно выполнения условий

$$\delta_0 = -1 + \sum_{\mu,\nu} \alpha_\mu^\nu = 0, \delta_1 = \sigma + \sum_{\mu,\nu} (\mu - \nu a\tau/h)\, \alpha_\mu^\nu = 0. \tag{15.4}$$

Для получения схем более высокого порядка аппроксимации необходимо использовать условия порядка с более высокими k.

Для построения разностных схем с заданными свойствами в [9] предложено ввести линейное пространство неопределеннных коэффициентов $\{\alpha_\mu^\nu\}$, в [10] на основе такого подхода предложена теория построения разностных схем повышенного порядка точности, в [6] – теория построения гибридных схем, наиболее близких в этом пространстве к монотонным по евклидовой норме.

Существует несколько определений монотонной схемы. Они, вообще говоря, не эквивалентны. Одно из определений приведено выше — это неотрицательность коэффициентов разностной схемы при записи в виде, разрешенном относительно точки (t^{n+1}, x_m). Монотонная схема по

Борису и Буку — схема, не увеличивающая число экстремумов в разностном решении задачи по сравнению с количеством экстремумов в точном решении задачи. Дадим еще одно определение монотонной схемы.

Определение. Схема называется монотонной, если из условия $u_{m+1}^n + u_m^n \geqslant 0$ следует $u_{m+1}^{n+1} + u_m^{n+1} \geqslant 0$ для всех m.

Для системы уравнений в частных производных гиперболического типа

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A}\frac{\partial \mathbf{u}}{\partial x} = 0,$$

где $\mathbf{A}$ — квадратная матрица с постоянными коэффициентами размера $n \times n$, $\mathbf{u} = (u_1, \ldots, u_n)^{\mathrm{T}}$ — вектор-столбец, разностную схему можно представить в виде [11] (см. также лекцию 14):

$$\mathbf{u}_m^{n+1} = \mathbf{u}_m^n - \frac{\tau}{2h}\mathbf{A}(\mathbf{u}_{m+1}^n - \mathbf{u}_{m-1}^n) + \frac{\tau}{2h}|\mathbf{A}|\,(\mathbf{u}_{m+1}^n - 2\mathbf{u}_m^n + \mathbf{u}_{m-1}^n),$$

или

$$\mathbf{u}_m^{n+1} = \mathbf{u}_m^n - \frac{\tau}{2h}\left\{\left(\mathbf{\Omega}^{-1}\mathbf{\Lambda}\mathbf{\Omega}\right)(\mathbf{u}_{m+1}^n - \mathbf{u}_{m-1}^n) + \right.$$
$$\left. + \left(\mathbf{\Omega}^{-1}|\mathbf{\Lambda}|\,\mathbf{\Omega}\right)(\mathbf{u}_{m+1}^n - 2\mathbf{u}_m^n + \mathbf{u}_{m-1}^n)\right\},$$

где $\mathbf{\Lambda} = \mathrm{diag}(\lambda_k)$ — диагональная матрица из собственных значений матрицы $\mathbf{A}$, $\mathbf{\Omega}$ — матрица размера $n \times n$, строками которой являются левые собственные векторы матрицы $\mathbf{A}$. Если матрица $\mathbf{A}$ невырожденная и все ее собственные числа действительны, то система имеет гиперболический тип, а матрица системы может быть представлена в виде произведения трех матриц $\mathbf{A} = \mathbf{\Omega}^{-1}\mathbf{\Lambda}\mathbf{\Omega}$. Матрица перехода в базис из собственных векторов в данном случае есть матрица из *левых* собственных векторов матрицы системы, так как матрица $\mathbf{A}$ несамосопряженная. Обратную матрицу к матрице перехода также надо вычислять непосредственно — как правило, в задачах механики сплошных сред матрица перехода не ортогональная, обратная матрица не равна транспонированной.

Эту же схему можно записать в виде

$$\mathbf{u}_m^{n+1} = \mathbf{u}_m^n - \frac{\tau}{h}\left\{\mathbf{A}^-(\mathbf{u}_{m+1}^n - \mathbf{u}_m^n) + \mathbf{A}^+(\mathbf{u}_m^n + \mathbf{u}_{m-1}^n)\right\},$$

где $\mathbf{A}^- = 0,5(\mathbf{A} + |\mathbf{A}|)$, $\mathbf{A}^+ = 0,5(\mathbf{A} - |\mathbf{A}|)$, или

$$\mathbf{u}_m^{n+1} = \mathbf{u}_m^n - \frac{\tau}{h}\left\{\left(\mathbf{\Omega}^{-1}\mathbf{\Lambda}^-\mathbf{\Omega}\right)(\mathbf{u}_{m+1}^n - \mathbf{u}_m^n) - \left(\mathbf{\Omega}^{-1}\mathbf{\Lambda}^+\mathbf{\Omega}\right)(\mathbf{u}_m^n + \mathbf{u}_{m-1}^n)\right\},$$

где $\mathbf{\Lambda}^+ = 0,5(\mathbf{\Lambda} - |\mathbf{\Lambda}|)$, $\mathbf{\Lambda}^- = 0,5(\mathbf{\Lambda} + |\mathbf{\Lambda}|)$,

Если по аналогии со скалярными потоками $f^{n+1/2}_{m+1/2}$, $f^{n+1/2}_{m-1/2}$ ввести векторные потоки $\mathbf{f}_{m+1/2}$, $\mathbf{f}_{m-1/2}$, то разностная схема (15.2) может быть представлена в виде:

$$\mathbf{u}^{n+1}_m = \mathbf{u}^n_m - \frac{\tau}{h}(\mathbf{f}_{m+1/2} - \mathbf{f}_{m-1/2}),$$

$$\text{где}\mathbf{f}_{m+1/2} = \frac{1}{2}\left\{\mathbf{A}(\mathbf{u}^n_{m+1} + \mathbf{u}^n_m) - |\mathbf{A}|\,(\mathbf{u}^n_{m+1} - \mathbf{u}^n_m)\right\} =$$

$$= \frac{1}{2}\left\{(\mathbf{\Omega}^{-1}\mathbf{\Lambda}\mathbf{\Omega})(\mathbf{u}^n_{m+1} + \mathbf{u}^n_m) - (\mathbf{\Omega}^{-1}\,|\mathbf{\Lambda}|\,\mathbf{\Omega})(\mathbf{u}^n_{m+1} - \mathbf{u}^n_m)\right\},$$

$$\mathbf{f}_{m-1/2} = \frac{1}{2}\left\{\mathbf{A}(\mathbf{u}^n_m - \mathbf{u}^n_{m-1}) - |\mathbf{A}|\,(\mathbf{u}^n_m - \mathbf{u}^n_{m-1})\right\} =$$

$$= \frac{1}{2}\left\{(\mathbf{\Omega}^{-1}\mathbf{\Lambda}\mathbf{\Omega})(\mathbf{u}^n_m - \mathbf{u}^n_{m-1}) - (\mathbf{\Omega}^{-1}\,|\mathbf{\Lambda}|\,\mathbf{\Omega})(\mathbf{u}^n_m - \mathbf{u}^n_{m-1})\right\}.$$

Для одномерной квазилинейной системы уравнений газовой динамики разностная схема выписана подробно в предыдущей лекции. В квазилинейном случае потоки $\mathbf{f}_{m+1}$, $\mathbf{f}_{m-1}$ можно представить в виде

$$\mathbf{f}_{m+1/2} = \frac{1}{2}\left\{\mathbf{A}^n_{m+1/2}(\mathbf{u}^n_{m+1} - \mathbf{u}^n_m) - \left|\mathbf{A}^n_{m+1/2}\right|(\mathbf{u}^n_{m+1} - \mathbf{u}^n_m)\right\},$$

$$\mathbf{f}_{m-1/2} = \frac{1}{2}\left\{\mathbf{A}^n_{m+1/2}(\mathbf{u}^n_m - \mathbf{u}^n_{m-1}) - \left|\mathbf{A}^n_{m-1/2}\right|(\mathbf{u}^n_m - \mathbf{u}^n_{m-1})\right\}.$$

Матрицы, входящие в приведенные выше формулы, выписаны в лекции 14.

15.4. Метод коррекции потоков Бориса–Бука

Метод коррекции потоков предложен в [12], как схема «предиктор-корректор». На этапе «предиктор» погрешность метода вносит в численное решение поток численной диффузии (вязкости), на этапе «корректор» водятся потоки искусственной антидиффузии, уменьшающие их. Пусть $\tilde{u}_m$ — численное решение, полученное после предиктора. Корректор представляется в виде

$$u^{n+1}_m = \tilde{u}_m - (\tilde{f}_{m+1/2} - \tilde{f}_{m-1/2}),$$

где определены антидиффузионные потоки $\tilde{f}_{m+1/2} = \mu(\tilde{u}_{m+1} - \tilde{u}_m)$, $\tilde{f}_{m-1/2} = \mu(\tilde{u}_m - \tilde{u}_{m-1})$ через границы $x_{m+1/2}, x_{m-1/2}$, (μ — коэффициент антидиффузии).

В [13] предложен общий вид корректора:

$$u_m^{n+1} = \tilde{u}_m + \mu(u_{m+1}^n - 2u_m^n + u_{m-1}^n) + \tilde{\mu}(\tilde{u}_{m+1} - 2\tilde{u}_m + \tilde{u}_{m-1}),$$

причем коэффициенты антидиффузии вычисляются с помощью подхода, предложенного в [6, 10].

15.5. TVD-схемы

Идею схем TVD (Total Variation Diminition), т. е. схем с уменьшением полной вариации, представим на примере схемы Лакса-Вендроффа [14]:

$$u_m^{n+1} = u_m^n - \frac{a\tau}{h}(u_m^n - u_{m-1}^n) - (f_{m+1/2}^n - f_{m-1/2}^n),$$

$$f_{m+1/2} = \frac{a\tau}{h}(1 - a\tau)(\bar{u}_{m+1} - \bar{u}_m), f_{m-1/2} = \frac{a\tau}{2h}(1 - a\tau)(\bar{u}_m - \bar{u}_{m-1}).$$

Эта схема немонотонная, но в отсутствии последнего слагаемого $(f_{m+1/2}^n - f_{m-1/2}^n)$ она была бы монотонной. Этот факт можно проинтерпретировать следующим образом: антидиффузионные потоки в схеме Лакса-Вендроффа слишком велики и приводят к появлению осцилляций. Следовательно, эти потоки необходимо ограничить, например, как

$$\bar{f}_{m+1/2} = \varphi(r_m)\frac{a\tau}{h}(1 - a\tau)(\tilde{u}_{m+1} - \tilde{u}_m).$$

Поток $f_{m+1/2}^n$ ограничивается некой функцией $\varphi(r_m)$, называемой ограничителем или лимитером. Параметр r_m вычисляется по формуле

$$r_m = \frac{u_m - u_{m-1}}{u_{m+1} - u_m},$$

его можно назвать показателем гладкости решения.

Для гладких решений $r_m \approx 1$, при больших же градиентах $r_m \approx 0$.

Функция $\varphi(r_m)$ выбирается так, чтобы схема относилась к классу TVD, т. е. чтобы уменьшалась полная вариация на следующем слое по времени, $\mathrm{TV}(u^{n+1}) \leqslant \mathrm{TV}(u^n)$. Выражение для полной вариации есть $\mathrm{TV}(u^n) = \sum_{m=-\infty}^{n=\infty} |u_{m+1}^n - u_m^n|$. Это условие более слабое, чем условие монотонности разностной схемы.

Для того чтобы полная вариация уменьшалась, достаточно выбрать лимитер следующим образом:

$$0 < \varphi(r_m) \leqslant \min(2r_m, 2), r_m > 0,$$
$$\varphi(r_m) = 0, r_m \leqslant 0,$$

причем для обеспечения второго порядка аппроксимации необходимо, чтобы $\varphi(1) = 1$.

Другой ограничитель имеет вид

$$\varphi(r_m) = \begin{cases} \min(2, r_m), r > 1, \\ \min(2r_m, 1), 0 < r \leqslant 1, \\ 0, r \leqslant 0. \end{cases}$$

Заметим, что вместо свободных параметров в этой схеме вводится функция-ограничитель, а сама схема является одношаговой. Иногда в расчетах полагают

$$r_m = \frac{u_m - u_{m-1} + \varepsilon}{u_{m+1} - u_m + \varepsilon},$$

где малая величина ε ($\varepsilon \approx 10^{-5} \div 10^{-10}$) играет роль шумового фильтра.

Вместо условия уменьшения полной вариации разностной схемы можно ввести более слабое ограничивающее условие $\mathrm{TVD}(u^n) \leqslant C$, причем $n\tau \leqslant C$ (схемы TVB).

В [15, 16] разностную схему для численного решения уравнения переноса предложено представить в виде

$$u_m^{n+1} = u_m^n - \frac{a\tau}{h}(u_m^n - u_{m-1}^n) + \frac{a\tau}{h}[\xi_{m+1/2}(u_{m+1}^n - u_m^n) - \xi_{m-1/2}(u_m^n - u_{m-1}^n)],$$

где $\xi_{m\pm1/2} \geqslant 0$; или

$$u_m^{n+1} = u_m^n + \frac{u_m^n - u_{m-1}^n}{h}\left(1 + \frac{\xi_{m+1/2}}{2} \cdot \frac{u_{m+1}^n - u_m^n}{u_m^n - u_{m-1}^n} - \frac{\xi_{m-1/2}}{2}\right) = 0, a > 0.$$

В соответствии с [16], эта схема будет монотонной, если выражение в скобках неотрицательно. Монотонность схемы может быть достигнута выбором коэффициента $\xi_{m\pm1/2}$, как функции от r. Как и ранее, $r_m = \frac{u_m - u_{m-1}}{u_{m+1} - u_m}$. Выбор весового множителя осуществим в соответствии с правилом

$$\xi(r_m) = \begin{cases} 0, & r_m \leqslant 0, \\ \frac{[c+b(1-\delta)]}{(c+b)(1-\delta)} r_m, & 0 < r_m < 1 - \delta, \\ \frac{c+br_m}{c+b}, & |r_m - 1| \leqslant \delta, \\ \frac{(c+b(1-\delta))-2c\delta}{(c+b)(1-\delta)} r_m, & 1 + \delta < r_{m<2}, \\ \leqslant 2, & r_m \geqslant 2. \end{cases}$$

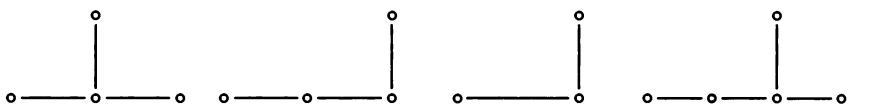
Здесь δ, c, b — константы, $0 < \delta < 1$; если $c + b \neq 0$ получаем схему второго порядка аппроксимации, причем, при $c = 1/3$ и $b = 2/3$ — третьего порядка везде, кроме точек разрыва функций.

Отметим, что, по-видимому, основные идеи, использованные при построении TVD и ENO схем, впервые были описаны В. П. Колганом в [17] и Р. П. Федоренко [5]. Схема с различными шаблонами, которую можно рассматривать как развитие идеи гибридных схем Р. П. Федоренко [5], предложенная Колганом, имеет вид

$$u_m^{n+1} = \begin{cases} u_m^n - \frac{a\tau}{2h}(\Delta + \Delta^-)u_m^n = 0, & |\Delta u_{m-2}| \geqslant |\Delta u_{m-1}| \geqslant |\Delta u_m|, \\ u_m^n - \frac{a\tau}{h}\Delta u_m^n - \frac{a\tau}{2}\frac{\Delta\Delta^-}{h^2}u_{m-1}^n, & |\Delta u_{m-2}| \leqslant |\Delta u_{m-1}| \leqslant |\Delta u_m|, \\ u_m^n - \frac{a\tau}{h}\Delta^- u_m^n, |\Delta u_{m-2}| \geqslant |\Delta u_{m-1}|, & |\Delta u_m| \geqslant |\Delta u_{m-1}|, \\ u_m^n - \frac{a\tau}{2h}(\Delta + \Delta^-)u_m^n - \frac{a\tau}{2}\frac{\Delta\Delta^-}{h^2}u_{m-1}^n, & \\ |\Delta u_{m-2}| \leqslant |\Delta u_{m-1}|, \quad |\Delta u_m| \leqslant |\Delta u_{m-1}| & \end{cases}$$

Здесь использованы обозначения $\Delta u_m = u_{m+1} - u_m$, $\Delta^- u_m = u_m - u_{m-1}$, $(\Delta + \Delta^-)u_m = u_{m+1} - u_{m-1}$, $(\Delta\Delta^-)u_m = \Delta(\Delta^-)u_m = u_{m+1} - 2u_m + u_{m-1}$.

Соответствующие шаблоны показаны на рисунках ниже.



Рассмотрим способ конструирования TVD-схемы. Произвольную четырехточечную схему (три точки на нижнем временном слое) можно представить в виде $u_m^{n+1} = u_m^n + C_{m+1/2}^+\Delta_{m+1/2}u - C_{m-1/2}^-\Delta_{m-1/2}u$, где введены потоки $\Delta_{m+1/2}u = u_{m+1} - u_m, \Delta_{m-1/2}u = u_m - u_{m-1}$. Положим, что коэффициенты схемы удовлетворяют условиям $C_{m+1/2}^+ \geqslant 0$, $C_{m-1/2}^- \geqslant 0$, для всех m. Тогда приведенная разностная схема является TVD-схемой. Покажем, что это так. Для этого вычислим полную вариацию

$$TV(u^n) = \sum_{m=-\infty}^{\infty} |u_{m+1}^n - u_m^n| = \sum_{m=-\infty}^{\infty} |\Delta_{m+1/2}u^n|.$$

Запишем разностную схему в операторном виде $u_m^{n+1} = \mathbf{L}u_m^n$, где $\mathbf{L}u_m^n = u_m^n + C_{m+1/2}^+\Delta_{m+1/2}u - C_{m-1/2}^-\Delta_{m-1/2}u$. Покажем, что $TV(u^{n+1}) \leqslant TV(u^n)$, или $TV(\mathbf{L}u^n) \leqslant TV(u^n)$. Оценим величину $\Delta_{m+1/2}u^{n+1}$, учитывая, что $u_{m+1}^{n+1} = u_{m+1}^n + C_{m+3/2}^+\Delta_{m+3/2}u - C_{m+1/2}^-\Delta_{m+1/2}u$; $u_m^{n+1} = u_m^n + C_{m+1/2}^+\Delta_{m+1/2}u - C_{m-1/2}^-\Delta_{m-1/2}u$. Тогда

$$|u_{m+1}^{n+1} - u_m^{n+1}| \leqslant |\Delta_{m+1/2}u^n|(1 - C_{m+1/2}^+ - C_{m-1/2}^-) + C_{m-1/2}^-|\Delta_{m-1/2}u| + C_{m+3/2}^+|\Delta_{m+3/2}u|,$$

откуда следует

$$TV(u^{n+1})_m = \sum_{-\infty}^{\infty} |\Delta_{m+1/2} u^{n+1}| \leqslant \sum_{-\infty}^{\infty} (1 - C^+_{m-1/2} - C^-_{m+1/2}) \, |\Delta_{m+1/2} u^n| +$$

$$+ \sum_{-\infty}^{\infty} C^-_{m-1/2} \, |\Delta_{m-1/2} u| + \sum_{-\infty}^{\infty} C^+_{m+3/2} \, |\Delta_{m+3/2} u| \leqslant \sum_{-\infty}^{\infty} |\Delta_{m+1/2} u^n| = TV(u^n).$$

15.6. ENO-схемы

Рассмотрим построение схемы типа ENO, предложенной в [4], для уравнения переноса. Основная идея этих схем заключается в использовании двух (или более) шаблонов для обеспечения двух традиционно противоположных свойств: второго (или третьего) порядка точности и монотонности без наличия аппроксимационной или искусственной вязкости. Обычно используются две базовые схемы: с центральными и односторонними разностями, а также с условием переключения, зависящим от значения производных (первой и второй) и знака скорости переноса.

Рассмотрим один из вариантов построения этого метода, для численного решения линейного уравнения переноса. Запишем схему в потоковой форме

$$u_1^{n+1} = u_m^n + \frac{a\tau}{h}(\bar{f}_{m+1/2} - \bar{f}_{m-1/2}),$$

а сами числовые потоки будем вычислять по формулам, в которые введем неопределенные пока коэффициенты:

$$\bar{f}_{m+1/2} = \begin{cases} \alpha u^n_{m-1} + (1-\alpha-\beta) u^n_m + \beta u^n_{m+1}, & a \geqslant 0 \\ \alpha u^n_{m+2} (1-\alpha-\beta) u^n_{m+1} + \beta u^n_m, & a < 0, \end{cases}$$

$$\bar{f}_{m-1/2} = \begin{cases} \alpha\, u^n_{m-2} + (1-\alpha-\beta)\, u^n_{m-1} + \beta\, u^n_m, & a \geqslant 0, \\ \alpha\, u^n_{m+1} + (1-\alpha-\beta)\, u^n_m + \beta\, u^n_{m-1}, & a < 0. \end{cases}$$

Выпишем *первое дифференциальное приближение (ПДП)* разностной схемы (лекция 13). Для этого выпишем проекцию точного решения уравнения переноса на введенную разностную сетку, разложим в ряды Тейлора в окрестности любой точки шаблона и приведем подобные слагаемые. В получившемся дифференциальном уравнении оставим главные члены невязки. Кроме того, как более информативную, рассматриваем П-форму, после всех вычислений получим

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = \left[\frac{h}{2}|a|(1 + 2\alpha - 2\beta) - \frac{\tau a^2}{2}\right] \frac{\partial^2 u}{\partial x^2}.$$

Из ПДП следует, что при $\alpha = \beta = 0$ получаем схему первого порядка, при $\alpha = 0, \beta = 1/2$ — второго порядка с центральными разностями, $\alpha = -\frac{1}{2}, \beta = 0$ — второго с ориентированными разностями.

Потребуем, чтобы коэффициент при второй производной в первом дифференциальном приближении схемы с центральными разностями обратился в нуль. Это означает, что в схеме отсутствует схемная (аппроксимационная) вязкость. Отсюда получим

$$\beta = 0,5\left(1 - \frac{|a|\,h}{\tau}\right).$$

Если ввести аналог числа Куранта $\sigma = \frac{|a|h}{\tau}$, то данное условие можно записать более компактно: $\beta = 0,5(1-\sigma)$.

Для схем с ориентированными разностями ($\beta = 0$) условие равенства нулю аппроксимационной вязкости дает

$$\alpha = -\frac{1 - |a/\tau|\,h}{2}.$$

Исследование обеих схем на монотонность, которое здесь опускается, приводит к условиям $\Delta_{m+1/2}u \geqslant \frac{1-\sigma}{2(2-\sigma)}\Delta_{m-1/2}u$ для схемы с центральными разностями, и $\Delta_{m+1/2}u \leqslant \frac{2(1+\sigma)}{\sigma}\Delta_{m-1/2}u$ для схемы с ориентированными разностями. Объединение этих условий дает условие монотонности обеих схем:

$$\frac{1-\sigma}{2(2-\sigma)}\Delta_{m-1/2}u \leqslant \Delta_{m+1/2}u \leqslant \frac{2(1+\sigma)}{\sigma}\Delta_{m-1/2}u.$$

Можно выделить монотонные разностные схемы без аппроксимационной вязкости, если реализовать переключение с одной схемы на другую.

Для такой монотонной схемы можно выписать окончательный вид потоков. Формулы для потоков выпишем через число Куранта.

$$\bar{f}_{m+1/2} = \begin{cases} \frac{3-\sigma}{2}u_m^n - \frac{1-\sigma}{2}u_{m-1}^n, a \geqslant 0 \\ \frac{3-\sigma}{2}u_{m+1}^n - \frac{1-\sigma}{2}u_{m+2}^n, a < 0 \end{cases}$$

если

$$(a \cdot \Delta u \cdot \Delta^2 u)_{m+1/2} \geqslant 0;$$

$$\bar{f}_{m+1/2} = \frac{1 - \sigma\mathrm{sign}a}{2}u_{m+1}^n + \frac{1+\sigma\mathrm{sign}a}{2}u_m^n,$$

если

$$(a \cdot \Delta u \cdot \Delta^2 u)_{m+1/2} < 0.$$

Здесь применяются обозначения $\Delta_{m+1/2}u = u_{m+1} - u_m$, $\Delta^2 u_{m+1/2} = \Delta u_{m+1} - \Delta u_m$.

15.7. Разностные схемы для квазилинейного уравнения переноса

Рассмотрим теперь нелинейное уравнение переноса, записанное в *дивергентной* форме

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0 \tag{15.5}$$

и соответствующей *характеристической* форме, которая легко получается из дивергентной:

$$\frac{\partial u}{\partial t} + a(u)\frac{\partial u}{\partial x} = 0, a(u) = \frac{\partial f}{\partial u}. \tag{15.6}$$

В дивергентной форме записи этого квазилинейного уравнения переноса величина f играет роль потока. Представим явную трехточечную разностную схему для численного решения (15.5) или (15.6) в потоковой форме

$$u_m^{n+1} = u_m^n - \bar{\sigma}(f_{m+1/2}^n - f_{m-1/2}^n), \tag{15.7}$$

где $\bar{\sigma} = \frac{\tau}{h}$,

$$f_{m+1/2}^n = \frac{1}{2}\left[f_m^n + f_{m+1}^n - \eta(a_{m+1/2}) \cdot \Delta u_{m+1/2}\right],$$

$$f_{m-1/2}^n = \frac{1}{2}\left[f_{m-1}^n + f_m^n - \eta(a_{m-1/2}) \cdot \Delta u_{m-1/2}\right],$$

$$\Delta u_{m+1/2} = u_{m+1} - u_m, \quad \Delta u_{m-1/2} = u_m - u_{m-1/2},$$

η — коэффициент при второй разности в разностной схеме, если $a = \mathrm{const}$.

В случае схемы Куранта–Изаксона–Риса (или схемы с разностями, ориентированными против потока — upwind scheme) имеем

$$f_{m+1/2} = \frac{1}{2}(f_{m+1} + f_m - |a_{m+1/2}|\,(u_{m+1} - u_m)),$$

$$f_{m-1/2} = \frac{1}{2}(f_m + f_{m-1} - |a_{m-1/2}|\,(u_m - u_{m-1})),$$

для сокращения записи введена скорость переноса, вычисляемая по формуле

$$\bar{a}_{m+1/2} = \begin{cases} \frac{f_{m+1} - f_m}{u_{m+1} - u_m} \approx \frac{\partial f}{\partial u}, (u_{m+1} - u_m) \neq 0, \\ a(u_m), (u_{m+1} - u_m) = 0. \end{cases}$$

В квазилинейном случае совпадавшие в линейном случае схемы будут приводить к разным разностным уравнениям, соответственно, будут различаться и численные решения из-за разности погрешностей аппроксимации. В случае квазилинейного уравнения для схемы Лакса–Вендроффа выражения для потоков будут иметь следующий вид:

$$f_{m+1/2} = \frac{1}{2}\left[f_{m+1} + f_m - \frac{\tau}{h}a^2_{m+1/2}(u_{m+1} - u_m)\right],$$
$$f_{m-1/2} = \frac{1}{2}\left[f_m + f_{m-1} - \frac{\tau}{h}a^2_{m-1/2}(u_m - u_{m-1})\right].$$

Для схем Куранта-Изаксона-Риса и Лакса-Вендроффа коэффициенты $\eta(a_{m\pm1/2})$ будут иметь значения $|a_{m\pm1/2}|$ и $\frac{\tau}{h}(a_{m\pm1/2})^2$ соответственно. Если, например, функция $\eta(x) = |x|$ при $|x| \approx \varepsilon$, где ε — малое число (порядка величины аппроксимационной вязкости для разностной схемы первого порядка Куранта-Изаксона-Риса), то полагают $\eta(x) = |x|$ при $|x| \geqslant \varepsilon, \eta(x) = \frac{x^2+\varepsilon^2}{2\varepsilon}$ при $|x| < \varepsilon$.

Введем обозначение $\gamma_{m+1/2} = 2\bar{\sigma}a_{m+1/2}$, тогда величины потоков $f_{m\pm1/2}$ можно переписать как

$$\bar{\sigma}f_{m+1/2} = \bar{\sigma}f_m - \frac{1}{2}\left[-\gamma_{m+1/2} + Q(\gamma_{m+1/2})\right]\Delta_{m+1/2}u,$$

$$\bar{\sigma}f_{m-1/2} = \bar{\sigma}f_m - \frac{1}{2}\left[\gamma_{m-1/2} + Q(\gamma_{m-1/2})\right]\Delta_{m-1/2}u.$$

Здесь введены обозначения $Q(\gamma_{m+1/2}) = 2\bar{\sigma}\eta(a_{m+1/2}), Q(\gamma_{m+1/2}) = 2\bar{\sigma}\eta(a_{m-1/2}), f_{m+1} = f_m + a_{m+1/2}\Delta_{m+1/2}u$. Подставляя эти выражения для численных потоков в (15.7), получаем

$$u^{n+1}_m = u^n_m + C^+_{m+1/2}\Delta_{m+1/2}\cdot u^n - C^-_{m-1/2}\Delta_{m-1/2}u^n, \qquad (15.8)$$

где $C^+_{m+1/2} = \frac{1}{2}\left[Q(\gamma_{m+1/2}) - \gamma_{m+1/2}\right], C^-_{m+1/2} = \frac{1}{2}\left[Q(\gamma_{m+1/2}) + \gamma_{m+1/2}\right]$, откуда следует

$$C^+_{m+1/2} + C^-_{m+1/2} = Q(\gamma_{m+1/2}).$$

В случае линейного уравнения переноса с постоянным коэффициентом (постоянной скоростью переноса) $Q_{m+1/2} = 2\sigma|a|, \gamma_{m+1/2} = 2a\sigma, \eta = |a|$. Условие устойчивости такой схемы имеет вид $\tau \leqslant \frac{h}{\max\limits_m |a^n_{m+1/2}|}$

Построенная разностная схема относится к классу TVD-схем, что непосредственно проверяется. Схема обеспечивает выполнение условия $TV(u^{n+1}) \leqslant TV(u^n)$ с учетом неравенств $C^+_{m+1/2} \geqslant 0$, $C^-_{m-1/2} \geqslant 0$, $C^+_{m+1/2} + C^-_{m+1/2} \leqslant 1$ для любых m.

15.8. Однопараметрическое семейство неявных схем

Рассмотрим однопараметрическое семейство неявных разностных схем для численного решения нелинейного уравнения переноса. Схемы, принадлежащие этому семейству, запишутся следующим образом:

$$u_m^{n+1} + \sigma\theta(f_{m+1/2}^{n+1} - f_{m-1/2}^{n+1}) = u_m^n - \sigma(1-\theta)(f_{m+1/2}^n - f_{m-1/2}^n),$$

весовой множитель меняется от нуля до единицы: $0 \leqslant \theta < 1$.

Потоковая форма записи этих квазилинейных уравнений будет

$$u_m^{n+1} = u_m^n - \sigma(\bar{f}_{m+1/2} - \bar{f}_{m-1/2}), \bar{f}_{m\pm1/2} = (1-\theta)f_{m+1/2}^n + \theta f_{m+1/2}^{n+1}.$$

Такая запись однопараметрического семейства схем включает в себя как явные (при $\theta = 0$), так и неявные (например, при $\theta = 1, \theta = 0,5$) разностные схемы.

Для вычисления числового потока будем использовать формулы

$$f_{m+1/2} = \frac{1}{2}(f_{m+1} + f_m + \varphi_{m+1/2}), f_{m-1/2} = \frac{1}{2}(f_m + f_{m-1} + \varphi_{m-1/2}),$$

где дополнительные слагаемые для вычисления потока в полуцелых точках есть

$$\varphi_{m+1/2} = -\Delta_{m+1/2}u\left[\eta(a_{m+1/2})(1-Q_{m+1/2}) + \sigma_{m-1/2}^2 Q_{m+1/2}\right],$$

$$\varphi_{m-1/2} = -\Delta_{m-1/2}u\left[\eta(a_{m-1/2})(1-Q_{m-1/2}) + \sigma_{m-1/2}^2 Q_{m-1/2}\right].$$

Подстановка выражений для потоков в исходную однопараметрическую разностную схему приводит к следующему выражению:

$$u_m^{n+1} + \frac{\sigma}{2}\theta\left\{f_{m+1} - \left[\eta(a_{m+1/2})(1-Q_{m+1/2}) + \sigma_{m+1/2}^2 Q_{m+1/2}\right]\Delta_{m+1/2}u\right\}^{n+1} -$$

$$-\frac{\sigma}{2}\theta\left\{f_{m-1} - \left[\eta(a_{m-1/2})(1-Q_{m-1/2}) + \sigma_{m-1/2}^2 Q_{m-1/2}\right]\Delta_{m-1/2}u\right\}^{n+1} = F_m^n,$$

или

$$u_m^{n+1} + \frac{\sigma}{2}\theta\left(\frac{f_{m+1} - f_m}{\Delta_{m+1/2}u} - \Psi_{m+1/2}\Delta_{m+1/2}u\right)^{n+1} -$$

$$-\frac{\sigma}{2}\theta\left(\frac{f_m - f_{m-1}}{\Delta_{m-1/2}u} - \Psi_{m-1/2}\Delta_{m-1/2}u\right)^{n+1} = F_m^n,$$

где F_m^n — величины, вычисляемые на n слое по времени, кроме того, в левой части прибавили и вычли величину f_m^{n+1}, введено обозначение $\psi_{m\pm1/2} = \eta(a_{m\pm1/2})(1 - Q_{m\pm1/2}) + \eta^2_{m\pm1/2}Q_{m\pm1/2}$.

При способе вычисления локальной скорости переноса в соответствии с правилами

$$a_{m+1/2} = \begin{cases} \frac{f_{m+1/2}-f_m}{\Delta_{m+1/2}u}, & \Delta_{m+1/2}u \neq 0 \\ a(u_m), & \Delta_{m+1/2}u = 0, \end{cases}$$

$$a_{m-1/2} = \begin{cases} \frac{f_m-f_{m-1}}{\Delta_{m-1/2}u}, & \Delta_{m-1/2}u \neq 0, \\ a(u_m), & \Delta_{m-1/2}u = 0, \end{cases}$$

получим разностную схему

$$u_m^{n+1} + \frac{\sigma}{2}\theta(a^{n+1}_{m+1/2} - \Psi^{n+1}_{m+1/2})\Delta_{m+1/2}u^{n+1} -$$

$$-\frac{\sigma}{2}\theta(-a^{n+1}_{m-1/2} - \Psi^{n+1}_{m-1/2})\Delta_{m-1/2}u^{n+1} = F_m^n,$$

или, в чуть сокращенной форме записи,

$$u_m^{n+1} - \sigma\theta(C^+_{m+1/2}\Delta_{m+1/2}u - C^-_{m-1/2}\Delta_{m-1/2}u)^{n+1} =$$

$$= u_m^n + \sigma(1-\theta)(C^+_{m+1/2}\Delta_{m+1/2}u - C^-_{m-1/2}\Delta_{m-1/2}u)^n$$

Алгоритм решения приведенного разностного уравнения — прогонка.

15.9. TVD-схемы для квазилинейного уравнения с антидиффузией.

Повысить порядок аппроксимации TVD-схем можно, введя в уравнение переноса антидиффузионный член

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(f + \frac{h}{\tau}\bar{f}\right) = 0,$$

где дополнительный антидиффузионный поток имеет порядок погрешности аппроксимации и вводится по формуле $\bar{f} = h\mu(\sigma, u)\frac{\partial u}{\partial x}$.

Очевидно, что это дополнительное слагаемое в исходном уравнении будет иметь вид

$$\frac{\partial}{\partial x}(h\mu\frac{\partial u}{\partial x}).$$

Таким образом, проведена коррекция потока, компенсирована аппроксиммационная вязкость. Хартен предложил модифицировать поток, введя в него функцию P:

$$\tilde{f}_{m+1/2} = \frac{1}{2}(f_m + f_{m+1} + P_{m+1/2}),$$

причем разные TVD-схемы различаются выбором этого слагаемого P. В исходном варианте разностной схемы, предложенной Хартеном, эта функция выбиралась следующим образом:

$$P_{m+1/2} = \frac{h}{\tau}(\bar{f}_m + \bar{f}_{m+1} - |\xi_{m+1/2} + \bar{\xi}_{m+1/2}|\,\Delta_{m+1/2}u),$$

где $\xi_{m+1/2} = \sigma a_{m+1/2}$, $\bar{f}_{m\pm1/2} = \frac{1}{2}\left(|\xi_{m+1/2}| - \xi^2_{m+1/2}\right)\Delta_{m+1/2}u$.

$$\xi_{m+1/2} = \begin{cases} \frac{\bar{f}_{m+1}-\bar{f}_m}{u_{m+1}-u_m}, & (u_{m+1}-u_m) \neq 0 \\ 0, & (u_{m+1}-u_m) = 0, \end{cases}$$

$$\bar{f}_{m\pm1/2} = \frac{1}{2}\left(|\xi_{m+1/2}| - \xi^2_{m+1/2}\right)\Delta_{m+1/2}u.$$

Эта схема, тем не менее, приводила к заметному размазыванию разрывов в решении.

Приведем также другие виды ограничительной функции. Общий вид записи лимитера будет один и тот же:

$$P_{m+1/2} = -\left[\sigma Q_{m+1/2} + \eta(a_{m+1/2})(\Delta_{m+1/2}u - Q_{m+1/2})\right],$$

а $Q_{m+1/2}$ может выбираться из следующих вариантов:

$$Q_{m+1/2} = \min\bmod(\Delta_{m+1/2}u, \Delta_{m-1/2}u) + \\ + \min\bmod(\Delta_{m+1/2}u, \Delta_{m+3/2}u) - \Delta_{m+1/2},$$

$$Q_{m+1/2} = \min\ \bmod\ (\Delta_{m-1/2}u,\ \Delta_{m+1/2}u, \Delta_{m+3/2}u),$$

$$Q_{m+1/2} = \min\ \bmod\ (2\Delta_{m-1/2}u, 2\Delta_{m+1/2}u), 2\Delta_{m+3/2}u, 0,5(\Delta_{m+3/2}u + \Delta_{m-1/2})).$$

Еще один вариант выбора лимитера в уравнении квазилинейного типа будет

$$P_{m+1/2} = -\left[\sigma(a^2_{m+1/2})\tilde{Q}_{m+1/2} + \eta(a_{m+1/2})(1 - \bar{Q}_{m+1/2})\right]\Delta_{m+1/2}u,$$

С согласованным выбором функции $\tilde{Q}_{m+1/2} = Q_{m+1/2} \cdot \Delta_{m+1/2}u$.

Введем обозначения показателей локальной гладкости решения

$$r^- = \frac{\Delta_{m-1/2}u}{\Delta_{m+1/2}u}, r^+ = \frac{\Delta_{m+3/2}u}{\Delta_{m+1/2}u},$$

и представим функции $Q(r^-, r^+)$, входящие в выражения для ограничения потока, в виде

$$Q(r^-, r^+) = \min \operatorname{mod}(1, r^-) + \min \operatorname{mod}(1, r^+) - 1,$$

$$Q(r^-, r^+) = \min \operatorname{mod}(r^-, r^+),$$

$$Q(r^-, r^+) = \min \operatorname{mod}(2, 2r^-, 2r^+, 0,5(r^- + r^+)).$$

Итак, напомним основные идеи, направленные на построение TVD-схем повышенного порядка точности:

— построение гибридных схем, аналогичных методу коррекции потоков [17];

— построение схем с модифицированным по Хартену потоком [18];

— построение схем, основанных на методе Годунова второго порядка аппроксимации [20, 21].

15.10. TVD-схемы для линейных систем уравнений гиперболического типа

Построим разностную схему типа TVD для случая одномерной системы линейных уравнений в частных производных гиперболического типа, к которым относятся, например, системы уравнений акустики или теории упругости.

Запишем систему как и ранее в матричной форме представления

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A}\frac{\partial \mathbf{u}}{\partial x} = 0,$$

где u — вектор-столбец искомых функций, $\mathbf{A}$ — квадратная матрица $n \times n$ с постоянными коэффициентами, t, x — независимые переменные. Пусть матрица $\mathbf{A}$ имеет n действительных собственных чисел λ_i и собственных векторов ω_i. Без ограничения общности можно считать, что среди этих собственных чисел нет кратных, а соответствующие собственные вектора образуют базис.

Тогда возможен переход к базису из собственных векторов, в котором матрица системы диагонализуется, сама система запишется как

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{\Omega}^{-1}\mathbf{\Lambda}\mathbf{\Omega}\frac{\partial \mathbf{u}}{\partial x} = 0,$$

где $\mathbf{\Lambda}$ — диагональная матрица из собственных чисел матрицы $\mathbf{A}$, $\mathbf{\Omega}$ — матрица, строками которой являются соответствующие левые собственные векторы. Последнее уравнение можно переписать следующим образом:

$$\frac{\partial \mathbf{R}}{\partial t} + \mathbf{\Lambda}\frac{\partial \mathbf{R}}{\partial x} = 0,$$

где $\mathbf{R} = \mathbf{\Omega u}$ — инварианты Римана.

Разностную схему для численного решения уравнения в инвариантах представим в виде:

$$\mathbf{R}_m^{n+1} = \mathbf{R}_m^n - \sigma(\tilde{\mathbf{f}}_{m+1/2} - \tilde{\mathbf{f}}_{m-1/2}),$$

$$\tilde{\mathbf{f}}_{m\pm1/2} = \frac{1}{2}(\mathbf{f}_m + \mathbf{f}_{m+1} + \mathbf{\Omega}_{m+1/2}\varphi_{m+1/2}).$$

Здесь $\sigma = \tau/h$, а $\varphi_{m+1/2}$ вычисляется аналогично скалярному случаю, с учетом того, что вместо $a_{m+1/2}$ необходимо брать $\lambda^i_{m+1/2}$, i — номер собственного значения, вместо $\Delta_{m+1/2}u$: $\mathbf{U}_{m+1/2} = \Omega^{-1}_{m+1/2}(\mathbf{u}_{m+1} - \mathbf{u}_m)$.

Явную схему TVD второго порядка точности типа предиктор-корректор для численного решения нелинейной системы уравнений гиперболического типа представим в виде:

$$\tilde{\mathbf{u}}_m = \mathbf{u}_m^n - \sigma(\mathbf{f}_m^n - \mathbf{f}_{m-1}^n),$$

$$\tilde{\tilde{\mathbf{u}}}_m = \frac{1}{2}(\tilde{\mathbf{u}}_m + \mathbf{u}_m^n) - \frac{\sigma}{2}(\tilde{\mathbf{f}}_{m+1} - \tilde{\mathbf{f}}_m),$$

$$\mathbf{u}_m^{n+1} = \tilde{\tilde{\mathbf{u}}}_m + \mathbf{\Omega}^n_{m+1/2}\varphi^n_{m+1/2} - \mathbf{\Omega}^n_{m-1/2}\varphi^n_{m-1/2},$$

где компоненты вектора $\varphi^i_{m+1/2} = -\frac{\eta(\lambda^i_{m+1/2}) + \gamma^2_{m+1/2}}{U^i_{m+1/2} - Q^i_{m+1/2}}$.

Функция, входящая в выражение для лимитера, вычисляется

$$Q_{m+1/2} = \begin{cases} \min\operatorname{mod}(\mathbf{U}_{m+1/2}, \mathbf{U}_{m-1/2}) + \\ + \min\operatorname{mod}(\mathbf{U}_{m+1/2}, \mathbf{U}_{m+3/2}) - \mathbf{U}_{m+1/2}, \\ \min\operatorname{mod}(\mathbf{U}_{m+1/2}, \mathbf{U}_{m-1/2}, \mathbf{U}_{m+3/2}), \\ \min\operatorname{mod}(2\mathbf{U}_{m-1/2}, 2\mathbf{U}_{m+1/2}, 2\mathbf{U}_{m+3/2}, 0,5(\mathbf{U}_{m-1/2} + \mathbf{U}_{m+3/2})). \end{cases}$$

15.11. Метод С. К. Годунова

Широко распространенный метод С. К. Годунова для получения численных решений с особенностями разрывного характера основан на решении задачи о распаде разрыва [2, 21–22]. В газовой динамике хорошо известно точное решение этой задачи и рассмотрены все возможные конфигурации решений. Положим, что начальные данные есть кусочно-постоянные функции на сетке $\{x_m\}_0^M$

$$\mathbf{u}(0,x) = \{\mathbf{u}^0_{m+1/2}\},$$

$$\mathbf{u}^n_{m+1/2} = \mathbf{u}[t_n, 0,5(x_m + x_{m+1})].$$

Например, для системы уравнений газовой динамики $\mathbf{u} = \{u, \rho, \varepsilon\}$.

Иногда начальные данные задают в ячейках с целыми номерами $\mathbf{u}(0,x) = \{\mathbf{u}^0_m\}$.

Решение задачи строится следующим образом. В окрестности каждого узла x_m (или $x_{m+1/2}$ при другой нумерации) решается задача о распаде разрыва независимо от других возмущений. Это решение используется до того момента, когда волна, образовавшаяся от разрыва в точке x_m, не встретится с волной, идущей от точки x_{m+1}. Далее полагаем, что и при $t = t_1 = \tau$ решение также приближается кусочно-постоянными функциями

$$\mathbf{u}^1_{m+1/2} = \frac{1}{h}\int\limits_{x_m}^{x_{m+1}} \mathbf{u}(t_1,x)dx,$$

$$h = x_{m+1} - x_m,$$

или, на сетке с отличной нумерацией узлов

$$\mathbf{u}^1_m = \frac{1}{h}\int\limits_{x_{m-1/2}}^{x_{m+1/2}} \mathbf{u}(t_1,x)dx,$$

$$h = x_{m+1/2} - x_{m-1/2}.$$

Представим систему дифференциальных уравнений в частных производных, записанную в дивергентном виде,

$$\frac{\partial \mathbf{S}}{\partial t} + \frac{\partial \mathbf{P}}{\partial x} = 0$$

в интегральной форме

$$\iint\limits_G \left(\frac{\partial \mathbf{S}}{\partial t} + \frac{\partial \mathbf{P}}{\partial x}\right)dtdx = \oint\limits_{\partial G} (\mathbf{S}dx - \mathbf{P}dt) = 0,$$

где G — некая односвязная область, ∂G — ограничивающий ее замкнутый контур, $\mathbf{S}, \mathbf{P}$ — функции от $\mathbf{u}$. Для этого выберем в качестве G ячейку $\{(t_n, t_{n+1}) \times (x_m, x_{m+1})\}$ и получим интегральное уравнение

$$\int_{x_m}^{x_{m+1}} \mathbf{S}(t_n, x)dx - \int_{t_n}^{t_{n+1}} \mathbf{P}(t, x_{m+1})dt - \int_{x_m}^{x_{m+1}} \mathbf{S}(t_{n+1}, x)dx + \int_{t_n}^{t_{n+1}} \mathbf{P}(t, x_m)dt = 0.$$

Первый и третий интегралы вычисляются просто по формуле средних — функции на отрезке $[x_m, x_{m+1}]$ кусочно-постоянны. Положив все функции кусочно-постоянными и на отрезке $[t_n, t_{n+1}]$ в силу автомодельности решения задачи Римана относительно переменных (x/t), получим равенство

$$\mathbf{S}^n_{m+1/2}h - \mathbf{P}^n_{m+1}\tau - \mathbf{S}^{n+1}_{m+1/2}h + \mathbf{P}^n_m\tau = 0,$$

или, разделив правую и левую части на произведение τh,

$$\frac{\mathbf{S}^{n+1}_{m+1/2} - \mathbf{S}^n_{m+1/2}}{\tau} + \frac{\mathbf{P}^n_{m+1} - \mathbf{P}^n_m}{h} = 0,$$

или $\frac{\mathbf{S}^{n+1}_m - \mathbf{S}^n_m}{\tau} + \frac{\mathbf{P}^n_{m+1/2} - \mathbf{P}^n_{m-1/2}}{h} = 0$ при другой индексации.

При этом потоки $\mathbf{P}^n_m$, $\mathbf{P}^n_{m+1}$ вычисляются при помощи решения задачи о распаде разрыва, которая сводится к решению системы нелинейных уравнений.

Повышение порядка аппроксимации схем типа Годунова, основанных на решении задачи распада разрыва (или солверов Римана) реализуется путем использования кусочно-линейной аппроксимации искомых величин внутри ячеек (в отличие от кусочно-постоянного их представления в методе Годунова) и различных алгоритмов пересчета по времени (алгоритмов типа предиктор-корректор) [19, 20].

Рассмотрим один из таких методов. Пусть внутри ячеек для всех сеточных функций заданы кусочно-линейные распределения

$$\mathbf{u}(t_n, x) = \mathbf{u}^n_m + \mathbf{Q}^n_m(x - x_m),$$

где $\mathbf{Q}^n_m$ — вектор наклонов функций u внутри ячейки. При этом изменение этой функции по времени внутри ячейки будет

$$\frac{\partial \mathbf{u}}{\partial t} = -\mathbf{A}\frac{\partial \mathbf{u}}{\partial x} = -\mathbf{A}\mathbf{Q}^n_m.$$

Предиктор (первый шаг) выглядит следующим образом:

$$\frac{\tilde{\mathbf{u}}_m - \mathbf{u}^n_m}{\tau} + \frac{\mathbf{f}(\mathbf{u}^n_m + \frac{h}{2}\mathbf{Q}^n_m) - \mathbf{f}(u^n_m - \frac{h}{2}\mathbf{Q}^n_m)}{h} = 0;$$

$\mathbf{u}_m^{n+1/2}$ на промежуточном слое $n+1/2$ вычисляем, как среднее арифметическое

$$\mathbf{u}_m^{n+1/2} = \tfrac{1}{2}(\tilde{\mathbf{u}}_m + \mathbf{u}_m^n).$$

На втором шаге — корректор — получим, используя метод конечных объемов

$$\frac{\mathbf{u}_m^{n+1} - \mathbf{u}_m^n}{\tau} + \frac{\mathbf{f}_{m+1/2} - \mathbf{f}_{m-1/2}}{h} = 0,$$

где, $\mathbf{f}_{m\pm1/2} = f(\mathbf{u}_{m\pm1/2})$, а функции $\mathbf{u}_{m\pm1/2}$ определяются из решения задачи Римана со следующими начальными данными:

$$\mathbf{u}_m^{n+1/2} + \frac{h}{2}\mathbf{Q}_m^n, x_{m+1/2} < 0, \mathbf{u}_{m+1}^{n+1/2} + \frac{h}{2}\mathbf{Q}_m^n, x_{m+1/2} > 0.$$

Эта схемы была получена в работах [27, 29]. В качестве начальных данных можно выбирать, не изменяя порядок точности, например, такие:

$$\mathbf{u}_m^n + \frac{h}{2}\mathbf{Q}_m^n, x_{m+1/2} < 0, \mathbf{u}_{m+1}^n + \frac{h}{2}\mathbf{Q}_{m+1}^n, x_{m+1/2} > 0$$

или $\mathbf{u}_m^n, x_{m+1/2} < 0, \mathbf{u}_{m+1}^n, x_{m+1/2} > 0$.

Простым способом вычисления наклона Q_m в ячейке с номером m для сеточной функции u_m, обеспечивающим устойчивость схемы, является использование функции $\min \bmod (y,z)$=$0,5(\mathrm{sign}y+\mathrm{sign}z)\min(|y|,|z|)$, которая выбирает наклон с минимальным значением модуля, при условии, что знаки обоих аргументов совпадают (при разных знаках аргументов эта функция равна нулю):

$$Q_m = \min \bmod\left(\frac{u_{m+1} - u_m}{h}, \frac{u_m - u_{m-1}}{h}\right).$$

Подобный алгоритм предиктор-корректор можно использовать и для системы уравнений, записанной в характеристической форме:

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{A}(\mathbf{u})\frac{\partial \mathbf{u}}{\partial x} = 0, \mathbf{A} = \mathbf{\Omega}^{-1}\mathbf{\Lambda}\mathbf{\Omega}.$$

При этом предиктор имеет вид

$$\frac{\tilde{\mathbf{u}}_m^{n+1} - \mathbf{u}_m^n}{h} + \mathbf{A}_m^n \cdot \mathbf{Q}_m^n = 0,$$

$$\mathbf{u}_m^{n+1/2} = \frac{\tilde{\mathbf{u}}_m^{n+1} + \mathbf{u}_m^n}{2} = \mathbf{u}_m^n - \frac{h}{2}\mathbf{A}_m^n\mathbf{Q}_m^n.$$

В качестве корректора используем, например, полученную ранее сеточно-характеристическую схему

$$\frac{\mathbf{u}_m^{n+1}-\mathbf{u}_m^n}{\tau} + (\mathbf{\Omega}^{-1}\mathbf{\Lambda}^-\mathbf{\Omega})_{m+1/2}^n \frac{\mathbf{u}_{m+1}^n-\mathbf{u}_m^n}{h} + (\mathbf{\Omega}^-\mathbf{\Lambda}^+\mathbf{\Omega})_{m-1/2}^n \frac{\mathbf{u}_m^n-\mathbf{u}_{m-1}^n}{h} = 0,$$

с учетом того, что начальные данные являются кусочно-постоянными:

$$\frac{u_m^{n+1}-u_m^n}{\tau}+(\Omega^{-1}\Lambda^-\Omega)_{m+1/2}^{n+1/2}\frac{u_{m+1}^{n+1/2}-\frac{h}{2}Q_{m+1}^n-u_m^{n+1/2}-\frac{h}{2}Q_m^n}{h}+$$
$$+(\Omega^{-1}\Lambda^+\Omega)_{m-1/2}^{n+1/2}\frac{u_m^{n+1/2}-\frac{h}{2}Q_{m-1}^{n+1/2}-u_{m-1}^{n+1/2}-\frac{h}{2}Q_{m-1}^n}{h}=0.$$

Литература

[1] *Courant T.R., Isacson E, Rees M.* On the solutions of nonlinear hyperbolic differential equations by finite differences. // Commun. Pure and Appl. Math. 1952. v. 5. № 5. PP. 243-254.

[2] *Годунов С.К., Забродин А.В., Прокопов Г.П.* Разностная схема для двумерных нестационарных задач газовой динамики и расчет обтекания с отошедшей ударной волной. // ЖВМиМФ. 1961. т. 1. № 6. С. 1020-1050.

[3] *Yee H.C.* Construction of Explicit and Implicit Symmetric TVD Schemes and Their Applications. // J. of Comp. Physics. 1987. Vol. 68. PP. 151-179.

[4] *Гущин В.А., Коньшин В.Н.* Численное моделирование волновых движений жидкости. Сообщения по прикладной математике. / Препринт ВЦ АН СССР. 1985. 36 с.

[5] *Федоренко Р.П.* Применение разностных схем высокой точности для численного решения гиперболических уравнений. // ЖВМиМФ. 1962. т. 2 № 6. С. 1122-1128.

[6] *Петров И.Б., Холодов А.С.* О регуляризации разрывных численных решений уравнений гиперболического типа. // ЖВМиМФ. 1984. т. 24. № 8. С. 1172-1188.

[7] *Leer B.Van.* Towards the ultimate conservative difference scheme. II. Monotonicity and conservation combined in a second-order scheme. // J. of Appl. Phys. 1974. v. 14. № 4. PP. 361-370.

[8] *Магомедов М.-К.М., Холодов А.С.* Сеточно-характеристические численные методы. М.: Наука, 1988. 288 с.

[9] *Холодов А.С.* О построении разностных схем с положительной аппроксимацией для уравнений гиперболического типа. // ЖВМиМФ. 1980. т. 20. № 6. С. 1601-1620.

[10] *Холодов А.С.* О построении разностных схем повышенного порядка точности для уравнений гиперболического типа. // ЖВМиМФ. 1980. т. 20. № 6. С. 1601-1620.

[11] *Магомедов К.М., Холодов А.С.* О построении разностных схем для уравнений гиперболического типа на основе характеристических соотношений // ЖВМиМФ 1969. т. 9. № 2. с. 373-386.

[12] *Boris J.P., Book D.L.* Flux-corrected transport. I. Shasta a fluid transport algorithm that works. // J. of C. Ph. 1973. Vol 11. № 1. PP. 38-69.

[13] *Воробьев О.В., Холодов А.С.* Об одном методе численного интегрирования одномерных задач газовой динамики. // Математическое моделирование. 1996. т. 8. № 1. С. 77-92.

[14] *Lax P.D., B.Wendroff* Difference schemes for hyperbolic equations with high orders of accuracy. // Comm. Pure. Appl. Math. 1964. v. 17. № 3. PP. 381-398.

[15] *Vyaznikov K.V., Tishkin V.F., Favorskii A.P.* One way to Construct Higher-Order Accurate Monotonic Difference Schemes for Systems of Hyperbolic Equations. // MMCE. 1994. v. 2. № 2. PP. 189-212.

[16] *Вязников К.В., Тишкин В.Ф., Фаворский А.П., Шашков М.Ю.* Квазимонотонные разностные схемы высокого порядка точности. / Препринт ИПМ АН СССР. 1987. № 36. 27 с.

[17] *Колган В.П.* Применение принципа минимальных значений производной к построению конечно-разностных схем для расчета разрывных решений газовой динамики. // Ученые записки ЦАГИ. 1972. т. 3. № 6. С. 68-77.

[18] *Harten A.J.* High resolution schemes for hyperbolic conservation laws. // J. Comput. Phys. 1983. v. 49. PP. 357-393.

[19] *Родионов А.В.* Повышение порядка аппроксимации схемы С.К. Годунова. // ЖВМиМФ. 1987. т. 27. № 12. С. 1853-1860.

[20] *Bovrel M., Montagne J.L.* Numerical study of a non-centered scheme with application to aerodynamics. // AIAA Paper. 1985. №. 85-1497. [Idem, in AIAA 7th Comput. Fluid Dyn. Conf. Cincinnati, Ohio, 1985, July 15-17. A Collect. Techn. Papers, 88-97, AIAA, New York].

[21] *Куликовский А.Г., Погорелов Н.В., Семенов А.Ю.* Математические вопросы численного решения гиперболических систем уравнений. М.: Физматлит, 2001. 608 с.

[22] *Рождественский Б.Л., Яненко Н.Н.* Системы квазилинейных уравнений и их приложения к газовой динамике. М.: Наука, 1978. 687 с.

[23] *Куропатенко В.Ф.* О разностных методах для уравнений гидродинамики. Тр. МИАН СССР, 1966. Т. 74. С. 107-137.

[24] *Родионов А.В.* Монотонная схема второго порядка аппроксимации для сквозного расчета неравновесных течений. // ЖВМиМФ. 1987. т. 27. № 4. С. 585-593.

[25] *Leer B.Van.* On the relation between the upwind-differencing schemes of Godunov, Engquist-Osher and Roe. // SIAM J. Sci. Stat. Comput. 1984. vol. 5. № 1, PP. 1-20.

[26] *Engqist B., Osher S.* One-sided difference approximations for nonlinear conservation laws. // Math. Comput. 1981. vol 36. № 154. PP. 321-351.

[27] *Osher S.* Numerical solution of singular perturbation problems and hyperbolic system of conservation laws. // In: North Holland Mathematical Studies. 1981. vol. 47. PP. 179-205.

[28] *Roe P.L.* The use of the Riemann problem in finite differences. / Lect. Notes Phys. 1981. Proc. 7th Int. Cont. Numer. Meth. Fluid Dynamics. June 23-27. 1980. vol. 141. PP. 354-359.

[29] *Roe P.L.* Approximate Riemann problem solvers, parameter vectors, and difference schemes. // J. Comput. Phys. vol. 43. № 2. PP. 357-372.

[30] *Miller G.N., Pucket E.G.* A high-order Godunov method for multiple condensed phases. // J. Comp. Phys. 1996. vol. 128. № 1. PP. 134-164.

[31] *Miller G.N., Colella P.* A high-order eulerian Godunov method for elastic-plastic flow in solids. // J. Comp. Phys. vol. 167. № 1. PP. 131-176.

[32] *Leer B.Van.* Towards the ultimate conservative difference scheme. V. A. second-order sequel to Godunov's method. // J. Comp. Phys. 1979. v. 32. № 1. PP. 101-136.

[33] *Меньшов И.С.* Повышение порядка аппроксимации схемы Годунова на основе решения обобщенной задачи Римана. // ЖВМиМФ. 1990. т. 30. № 9. С. 1357-1371.

[34] *Моисеев Н.Я.* Об одном способе повышения точности решений в разностных схемых, построенных на основе метода С.К. Годунова. // Вопросы атомной науки и техники. Сер.: Методики и программы численного решения задач математической физики. 1988. вып. 1. С. 38-45.

[35] *A. Harten* ENO Schemes with Subcell Resolution. // Journal of Computational Physics. 1989. v. 83. pp. 148-184.

[36] *Harten A.* Uniformly High Order Accurate Essentially Non-oscillatory Schemes. // J. of Comp. Ph. 1987. vol. 71. PP. 231-303.

[37] *Jee N.S.* Construction of explicit and implicit symmetric TVD schemes and their application. // J. Comp. 1987. v. 68. № 1. PP. 151-179.

[38] *Ершов С.В.* Монотонная ENO-схема повышенной точности для интегрирования уравнений Эйлера и Навье-Стокса. // Математическое моделирование. 1994. Т. 6. № 11. С. 63-75.

[39] *Ильин С.А., Тимофеев Е.В.* Сравнение квазимонотонных разностных схем сквозного счета на задаче Коши для одномерного линейного уравнения переноса. // Математическое моделирование. 1992. Т. 4. № 3. С. 62-75.

Лекция 16. Численное решение уравнений в частных производных эллиптического типа на примере уравнений Лапласа и Пуассона

В лекции разбираются постановка простейшей разностной задачи для уравнений Лапласа и Пуассона в прямоугольной области (схема «крест»). Дается обзор методов решения сеточных уравнений. Вкратце описываются идеи современных методов решения эллиптических уравнений в области произвольной геометрии — многосеточный метод и метод построения мажорантных разностных схем в пространстве неопределенных коэффициентов.

Ключевые слова: схема крест, разностный принцип максимума, методы решения сеточных уравнений, метод простых итераций, чебышевский набор параметров, методы Якоби, Зейделя, верхней релаксации, трехслойный итерационный метод, метод переменных направлений, коэффициенты Фурье, попеременно-треугольный итерационный метод, многосеточный метод Р. П. Федоренко, мажорантные разностные схемы.

Среди всех типов уравнений математической физики эллиптические уравнения с точки зрения вычислителей стоят особняком. С одной стороны, имеется хорошо развитая теория решения эллиптических уравнений и систем. Достаточно легко доказываются теоремы об устойчивости разностных схем для эллиптических уравнений. Во многих случаях получаются априорные оценки точности расчетов и числа итераций при решении возникающих систем сеточных уравнений. С другой стороны, системы сеточных уравнений, возникающие при решении уравнений методами сеток, имеют большую размерность и плохо обусловлены. Для решения таких систем разработаны специальные итерационные методы.

16.1. Постановка задачи. Простейшая разностная схема «крест». Устойчивость схемы «крест»

Будем рассматривать двухмерное уравнение Пуассона

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

в единичном квадрате $0 \leqslant x \leqslant 1, 0 \leqslant y \leqslant 1$ с краевыми условиями первого рода на границе расчетной области Г:

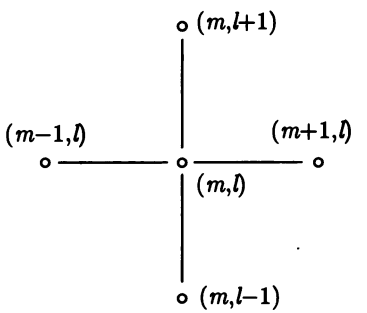
$$u_\Gamma = \varphi$$

(φ — заданная на границе функция).

В случае прямоугольной области граничные условия удобно записать в следующем виде:

$$u(0,y)=\varphi_1(y),$$
$$u(1,y)=\varphi_2(y),$$
$$u(x,0)=\varphi_3(x),$$
$$u(x,1)=\varphi_4(x).$$

Для простоты выкладок введем равномерную расчетную сетку с узлами $\{x_m, y_l\}$, $m,l=0,1,\ldots,M$ с равным количеством шагов по каждому пространственному направлению, сеточную область D — совокупность всех узлов сетки, включая граничные, и сеточную функцию u_{ml}. В этом случае шаги по координатам предполагаются равными. В случае неравных шагов по каждому направлению полученные результаты не изменятся, а запись уравнений станет более громоздкой.



Выбираем простейший пятиточечный шаблон разностной схемы «крест». На этом шаблоне аппроксимирующее разностное уравнение легко выписать. Для этого производные заменим вторыми разностями:

$$\frac{u_{m-1,l}-2u_{ml}+u_{m+1,l}}{h^2}+\frac{u_{m,l-1}-2u_{ml}+u_{m,l+1}}{h^2}=f_{ml},$$

где h — шаг по координатам, или в операторной форме

$$\Lambda_1 u_{ml}+\Lambda_2 u_{ml}=f_{ml},$$

здесь $\Lambda_1 u_{ml}=\frac{u_{m-1,l}-2u_{ml}+u_{m+1,l}}{h^2}$, $\Lambda_2 u_{ml}=\frac{u_{m/l-1}-2u_{ml}+u_{m,l+1}}{h^2}$,
$u_{0l}=\varphi_1(y_l), u_{m0}=\varphi_3(x_m), u_{ml}=\varphi_2(y_l), u_{mM}=\varphi_4(x_m)$.

Эту же разностную схему можно записать в каноническом виде для разностных схем для эллиптических уравнений:

$$u_{ml}=\frac{1}{4}(u_{m-1,l}+u_{m+1,l}+u_{m,l-1}+u_{m,l+1})+h^2 f_{ml}.$$

Такую каноническую запись не следует путать с канонической формой записи итерационного метода, которая встретится ниже.

Такая схема обладает вторым порядком аппроксимации по обеим координатам. Это легко показать, применяя разложение в ряд Тейлора функции — проекции точного решения на сетку — вплоть до членов четвертого порядка включительно. Проведем такое разложение для одного из операторов, стоящих в данном разностном уравнении:

$$\Lambda_1 u_m = \frac{u_{m-1} - 2u_m + u_{m+1}}{h^2} = \left(\frac{\partial^2 u}{\partial x^2}\right)_m + \frac{h^2}{12}\left(\frac{\partial^4 u}{\partial u^4}\right)_m + O(h^4).$$

Здесь учтено разложение проекции точного решения в ряд Тейлора

$$u_{m+1} = u(x_{m+1}) = u_m + h\left(\frac{\partial u}{\partial x}\right)_m + \frac{h^2}{2!}\left(\frac{\partial^2 u}{\partial x^2}\right)_m +$$

$$+\frac{h^3}{3!}\left(\frac{\partial^3 u}{\partial x^3}\right)_m + \frac{h^4}{4!}\left(\frac{\partial^4 u}{\partial x^4}\right)_m + \frac{h^5}{5!}\left(\frac{\partial^5 u}{\partial x^5}\right)_m + O(h^6)$$

и аналогичное разложение для u_{m-1}.

Для рассматриваемого двухмерного уравнения получим выражение для главного члена невязки

$$\Lambda_1 u_{ml} + \Lambda_2 u_{ml} = \left[\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right]_{ml} + \frac{h^2}{12}\left[\frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4}\right]_{ml} + O(h^4).$$

Рассмотрим устойчивость полученной схемы. Отметим, что методы исследования на устойчивость, применяемые для эволюционных (зависящих от времени) уравнений, здесь не работают. Действовать приходится на основе определения устойчивости.

Сформулируем и докажем две леммы, которые облегчат процедуру доказательства устойчивости разностной схемы.

Лемма 1. *Пусть сеточная функция u_{ml} определена на сетке $(x_m, y_l)(m, l = 0, \ldots, M)$ и во всех внутренних узлах сетки удовлетворяет уравнению $\frac{u_{m-1,l} - 2u_{ml} + u_{m+1,l}}{h^2} + \frac{u_{m,l-1} - 2u_{ml} + u_{m,l+1}}{h^2} = f_{ml}$, причем во всех внутренних узлах сетки правая часть этого уравнения неотрицательная: $f_{ml} \geqslant 0$. Тогда наибольшее значение сеточная функция u_{ml} достигает хотя бы в одной точке границы сеточной области.*

Доказательство.

Предположим, что существует такая внутренняя точка области, в которой сеточная функция принимает наибольшее значение. Пусть это точка с индексами i, j. Перепишем разностное уравнение, несколько перегруппировав члены в левой части:

$$\frac{1}{h^2}\left((u_{i-1,j} - u_{ij}) + (u_{i+1,j} - u_{ij}) + (u_{i,j-1} - u_{ij}) + (u_{i,j+1} - u_{ij})\right) = f_{ij}.$$

В силу сделанного предположения, хотя бы одна из скобок в левой части отрицательна, а остальные неположительны. Тогда в левой части стоит отрицательное число, в то время как в правой — положительное (по условиям леммы). Получившееся противоречие возникло из-за предположения о том, что максимальное значение сеточная функция принимает во внутренней точке сеточной области. Лемма доказана. ■

Лемма 2. *Пусть сеточная функция* u_{ml} *определена на сетке* (x_m, y_l) $(m, l = 0, \ldots, M)$ *и во всех внутренних узлах сетки удовлетворяет уравнению* $\frac{u_{m-1,l}-2u_{ml}+u_{m+1,l}}{h^2} + \frac{u_{m,l-1}-2u_{ml}+u_{m,l+1}}{h^2} = f_{ml}$, *причем во всех внутренних узлах сетки правая часть этого уравнения неположительная:* $f_{ml} \leqslant 0$. *Тогда наименьшее значение сеточная функция* u_{ml} *достигает хотя бы в одной точке границы сеточной области.*

Доказательство от противного практически повторяет доказательство предыдущей леммы.

Лемма 3 (сеточный принцип максимума). *Каждое решение разностного уравнения Лапласа* $\frac{u_{m-1,l}-2u_{ml}+u_{m+1,l}}{h^2} + \frac{u_{m,l-1}-2u_{ml}+u_{m,l+1}}{h^2} = 0$ *достигает своего минимального и максимального значения на границе сеточной области.*

Доказательство очевидно — это объединение утверждений леммы 1 и леммы 2.

Введем норму сеточной функции как

$$\|u_{ml}\| = \max_{m,l \in D} |u_{ml}| .$$

Для доказательства устойчивости теперь надо доказать однозначную разрешимость разностной задачи для уравнения Пуассона с любой правой частью и любыми граничными условиями, получить оценку

$$\|u_{ml}\| \leqslant C \|f\| .$$

Здесь в правой части стоит норма правой части задачи, записанной в операторном виде, $\|f\| = \max\limits_{m,l \in D} |f_{ml}| + \max\limits_{m,l \in \partial D} |\varphi_{ml}|$. Первый максимум в этой сумме берется по всем внутренним точкам, второй — по всем точкам сеточной границы.

Докажем однозначную разрешимость разностной задачи. Рассмотрим сеточное уравнение Лапласа с нулевыми граничными условиями. В силу принципа максимума такая задача имеет лишь тривиальное решение. Но сеточная система — это система линейных уравнений. Если система с нулевой правой частью имеет лишь тривиальное решение, то она однозначно разрешима при любой правой части.

Заметим, что в точной арифметике действие разностного оператора, приближающего дифференциальный оператор Лапласа, на произвольный полином второй степени совпадает по результату с действием дифференциального оператора — погрешность аппроксимации, как следует из приведенной выше оценки, будет нулевая. Рассмотрим вспомогательную функцию

$$P^h = \frac{1}{4}\left[R^2 - (x^2+y^2)\right] \max_{m,l\in D} |f_{ml}| + \max_{m,l\in \partial D} |\varphi_{ml}|,$$

где R — радиус окружности с центром в точке $(0,0)$ и включающей в себя рассматриваемую область. В данном случае $R=\sqrt{2}$. Эта функция иногда называется мажорантой Гершгорина.

Индекс h означает, что рассматривается сеточная проекция мажоранты. Обратимся к сеточной функции $w_{ml} = u_{ml} - P^h_{ml}$ и применим к ней разностный оператор Лапласа. Получим $\Lambda_1 w_{ml} + \Lambda_2 w_{ml} = f_{ml} + \max\limits_{m,l\in D} |f_{ml}|$ во всех внутренних точках области. Отсюда следует, что свое наибольшее значение рассматриваемая сеточная функция достигает на границе сеточной области в соответствии с доказанной леммой 1. Но, как легко убедиться, на границе области сеточная функция $w_{ml} = u_{ml} - P^h_{ml}$ принимает только отрицательные значения. Тогда $u_{ml} - P^h_{ml} \leqslant 0$ во всех точках сеточной области, включая граничные. Рассмотрим сеточную функцию $v_{ml} = u_{ml} + P^h_{ml}$. Проведя такие же рассуждения, придем к неравенству $u_{ml} + P^h_{ml} \geqslant 0$ во всех точках сеточной области, включая граничные.

Объединяя полученные результаты, находим

$$|u_{ml}| \leqslant \left|P^h_{ml}\right| \leqslant \frac{1}{4}R^2 \max_{m,l\in D} |f_{ml}| + \max_{m,l\in \partial D} |\varphi_{ml}|,$$

откуда следует неравенство $\|u_{ml}\| \leqslant \max(1, R^2/4)\,\|f\|$. Таким образом, устойчивость самой разностной схемы доказана.

16.2. Методы решения сеточных уравнений

При решении эллиптических систем существенная часть вычислительной работы — решение возникающих сеточных уравнений. Фактически для нахождения сеточной функции решения надо получить решение системы линейных уравнений большой размерности с разреженной матрицей специального вида. При аппроксимации уравнения Лапласа или Пуассона на регулярных сетках матрица системы самосопряженная. Методы решения таких систем рассмотрим ниже. Перед чтением данного раздела рекомендуется просмотреть лекцию 2.

16.2.1. Метод простых итераций

Наиболее эффективные алгоритмы для численного решения полученной СЛАУ — итерационные. Действительно, прямые методы требуют вычисления обратной матрицы, обратная матрица получается заполненной. Пусть u^0_{ml} — начальное приближение, выбирать которое желательно как можно ближе к искомому решению, $u^1_{ml}, u^2_{ml}, \ldots$ — последующие приближения. Верхний индекс в данных обозначениях указывает номер итерации.

Если выполняется условие

$$\left\|u^i_{ml} - u_{ml}\right\| \to 0$$

где u_{ml} — проекция точного решения на сетку, то итерационный метод является сходящимся. Оценка сходимости при этом может быть получена в виде

$$\left\|u^{i+1}_{ml} - u^i_{ml}\right\| \leqslant cq^i,$$

где c — константа, $0 < q < 1$. Итерации продолжаются до тех пор, пока не выполнено условие $\left\|u^{i+1}_{ml} - u^i_{ml}\right\| \leqslant \varepsilon$, где ε — заданная точность. В таком случае можно оценить количество итераций, необходимое для достижения этой точности $i \approx [\ln(\varepsilon/c)/\ln q] + 1$. Квадратные скобки в этой записи — операция взятия целой части числа.

Метод простых итераций записывается для системы сеточных уравнений в следующем виде: $u^{i+1}_{ml} = u^i_{ml} + \tau(\Lambda u^i_{ml} - f_{ml})$, если точка принадлежит внутренней части сеточной области, и $u^i_{ml} = \varphi_{ml}$ если точка с индексами ml принадлежит границе сеточной области. Здесь $\Lambda = \Lambda_1 + \Lambda_2$, τ — итерационный параметр. Количество арифметических операций при реализации метода простых итераций $\sim O(N^2)$.

Получим формулу для эволюции погрешности. Вычтем из итерационной формулы $\mathbf{u}^{i+1}_{ml} = \mathbf{u}^i_{ml} + \tau(\Lambda\mathbf{u}^i - \mathbf{f})_{ml}$ очевидное тождество $\mathbf{u}_{ml} = \mathbf{u}_{ml} + \tau(\Lambda\mathbf{u} - \mathbf{f})_{ml}$ во внутренних точках, а из равенства $\mathbf{u}^{i+1}_{ml} = \mathbf{u}^i_{ml}$ вычтем тождество $\mathbf{u}_{ml} = \mathbf{u}_{ml}$ в граничных узлах. Тогда получим $\mathbf{r}^{i+1}_{ml} = \mathbf{r}^i_{ml} + \tau\Lambda\mathbf{r}^i_{ml}, \mathbf{r}^{i+1}_{ml} = 0$, во внутренних и граничных узлах соответственно. Здесь $\mathbf{r}^i_{ml} = \mathbf{u}^i_{ml} - \mathbf{u}_{ml}$ — невязка на i итерации.

В таком случае для сеточной функции невязки r_{ml}, равной нулю на границе, ее эволюция описывается уравнением

$$r^{i+1}_{ml} = (\mathbf{E} + \tau\Lambda)r^i_{ml}.$$

Для оценки сходимости итерационного процесса необходимо перейти к неравенству для норм, например евклидовых, и оценить норму оператора перехода

$$\left\|\mathbf{r}^{i+1}_{ml}\right\| \leqslant \left\|\mathbf{E} + \tau\Lambda\right\| \left\|\mathbf{r}^i_{ml}\right\|,$$

откуда получим

$$\left\|\mathbf{r}_{ml}^{i}\right\| \leqslant \|\mathbf{E}+\tau\Lambda\|^{i}\left\|\mathbf{r}_{ml}^{0}\right\|.$$

Наиболее простым для этих целей является метод Фурье (или спектральный анализ оператора перехода).

Непосредственной проверкой доказываются два следующих утверждения. Семейство функций $\varphi_m^p = \sin\left(\frac{mp\pi}{M}\right)$ являются собственными функциями оператора Λ_1; им соответствуют собственные значения $\lambda^p = \frac{4}{h^2}\sin^2\left(\frac{p\pi}{2M}\right)$.

Здесь p — номер собственного значения (собственной функции), m — номер сеточного узла; $p = 1, \ldots, M-1$.

Для этого необходимо непосредственной подстановкой убедиться в истинности равенства

$$\Lambda_1\varphi_m^p = -\lambda^p\varphi_m^p,$$

или

$$\frac{1}{h^2}\left[\sin\frac{(m-1)p\pi}{M} - 2\sin\frac{mp\pi}{M} + \sin\frac{(m+1)p\pi}{M}\right] = \left(-\frac{4}{h^2}\sin^2\frac{p\pi}{2M}\right)\sin\left(\frac{mp\pi}{M}\right).$$

Семейство функций $\Psi_{ml}^{pq} = \varphi_m^p\varphi_l^q$ является собственными функциями оператора Λ, соответствующие собственные значения есть $\lambda^{pq} = \frac{4}{h^2}\left(\sin^2\frac{p\pi}{2M} + \sin^2\frac{q\pi}{2M}\right)$.

Равенство $\Lambda\Psi_{ml}^{pq} = -\lambda^{pq}\Psi_{ml}^{pq}$ также проверяется непосредственной подстановкой.

В дальнейшем необходимо будет знать границы спектра $\lambda_{\min} = \lambda^1 = l = \frac{4}{h^2}\sin^2\frac{\pi}{2M} \approx \pi^2$, так как $M \gg 1, h = 1/M$,

$$\lambda_{\max} = \lambda^{(M-1)} = L = \frac{4}{h^2}\sin^2\frac{(M-1)\pi}{2M} \approx \frac{4}{h^2}\sin^2\frac{\pi}{2} = \frac{4}{h^2} = 4M^2.$$

Для λ^{pq} имеем $l \leqslant \lambda^{pq} \leqslant L$, где $l \approx 2\pi^2, L \approx 8/h^2 = 8M^2$. Легко также оценить число обусловленности системы сеточных уравнений.

16.2.2. Метод простых итераций с оптимальным параметром

Представим сеточную функцию невязки r_{ml}^0, равную нулю на границе, в виде разложения по базису из собственных функций разностного оператора (Ψ_{ml}^{pq} — собственные функции оператора Λ)

$$r_{ml}^0 = \sum_{pq} c_{pq}\Psi_{ml}^{pq},\ c_{pq} = (r_{ml}^0, \Psi_{ml}^{pq}),$$

при этом выполняется равенство Парсеваля

$$\left\|r_{ml}^0\right\| = (r^0, r^0) = \sqrt{\sum_{pq} c_{pq}^2} = \|c\|.$$

Далее, используя это разложение, получим

$$r^1 = (\mathbf{E} + \tau\Lambda)r^0 = (\mathbf{E} + \tau\Lambda)\sum_{pq} c_{pq}\Psi^{pq} =$$

$$= \sum_{pq} c_{pq}(\mathbf{E} + \tau\Lambda)\Psi^{pq} = \sum_{pq} c_{pq}(1 - \tau\lambda^{pq})\Psi^{pq}.$$

Здесь используется то, что $(1-\tau\lambda^{pq})$ являются собственными числами оператора $(\mathbf{E}+\tau\Lambda)$. При сложении равенств $\tau\Lambda\varphi^p_m = -\tau\lambda\varphi^p_m$ и $\mathbf{E}\varphi^p_m = \varphi^p_m$ получаем $(\mathbf{E}+\tau\Lambda)\varphi^p_m = (1-\tau\lambda)\varphi^p_m$. Легко получается оценка нормы этой сеточной функции на первой итерации:

$$\left\|v^1\right\| = \sqrt{\sum c^2_{pq}(1-\tau\lambda^{pq})^2} \leqslant \max_{\lambda\in[l,L]}|1-\tau\lambda|\sqrt{\sum c^2_{pq}} = \max_{\lambda\in[l,L]}|1-\tau\lambda|\left\|\mathbf{r}^0\right\|.$$

Для последовательности итераций также легко получается стандартная оценка нормы

$$\left\|\mathbf{r}^i\right\| \leqslant (\max_{\lambda\in[l,L]}|1-\tau\lambda|)^i\left\|\mathbf{r}^0\right\|.$$

Отсюда видно, что значение q вычисляется, как $\max\{|1-\tau l|, |1-\tau L|\}$, а условие сходимости $q < 1$ выполняется при $0 < \tau < 2/L$. Границы спектра разностного оператора уже оценены в предыдущем пункте.

Для определения параметра τ, обеспечивающего максимальную скорость сходимости, необходимо решать следующую оптимизационную задачу:

$$\min_{\tau}(\max_{\lambda\in[l,L]}|1-\tau\lambda|).$$

Так как $\max|1-\tau\lambda|$ достигается на правой или левой границе интервала [l, L], то выполняется равенство $\max\limits_{\lambda\in[l,L]}|1-\tau\lambda| = \max\{|1-\tau l|, |1-\tau L|\}$. В таком случае необходимо определить τ, при котором достигается $\min\limits_{\tau}\{\max(|1-\tau l|, |1-\tau L|)\}$, или $\tau_0 = \arg[\min\limits_{\tau}\max(|1-\tau l|, |1-\tau L|)]$. где τ_0 — оптимальный итерационный параметр.

Как показано на рис. 16.1, $\min\limits_{\tau}\max(|1-\tau l|, |1-\tau L|)$ достигается при $|1-\tau l| = |1-\tau L|$. Справа от точки B при любых τ максимальна функция $|1-\tau L|$, слева — функция $|1-\tau l|$, и тогда минимум от искомого максимума достигается в точке B. Отсюда получим $1-\tau l = -(1-\tau L)$. Следовательно, значение оптимального итерационного параметра τ_0 равно $\tau_0 = \frac{2}{l+L}$.

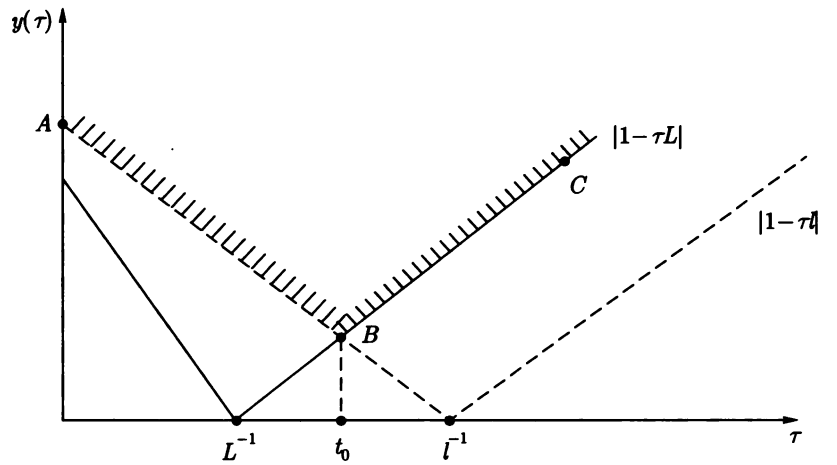


Рис. 16.1

Оптимальное значение функции, отвечающей за скорость сходимости, будет

$$q_0 = q(\tau_0) = 1 - \tau_0 l = 1 - \frac{2}{L+l} \cdot l = \frac{L-l}{L+l} = \\ = \frac{1-\mu^{-1}}{1+\mu^{-1}} = \frac{1+\mu^{-1}-2\mu^{-1}}{1+\mu^{-1}} \approx 1 - 2\mu^{-1},$$

где $\mu = L/l$ — число обусловленности системы сеточных уравнений.

Количество итераций, соответствующее этому методу, легко оценивается:

$$i = \left[\frac{\ln\varepsilon}{\ln q}\right] + 1 = \left[\frac{\ln\varepsilon}{\ln(1-2l/L)}\right] + 1 \approx \left[\frac{\ln\varepsilon}{(-2l/L)} + 1 \approx \frac{L}{2l}\ln\varepsilon^{-1} + 1\right].$$

Пусть расчеты приводятся с точностью $\varepsilon = 10^{-5}$ на сетке 100×100, тогда оценка числа итераций дает при $l \approx 2\pi^2$ и $L = 8N^2$ $i \approx \frac{8N^2}{2\cdot 2\pi^2}\ln 10^5 \approx$ $\approx 2 \cdot 10^4$ итераций.

Показатель сходимости $q_0 = 1 - 2\mu^{-1} \approx 0,9995$.

Параметр $\mu = L/l$ — число обусловленности матрицы (лекция 2); чем оно больше, тем медленнее сходятся итерации. Напомним (лекция 2), что в n-мерном линейном нормированном пространстве L_n вводятся три наиболее употребительных нормы вектора:

$$\|\mathbf{u}\|_1 = \max_i |\mathbf{u}_i|, i = 1, \ldots, n,$$

$$\|\mathbf{u}\|_2 = \sum_{i=1}^{N} |\mathbf{u}_i|,$$

$$\|\mathbf{u}\|_3 = \sqrt{\sum_{i=1}^{n} |\mathbf{u}_i|^2} = \sqrt{(\mathbf{u}, \mathbf{u})},$$

которым, в соответствии с определением согласованной нормы матрицы $\mathbf{A}$

$$\|A\| = \sup_{u \in} \frac{\|A\mathbf{u}\|}{\|\mathbf{u}\|},$$

сопоставляются нормы матрицы $\mathbf{A}$ с элементами a_{ij}:

$$\|\mathbf{A}\|_1 = \max_i \sum_{j=1}^{n} |a_{ij}|,$$

$$\|\mathbf{A}\|_2 = \max_j \sum_{i=1}^{n} |a_{ij}|,$$

$$\|\mathbf{A}\|_3 = \sqrt{\max_i \lambda^i(\mathbf{A} * \mathbf{A})}.$$

Показывается, что для симметричной матрицы $\mathbf{A}$ число обусловленности μ может быть представлено в третьей норме:

$$\mu = \frac{\max_i |\lambda^i_{\mathbf{A}}|}{\min_i |\lambda^i_{\mathbf{A}}|} = \frac{L}{l}.$$

Таким образом, доказана следующая теорема.

Теорема. *Рассмотрим итерационный метод* $u^{i+1}_{ml} = u^i_{ml} + \tau(\Lambda u^i_{ml} - f_{ml})$, *с оператором* $\Lambda = \Lambda^* > 0$ *для численного решения разностного аналога уравнения Пуассона* $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y)$ *с помощью аппроксимирующей его разностной схемы*

$$\Lambda u_{ml} = f_{ml}.$$

Пусть l *и* L *— минимальное и максимальное собственные числа оператора* Λ *соответственно.*

Если итерационный параметр τ *удовлетворяет условию* $0 < \tau < 2/L$, *то последовательность итераций* v^i *сходится к проекции решения исходного дифференциального уравнения, причем выполнено неравенство* $\|v^i\| \leqslant q^i \|v^0\|$, *где параметр* $0 < q < 1$ *определяется следующим образом:*

$$q = \max\{|1 - \tau l|, |1 - \tau L|\}.$$

Параметр q принимает наименьшее значение q_0 при $\tau = \tau_0 = \frac{2}{L+l}$.
При этом $q_0 = \frac{1-l/L}{1+l/L} \approx 1 - 2\frac{l}{L}$.

16.2.3. Чебышёвское ускорение метода простых итераций

Рассмотрим итерационный метод с выбором параметра τ на каждой итерации:

$$\mathbf{u}^{i+1} = \mathbf{u}^i + \tau_{i+1}(\Lambda u^i - \mathbf{f}).$$

Соответствующие соотношения для эволюции невязки имеют вид:

$$\mathbf{r}^{i+1} = (\mathbf{E} + \tau_{i+1}\Lambda)\mathbf{r}^i,$$

$$\mathbf{r}^i = \prod_{j=1}^{i} (\mathbf{E} + \tau_j \Lambda)\mathbf{r}^0.$$

После разложения невязок на двух соседних итерациях по базису из собственных функций сеточного оператора получим равенство

$$\sum c_{pq}^{i+1}\Psi^{pq} = \sum c_{pq}^i(\mathbf{E} + \tau_{i+1}\Lambda)\Psi^{pq} = \sum c_{pq}^i(1 - \tau_{i+1}\lambda^{pq})\Psi^{pq}.$$

Для коэффициентов разложения и компонентов невязки справедливы следующие равенства:

$$c_{pq}^{i+1} = c_{pq}^i(1 - \tau_{i+1}\lambda^{pq}),$$

$$c_{pq}^i = \prod_{j=1}^{i} (1 - \tau_j\lambda^{pq})\, c_{pq}^0,$$

$$\mathbf{r}^i = \sum c_{pq}^i \Psi^{pq} = \sum_{pq} c_{pq}^0 \prod_{j=1}^{i} (1 - \tau_j\lambda^{pq}).$$

Оценим погрешности на i шаге итераций

$$\left\|\mathbf{r}^i\right\| \leqslant \max_{\lambda\in[l,L]} \left|\prod_{j=1}^{i}(1 - \tau_j\lambda^{pq})\right| \left\|\sum c_{pq}\Psi^{pq}\right\| \leqslant \max_{\lambda\in[l,L]} \left|\prod_{j=1}^{i}(1 - \tau_j\lambda^{pq})\right| \left\|\mathbf{r}^0\right\|.$$

Вновь приходим к минимаксной задаче: найти такую последовательность итерационных параметров $\{\tau_j\}_{j=1}^i$, чтобы выполнялось

$$\min_{\{\tau_j\}} \max_{\lambda\in[l,L]} \left|\prod_{j=1}^{i}(1 - \tau_j\lambda^{pq})\right|.$$

Заметим, что $\prod_{j=1}^{i}(1-\tau_j\lambda^{pq})$ есть полином (относительно τ) степени i. Задача — сделать его наименее уклоняющимся от нуля на отрезке $[l, L]$.

Эта задача, как известно, решена Чебышевым, а корни этого полинома являются нулями полинома Чебышева (лекция 6):

$$\tau_j = \left[\frac{L+l}{2} + \frac{L-l}{2}\cos\frac{\pi(2j-1)}{2i}\right]^{-1}, j = 1, 2, \ldots, i.$$

Достаточно громоздкие выкладки, которые опускаются, дают в результате оценку скорости сходимости метода с оптимальным набором параметров $q \approx 1-2\sqrt{1/\mu}$, и числа итераций, необходимого для достижения заданной точности $i \approx \left[\frac{\sqrt{\mu}}{2}\ln\cdot\varepsilon^{-1}\right] + 1$.

Пусть расчеты приводятся с точностью $\varepsilon = 10^{-5}$ на сетке 100×100, тогда оценка числа итераций и скорости сходимости есть $q \approx 0,968, \quad i \approx 360$. «Цена» каждой итерации — приблизительно $10M^2$ операций.

Однако метод итераций с чебышевскими параметрами в таком виде оказывается неустойчивым по двум причинам: рост ошибок округления в расчетах и некоторые свойства нулей полиномов Чебышева, в частности, сгущение τ_k^{-1} — величин, обратных корням полинома — к границам спектра. Не останавливаясь на доказательстве неустойчивости чебышевского итерационного процесса, заметим, что для ее устранения необходимо переставить итерационные параметры не в их естественном порядке, а так, чтобы все частичные произведения $\prod_{j=1}^{i}(1-\tau_j\lambda)$ не возрастали бы вблизи границ спектра при любом i. Эта необходимо, поскольку частичное произведение, относящееся к правой части спектра со сгущающейся чебышевской сеткой, очень быстро растет из-за больших величин $(1-\tau_j\lambda)$. Задача упорядочения итерационных параметров достаточно сложна, ее решение связано с именами В. И. Лебедева, В. П. Финогенова, А. А. Самарского и Е. С. Николаева [4]. Приведем результат ее решения для $i = 2^r$, где i — количество сомножителей в произведении (число итераций) $\prod_{j=1}^{i}(1-\tau_j\lambda)$, r — натуральное число.

При $i = 2$ перебираем корни полинома Чебышева в их естественном порядке (в фигурных скобках указываем номер корня) $\{1,2\}$ или в порядке убывания номера $\{2,1\}$. Далее последовательность номеров корней получаем следующим образом. Каждый номер корня меняется на пару чисел: первое число — номер корня, второе — дополняет сумму в каждой паре до значения $i + 1(2^r + 1)$. Таким образом, при $i = 4$ получаем два упорядоченных набора. Из последовательности $\{1,2\}$ получаем $\{1,4,2,3\}$, а из $\{2,1\}$ - $\{2,3,1,4\}$. Действуя аналогично далее, имеем при i = 8 $\{1,8,4,5,2,7,3,6\}$ в первой последовательности чебышевских параметров или $\{2,7,3,6,1,8,4,5\}$ во второй последовательности. Следующий шаг дает $i = 16\{1,16,8,9,4,13,5,12,2,15,7,10,3,14,6,11\}$ в первой

последовательности чебышевских параметров или {2, 15, 7, 10, 3, 14, 6, 11, 1, 16, 8, 9, 4, 13, 5, 12} — во второй. Построение таких упорядоченных наборов легко можно продолжить. Приведенное упорядочение является универсальным — оно обеспечивает устойчивость любых методов, где необходим чебышевский набор итерационных параметров.

Обычно при реализации чебышевских методов задаются последовательностью из 32 или 64 параметров и делают серию итераций. Если невязка велика, то результат серии используют в качестве начального приближения для следующей серии итераций, если невязка мала, то решение считается найденным. Такая упрощенная реализация работает несколько медленнее, чем «честное» чебышевское ускорение, но весьма эффективно.

Приведем итерационные формулы трехслойного метода Чебышева, который не уступает предыдущему по скорости сходимости, однако не требует изменения порядка выбора итерационных параметров. В качестве платы за удобство использования метода затраты памяти становятся несколько большими. Формулы трехслойного метода будут

$$\mathbf{u}^1 = (\mathbf{E} - \tau\mathbf{A})\mathbf{u}^0 + \tau\mathbf{f}, \mathbf{u}^{i+1} = \\ = \alpha_{i+1}(\mathbf{E} - \tau\mathbf{A})\mathbf{u}^i + (1 - \alpha_{i+1})\mathbf{u}^{i-1} + \tau\alpha_{i+1}\mathbf{f}, i = 1, 2, \ldots$$

где $\tau = 2/(l + L), \alpha_1 = 2, \alpha_{i+1} = 4/(4 - \rho^2\alpha_i), \rho = (L - l)/(L + l)$.

Можно показать, что погрешность r^i удовлетворяет оценке

$$\|\mathbf{r}^i\| \leqslant \frac{2q^i}{1 + q^{2i}} \|\mathbf{r}^0\|,$$

где $q = \frac{1 - 1/\sqrt{\mu}}{1 + 1/\sqrt{\mu}}$.

Трехслойный метод Чебышева можно также представить в следующем виде:

$$\mathbf{u}^1 = (\mathbf{E} - \tau\mathbf{A})\mathbf{u}^0 + \tau\mathbf{f}, \mathbf{u}^{i+1} = \frac{2\gamma_1\gamma_i}{\gamma_{i+1}}(\mathbf{E} - \tau\mathbf{A})\mathbf{u}^i - \frac{\gamma_{i-1}}{\gamma_{i+1}}\mathbf{u}^{i-1} + \frac{2\gamma_1\gamma_i}{\gamma_{i+1}}\tau\mathbf{f},$$

где $i = 1, 2, \ldots, \tau = \frac{2}{l+L}, \gamma_i = \gamma_i(1/\mu), \gamma_0 = 1, \gamma_1 = \frac{\mu+1}{\mu-1}, \gamma_{i+1} = 2\gamma_1\gamma_i - \gamma_{i-1}$.

Трехслойный метод Чебышева в настоящее время применяется значительно чаще двухслойного при численном решении уравнений эллиптического типа.

Каноническая форма записи трехслойного итерационного метода (к которым и относится приведенный трехслойный метод Чебышева) есть

$$\mathbf{B}\mathbf{u}^{i+1} = \alpha_{i+1}(\mathbf{B} - \tau_{i+1}\mathbf{A})\mathbf{u}^i + (1 - \alpha_{i+1})\mathbf{B}\mathbf{u}^{i-1} + \alpha_{i+1}\tau_{i+1}\mathbf{f},$$

$$\mathbf{B}u^1 = (\mathbf{B} - \tau_1\mathbf{A})u^0 + \tau_1\mathbf{f},$$

$$i = 1, 2, \ldots.$$

Если оператор $\mathbf{B}$ — единичный, то трехслойный итерационный метод называется *явным*, в противном случае — *неявным*. Заметим, что *каноническая форма* записи *двухслойного* итерационного метода имеет вид

$$\mathbf{B}\frac{\mathbf{u}^{i+1} - \mathbf{u}^i}{\tau_{i+1}} + \mathbf{A}u^i = \mathbf{f}.$$

Если оператор $\mathbf{B}$ — единичный, то двухслойный итерационный метод называется *явным*, в противном случае — *неявным*.

Каноническая форма записи трехслойного итерационного метода получается из двухэтапного итерационного процесса. На первом этапе (предиктор) используется метод

$$\mathbf{B}\frac{\tilde{\mathbf{u}} - \mathbf{u}^i}{\tau_{i+1}} + \mathbf{A}\mathbf{u}^i = \mathbf{f},$$

где $\tilde{\mathbf{u}}$ — промежуточное значение. Второй этап — корректор:

$$\mathbf{u}^{i+1} = \alpha_{i+1}\tilde{\mathbf{u}} + (1 - \alpha_{i+1})\mathbf{u}^{i-1}.$$

В трехслойных схемах используются два итерационных параметра: τ_i и α_i причем при $\alpha_i = 1$ трехслойная схема переходит в двухслойную. Рассмотренные методы основываются на следующих свойствах оператора $\mathbf{A}$:

— самосопряженность: $\mathbf{A} = \mathbf{A}*$,

— положительная определенность: $\mathbf{A} > 0, 0 < l < \lambda_i < L$; для реализации алгоритма необходимо знание только границ спектра оператора.

16.2.4. Метод переменных направлений

Еще большие успехи при попытках ускорить итерационные методы были достигнуты при использовании методов *переменных направлений*. Можно показать, что решение нестационарной задачи

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} - f,$$

со стационарными граничными условиями $u|_\Gamma = \varphi$ будет стремиться к некоторому стационарному пределу при $t \longrightarrow \infty$ или при $n \longrightarrow \infty$. В этом

пункте n — количество шагов по времени при решении рассматриваемого дифференциального уравнения в частных производных разностным методом. В этом случае итерационный параметр τ играет роль шага по времени, отличие состоит лишь в том, что итерационный параметр не обязан быть малым, хорошая аппроксимация нестационарного уравнения не требуется.

Представим метод *переменных направлений* в следующем виде:

$$\frac{\tilde{u}_{ml}-u^i_{ml}}{\tau}=\Lambda_1\tilde{u}_{ml}+\Lambda_2 u^i_{ml}-f_{ml},\ \frac{u^{i+1}_{ml}-\tilde{u}_{ml}}{\tau}=\Lambda_1\tilde{u}_{ml}+\Lambda_2 u^{i+1}_{ml}-f_{ml}$$

Уравнение для эволюции погрешности получается, если из приведенных формул вычесть тождество $\frac{u_{ml}-u_{ml}}{\tau}=\Lambda_1 u_{ml}+\Lambda_2 u_{ml}-f$: $\frac{\tilde{r}_{ml}-r^i_{ml}}{\tau}=\Lambda_1\tilde{r}_{ml}+\Lambda_2 r^i_{ml}$, $\frac{r^{i+1}_{ml}-\tilde{r}_{ml}}{\tau}=\Lambda_1\tilde{r}_{ml}+\Lambda_2 r^{i+1}_{ml}$.

Представим r_{ml} и $\tilde{r}_{ml}$ в виде разложения по базису из собственных функций:

$$r^i_{ml}=\sum_{p\cdot q} c^i_{pq}\Psi^{pq}_{ml};\tilde{r}_{ml}=\sum_{p\,q}\tilde{c}_{pq}\Psi^{pq}_{ml}$$

и рассмотрим эволюцию погрешности за одну итерацию. Ψ^{pq}, как и ранее, собственные векторы разностных операторов Λ_1 и Λ_2, c_{pq} — коэффициенты Фурье; невязка r^i_{ml} обращается в нуль на границах области интегрирования.

Используя разностное уравнение для $\tilde{r}_{ml}$ представим связь между невязкой на предыдущей итерацией и невязкой на промежуточном слое в виде

$$(\mathbf{E}-\tau\Lambda_1)\tilde{r}_{ml}=(\mathbf{E}+\tau\Lambda_2)r^i_{ml}.$$

Используя представление Фурье, получим

$$(\mathbf{E}-\tau\Lambda_1)\sum_{pq}\tilde{c}_{pq}\Psi^{pq}_{ml}=(\mathbf{E}+\tau\Lambda_2)\sum_{pq}c^i_{pq}\Psi^{pq}_{ml},$$

откуда после действия операторов Λ_1 и Λ_2 на суммы, получаем

$$\sum_{pq}\tilde{c}_{pq}(1+\tau\lambda^p)\Psi^{pq}=\sum_{pq}c^i_{pq}(1-\tau\lambda^p)\Psi^{pq}.$$

Здесь λ^p и λ^q — собственные значения разностных операторов Λ_1 и $\Lambda_2(0<l\leqslant\lambda^p,0<\lambda^q\leqslant L,l\approx\pi^2,L\approx 4N^2)$, N — число шагов по каждому пространственному измерению рассматриваемой системы разностных уравнений. В силу единственности разложения сеточной функции по базису Ψ^{pq}, получаем соотношение для коэффициентов Фурье

$$\tilde{c}_{pq}=\frac{1-\tau\lambda^q}{1+\tau\lambda^p}c^i_{pq}.$$

Аналогично, после спектрального анализа второго этапа расчета, получим

$$c_{pq}^{i+1} = \frac{1-\tau\lambda^p}{1+\tau\lambda^q}\tilde{c}_{pq} = \frac{1-\tau\lambda^p}{1+\tau\lambda^q}\frac{1-\tau\lambda^q}{1+\tau\lambda^p}c_{pq}^{i}$$

Введем обозначение

$$\mu(\tau) = \max_{\lambda\in[l,L]} \left|\frac{1-\tau\lambda}{1+\tau\lambda}\right|,$$

тогда для коэффициентов разложения справедливо неравенство $\left|c_{pq}^{i+1}\right| \leqslant \mu^2(\tau)\left|c_{pq}^{i}\right|$, следовательно, для нормы невязки имеем условие $\left\|\mathbf{r}^{i+1}\right\| \leqslant \mu^2(\tau)\left\|\mathbf{r}^{i}\right\|$, так как $\left\|\mathbf{r}^{i+1}\right\| = \left\|\sum c_{pq}^{i+1}\Psi^{pq}\right\| \leqslant \left\|\sum \mu^2 c_{pq}^{i}\Psi^{pq}\right\| \leqslant \mu^2\left\|\sum c_{pq}^{i+1}\Psi^{pq}\right\| = \mu^2\left\|\mathbf{r}^{i}\right\|$, поскольку справедливо равенство Парсеваля. Необходимо обеспечить наиболее быструю сходимость, минимизировав коэффициент $\mu(\tau)$.

Найдем оптимальное значение параметра τ, обеспечивающее $\min\limits_{\tau}\mu(\tau)$. Для этого необходимо решить задачу

$$\tau = \arg\left\{\min_{\tau}\left[\max_{\lambda\in[l,\,L]}\left|\frac{1-\tau\lambda}{1+\tau\lambda}\right|\right]\right\},$$

которая решается так же, как и ранее. Простой графический анализ функции $\left|\frac{1-\tau\lambda}{1+\tau\lambda}\right|$ приводит к результату $\max\limits_{\lambda\in[l,L]}\left|\frac{1-\tau\lambda}{1+\tau\lambda}\right| = \max\limits_{\lambda\in[l,L]}\left\{\left|\frac{1-\tau l}{1+\tau l}\right|, \left|\frac{1-\tau L}{1+\tau L}\right|\right\}$.

Минимум $\mu(\tau)$ достигается, как и в предыдущем случае, при $\frac{1-\tau_0 l}{1+\tau_0 l} = -\frac{1-\tau_0 L}{1+\tau_0 L}$, откуда сразу получается $2 = 2\tau_0^2 lL$, $\tau_0 = 1\Big/\sqrt{lL}$.

Вычислим значение $\mu(\tau_0) = \frac{1-l/\sqrt{lL}}{1+l/\sqrt{lL}} \approx 1 - 2\sqrt{l/L}$. Погрешность за одну двухэтапную итерацию убывает, таким образом, в $q = \mu^2 \approx 1 - 4\sqrt{l/L}$ раз, $l \ll L$.

Количество итераций, необходимое для достижения заданной точности, можно оценить как $i \approx \frac{\ln\varepsilon^{-1}}{4\sqrt{l/L}} = \frac{1}{4}\sqrt{\frac{L}{l}}\ln\varepsilon^{-1}$.

Скорость сходимости итерационного процесса для приведенного метода приблизительно такая же, как и для процесса с чебышевским набором итерационных параметров. Эти результаты могут быть сформулированы в виде теоремы.

Естественным обобщением приведенного итерационного процесса представляется замена одного итерационного параметра τ набором τ_i. Выбор этих параметров приводит к минимаксной задаче

$$\min_{\tau}\left\{\max_{\lambda\in[l,L]}\prod_{j=1}^{i}\left|\frac{1-\tau_j\lambda}{1+\tau_j\lambda}\right|^2\right\},$$

решение которой представляется в виде громоздкого алгоритма, который здесь не приводится. Для этого метода имеем оценку числа итераций для достижения заданной точности $i \approx \ln \varepsilon^{-1} \ln L/l$.

16.2.5. Методы Якоби, Зейделя, верхней релаксации

Для системы сеточных уравнений, полученных при использовании схемы «крест»

$$\frac{u_{m-1,\,l} - 2u_{ml} + u_{m+1,\,l}}{h^2} + \frac{u_{m,l-1} - 2u_{ml} + u_{m,l+1}}{h^2} = f_{ml}$$

запишем итерационный метод Якоби. Расчетные формулы (верхний индекс, как обычно, показывает номер итерации) для этого метода будут

$$\frac{u^i_{m-1,l} - 2u^{i+1}_{ml} + u^i_{m+1,l}}{h^2} + \frac{u^i_{m,l-1} - 2u^{i+1}_{ml} + u^i_{m,l+1}}{h^2} = f_{ml},$$

$m, l = 1, \ldots, M-1, hM = 1$, с условиями на сеточной границе $u^{i+1}_{ml}/\text{г} = U_{ml}$.

Если в явном виде выразить u^{i+1}_{ml}, получим каноническую форму записи сеточного метода, правую часть берем на предыдущей итерации, левую — на текущей:

$$u^{i+1}_{ml} = \frac{1}{4}(u^i_{m-1,l} + u^i_{m+1,l} + u^i_{m,l-1} + u^i_{m,l+1}) + \frac{h^2}{4} f_{ml}.$$

Количество итераций, требуемое для вычисления решения с точностью ε, оценивается по формуле

$$i = \frac{2N^2}{\pi^2} \ln \varepsilon^{-1}.$$

Метод Зейделя, учитывающий результаты вычислений на $i+1$ итерации, записывается для рассматриваемого уравнения Пуассона

$$\frac{u^{i+1}_{m-1,l} - 2u^{i+1}_{ml} + u^i_{m+1,l}}{h^2} + \frac{u^{i+1}_{m,l-1} - 2u^{i+1}_{ml} + u^i_{m,l+1}}{h^2} = f_{ml},$$

$m, l = 1, \ldots, M-1, hM = 1$, с условиями на сеточной границе $u^{i+1}_{ml}/\text{г} = U_{ml}$.

Напомним, что хотя метод Зейделя неявный, но его реализация оказывается простой, если правильно установить последовательность вычислений. В каноническом виде формулы для метода Зейделя есть

$$u^{i+1}_{ml} = \frac{1}{4}(u^{i+1}_{m-1,l} + u^i_{m+1,l} + u^{i+1}_{m,l-1} + u^i_{m,l+1}) + \frac{h^2}{4} f_{ml}.$$

Сначала из последнего уравнения, используя граничные условия $u_{01}^{i+1} = U_{01}$ и $u_{10}^{i+1} = U_{10}$, находим u_{11}^{i+1}. Затем, зная u_{11}^{i+1}, можно аналогично найти u_{12}^{i+1} и так далее. Значения сеточной функции вычисляются в следующем порядке изменения индексов: $(1,1),(1,2),\ldots,(1,M-1),(2,1),(2,2),\ldots,(2,M-1),(M-1,1),(M-1,2),\ldots,(M-1,M-1)$. Оценка количества итераций, необходимых для достижения точности ε, есть $i \approx \frac{N^2}{\pi^2}\ln\varepsilon^{-1}$.

Метод Зейделя сходится быстрее метода Якоби, однако число итераций также оценивается как $O(N^2) = O(L/l)$. В случае метода *верхней релаксации* система представляется в виде

$$(\mathbf{L}+\mathbf{U}+\mathbf{D})\mathbf{u} = \mathbf{f},$$

где $\mathbf{L}$ и $\mathbf{U}$ — нижняя и верхняя треугольные матрицы с нулевыми диагоналями, $\mathbf{D}$ — диагональная матрица. Вводится параметр $1 < \tau < 2$ и итерационные формулы записываются в виде $\mathbf{L}\mathbf{u}^{i+1} + \mathbf{U}\mathbf{u}^{i} + \mathbf{D}\left[\frac{\mathbf{u}^{i+1}}{\tau} + (1-\frac{1}{\tau})\mathbf{u}^{i}\right] = \mathbf{f}$. При $\tau = 1$ получаем метод Зейделя.

Рассмотрим реализацию метода верхней релаксации для разностной аппроксимации уравнения Пуассона. Переписав разностную схему в виде

$$\frac{u_{m-1,l}+u_{m,l-1}}{h^2} + \frac{u_{m+1,l}+u_{m,l+1}}{h^2} - \frac{4u_{ml}}{h^2} = f_{ml}u_{ml} = U_{ml},$$

выпишем расчетные формулы для метода релаксаций:

$$\frac{u_{m-1,l}^{i+1}+u_{m,l-1}^{i+1}}{h^2} + \frac{u_{m+1,l}^{i}+u_{m,l+1}^{i}}{h^2} - \\ - \frac{4}{h^2}\left[\frac{u_{ml}^{i+1}}{\tau} + (1-\frac{1}{\tau})u_{ml}^{i}\right] = f_{ml}, u_{ml}^{i+1} = U_{ml}.$$

Последовательность вычислений в методе релаксаций такая же, как и в методе Зейделя. Уравнения представляются в виде

$$u_{m-1,l}^{i+1} + u_{m,l-1}^{i+1} - \frac{4}{\tau}u_{ml}^{i+1} = -(u_{m+1,l}^{i} + u_{m,l+1}^{i}) + 4(1-\frac{1}{\tau})u_{ml}^{i} - h^2 f_{ml},$$

решение u_{ml}^{i+1} вычисляется так же с левого нижнего угла области интегрирования. Основное преимущество метода верхней релаксации перед методом Зейделя состоит в существенном ускорении скорости сходимости при соответствующем выборе параметра τ. Не проводя исследований на сходимость, приведем их окончательный результат. Необходимое количество итераций для достижения точности ε равно $i \approx \frac{2M}{\pi}\ln\varepsilon^{-1}, h = 1/M$.

Отметим также, что для реализации этих трех методов не требуется знания спектра задачи. При оптимальном выборе итерационного параметра в методе верхней релаксации знание границ спектра позволяет ускорить сходимость метода.

16.3. Попеременно-треугольный итерационный метод

При аппроксимации уравнений Пуассона или Лапласа получается система сеточных уравнений

$$\mathbf{Au} = \mathbf{f},$$

без ограничения общности считаем оператор (матрицу системы) самосопряженным и положительно определенным. Сеточный оператор Лапласа самосопряжен — это легко доказать, но отрицателен. Для того чтобы получить систему уравнений с положительным оператором, достаточно правую и левую части системы умножить на -1. Здесь $\mathbf{A}$ — квадратная матрица размером $N \times N$.

Зададим матрицу $\mathbf{R} = \{r_{ij}\}$:

$$r_{ij} = \begin{cases} a_{ij}, & i > j \\ \frac{a_{ij}}{2}, & i = j \\ 0, & i < j. \end{cases}$$

В таком случае $\mathbf{A}$ можно представить в виде суммы двух треугольных матриц $\mathbf{A} = \mathbf{R} + \mathbf{R}^*$.

$\mathbf{R}$ и $\mathbf{R}^*$ — нижняя и верхняя треугольные матрицы, а их диагональные элементы совпадают. Далее будем рассматривать систему сеточных уравнений как операторное уравнение в конечномерном евклидовом пространстве (унитарном в комплексном случае).

Попеременно-треугольный итерационный метод (ПТИМ) может быть представлен в каноническом виде

$$\mathbf{B}\frac{u^{i+1} - u^i}{\tau} + \mathbf{A}u^i = f.$$

Здесь $\mathbf{B}$ — самосопряженный положительный оператор. Его часто называют оператором предобуславливания. Для рассматриваемого метода оператор предобуславливания представляется в виде произведения

$$\mathbf{B} = (\mathbf{E} + \omega\mathbf{R}^*)(\mathbf{E} + \omega\mathbf{R}),$$

где $\mathbf{E}$ — единичный оператор, ω — параметр (действительное число).

При известных ω и τ значение u_{k+1} находится в два этапа. Сначала вычисляется невязка на итерации $r^i = \mathbf{A}u^i - f$, а затем решается система матричных уравнений

$$(\mathbf{E} + \omega\mathbf{R}^*)\tilde{u} = \mathbf{B}u^i - \tau r^i,$$

$$(\mathbf{E} + \omega\mathbf{R})u^{i+1} = \tilde{u}.$$

Поскольку матрицы $(\mathbf{E} + \omega\mathbf{R}^*)$ и $(\mathbf{E} + \omega\mathbf{R})$ являются треугольными, то эти уравнения решаются значительно проще, чем исходное.

Для формулирования теоремы о ПТИМ введем понятие операторного неравенства: будем полагать, что $\mathbf{A} \leqslant \mathbf{B}$, если для любой сеточной функции, не равной нулю тождественно, выполнено неравенство

$$((\mathbf{A} - \mathbf{B})\mathbf{u}, \mathbf{u}) \geqslant 0.$$

Теорема (без доказательства). *Пусть* $\mathbf{A} = \mathbf{R} + \mathbf{R}^*$ *и существуют положительные постоянные* δ *и* Δ, *при которых выполнены неравенства*

$$\mathbf{A} \geqslant \delta\mathbf{E}, 4\mathbf{R}^*\mathbf{R} \leqslant \Delta\mathbf{A}.$$

Пусть также $\omega = \frac{2}{\sqrt{\delta\Delta}}$, $\tau = \frac{2}{\gamma_1+\gamma_2}$, *где* $\gamma_1 = \frac{\delta}{2\left(1+\sqrt{\frac{\delta}{\Delta}}\right)}$, $\gamma_2 = \frac{\sqrt{\delta\Delta}}{4}$.

Тогда двухэтапный итерационный метод

$$r^i = \mathbf{A}u^i - f,$$

$$(\mathbf{E} + \omega\mathbf{R}^*)\tilde{u} = \mathbf{B}u^i - \tau r^i,$$

$$(\mathbf{E} + \omega\mathbf{R})u^{i+1} = \tilde{u}$$

сходится, причем для его погрешности справедлива оценка

$$\|u_i - u\|_A \leqslant \rho^i \|u_0 - u\|_A,$$

где $\rho = \frac{1-\sqrt{\delta/\Delta}}{1+3\sqrt{\delta/\Delta}}$, $\|x\|_{\mathbf{A}} = \sqrt{(\mathbf{A}x, x)}$.

В качестве δ можно взять минимальное собственное значение $\lambda_{\min}$ оператора $\mathbf{A}$, либо любую положительную постоянную $s \leqslant \lambda_{\min}$. Можно показать, что $\Delta \geqslant \lambda_{\max}$, $\Delta > \delta$, где $\lambda_{\max}$ — максимальное собственное значение оператора $\mathbf{A}$.

Рассмотрим применение ПТИМ к численному решению уравнения Пуассона (знак минус ставим для удобства записи):

$$-(\Lambda_1 + \Lambda_2)u_{ml} = f_{ml},$$

с нулевыми граничными условиями

$$u_{m0} = u_{mN} = u_{0l} = u_{Nl} = 0.$$

Эта задача может быть представлена как операторное уравнение, где $\mathbf{A}u_{ml} = -(\Lambda_1 + \Lambda_2)u_{ml} m, l = 1, \ldots, N-1$. Оператор будет самосопряженным и положительным.

Далее необходимо представить матрицу $\mathbf{A}$ в виде $\mathbf{A} = \mathbf{R} + \mathbf{R}^*$ и определить константы δ и Δ.

Рассмотрим разностную аппроксимацию уравнения Пуассона по схеме «крест», переписав схему в эквивалентном виде:

$$\mathbf{A}u_{ml} = \frac{1}{h}\left(\frac{u_{ml} - u_{m-1,l}}{h} + \frac{u_{ml} - u_{m,l-1}}{h}\right) - \frac{1}{h}\left(\frac{u_{m+1,l} - u_{ml}}{h} + \frac{u_{m,l+1} - u_{ml}}{h}\right).$$

Тем самым представили оператор $\mathbf{A}$ как сумму двух операторов $\mathbf{A} = \mathbf{R} + \mathbf{U}$, где

$$\mathbf{R}u_{ml} = \frac{1}{h}\left(\frac{u_{ml} - u_{m-1,l}}{h} + \frac{u_{ml} - u_{m,l-1}}{h}\right),$$

$$\mathbf{R}^*u_{ml} = -\frac{1}{h}\left(\frac{u_{m+1,l} - u_{ml}}{h} + \frac{u_{m,l+1} - u_{ml}}{h}\right).$$

При этом матрица оператора $\mathbf{R}$ является нижней треугольной, а матрица $\mathbf{R}^*$ — верхней треугольной (в этом легко убедиться, записав рассматриваемую систему в виде СЛАУ). Можно также показать, что оператор $\mathbf{U}$ является сопряженным оператором к $\mathbf{R}$, т. е. $\mathbf{A} = \mathbf{R} + \mathbf{R}^*$.

Также показывается, что $\delta = \frac{8}{h^2}\sin^2\frac{\pi h}{2} = \lambda_{\min}(\mathbf{A}), \Delta = \frac{8}{h^2} \geqslant \frac{8}{h^2}\cos^2\frac{\pi h}{2} = \lambda_{\max}(\mathbf{A})$.

Воспользовавшись формулами для τ и ω, получим

$$\omega = \frac{h^2}{4\sin\frac{\pi h}{2}} \approx \frac{h}{2\pi}, \tau = \frac{h^2(1+\sin\frac{\pi h}{2})}{\sin\frac{\pi h}{2}(1+3\sin\frac{\pi h}{2})} \approx \frac{2h}{\pi}.$$

Оценка количества итераций для этого метода дает

$$i \approx \frac{\ln\varepsilon^{-1}}{2\pi h} = \frac{N}{2\pi}\ln\varepsilon^{-1}.$$

Напомним, что границы спектра для рассматриваемого оператора $L \approx 8N^2, l \approx 2\pi^2$ и $N \approx \frac{\pi}{2}\sqrt{L/l}$.

Алгоритм вычисления u_{ml}^{i+1}, в соответствии с ранее приведенными формулами двух этапов ПТИМ будет таков.

Первый этап:

$$r^i = \mathbf{A}u^i - f,$$

$$\tilde{u}_{ml} - \frac{\omega}{h}\left(\frac{\tilde{u}_{m+1,l} - \tilde{u}_{ml}}{h} + \frac{\tilde{u}_{m,l+1} - \tilde{u}_{ml}}{h}\right) = \mathbf{B}u^i_{ml} - \tau r^i_{ml},$$

второй этап:

$$u^{i+1}_{ml} + \frac{\omega}{h}\left(\frac{u^{i+1}_{ml} - u^{i+1}_{m-1,l}}{h} + \frac{u^{i+1}_{ml} - u^{i+1}_{m,l-1}}{h}\right) = \tilde{u}_{ml},$$

$$u^{i+1}_{0l} = 0, \quad u^{i+1}_{m0} = 0.$$

Уравнение для $\tilde{u}_{ml}$ нужно начинать решать от точки $m = N-1, l = N-1$, учитывая, что на границе сеточной области сеточная функция равна нулю. Таким образом, вычисления ведутся рекуррентно. Система решается, начиная с правого верхнего угла области до левого нижнего угла. Система линейных уравнений, соответствующая второму этапу, решается аналогично, но вычисления здесь начинаются в точке $m = 1, l = 1$ и заканчивается в точке $m = N-1, l = N-1$.

Метод по своим идеям напоминает схему Саульева, рассмотренную ранее в одномерном варианте при решении уравнений параболического типа.

Использование в ПТИМ набора чебышевских итерационных параметров приводит к следующему результату. Расчетные формулы для метода таковы:

$$\mathbf{B}\frac{u^{i+1} - u^i}{\tau} + \mathbf{A}u^i = f,$$

$$\mathbf{A} = \mathbf{R} + \mathbf{R}^*, \mathbf{B} = (\mathbf{E} + \omega\mathbf{R}^*)(\mathbf{E} + \omega\mathbf{R}),$$

$$\tau_j = \frac{\tau_0}{1 + \eta t_j}, \quad \tau_0 = \frac{2}{\gamma_1 + \gamma_2},$$

$$\eta = \frac{1 - \gamma_1/\gamma_2}{1 + \gamma_1/\gamma_2},$$

$$t_j = \cos\frac{(2j-1)\pi}{2i}, j = 1, \ldots, i.$$

Для проведенного примера $\gamma_1 \approx 2\sin\frac{\pi h}{2}$, $\gamma_2 \approx 1 + \sin\frac{\pi h}{2}$, $\frac{\gamma_1}{\gamma_2} \approx \pi h$. Оценка количества итераций будет $i = \frac{\sqrt{N}}{\sqrt{2\pi}}\ln\varepsilon^{-1}$.

16.4. Сводка результатов по итерационным методам решения сеточных уравнений

После рассмотрения самых распространенных методов решения сеточных уравнений получена возможность сравнить итерационные мето-

ды по количеству итераций, необходимых для достижения точности ε. Приведем только выражения для сомножителя, пропорционального числу узлов сетки.

$\frac{2M^2}{\pi^2}$ — методы Якоби и простых итераций с оптимальным выбором параметра,

$\frac{M^2}{\pi^2}$ — метод Зейделя,

$\frac{2M}{\pi}$ — метод верхней релаксации,

$\frac{M}{\pi}$ — метод простых итераций с чебышевским набором параметров,

$\frac{1}{2}\frac{M}{\pi}$ — методы переменных направлений и попеременно-треугольный,

$\frac{\sqrt{M}}{2\sqrt{\pi}}$ — попеременно-треугольный метод с чебышевским набором параметров,

$\frac{2}{\gamma}\ln(\frac{2}{\pi}M)$ — метод переменных направлений с чебышевским набором параметров, $\gamma \approx 3,2$.

16.5. Основные идеи многосеточного метода Р. П. Федоренко

Вернемся к рассмотренному выше вопросу о сходимости итерационных методов решения систем сеточных уравнений, получающихся при решении уравнений Лапласа или Пуассона. Выше был проведен спектральный анализ итерационных методов решения такой системы. Вернемся, например, к результатам исследования метода переменных направлений. Для него установлено равенство для коэффициентов разложения в конечный ряд Фурье невязки

$$c_{pq}^{i+1} = \frac{1-\tau\lambda^p}{1+\tau\lambda^q}\tilde{c}_{pq} = \frac{1-\tau\lambda^p}{1+\tau\lambda^q}\frac{1-\tau\lambda^q}{1+\tau\lambda^p}c_{pq}^i$$

Из этого выражения следует, что коэффициенты при высокочастотных составляющих (т. е. при достаточно больших λ при фиксированном τ) убывают достаточно быстро, в то время как при низкочастотных гармониках ($\lambda \approx 0$) коэффициенты убывают медленно. Такой же эффект существует и для других итерационных методов.

Рассмотрим сеточное уравнение Пуассона

$$\Lambda_1 u_{ml} + \Lambda_2 u_{ml} = f_{ml}, \tag{16.1}$$

и, кроме того, уравнение

$$\Lambda_1 e_{ml} + \Lambda_2 e_{ml} = r_{ml}, \tag{16.2}$$

где r — невязка, e — погрешность решения. Если известно приближенное решение задачи (16.1), то, решая (16.2), можно найти решение исходной системы по формуле

$$u_{ml} = \tilde{u}_{ml} + e_{ml},$$

но решение задачи (16.2) по сложности точно такое же, как и решение исходной задачи.

Если применять итерационный метод (16.1), сделав несколько итераций, то приближенное решение задачи будет известно, а невязка станет плавно меняющейся функцией — в разложении в конечный ряд Фурье будут отсутствовать высокочастотные составляющие. Невязка будет хорошо представляться на более грубой сетке, тогда размерность вспомогательной системы (16.2) понизится.

Введем сетку с шагом $2h$ и рассмотрим на ней задачу (16.2). Размерность системы сеточных уравнений получилась в четыре раза меньше, чем размерность исходной системы. Предположим, что есть средство эффективно решать такие задачи. После этого для восстановления решения надо найти невязку на более подробной сетке. Строим оператор интерполяции с грубой сетки на подробную. Очевидно, что ошибка интерполяции будет иметь высокочастотный компонент. Тогда используем полученное решение (низкочастотные компоненты погрешности погашены на грубой сетке, остались только высокие частоты) в качестве начального приближения и делаем несколько сглаживающих итераций. Этот этап называется *коррекцией с грубой сетки.*

Схема приближенного построения решения получается следующей.

1. Решается уравнение (16.) на подробной сетке, проводится несколько сглаживающих итераций,
2. По приближенному решению находится невязка $r_{ml} = \Lambda_1 \tilde{u}_{ml} + + \Lambda_2 \tilde{u}_{ml} - f_{ml}$,
3. Ищется ограничение (проекция) невязки на грубую сетку.
4. На грубой сетке решается уравнение $\Lambda_1 e_{ml} + \Lambda_2 e_{ml} = r_{ml}$.
5. Строится оператор интерполяции решения на подробную сетку, значения невязки пересчитываются на подробную сетку.
6. На подробной сетке проводится несколько сглаживающих итераций для уравнения $\Lambda_1 e_{ml} + \Lambda_2 e_{ml} = r_{ml}$.
7. Ищется решение по формуле $u_{ml} = \tilde{u}_{ml} + e_{ml}$.

Выполнение пункта 4 — решения сеточной системы — может оказаться трудоемким, поэтому внутри этого шага алгоритма можно еще раз

перейти к более грубой сетке. Многосеточный метод состоит в рекурсивном повторении данной процедуры до тех пор, пока не будет получена система малой размерности, которая может быть эффективно решена прямым методом (например, Гаусса или сопряженных градиентов).

Простейшая реализация многосеточного метода — *каскадный алгоритм*. Он предусматривает последовательное решение задач на сетке со все большим числом узлов, а интерполяция решения с более грубой сетки используется в качестве начального приближения. Для самосопряженных эллиптических задач часто используется V-цикл. Сначала ищется решение на подробной сетке, затем проводится последовательно вычисление невязок на все более грубых сетках, затем — несколько коррекций с грубой сетки и сглаживающие итерации на все более и более подробных сетках. Для несамосопряженных задач лучше зарекомендовал себя W-цикл, когда сначала вспомогательная система решается на последовательности все более и более грубых сеток, затем идут итерации на все более подробных сетках, затем снова итерации на грубых сетках и затем — на подробных вплоть до сетки с максимальным числом разбиений.

Впервые многосеточный метод предложен в [10]. Более подробно о многосеточных методах и реализации многосеточных алгоритмов можно прочитать в [8] и статьях Н. С. Бахвалова с соавторами и Ю. В. Василевского с соавторами в [9].

16.6. Построение разностных схем для эллиптических уравнений на нерегулярных сетках. Монотонные схемы (подход А.С.Холодова)

В этом разделе будем следовать статье [11]. Кроме уравнений Лапласа и Пуассона, в этом разделе будем рассматривать произвольные уравнения эллиптического типа. Рассмотрим задачу

$$u_{xx} + e_{12}u_{xy} + e_{22}u_{yy} + e_1u_x + e_2u_y = f(x, y, u)$$

с условиями на коэффициенты

$$e_{ml}(x, y, u) = \{e_{mlj}\}, \quad e_m(x, y, u) = \{e_{mj}\},$$
$$m, l = 1, 2, \quad j = 1, \ldots, J, \quad e_{22} > e_{12}^2/4.$$

Последнее условие обеспечивает эллиптичность задачи. Задача решается в произвольной замкнутой области с несколькими несвязными границами. Предположим сначала, что область не содержит входящих

углов (т. е. таких, величина которых больше π). В случае со входящими углами в решении эллиптических задач возникают особенности, во многих случаях необходимо сочетание численных методов и аналитических (асимптотических методов в окрестности входящего угла), подробнее в [12]. Индекс j — номер узла сетки; для простоты выкладок узлы сетки нумеруются одним индексом.

Запишем разностную схему для аппроксимации приведенного выше уравнения в каноническом виде

$$u_k = \sum_i \alpha_{ki} u_i + f_k, k = 1, \ldots, K, i = i_1, \ldots, i_I$$

— точки, принадлежащие *шаблону* схемы (см. ниже), $I \geqslant 5$, причем *неопределенные коэффициенты* удовлетворяют условию монотонности $\alpha_{ki} \geqslant 0$. Здесь K — общее число внутренних сеточных узлов с одноиндексной нумерацией, $i_1, \ldots, i_I$ — номера расположенных достаточно произвольно в области интегрирования внутренних и граничных сеточных узлов, α_{ki} — неопределенные коэффициенты. Часть этих коэффициентов (или все, если $I = 5$) определяются условиями аппроксимации первого порядка или второго порядка на решениях исходного уравнения. Эти условия получаются стандартно — проекция точного решения на нерегулярную сетку раскладывается в ряд Тейлора в окрестности точки u_k. После несложных, но громоздких выкладок получаем, что следующие условия обеспечивают первый порядок аппроксимации:

$$\sum_i \alpha_{ki} = 1,$$

$$\sum_i \alpha_{ki} X_i (1 - 0.5 X_i (e_1 - X_i (e_1^2 + f_u - e_{1x})/3 + Y_i e_{1y})) = 0,$$

$$\sum_i \alpha_{ki} (Y_i - 0.5 X_i^2 (e_2 + X_i (e_2 f_u - e_1 e_2 + e_{2x})/3 - Y_i (f_u - e_{2y}))) = 0,$$

$$\sum_i \alpha_{ki} X_i (Y_i - 0.5 X_i (e_{12} + X_i (e_2 + e_{12x} - 2 e_{12} e_1)/3 + Y_i (e_1 + e_{12y}))) = 0,$$

$$\sum_i \alpha_{ki} (Y_i^2 - X_i^2 (e_{22} - X_i (e_{12} e_2 + e_{22} e_1 - e_{22x})/3 + Y_i (e_2 + e_{22y}))) = 0,$$

$$\beta_0 = -\sum_i \alpha_{ki} X_i^2 (1 - e_1 X_i)/3)/2 = 0.$$

Суммирование ведется по всем точкам, включенным в шаблон. Если к приведенным выше условиям добавить условия

$$\sum_i \alpha_{ki} X_i (3Y_i(Y_i - e_{12}X_i) - X_i^2(e_{22} - e_{12}^2)) = 0,$$

$$\sum_i \alpha_{ki}(Y_i^3 - e_{22}X_i^2(3Y_i - e_{12}X_i)) = 0,$$

$$\beta_1 = -\sum_i \alpha_{ki} X_i^3/6 = 0, \beta_2 = -\sum_i \alpha_{ki} X_i^2 Y_i/2 = 0,$$

то схема имеет второй порядок аппроксимации на решениях. Здесь введены обозначения $X_i = x_i - x_k, Y_i = y_i - y_k$.

В обычных разностных методах шаблон фиксирован (т. е. все внутренние точки сетки имеют одинаковое число соседей). В данном подходе для каждой рассчитываемой точки $k = 1, \ldots, K$ специальным образом подбираются соседи (сеточный шаблон) так, чтобы выполнялись условия неотрицательности коэффициентов α_{ki} в канонической форме записи разностной схемы. Эти условия обеспечивают неотрицательность разностного оператора (мажорантность схемы).

Получающийся в результате решения уравнений для условий аппроксимации первого порядка и неравенств, обеспечивающих монотонность схемы, для каждого $k = 1, \ldots, K$ набор коэффициентов α_{ki} приводит к знакопостоянной линейной (или нелинейной для квазилинейных уравнений) системе уравнений

$$\mathbf{Au} = \mathbf{b},$$

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & \ldots & 0 & -\alpha_{1i_1} & 0 & \ldots & 0 & -\alpha_{1i_l} & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 & -\alpha_{Ki_1} & 0 & \ldots & 0 & -\alpha_{Ki_l} & 1 \end{pmatrix}$$

$$\alpha_{ki} \geqslant 0, \quad \sum_i \alpha_{ki} \leqslant 1.$$

Условия неотрицательности неопределенных коэффициентов обеспечивают выполнение достаточных условий сходимости и принципа максимума, в том числе, при разрывных граничных условиях. Тогда простейший итерационный метод Якоби

$$u_k^{n+1} = \sum_i \alpha_{ki} u_i^n$$

является сходящимся.

Можно показать, что при весьма слабых ограничениях на расположение сеточных узлов в области интегрирования такие схемы могут быть конструктивно построены. Этот же подход распространен на случай схем со вторым порядком аппроксимации на решениях.

16.7. Задачи

1. В современных формулировках метод релаксации рассматривается с черно-белым (шахматным, красно-черным) упорядочением узлов. Назовем все внутренние узлы сетки черными, если для них сумма значений индексов четная, все прочие внутренние узлы назовем белыми. Получить расчетные формулы метода верхней релаксации для сетки с черно-белым (шахматным, красно-черным) упорядочением узлов.

   Решение. Заметим, что при расчетах белые узлы соседствуют только с черными, и наоборот. Тогда с учетом этого расчетные формулы будут для всех белых узлов

$$\frac{u^i_{m-1,l}+u^i_{m,l-1}}{h^2}+\frac{u^i_{m,l+1}+u^i_{m+1,l}}{h^2}-\frac{4}{h^2}\left[\frac{u^{i+1}_{ml}}{\tau}+(1-\frac{1}{\tau})u^i_{ml}\right]=f_{ml},$$

   а для всех черных

$$\frac{u^{i+1}_{m-1,l}+u^{i+1}_{m,l-1}}{h^2}+\frac{u^{i+1}_{m,l+1}+u^{i+1}_{m+1,l}}{h^2}-\frac{4}{h^2}\left[\frac{u^{i+1}_{ml}}{\tau}+(1-\frac{1}{\tau})u^i_{ml}\right]=f_{ml},$$

   Последовательность вычислений для такого варианта будет несколько отличаться от первоначальной формулировки метода релаксаций. Сначала ищется значение на следующей итерации для всех белых узлов, затем — для черных. Такой итерационный метод очевидным образом связан со схемой «классики» для решения параболических уравнений.

2. Пусть число внутренних узлов равно 9. Выписать в матричном виде сеточные уравнения при классической формулировке схемы «крест» и для случая черно-белого упорядочения узлов. Для простоты рассмотреть вариант, когда значения функции на границе области равны нулю.

   Решение.

   В «классическом» варианте сеточная система есть

$$-\mathbf{Au}=\mathbf{f},$$

где

$$\mathbf{A} = \begin{pmatrix} 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 & -1 & 4 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 \end{pmatrix}$$

$$\mathbf{u} = (u_{11}, u_{12}, u_{13}, u_{21}, u_{22}, u_{23}, u_{31}, u_{32}, u_{33})^{\mathrm{T}},$$

$$\mathbf{f} = -h^2(f_{11}, f_{12}, f_{13}, f_{21}, f_{22}, f_{23}, f_{31}, f_{32}, f_{33})^{\mathrm{T}},$$

а в случае черно-белого упорядочения узлов сетки

$$-\mathbf{A}\mathbf{u} = \mathbf{f},$$

где

$$\mathbf{A} = \begin{pmatrix} 4 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 & -1 & 0 & -1 & 0 \\ 0 & 0 & 4 & 0 & 0 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 4 & 0 & 0 & -1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 4 & 0 & 0 & -1 & -1 \\ -1 & -1 & -1 & 0 & 0 & 4 & 0 & 0 & 0 \\ -1 & 0 & -1 & -1 & 0 & 0 & 4 & 0 & 0 \\ 0 & -1 & -1 & 0 & -1 & 0 & 0 & 4 & 0 \\ 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 & 4 \end{pmatrix}$$

$$\mathbf{u} = (u_{11}, u_{13}, u_{22}, u_{31}, u_{33}, u_{12}, u_{21}, u_{23}, u_{32})^{\mathrm{T}},$$

$$\mathbf{f} = -h^2(f_{11}, f_{13}, f_{22}, f_{31}, f_{33}, f_{12}, f_{21}, f_{23}, f_{32})^{\mathrm{T}}.$$

16.8. Задачи для самостоятельного решения

1. Будем рассматривать только частные типы краевых задач для поля φ, зависящего от двух пространственных переменных (x, y), удовлетворяющего уравнению

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\varphi = -S(x, y).$$

В задачах электростатики φ — это потенциал, а S соответствует плотности заряда; в стационарной тепловой задаче φ — температура, S — локальная скорость выделения или поглощения тепла. Будут рассматриваться граничные условия Дирихле, в которых значения φ задаются на некоторой замкнутой кривой в плоскости (x, y) и, возможно, на некоторых дополнительных кривых внутри области.

Реализовать численные алгоритмы, основанные на:

(a) непосредственной аппроксимации дифференциального оператора и решении системы сеточных уравнений методом Гаусса;

(b) применении итерационного алгоритма.

а) Для уравнения Лапласа, $S(x, y) \equiv 0$, рассмотреть численное решение для простейших граничных условий (типа констант или линейных функций).

б) Для уравнения Пуассона вычислить разность потенциалов между двумя зарядами как функцию расстояния между ними и сравнить полученные значения с аналитическими.

в) Изменить программу так, чтобы можно было задавать на некоторых внешних и внутренних границах условия Неймана. Изучить решения с такими граничными условиями.

г) Вместо граничных условий Дирихле задаются периодические граничные условия. Тогда потенциалы на левой и правой, а также на верхней и нижней границах области произвольные, но равные по величине друг другу. Т. е. для всех i и j $\varphi_{i1} = \varphi_{iN}$; $\varphi_{1j} = \varphi_{Nj}$. Уравнения с такими условиями описывают пространственно-периодическое распределение плотности заряда в кристалле. Модифицировать программу и решить уравнение Пуассона с этими граничными условиями.

2. Для решения приведенной выше задачи с различными граничными условиями реализовать алгоритм быстрого дискретного преобразования Фурье (алгоритмы такого преобразования описаны, например, в [4, 8]).

3. **Модель малярной кисти** Подробнее об этой задаче и других примерах установившихся течений жидкости в [13, С. 232–240].

   При окраске стены кистью, часть краски остается на стенке в виде слоя за кистью. Рассмотрим приближенную модель процесса. Предположим, что кисть состоит из большого числа параллельных

и равноотстоящих друг от друга пластин, которые совместно скользят по плоской стенке, в направлении их контакта со стенкой вдоль оси x. Предположим, что пластины имеют бесконечные размеры в направлениях осей x и z, так что результирующее движение представляет собой установившееся течение одного направления. Уравнение движения имеет вид

$$\frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0.$$

Здесь u — скорость течения жидкости.

Оси координат удобно связать с пластинами, тогда граничные условия для течения в канале между двумя соседними пластинами будут

$$u = 0 \text{ при } y = 0 \text{ и } y = b, \quad 0 < z < \infty,$$

$$u = U \text{ при } z = 0, \quad 0 < y < b.$$

а) Получить численное решение поставленной задачи. Сравнить результат с точным решением

$$u(y,z) = \frac{4U}{\pi} \sum_{\substack{n \\ \text{нечетное}}} \frac{1}{n} e^{-n\pi z/b} \sin \frac{n\pi y}{b}.$$

Объяснить, как реализован алгоритм для вычисления значений функции точного решения.

б) Получить оценку толщины слоя жидкости, который будет оставаться на стенке позади кисти, при предположении, что все пластины имеют заднюю кромку при одном и том же значении x. Использовать формулу для объемного расхода жидкости, вытекающей из одного канала: $Q = \int_0^b \int_0^\infty u\, dydz$. Сравнить со значением для точного решения $Q = AUb^2$, $A \approx 0,27$.

в) Указать недостатки рассмотренной модели. Определить характер их влияния на решение.

4. **Стационарное движение несжимаемой вязкой жидкости в цилиндрических трубах**

Рассмотрим движение несжимаемой вязкой жидкости в цилиндрической трубе произвольного поперечного сечения. Обозначая градиент давления $\frac{\partial p}{\partial x} = -G(t)$, получим уравнение движения в виде

$\frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = -\frac{G}{\mu}$, где μ — коэффициент вязкости, u — скорость течения жидкости. Требуется решить уравнение, подчиняющееся граничным условиям, с помощью которых задаются градиент давления и значения u при определенных y и z.

(a) Решить задачу для трубы круглого поперечного сечения, для которой $u = 0$ на границе трубы, т. е. при $r = \sqrt{y^2 + z^2} = a$. Сравнить численные результаты с точным решением $u = \frac{G}{4\mu}(a^2 - r^2)$. Получить численно величину объемного расхода жидкости через произвольное сечение $Q = \int\limits_0^a u\, 2\pi r dr$, сравнить ее с точным значением $Q = \frac{\pi a^4 G}{8\mu}$. Предложить, как можно использовать данную величину для контроля точности численного расчета.

(b) Получить решение задачи для трубы эллиптического поперечного сечения с полуосями b (по оси y) и c (вдоль оси z). Сравнить результат численного решения с точным распределением скорости

$$u(y,\ z) = \frac{G}{2\mu(b^{-2} + c^{-2})}\left(1 - \frac{y^2}{b^2} - \frac{z^2}{c^2}\right).$$

Самостоятельно вывести величину объемного расхода жидкости и сравнить ее точное значение с численными данными.

(c) Вычислить распределение скорости в трубе прямоугольного поперечного сечения со сторонами $y = mb$, $z = mc$ (для определенности пусть $c > b$). Получить точное решение для скорости и расхода, сравнить его с численным значением.

Указание. Для получения точного решения учесть, что разность $u - \frac{G}{2}\frac{b^2 - y^2}{\mu}$ представляет собой четную функцию как от y, так и от z, которая удовлетворяет уравнению Лапласа и равна 0 при $y = mb$.

5. Гравитационные волны

Примером нестационарных течений жидкости являются волновые движения с колебаниями отдельных частиц. Рассмотрим волны на поверхности жидкости, возникающие в результате того, что поверхность выведена из состояния равновесия и колеблется под действием силы тяжести. Такие волны называются гравитационными. Гравитационные волны описываются уравнениями нестационарных течений идеальной несжимаемой жидкости.

Пусть начальное возмущение заключается в отклонении жидкости от состояния равновесия. Предположим, что этим начальным возмущением являются мгновенные добавочные давления, вызванные, например, порывом ветра. Возникающие при этом движения будут потенциальны: $\mathbf{V} = \nabla\varphi$. Уравнение неразрывности обращается в уравнение Лапласа:

$$\Delta\varphi = 0.$$

Уравнения движения Эйлера приводятся к интегралу Коши–Лагранжа следующего вида:

$$\frac{\partial\varphi}{\partial t} + \frac{v^2}{2} + \frac{p}{\rho} + gz = f(t),$$

где gz представляет собой потенциал сил тяжести.

Пусть течение медленное, тогда квадратом скорости в последнем уравнении можно пренебречь. Кроме того, так как φ определяется с точностью до произвольной функции, зависящей от времени, то это уравнение можно переписать в виде

$$\frac{p}{\rho} = -\frac{\partial\varphi}{\partial t} - gz \text{ при } z = 0.$$

Граничные условия. Предположим, что жидкость ограничена снизу непроницаемой поверхностью. На этой поверхности ставим условие непроницаемости на нормальный компонент скорости: $V_n = \frac{\partial\varphi}{\partial n} = 0$. Свободная поверхность (граница жидкости с газом) будет плоскостью, которую примем за координатную плоскость xy. На свободной поверхности жидкости давление p равно давлению газа над жидкостью (p_0). Граничное условие для скорости на свободной поверхности

$$V_z = \frac{\partial\varphi}{\partial z} = -\frac{1}{g}\frac{\partial^2\varphi}{\partial t^2} \text{ при } z = 0.$$

Начальные условия. Пусть возмущенная поверхность в начальный момент ($t = 0$) определяется уравнением $z = f(x, y)$. Тогда при $t = 0, z = 0$ справедливо соотношение $\frac{\partial\varphi}{\partial t} = -gf(x, y)$. Начальные скорости возникают в результате действия импульса давления, равного $\int\limits_0^\tau p d\tau$. Потенциал скорости в начальный момент можно представить в виде

$$\varphi = -\frac{1}{\rho}f_1(x, y).$$

Плоские волны. Рассмотрим волновое движение, называемое плоскими волнами. В этом случае свободная поверхность будет представлять собой цилиндрическую поверхность с образующими, параллельными оси y. Пусть жидкость ограничена плоским горизонтальным дном, отстоящим от свободной поверхности на расстояние h. Тогда для искомого потенциала скорости $\varphi(x, z, t)$ справедливы соотношения

$$\frac{\partial^2\varphi}{\partial x^2} + \frac{\partial^2\varphi}{\partial z^2} = 0,$$

$$\left(\frac{\partial\varphi}{\partial z}\right)_{z=-h} = 0,$$

$$\frac{\partial^2\varphi\,(x,0,t)}{\partial t^2} + g\frac{\partial\varphi\,(x,0,t)}{\partial z} = 0.$$

Начальные условия заменим требованием периодичности по времени t и координате x искомого решения.

(a) Построить аналитическое решение поставленной задачи, представив искомую функцию в виде $\varphi = \theta(t)X(x)Z(z)$. Показать, что полученное решение — результат наложения четырех колебаний.

(b) Получить численное решение и сравнить его с аналитическим. Определить профиль волны $\zeta(x, y, t)$, используя соотношение

$$\zeta = -\frac{1}{g}\frac{\partial\varphi\,(x,y,t)}{\partial t}.$$

Прогрессивные волны. Провести численное исследование частного случая плоских волновых движений, который определяется потенциалом скорости вида

$$\varphi = Ae^{kz}\cos(kx - \sigma t).$$

Профиль волны в этом случае имеет вид

$$\zeta = -\frac{A\sigma}{g}\sin(kx - \sigma t).$$

Описать основные закономерности рассматриваемого движения жидкости. Чему равен период волны, частота колебаний? Вывести формулу для траекторий частиц жидкости и сравнить ее на графике с траекторией, получаемой в численном решении.

Литература

[1] *Рябенький В.С.* Введение в вычислительную математику. М.: Физматлит, 2000. 294 с.

[2] *Федоренко Р.П.* Введение в вычислительную физику. М.: Изд-во МФТИ, 1994. 526 с.

[3] *Самарский А.А.* Теория разностных схем. М.: Наука, 1983. 656 с.

[4] *Самарский А.А., Николаев Е.С.* Методы решения сеточных уравнений. М.: Наука, 1978. 590 с.

[5] *Марчук Г.И.* Методы вычислительной математики. М., Наука, 1989. 608 с.

[6] *Иванов В.Д., Косарев В.И. и др.* Лабораторный практикум "Основы вычислительной математики". М.: МЗ Пресс, 2003. 193 с.

[7] *Самарский А.А., Гулин А.В.* Численные методы математической физики. М.: Научный мир, 2003. 316 с.

[8] *Деммель Дж.* Вычислительная линейная алгебра. Теория и приложения. М., Мир, 2001. 429 с.

[9] Современные проблемы вычислительной математики и математического моделирования, Том 1. М. Наука, 2005. 343 с.

[10] *Федоренко Р.П.* Релаксационный метод решения разностных эллиптических уравнений. // ЖВМ и МФ, 1961, т. 1, № 5. с. 922-927.

[11] *Холодов А.С.* Монотонные разностные схемы на нерегулярных сетках для эллиптических уравнений в области со многими несвязными границами. // Математическое моделирование. 1991. Т. 3. № 9. С. 104-113.

[12] *Марчук Г.И., Шайдуров В.В.* Повышение точности решений разностных схем. М.: Наука, 1979. 320 с.

[13] *Бэтчелор Дж.* Введение в динамику жидкости. М.: Мир, 1973. 758 с.

Лекция 17. Понятие о методах конечных элементов

Лекция дает первое представление о классе методов конечных элементов. Приводятся вариационная и проекционная постановки задачи. Рассматривается применение МКЭ к стационарным и нестационарным задачам. Вкратце обсуждаются вопросы устойчивости методов конечных элементов при решении нестационарных задач. Рассматривается общая схема применения методов конечных элементов к решению многомерных задач математической физики.

Ключевые слова: метод конечных элементов, вариационная формулировка, метод Ритца, проекционная формулировка, метод Галеркина, базисные функции, функции с финитным носителем, согласованный базис.

Основная идея метода конечных элементов, базирующая на методах Бубнова, Галеркина и Ритца, была предложена Р. Курантом в 1943 г., но осталась незамеченной, опередив потребности практики. В 50-х годах прошлого века с появлением первых компьютеров возникла необходимость в разработке новых инженерных подходов к численному решению задач со сложной геометрией, в которых области интегрирования разбивались на подобласти. Такие подобласти (носители финитных базисных функций, об этом ниже) и получили название *конечных элементов*.

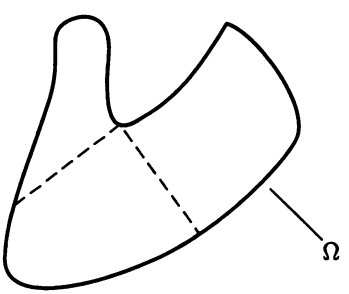


Рис. 17.1

Методы конечных элементов (МКЭ) в настоящее время, пожалуй, самые распространенные в мире численные методы. К их достоинствам относятся:

1. возможность счета на неравномерных сетках, в двумерном и трехмерном случаях для областей сложной геометрии;

2. «технологичность» методов (уточнение далее).

Современные МКЭ возникли в 50-е годы XX века при решении задач теории упругости.

Самая распространенная статическая задача — задача о нагруженной конструкции

$$\Delta u = -2, \quad u\,|_{\partial\Omega} = 0,$$

а область Ω — сложная. Например, область может иметь вид, представленный на рис. 1. Каждая простая подобласть — конечный элемент.

В настоящее время под МКЭ понимают целые семейства вариационных (Ритца) и проекционных (Галеркина или Бубнова–Галеркина) методов.

17.1. Вариационный подход Ритца

Рассмотрим две задачи:

$$\hat{L}_1 u(x) \equiv -(k(x)\, u'_x(x))'_x + p(x)\, u(x) = f(x), \tag{17.1}$$

$$u(0) = a; \quad u(X) = b;$$

$$k(x) \geqslant k_0 > 0; \quad p(x) \geqslant 0.$$

$$\hat{L}_2 u(x,\, y) \equiv -\mathrm{div}\,(k(x,\, y)\,\mathrm{grad}\, u(x,\, y)) + p(x,\, y)\, u(x,\, y) = g(x,\, y), \tag{17.2}$$

$$u\,|_{\partial\Omega} = \psi(l);$$

$$k(x,\, y) \geqslant k_0 > 0; p(x,\, y) \geqslant 0.$$

Эти задачи похожи: (17.1) является одномерным случаем более общей задачи (17.2). Уравнения (17.1) и (17.2) записаны в самосопряженной форме. Поставим задачам (17.1) и (17.2) в соответствие функционалы

$$I_1(u) = \int_0^X (k(u'_x)^2 + pu^2 - 2fu)\, dx \tag{17.3}$$

и

$$I_2(u) = \iint_\Omega (k(\nabla u, \nabla u) + pu^2 - 2gu) dx dy. \tag{17.4}$$

Будем рассматривать пространство функций $w \in W_2^1$ (пространство Соболева) с нормой

$$||w||^2_{W_2^1} = \int_0^X (w^2 + (w'_x)^2)\, dx \quad \text{для одномерного случая,}$$

$$||w||^2_{W_2^1} = \iint_\Omega w^2 dxdy + \iint_\Omega \left(\frac{\partial w}{\partial x}\right)^2 + \left(\frac{\partial w}{\partial y}\right)^2 dxdy \quad \text{для двумерного случая.}$$

Это — функции с ограниченным интегралом.

Теорема 5. *Среди всех функций $w \in W_2^1$, удовлетворяющих граничным условиям, решение задачи* (17.1) *придает наименьшее значение функционалу* (17.3), *а решение* (17.2) *— функционалу* (17.4).

Доказательство.

Докажем это утверждение для одномерного случая, а доказательство для уравнений (17.2, 17.4) оставим в качестве упражнений.

Введем $\xi(x) \equiv w(x) - u(x)$. Поскольку $w(x) \in W_2^1$, а $u(x)$ — дважды непрерывно дифференцируемая функция, то $\xi(x) \in W_2^1$ и $\xi(0) = \xi(X) = 0$.

$$I_1(w) = I_1(u(x) + \xi(x)) =$$

$$= I_1(u) + \int_0^X (k(\xi'_x)^2 + p\xi^2 - 2f\xi)\, dx + \int_0^X 2(ku'_x\xi'_x + p\xi u)\, dx =$$

$$= I_1(u) + \int_0^X (k(\xi'_x)^2 + p\xi^2)\, dx + \int_0^X 2\xi(pu - f)\, dx + 2\int_0^X ku'_x\xi'_x\, dx =$$

$$= I_1(u) + J(\xi) + 2ku'_x\xi \Big|_0^X + \int_0^X 2\xi(-(ku'_x)'_x + pu - f)\, dx,$$

$$\text{где} J(\xi) \equiv \int_0^X (k(\xi'_x)^2 + p\xi^2)\, dx \geqslant 0. \qquad (17.5)$$

Третье слагаемое в (17.5) равно нулю в силу граничных условий для функции ξ; последнее слагаемое равно нулю, так как u — решение (17.1); второе слагаемое — неотрицательное. Следовательно, минимум функционала $I_1(w)$ достигается, когда $J(\xi) = 0$, т. е. $\xi \equiv 0$ или, что то же самое, $w(x) = u(x)$. ■

Чуть сложнее эта теорема доказывается для двумерного случая, где надо воспользоваться теоремой Остроградского–Гаусса. Таким образом, решение соответствующей задачи в частных производных (17.2) или краевой задачи для ОДУ (17.1) сводится к задаче минимизации некоторого функционала.

В том случае, если функционал (17.3) или (17.4) ограничен снизу, то экстремаль функционала — минимум, и численный метод, который будет построен ниже, носит название метода Ритца. Чаще, когда нет необходимости тщательно исследовать постановку задачи, говорят об экстремальной точке, стационарной точке функционала и т.д.

17.2. Общая схема метода Ритца

Решение задачи (17.1) ищут в виде

$$u^N(x) = \psi_0^N(x) + \sum_{k=1}^{N} C_k \psi_k^N(x), \tag{17.6}$$

где $\psi_0^N(x), \ldots, \psi_N^N(x)$ — базисные функции в W_2^1; $\psi_0^N(x)$ удовлетворяет граничным условиям, а $\psi_k^N(x)$ при $k \geqslant 1$ такие, что $\psi_k^N(0) = \psi_k^N(X) = 0$. Если суммирование в (17.6) происходит до бесконечности, то эта формула дает точное решение задачи (17.1). Так как рассматривается конечное число базисных функций, то получаем лишь приближенное решение. Примером базисных функций для метода Ритца может служить тригонометрический базис, а в качестве приближенного решения получим конечный отрезок ряда Фурье.

Подставив (17.6) в (17.3), получаем

$$I_1(u^N) = \sum_{p=1}^{N} \sum_{q=1}^{N} C_p C_q \cdot \left\{ \int_0^X \left(k(x) \frac{\partial \psi_p^N}{\partial x} \frac{\partial \psi_q^N}{\partial x} + p(x) \psi_p^N \psi_q^N \right) dx \right\} -$$

$$-2 \sum_{r=1}^{N} \left\{ C_r \int_0^X \left(f(x) \psi_r^N - p(x) \psi_0^N \psi_r^N - k(x) \frac{\partial \psi_0^N}{\partial x} \frac{\partial \psi_r^N}{\partial x} \right) dx \right\} + I_1(\psi_0^N). \tag{17.7}$$

Находим минимум функционала (17.7) из условия $\frac{\partial I_1(u^N)}{\partial C_i} = 0$, получаем систему из N линейных уравнений для определения коэффициентов C_k. Затем объявляем (17.6) решением задачи.

Точно так же поступаем и для функционала (17.4). Число уравнений в системе для определения коэффициентов тоже будет N (у базисной функции только один индекс!). Вид функционала будет аналоги-

чен (17.7), но вместо интегралов по отрезку будут стоять двойные интегралы по рассматриваемой области пространства Ω, а вместо производных — градиенты.

Первая проблема, которая возникает в методе Ритца — выбор подходящего базиса. Как от набора функций $\psi_0^N(x), \ldots, \psi_N^N(x)$ зависит решение? Как оценить ошибки?

Существуют два типа базиса: *глобальный базис* для метода Ритца и *базис из функций с финитным носителем.*

Для того чтобы решения по методу Ритца сходились к точному, необходимо и достаточно, чтобы $\forall\ g \in W_2^1$ и $\forall\ \varepsilon > 0$ существовала линейная комбинация

$$g^N(x) \equiv \psi_0^N(x) + \sum_{j=1}^{N} C_j \psi_j^N(x), \text{такая, что} \|g^N - g\|_{W_2^1} \leqslant \varepsilon,$$

если вычисления проводятся точно.

Допустимый базис для применения в методе Ритца $\sin\left(\frac{\pi q x}{X}\right)$, $q = 1, \ldots, N$.

На отрезке $[0, 1]$ допустимые базисы: $\psi_j^N = x(1-x)T_j(2x-1)$, где $T_j(x)$ — j-й полином Чебышева; $\psi_j^N = x^j(1-x)$.

Матрица системы линейных уравнений для определения коэффициентов разложения по базису метода Ритца получается заполненной. В случае использования «неудачных» базисов ее число обусловленности достаточно велико.

Технологичность метода Ритца заключается в следующем. Матрица соответствующей системы является самосопряженной с диагональным преобладанием при правильном выборе базиса. Можно решать систему быстро сходящимися итерационными методами.

Рассмотрим простейший вариант метода Ритца с использованием базиса функций с финитным носителем. Напомним, что носитель функции — множество точек x, для которых $\psi_j^N(x) \neq 0$. Введем разбиение отрезка $[0,\ \ X]$ точками x_j (сетку): $0 = x_0 < x_1 < \ldots < x_N = X$. Строим базисные функции:

$$\psi_0^N = \begin{cases} a \cdot \frac{x - x_1}{x_0 - x_1}, & 0 \leqslant x \leqslant x_1; \\ 0, & x_1 \leqslant x \leqslant x_{N-1}; \\ b \cdot \frac{x - x_{N-1}}{x_N - x_{N-1}}, & x_{N-1} \leqslant x \leqslant x_N; \end{cases}$$

$$\psi_j^N = \begin{cases} \frac{x - x_{j-1}}{x_j - x_{j-1}}, & x_{j-1} \leqslant x \leqslant x_j; \\ 0, & x > x_{j+1},\quad x < x_{j-1}; \\ \frac{x_{j+1} - x}{x_{j+1} - x_j}, & x_j < x \leqslant x_{j+1}. \end{cases}$$

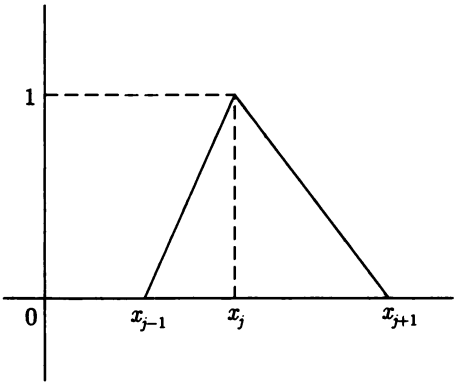


Рис. 17.2

Можно проверить, что $\psi_j^N \in W_2^1[0, \quad X]$. Интегралы и производные определяются в смысле обобщенных функций — недостаток базиса! Достоинством этого базиса является то, что базисные функции почти ортогональны.

Пусть

$$(\psi_j^N,\ \psi_k^N) = \int\limits_0^X \psi_j^N(x)\ \psi_k^N(x)\ dx,$$

тогда

$$(\psi_j^N,\ \psi_j^N) = \int\limits_{x_{j-1}}^{x_j} \frac{(x-x_{j-1})^2}{(x_j-x_{j-1})^2}\ dx + \int\limits_{x_j}^{x_{j+1}} \frac{(x-x_{j+1})^2}{(x_{j+1}-x_j)^2}\ dx =$$

$$= (x_j - x_{j-1}) \int\limits_0^1 t^2 dt + (x_{j+1} - x_j) \int\limits_0^1 t^2 dt = \frac{(x_j - x_{j-1})}{3}\, t^3 \Big|_0^1 + \ldots =$$

$$= \frac{x_j - x_{j-1}}{3} + \frac{x_{j+1} - x_j}{3} = \frac{x_{j+1} - x_{j-1}}{3}, (\psi_j^N,\ \psi_{j+1}^N) =$$

$$= \int\limits_{x_j}^{x_{j+1}} \frac{x - x_j}{x_{j+1} - x_j}\ \frac{x_{j+1} - x}{x_{j+1} - x_j}\ dx = \frac{x_{j+1} - x_j}{6}, (\psi_{j-1}^N,\ \psi_j^N) = \frac{x_j - x_{j-1}}{6},$$

а все остальные скалярные произведения равны нулю.

Также достаточно легко берутся интегралы, включающие в себя производные базисных функций. Носитель каждой такой базисной функции

называется конечным элементом, а метод Ритца с использованием такого базиса — первый метод из семейства МКЭ. Иногда конечным элементом также называют саму базисную функцию с финитным носителем.

17.3. Формулировка проекционного метода Галеркина

По-прежнему рассматриваем задачи (17.1) и (17.2).

В дальнейшем будет рассмотрен класс дифференциальных операторов. Главный недостаток метода Ритца — применимость лишь к дифференциальным задачам, допускающим вариационную формулировку, т. е. в линейном случае $\hat{L}$ — самосопряженный положительно определенный оператор (все собственные числа $\hat{L}$ положительны).

Наряду с формулировкой (17.1) и (17.2) будем использовать запись, определяющую *слабое* (обобщенное) решение:

$$(\hat{L}u,\ v) - (f,\ v) = 0, \tag{17.8}$$

где v — *любая* функция из рассмотренного ранее функционального пространства W_2^1, а скалярное произведение определено как

$$(u,\ v) = \int_0^X u(x)\ v(x)\ dx \quad \text{в одномерном случае;}$$

$$(u, v) = \iint_\Omega u(x,y)v(x,y)dxdy \quad \text{в двумерном случае.}$$

Равенство (17.8) определяет обобщенное решение задачи. Известно, что если u — классическое решение задачи, то оно является обобщенным решением в смысле (17.8). Обратное, по понятным причинам, неверно — в W_2^1 «больше» функций, чем в C^1 или C^2. У задачи может существовать обобщенное решение, но не существовать классического.

Рассмотрим конечномерное подпространство пространства W_2^1 с введенным базисом:

$$u^N = \psi_0^N + \sum_{k=1}^{N} C_k \psi_k^N,$$

ψ_k^N — базисные функции в W_2^1; они обязаны обладать теми же свойствами, что и базисные функции для метода Ритца. Рассмотрим теперь для (17.8) *конечную* систему *весовых* функций из W_2^1 : $v_1^N, \ldots, v_N^N$. Вместо (17.8) рассмотрим конечную систему проекций на весовые функции.

Введем также обозначение

$$R \equiv \hat{L}u^N - f \tag{17.9}$$

здесь R — невязка. Тогда, после подстановки разложения по базисным функциям в (17.8), получим систему соотношений

$$(R,\ v_k^N) = (\hat{L}u^N,\ v_k^N) - (f,\ v_k^N). \tag{17.10}$$

Минимум невязки в пространстве, определяемом функциями $v_1^N, \ldots, v_N^N$ достигается тогда, когда невязка принадлежит его ортогональному дополнению: $(R,\ v_k^N) = 0$ для всех k. Теперь надо потребовать, чтобы весовые функции образовывали базис в W_2^1. Естественно в качестве весовых функций использовать уже имеющиеся базисные $\psi_1^N, \ldots, \psi_N^N$. Тогда получаем проекционный метод Галеркина.

В итоге для определения коэффициентов разложения по базису из конечных элементов имеем систему соотношений вида

$$\left(\hat{L}\left(\psi_0^N + \sum_{j=1}^N C_j\psi_j^N\right), \psi_k^N\right) = \left(\psi_0^N + \sum_{j=1}^N C_j\psi_j^N,\ \hat{L}\psi_k^N\right) =$$

$$= (\psi_0^N,\ \hat{L}\psi_k^N) + \sum_{j=1}^N C_j(\hat{L}\psi_j^N,\ \psi_k^N);$$

$$(\psi_0^N,\ \hat{L}\psi_k^N) + \sum_{j=1}^N C_j(\hat{L}\psi_j^N,\ \psi_k^N) = (f,\ \psi_k^N)$$

или в матричной форме:

$$\mathbf{AC} = \eta\ , a_{jk} = (\hat{L}\psi_j^N,\ \psi_k^N);$$
$$\eta_k = (f,\ \psi_k^N) - (\hat{L}\psi_0^N,\ \psi_k^N) = (f - \hat{L}\psi_0^N, \psi_k^N).$$

Это же соотношение получается и при выводе системы уравнений для коэффициентов в методе Ритца.

При вычислении скалярных произведений использовалась самосопряженность линейного дифференциального оператора $\hat{L}$. Но при выводе соотношения (17.10) самосопряженность оператора не использовалась! Значит, метод Галеркина можно обобщать и на случай несамосопряженного (и нелинейного!) дифференциального оператора. При использовании в качестве базисных функций «функций-крышечек», введенных выше, получаем вариант МКЭ. Для задач (17.1) и (17.2) метод будет давать те же соотношения, что и метод Ритца.

17.4. Пример построения схемы конечных элементов

Для уменьшения числа выкладок считаем, что

$$h_i = h_{i+1} \equiv h \quad \text{для всех} \quad i(h_i \equiv x_i - x_{i-1}).$$

Рассмотрим несамосопряженный аналог задачи (17.1):

$$-k(x)\, u''_x + q(x)\, u'_x + p(x)\, u(x) = f(x),$$

$$u(0) = a; \qquad u(1) = b;$$

$$k(x) \geqslant k_0 > 0; \quad p(x) \geqslant 0.$$

Найдем сопряженное уравнение:

$$\int_0^1 (-ku'' + qu' + pu - f) \cdot v \, dx =$$

$$= -ku' \cdot v \Big|_0^1 + \int_0^1 ukv' \, dx + quv \Big|_0^1 - \int_0^1 u(qv)' \, dx + \int_0^1 puvdx - \int_0^1 fvdx =$$

$$= \left\{ \begin{array}{c} \text{члены определяемые} \\ \text{граничными условиями} \end{array} \right\} - \int_0^1 fvdx + \int_0^1 ((kv)'' - qv' + pv \cdot udx.$$

Из этого соотношения легко получить условия, при которых $\hat{L} \neq \hat{L}^*$.
Теперь запишем разложение по базису:

$$u^N = \psi_0^N + \sum_{j=1}^{N} C_j \psi_j^N$$

подставляем в исходное дифференциальное уравнение:

$$-\left(k(x)\left(\psi_0^N + \sum_{j=1}^{N} C_j \psi_j^N\right)''_{xx}, \psi_k^N\right) + \left(q(x)\left(\psi_0^N + \sum_{j=1}^{N} C_j \psi_j^N\right)'_x, \psi_k^N\right)$$

$$+ \left(p(x)\left(\psi_0^N + \sum_{j=1}^{N} C_j \psi_j^N\right), \psi_k^N\right) = (f, \psi_k^N).$$

Рассмотрим отдельно каждое слагаемое в левой части:

$$-\left(k(x)\left(\psi_0^N+\sum_{j=1}^{N}C_j\psi_j^N\right)''_{xx},\psi_k^N\right)=$$

$$=-(k({\psi_0^N}''_{xx}),\psi_k^N)+\left(-k(x)\sum_{j=1}^{N}C_j({\psi_j^N}''_{xx}),\psi_k^N\right)= \tag{17.11}$$

$$=-(k({\psi_0^N}''_{xx}),\psi_1^N)-(k({\psi_0^N}''_{xx}),\psi_N^N)-\left(k(x)\sum_{j=1}^{N}C_j{\psi_j^N}'_x,{\psi_k^N}'_x\right)$$

Первые два слагаемые в правой части получаются в силу того, что носители базисных функций финитны. Последнее слагаемое в правой части получается интегрированием по частям. Зачем необходимо интегрирование по частям? На первый взгляд $(\Psi_j^N)''_{xx}=0$. Но эта производная базисной функции — обобщенная функция, следовательно, в скалярных произведениях появятся δ-функции, при интегрировании возникнут сложности.

В итоге после всех необходимых вычислений коэффициент перед C_k:

$$-C_{k-1}\cdot k(x_{k-1/2})\,\frac{1}{h}+C_k\cdot k(x_k)\,\frac{2}{h}-C_{k+1}\cdot k(x_{k+1/2})\,\frac{1}{h}.$$

Функция $k(x)$ считается кусочно-постоянной на соответствующих отрезках, можно использовать какую-либо другую аппроксимацию $(k\varphi_j^N)'_x$, учитывая что $k(x)$ — заданная функция.

Первые два слагаемые в правой части (17.11) зависят от граничных условий и относятся к правой части системы уравнений для определения C_k.

Рассмотрим теперь

$$\left(q(x)\left(\psi_0^N+\sum_{j=1}^{N}C_j\psi_j^N\right)'_x,\psi_k^N\right)=(q(x)\,{\psi_0^N}'_x,\psi_1^N)+(q(x){\psi_0^N}'_x,\psi_N^N)+$$

$$+\sum_{j=1}^{N}C_j(q(x)\,{\psi_j^N}'_x,\,\psi_k^N).$$

Коэффициенты при C_k будут следующие:

$$-C_{k-1}\cdot q(x_{k-1/2})\cdot\frac{1}{2}+C_k\cdot q(x_k)\cdot 0+C_{k+1}\cdot q(x_{k+1/2})\cdot\frac{1}{2}$$

Здесь опять предполагается, что функция $q(x)$ кусочно-постоянная. Последнее слагаемое с $p(x)$ дает выражение

$$C_{k-1} \cdot p(x_{k-1/2}) \frac{h}{6} + C_k \cdot p(x_k) \frac{4h}{6} + C_{k+1} \cdot p(x_{k+1/2}) \frac{h}{6},$$

т. е. при «плохом» способе вычисляемых интегралов фактически получаем конечно-разностное соотношение, похожее на аппроксимацию Нумерова.

Вместе с тем, существует значительное отличие. Сеточная функция — это функция, заданная таблично. Решение (приближенное) МКЭ — это не сеточная функция, а элемент W_2^1.

17.5. Построение базисных функций

Математическая основа МКЭ — метод Галеркина и вариационный метод Ритца — развиваются, начиная со второго десятилетия XX века. Прогресс в МКЭ последних лет заключается именно в построении наборов базисных функций, обладающих достаточной гладкостью — так называемых согласованных базисов.

Базис из «крышечек» в двумерном случае. Процесс построения базисных функции включает в себя:

а) триангуляцию области — разбиение на треугольники, каждый из которых является носителем своей базисной функции;

б) построение базисных функций.

Требования к триангуляции (обозначения на рис. 17.3).

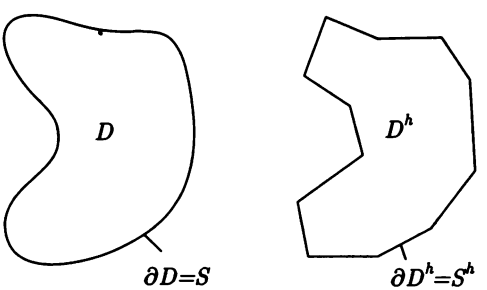


Рис. 17.3

1. Между точками S и S^h с помощью нормалей к S устанавливается взаимно-однозначное соответствие, расстояние между соответствующими точками не превосходит $\delta_1\ h^2$ (h — сеточный параметр).

2. Длины сторон треугольников и их площади лежат в пределах $[hl_1, hl_2]$ и $[h^2\gamma_1,\ h^2\gamma_2]$, где $l_1, l_2, \gamma_1, \gamma_2$ — положительные константы, не зависящие от h.

3. Существует непрерывное взаимно-однозначное преобразование D^h на область, границы которой параллельны осям координат, или составляют с ними угол $\pi/4$. Преобразование линейно внутри каждого треугольника и переводит последний в равнобедренный прямоугольный треугольник с катетами, равными h.

Простейший пример построения триангуляции.

1. Область D вписываем в прямоугольник.

2. Строим в прямоугольнике равномерную сетку с шагом h (рис. 17.4).

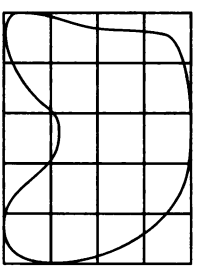


Рис. 17.4

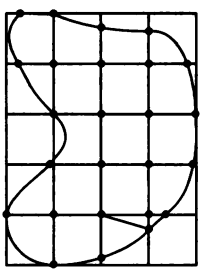


Рис. 17.5

3. Ближайшие к границе D узлы сетки сдвигаем на границу D (рис. 17.5).

4. Разбиваем четырехугольники внутри D^h диагоналями (рис. 17.6).

5. Убираем все ячейки, пересечение которых с D^h пусто (рис. 17.7).

Построение базисной функции — «крышечки». Фиксируем вершину P_1 какого-либо треугольника. Составляем список соседей — вершин, принадлежащих треугольникам, имеющим вершину P_1. Пусть в списке

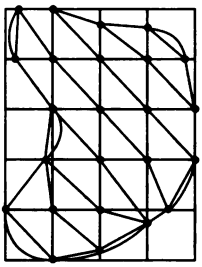


Рис. 17.6

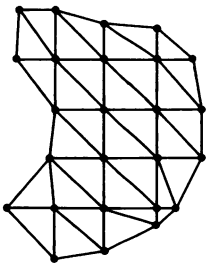


Рис. 17.7

есть вершины Q_1 и Q_2, принадлежащие треугольнику 1 (рис. 17.8). В этом треугольнике представляем

$$\varphi_1(x,\ y) = \frac{1 - \frac{y-y_1}{y_2-y_1} - \frac{x-y_2}{x_1-y_2}}{1 - \frac{y_{P_1}-y_1}{y_2-y_1} - \frac{x_{P_1}-y_2}{x_1-y_2}}.$$

Тогда для точки

$$P_1\psi_{P_1}^N(x,\ y) = \sum_{k=1}^{6} \varphi_k(x,\ y).$$

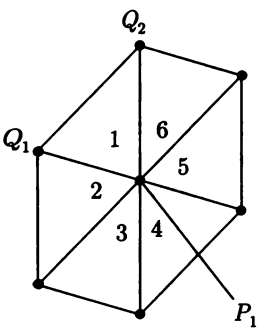


Рис. 17.8

К сожалению, базисных функций типа «крышечек» может не хватить для решения уравнений второго (по пространственной производной) порядка. До сих пор рассматривались уравнения второго порядка. Перейдем теперь к модельному уравнению

$$\frac{d^4u}{dx^4} + au = f(x); x \in [0,\ X] \qquad (17.12)$$

с какими-либо граничными условиями.

Будем искать решение в соответствии с методами МКЭ:

$$u^N = \psi_0^N + \sum_{j=1}^{N} C_j\psi_j^N, \qquad (17.13)$$

где ψ_j^N обладают финитным носителем. Подставляем разложение (17.13) в (17.12). Отвлекаясь от членов с граничными условиями, отнесенными к

ψ_0^N, при умножении на ψ_l^N имеем

$$\left(\frac{d^4}{dx^4}\sum_{j=1}^{N} C_j\psi_j^N,\ \psi_l^N\right) = \int_0^X \sum_{j=1}^{N} C_j\,\frac{d^4\psi_j^N}{dx^4}\,\psi_l^N\,dx =$$

$$= \int_0^X \sum_{j=1}^{N} C_j\,\frac{d^2\psi_j^N}{dx^2}\,\frac{d^2\psi_l^N}{dx^2}\,dx = \left(\sum_{j=1}^{N} C_j\,\frac{d^2\psi_j^N}{dx^2},\ \frac{d^2\psi_l^N}{dx^2}\right)\times$$

$$\times\left(\sum_{j=1}^{N} C_j\,\frac{d^2\psi_j^N}{dx^2},\ \frac{d^2\psi_l^N}{dx^2}\right) + a\left(\sum_{j=1}^{N} C_j\psi_j^N,\ \psi_l^N\right) = \eta(x). \qquad (17.14)$$

Отсюда следует, чтобы первая сумма в (17.14) вычислялась, желательно, чтобы базисы ψ_l^N были гладкими:

$$\psi_l^N \in W_2^2[0,\ X],$$

$$\|\psi_l^N\|^2_{W_2^2} = \int_0^X \left[(\psi_l^N)^2 + \left(\frac{d\psi_l^N}{dx}\right)^2 + \left(\frac{d^2\psi_l^N}{dx^2}\right)^2\right] dx,$$

а сходимость МКЭ следует понимать в норме $W_2^2[0,\ X]$.

Примеры *согласованных* базисных функций. Если используется базис из «крышечек», то в каждом узле (при стыковке конечных элементов) решение МКЭ будет иметь разрыв первой производной. Это происходит из-за выбора базиса МКЭ. Сама искомая функция непрерывна.

Допустим, необходимо найти решение, обладающее непрерывной первой производной.

Строим набор функций базиса:

$$\{\varphi_i(x)\}^m; \quad m = \frac{p+1}{2}; \quad (p \text{ — нечетное положительное число}).$$

Считаем, что размер конечного элемента равен 1. Для одномерной сетки всегда найдется линейное преобразование (свое для каждого элемента!), переводящее данный элемент в отрезок длины 1. Положим, что базисная функция есть

$$\varphi_i(x) \equiv 0, \text{ если } x \notin [-1;\ 1],$$

а на каждом отрезке $[-1;\ 0], [0;\ 1]$ — полином степени p. В точках $x = \pm 1$ $\varphi_i(x)$ и все ее производные до порядка $m-1$ равны нулю. В точке $x = 0$ $\frac{d^{i-1}\varphi_i(x)}{dx^{i-1}} = 1$.

Введем

$$\varphi_{ij}^h(x) = \varphi_i\left(\frac{x-a}{h} - j\right)$$

(в случае равномерного разбиения отрезка на конечные элементы). Тогда $\{\varphi_{ij}^h\}\ j = 0, \ldots, N; i = 1, \ldots, m$ - базис.

Рассмотрим случай $p = 1$. Тогда $m = p = 1$, на каждом отрезке функция линейна. Приходим к набору из «крышечек».

Возьмем $p = 3$, тогда $m = 2$. Строим набор базисных функций.

Фиксируем $i = 1$. На отрезках $[-1;\ 0]$, $[0;\ 1]$ получаем полином степени 3.

$$\varphi_1(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 (\text{на} [-1;\ 0]).$$

Условия: $\varphi_1'(-1) = 0, \varphi_1(0) = 1, \varphi_1(-1) = 0, \varphi_1'(0) = 0$ определяют коэффициенты

$$a_0 = 1; a_1 = 0; a_2 = -3; a_3 = -2.$$

В итоге на отрезке $[-1;\ 0]$

$$\varphi_1(x) = 1 - 3x^2 - 2x^3.$$

Аналогично поступаем на отрезке $[0;\ 1]$, там имеем

$$\varphi_1(x) = 1 - 3x^2 + 2x^3.$$

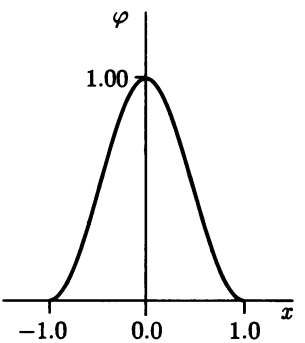


Рис. 17.9

График базисной функции $\varphi_1(x)$ представлен на рис. 17.9.

Пусть теперь $i = m = 2$. Строим набор $\varphi_2(x)$ такой, что

$$\varphi_2(x) = b_0 + b_1 x + b_2 x^2 + b_3 x^3.$$

Из условий $\varphi_2'(1) = 0, \varphi_2(1) = 0, \varphi_2(0) = 0, \varphi_2'(0) = 1$ получается

$$\varphi_2(x) = (1-x)^2 x \text{ при } 0 \leqslant x \leqslant 1.$$

Аналогично, $\varphi_2(x) = (1-x)^2 x$ при $-1 \leqslant x \leqslant 0$. График функции изображен на рис. 17.10.

Базис является согласованным, если для уравнения степени не выше $p+1$ все базисные функции непрерывны (принадлежат C^m).

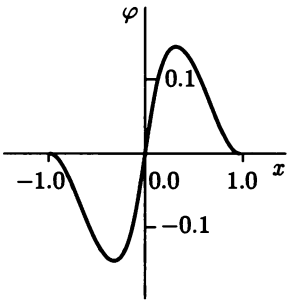


Рис. 17.10

Что представляет собой метод Галеркина при использовании такого базиса? Теперь в точках сетки (межэлементных) необходимо знать не только функцию u, но и ее первую, вторую, ..., $(m-1)$-ю производную по x:

$$u^h = \sum_{j=1}^{N} [u(a+jh)\varphi_{1j}^h(x) + u'_x(a+jh)\varphi_{2j}^h(x)],$$

$$u^h = \sum_{j=1}^{N} (c_j\psi_{1j}^N + b_j\varphi_{2j}^N(x)].$$

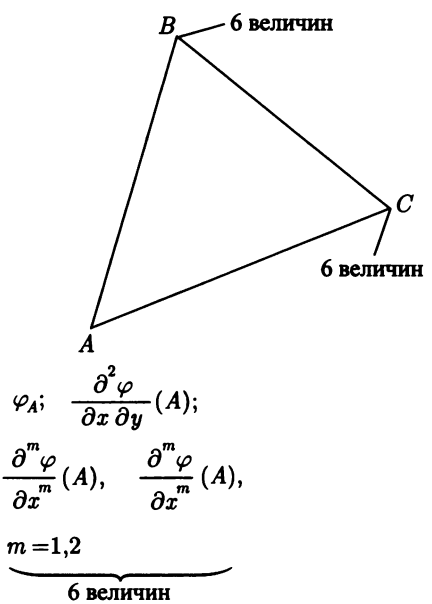


Рис. 17.11

Отметим, что $u(a+jh)$ и $u'_x(a+jh)$ определяются численно при решении уравнений методом Галеркина.

Увеличилось число базисных функций и коэффициентов разложения.

Заметим также, что матрица системы — разреженная, но уже не трехдиагональная (если порядок системы выше второго).

Согласование в двумерном случае. Надо сшивать следующие величины (рис. 17.11): 18 величин в узлах плюс 3 значения нормальных производных на гранях.

Получается 21 условие, значит необходимо иметь 21 произвольную константу. Полином должен иметь достаточно высокую степень (члены до x^5, y^5). Поэтому в многомерном случае, как правило, используются несогласованные базисные функции или с низким ($m = 1$) порядком согласования.

17.6. МКЭ для нестационарных уравнений

Рассмотрим простейшую МКЭ-аппроксимацию уравнения теплопроводности:

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2}$$

с соответствующими граничными и начальными условиями. Будем искать решение в виде

$$u = \sum_{j=1}^{N} C_j(t)\psi_j^N.$$

Используя подход Галеркина, получаем (в базисе из «крышечек»)

$$\frac{1}{6}\frac{dC_{j-1}}{dt} + \frac{2}{3}\frac{dC_j}{dt} + \frac{1}{6}\frac{dC_{j+1}}{dt} = \frac{D}{h^2}(C_{j-1} - 2C_j + C_{j+1}).$$

Это — система дифференциально-разностных уравнений. Теперь необходимо заменить производные по времени разностными отношениями.

Заметим, что «явная» схема (когда в правой части стоят коэффициенты разложения на предыдущем слое по времени C_j^n) уже не является явной, в соответствии с определением явных методов, данном выше:

$$\frac{1}{6}\frac{C_{j-1}^{n+1} - C_{j-1}^n}{\tau} + \frac{2}{3}\frac{C_j^{n+1} - C_j^n}{\tau} + \frac{1}{6}\frac{C_{j+1}^{n+1} - C_{j+1}^n}{\tau} = \frac{D}{h^2}(C_{j-1}^n - 2C_j^n + C_{j+1}^n)$$

и на $n+1$-м слое все равно необходимо решать систему уравнений методом прогонки. Причиной этого вычислительного неудобства является то, что система дифференциальных уравнений для определения зависимости коэффициентов разложения — это система обыкновенных дифференциальных уравнений, но не записанная в нормальной форме Коши.

Попытаемся исследовать схему на устойчивость спектральному признаку фон Неймана. Подставив в приведенное выше разностное уравнение

$$C_j^n = \lambda^n e^{ij\varphi},$$

получаем выражение для спектра оператора послойного перехода $\lambda(\varphi)$:

$$\frac{\lambda - 1}{6}[e^{i\varphi} + 4 + e^{-i\varphi}] = k\,[e^{i\varphi} + 2 + e^{-i\varphi}],$$

где $k = D\,\frac{\tau}{h^2}$. Отсюда видно, что устойчивость метода конечных элементов опять определяется безразмерной комбинацией параметров разбиения (размера конечного элемента, шага по времени) и коэффициента теплопроводности — параболическим аналогом числа Куранта. Преобразуем уравнение для λ:

$$\frac{\lambda - 1}{6} = \frac{k\,(2\,\cos\,\varphi - 2)}{4 + 2\,\cos\,\varphi},$$

$\lambda = 1 - 6k\,\dfrac{\cos\,\varphi - 1}{\cos\,\varphi + 2}$, откуда условием устойчивости будет $k \leqslant \dfrac{1}{6}$.

Имеет смысл пользоваться «неявной схемой» (правая часть берется с верхнего слоя по времени) или аппроксимацией типа Кранка-Никольсон.

Продолжим рассмотрение применения МКЭ к нестационарным уравнениям. Как и ранее, смотрим задачу для уравнения теплопроводности

$$\frac{\partial u}{\partial t} = D \frac{\partial^2 u}{\partial x^2}.$$

Выбирая базис из «крышечек», представляем решение в виде

$$u = \sum_{j=1}^{N} C_j(t) \psi_j^N.$$

Подставляя последнее уравнение в исходное и применяя стандартную процедуру метода Галеркина, получаем систему дифференциальных уравнений

$$\frac{1}{6} \frac{dC_{j-1}}{dt} + \frac{2}{3} \frac{dC_j}{dt} + \frac{1}{6} \frac{dC_{j+1}}{dt} = \frac{D}{h^2} (C_{j-1} - 2C_j + C_{j+1})$$

(шаг сетки считается постоянным), или, в матричном виде

$$\mathbf{B} \frac{d\mathbf{C}}{dt} = \frac{D}{h^2} \mathbf{A}\mathbf{C}. \tag{17.15}$$

При этом, в любом базисе

$$\mathbf{B} = \mathbf{B}^* > 0, \quad \mathbf{A} = \mathbf{A}^* > 0.$$

Тогда

$$\exists\, \mathbf{B}^{1/2}: \quad \mathbf{B}^{1/2}\mathbf{B}^{1/2} = \mathbf{B}.$$

Матрица $\mathbf{B}^{1/2}$ — самосопряженная положительно определенная. Можно записать последнее уравнение (17.15) в виде

$$\mathbf{B}^{1/2}\mathbf{B}^{1/2} \frac{d\mathbf{C}}{dt} = \frac{D}{h^2} \mathbf{A}\mathbf{B}^{-1/2}\mathbf{B}^{1/2}\mathbf{C}.$$

Введем вектор $\mathbf{z} \equiv \mathbf{B}^{-1/2}\mathbf{C}$ и умножим последнее соотношение слева на $\mathbf{B}^{-1/2}$, тогда получаем

$$\mathbf{B}^{1/2} \frac{d\mathbf{z}}{dt} = \frac{D}{h^2} \mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}\mathbf{z} = \frac{D}{h^2} \mathbf{P}\,\mathbf{z}. \tag{17.16}$$

Таким образом, из неявной системы (17.15) получена «явная» система (17.16) — перед вектором производных нет матричного множителя.

Запишем для (17.16) схему Кранка–Николсон:

$$\mathbf{z}^{n+1} - \mathbf{z}^n = \frac{D\tau}{2h^2}\,\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}(\mathbf{z}^{n+1} + \mathbf{z}^n). \qquad (17.17)$$

Вопрос об устойчивости схемы (17.17) можно решить следующим образом. Умножим (17.17) на $\mathbf{z}^{n+1} + \mathbf{z}^n$. Получаем соотношение:

$$(\mathbf{z}^{n+1},\ \mathbf{z}^{n+1}) - (\mathbf{z}^n,\ \mathbf{z}^n) = \frac{\sigma}{2}\,(\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}(\mathbf{z}^{n+1} + \mathbf{z}^n),\ (\mathbf{z}^{n+1} + \mathbf{z}^n));$$

$$||\mathbf{z}^{n+1}||^2 = ||\mathbf{z}^n||^2 + \frac{\sigma}{2}(\mathbf{P}(\mathbf{z}^{n+1}+\mathbf{z}^n), (\mathbf{z}^{n+1}+\mathbf{z}^n)) \leqslant ||\mathbf{z}^n||^2 - \frac{\sigma\lambda}{2}\,||\mathbf{z}^{n+1}+\mathbf{z}^n||^2$$

в силу того, что $\mathbf{A} < 0$ (спектр оператора $\mathbf{A}$ уже известен). Последнее неравенство и означает безусловную устойчивость метода.

17.7. Решение нелинейных уравнений с помощью МКЭ

Рассмотрим в качестве простейшего примера уравнение Хопфа

$$\frac{\partial u}{\partial t} + 6u\,\frac{\partial u}{\partial x} = 0. \qquad (17.18)$$

Его решение, как и ранее, ищем в виде (17.14), при этом по-прежнему используем базис из «крышечек». После вычислений получаем

$$\frac{1}{6}\,\frac{dC_{j-1}}{dt} + \frac{2}{3}\,\frac{dC_j}{dt} + \frac{1}{6}\,\frac{dC_{j+1}}{dt} - C_{j-1}^2 - C_jC_{j-1} + C_jC_{j+1} + C_{j+1}^2 = 0. \qquad (17.19)$$

Если написать дискретизацию (17.19) неявным образом (по величинам на $n+1$-м слое по времени или по аналогии со схемой Кранка–Николсон), то получается нелинейная относительно C_j^{n+1} система. Ее необходимо решать с помощью метода Ньютона. Можно линеаризовать (17.19) в окрестности C_j^n, считая, что

$$C_j^{n+1} \approx C_j^n + \tau\,\frac{dC_j}{dt}.$$

Тогда (17.19) преобразуется в следующую линейную относительно величин на $n+1$-м слое по времени запись:

$$\frac{1}{6}\,\frac{C_{j-1}^{n+1} - C_{j-1}^n}{\tau} + \frac{2}{3}\,\frac{C_j^{n+1} - C_j^n}{\tau} + \frac{1}{6}\,\frac{C_{j+1}^{n+1} - C_{j+1}^n}{\tau} - (C_{j-1}^n)^2 -$$
$$-2C_{j-1}^nC_{j-1}^{n+1} - C_j^nC_{j-1}^n - C_j^nC_{j-1}^{n+1} - C_j^{n+1}C_{j-1}^n +$$
$$+C_j^nC_{j+1}^n + C_j^nC_j^{n+1} + C_j^nC_{j+1}^{n+1} + (C_{j+1}^n)^2 + 2C_{j+1}^nC_{j+1}^{n+1} = 0.$$

Эту систему можно решать, используя метод немонотонной прогонки — гибрид метода прогонки и алгоритма Гаусса с выбором ведущего элемента.

Возможен и другой подход — линеаризация исходного уравнения (17.18) и решение получившейся последовательности линейных уравнений для определения приближенного решения задачи.

За рамками этой лекции, кроме технических подробностей, остались численные методы, основанные на применении минимизации функционала квадрата невязки — методы наименьших квадратов, а также методы граничных элементов для решения эллиптических задач.

17.8. Задачи для самостоятельного решения

1. При решении задачи с использованием метода Ритца на отрезке [0,4] используются глобальные базисы

   1) $\frac{x}{4}, x\,(4-x), x\,(16-x^2), \ldots, x\,(4^N - x^N)$;
   2) $\frac{x}{4}, x\,(4-x), x^2(4-x), \ldots, x^N(4-x)$;
   3) $\frac{x}{4}, \frac{x}{4}\,\left(1-\frac{x}{4}\right), \frac{x}{4}\,\left(1-\frac{x^2}{16}\right), \ldots, \frac{x}{4}\,\left(1-\frac{x^N}{4^N}\right)$.

   Описать достоинства и недостатки каждого из этих базисов.

Литература

[1] *Марчук Г.М., Агошков В.И.* Введение в проекционно-сеточные методы. М.: Наука, 1981. 414 с.

[2] *Стренг Г., Фикс Дж.* Теория метода конечных элементов. М.: Мир, 1977.

[3] *Самарский А.А., Гулин А.В.* Численные методы математической физики М.: Научный мир, 2003. 316 с.

[4] *Ши Д.* Математическое моделирование задач тепло- и массообмена. М.: Мир, 1988. 544 с.

[5] *Ректорис К.* Вариационные методы в математической физике и технике. М.: Мир, 1985. 590 с.

Лекция 18. Методы расщепления

Лекция знакомит с идеями построения экономичных разностных схем для уравнений математической физики, основанных на методах покомпонентного расщепления (локально-одномерные схемы) и на принципах расщепления по физическим процессам.

Ключевые слова: методы расщепления, схема Кранка–Никольсон, локально-одномерные схемы, расщепление по физическим процессам, расщепление с факторизацией оператора.

18.1. Понятие о методах расщепления

Рассмотрим дифференциальную задачу для уравнения в частных производных с постоянными коэффициентами:

$$\frac{\partial u}{\partial t} + \mathbf{A}u = 0; x \in \Omega_x,\ t \in \Omega_t,\ u|_\Gamma = u_\Gamma, u(t_0) = u_0. \tag{18.1}$$

Здесь оператор $\mathbf{A} \geqslant 0$ — положительный дифференциальный оператор с постоянными коэффициентами. В запись оператора $\mathbf{A}$ входят производные по пространственным переменным. Для любого ненулевого элемента выполнено $(\mathbf{A}\varphi, \varphi) \geqslant 0$. Γ — граница области интегрирования Ω_x; Λ — разностный оператор, аппроксимирующий $\mathbf{A}$. Можно проверить, что разностное уравнение

$$\frac{u^{n+1} - u^n}{\tau} + \Lambda\frac{u^{n+1} + u^n}{2} = 0, u^0 = u_0 \tag{18.2}$$

аппроксимирует (18.1) со вторым порядком по τ (схема Кранка-Никольсон). Заметим, что (18.2) можно трактовать как результат попеременного применения явной и неявной схем первого порядка аппроксимации, записанных на интервалах $[t^n, t^{n+1/2}], [t^{n+1/2}, t^{n+1}]$

$$\frac{u^{n+1/2} - u^n}{\tau/2} + \Lambda u^n = 0,$$

$$\frac{u^{n+1} - u^{n+1/2}}{\tau/2} + \Lambda u^{n+1} = 0. \tag{18.3}$$

Исключая из уравнений (18.3) значения функции на промежуточном слое по времени (с полуцелым индексом), получим (18.2). Если $\mathbf{A} = \mathbf{A}(t)$, то

$$\frac{u^{n+1}-u^n}{\tau}+\Lambda^n\frac{u^{n+1}+u^n}{2}=0 \tag{18.4}$$

при этом разностный оператор также является положительным:

$$(\Lambda_n u, u) \geqslant 0,$$

а решение на следующем слое по времени может быть записано в операторном виде следующим образом:

$$u^{n+1}=(\mathbf{E}+\frac{\tau}{2}\Lambda^n)^{-1}(\mathbf{E}-\frac{\tau}{2}\Lambda^n)u^n,$$

или

$$u^{n+1}=\mathbf{T}^n u^n,$$

где $\mathbf{T}^n=(\mathbf{E}+\frac{\tau}{2}\Lambda^n)^{-1}(\mathbf{E}-\frac{\tau}{2}\Lambda^n)$.

Для доказательства устойчивости полученного разностного уравнения умножим скалярно (18.4) на $(u^n+u^{n+1})/2$, получим

$$\frac{(u^{n+1},u^{n+1})-(u^n,u^n)}{2\tau}+\left(\Lambda^n\frac{u^{n+1}+u^n}{2},\frac{u^{n+1}+u^n}{2}\right)=0 \tag{18.5}$$

Так как в силу положительности разностного оператора $(\Lambda^n u, u) \geqslant 0$, то из (18.5) следует, что $\left\|u^{n+1}\right\| \leqslant \|u^n\|$, чем и обеспечена устойчивость схемы. Если разностный оператор Λ (пространственные разности) выбран в виде полусуммы разностных операторов на верхнем и нижнем слоях по времени $\frac{1}{2}(\Lambda^{n+1}+\Lambda^n)=\Lambda^{n+1/2}$, то схема имеет второй порядок аппроксимации по τ.

18.2. Метод расщепления первого и второго порядка точности по τ

18.2.1. Локально-одномерные схемы

Положим, что дифференциальный оператор $\mathbf{A}$ и соответствующий ему разностный оператор Λ можно представить в виде суммы операторов, каждый из которых включает производные лишь по одной пространственной переменной и разности лишь вдоль одного направления соответственно. Всего пространственных направлений N. Такие дифференциальные и разностные операторы будем называть *локально-одномерными*.

И дифференциальный, и разностный операторы записываются в виде суммы локально-одномерных:

$$\mathbf{A} = \sum_{i=1}^{N} \mathbf{A}_i, \Lambda = \sum_{i}^{N} \Lambda_i.$$

Для однородной задачи можно выписать схему *расщепления по направлениям*:

$$\frac{u^{n+1/N} - u^n}{\tau} + \Lambda_1 u^{n+1/N} = 0,$$

$$\frac{u^{n+2/N} - u^{n+1/N}}{\tau} + \Lambda_2 u^{n+2/N} = 0,$$

$$\dots$$

$$\frac{u^{n+1} - u^{n+(N-1)/N}}{\tau} + \Lambda_N u^{n+1} = 0.$$

Получена система разностных уравнений, каждое из которых не аппроксимирует исходное дифференциальное, но может быть легко решено (методом прогонки вдоль соответствующего направления, если разностные операторы содержат лишь первые и вторые разности). Тем не менее, последовательно примененные друг за другом, они дают на следующем слое по времени решение с разумной точностью. Говорят, что имеет место *суммарная аппроксимация* — результирующий оператор послойного перехода получился аппроксимирующим. Описанный выше способ называется иногда методом дробных шагов, и уже встречался при решении многомерного уравнения теплопроводности.

Для неоднородной задачи один из возможных вариантов схемы расщепления имеет вид

$$\frac{u^{n+1/N} - u^n}{\tau} + \Lambda_1 u^{n+1/N} = 0,$$

$$\frac{u^{n+2/N} - u^{n+1/N}}{\tau} + \Lambda_2 u^{n+2/N} = 0,$$

$$\dots$$

$$\frac{u^{n+1} - u^{n+(N-1)/N}}{\tau} + \Lambda_N u^{n+1} = f^n.$$

Возможны и другие способы учета правой части, например, введение ее во все уравнения с весовыми множителями, которые подбираются из условий наилучшей суммарной аппроксимации (минимизации ошибки аппроксимации на следующем слое по времени).

Приведенные выше схемы расщепления по направлениям абсолютно устойчивы.

18.2.2. Схемы Кранка–Никольсон

Рассмотрим обобщение схемы Кранка–Никольсон на случай многомерных уравнений с локально-одномерными операторами. Положим, как и ранее, $\Lambda = \sum_{i}^{N} \Lambda_i$. Если коэффициенты разностного оператора явно зависят от времени, они берутся на промежуточном временном слое $\Lambda = \Lambda(t^{n+1/2})$. Для простоты изложения рассмотрим двумерный случай.

Схему расщепления по направлениям представим в виде

$$\frac{u^{n+1/2} - u^n}{\tau} + \Lambda_1 \frac{u^{n+1/2} + u^n}{2} = 0$$

$$\frac{u^{n+1} - u^{n+1/2}}{\tau} + \Lambda_2 \frac{u^{n+1} + u^{n+1/2}}{2} = 0,$$

а решение на следующем слое по времени в операторной форме выписывается как $u^{n+1} = \mathbf{T}^n u^n$. Для оператора послойного перехода получается следующая формула:

$$\mathbf{T}^n = (\mathbf{E} + \frac{\tau}{2}\Lambda_2)^{-1}(\mathbf{E} - \frac{\tau}{2}\Lambda_2)(\mathbf{E} + \frac{\tau}{2}\Lambda_1)^{-1}(\mathbf{E} - \frac{\tau}{2}\Lambda_1).$$

При выполнении условия

$$\frac{\tau}{2}\|\Lambda_i\| < 1$$

схема устойчива, обладает вторым порядком аппроксимации по времени, если операторы Λ_1, Λ_2 коммутативны, и первым — если нет.

18.2.3. Общая формулировка методов расщепления

Заменим локально-одномерные дифференциальные операторы $\mathbf{A}_i$ разностными операторами на каждом шаге по времени $t_n \leqslant t \leqslant t_{n+1}$.

Представим схему расщепления в следующем общем виде:

$$\frac{u^{n+1/N} - u^n}{\tau} + \Lambda_{10} u^n + \Lambda_{11} u^{n+1/N} = 0,$$

$$\frac{u^{n+2/N} - u^{n+1/N}}{\tau} + \Lambda_{20} u^n + \Lambda_{21} u^{n+1/N} + \Lambda_{22} u^{n+2/N} = 0,$$

$$\ldots$$

$$\frac{u^{n+1} - u^{n+\frac{N-1}{N}}}{\tau} + \Lambda_{N0} u^n + \Lambda_{N1} u^{n+1/2} + \ldots + \Lambda_{NN} u^{n+1} = 0.$$

Условие устойчивости такой схемы расщепления будет

$$\|C_i C_{i-1} \dots C_1\| \leqslant 1 + c\tau, \quad c = \text{const},$$

где $C_i = (\mathbf{E} + \tau\Lambda_{ii})^{-1}(\mathbf{E} + \tau\Lambda_{i,i-1}), i = 1, 2, \dots, N$

Двухслойная схема расщепления с весовыми коэффициентами представлена в виде

$$\frac{u^{n+1/N} - u^n}{\tau} + \Lambda_1 \left[(1-\sigma)u^n + \sigma u^{n+1/N}\right] = 0,$$

$$\frac{u^{n+2/N} - u^{n+1/N}}{\tau} + \Lambda_2 \left[(1-\sigma)u^{n+1/N} + \sigma u^{n+2/N}\right] = 0,$$

$$\dots$$

$$\frac{u^{n+1} - u^{n+\frac{N-1}{N}}}{\tau} + \Lambda_N \left[(1-\sigma)u^{n+\frac{N-1}{N}} + \sigma u^{n+1}\right] = 0.$$

Если в этой схеме расщепления положить веса верхнего и нижнего слоев по времени равными, $\gamma = 0,5$, то в случае коммутирующих операторов Λ_i (каждый такой разностный оператор аппроксимирует со вторым порядком соответствующий локально-одномерный дифференциальный оператор) схема будет иметь второй порядок аппроксимации и по времени. Если при этом каждый оператор $\Lambda_i \geqslant 0$, то схема будет абсолютно устойчивой.

18.2.4. Схемы расщепления для уравнения теплопроводности

Рассматриваем нестационарное уравнение теплопроводности

$$\frac{\partial u}{\partial t} - \Delta u = 0, t \in \Omega_t, \{x, y, z\} \in \Omega.$$

Здесь оператор Лапласа определен как $\Delta = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$. Его также можно записать в виде суммы трех локально-одномерных операторов $\mathbf{A} = \mathbf{A}_x + \mathbf{A}_y + \mathbf{A}_z$. Соответствующие разностные операторы будут $\Lambda = \Lambda_{xx} + \Lambda_{yy} + \Lambda_{zz}$, где $\Lambda_{xx} = \frac{u_{m-1,jk} - 2u_{mjk} + u_{m+1,jk}}{h_x^2}$ аналогично определяются операторы вычисления второй разностной производной и по остальным направлениям $\Lambda_{yy}, \Lambda_{zz}$

Локально-одномерная схема для уравнения теплопроводности будет

$$\frac{u^{n+1/3} - u^n}{\tau} + \Lambda_{xx} u^{n+1/3} = 0,$$

$$\frac{u^{n+2/3} - u^{n+1/3}}{\tau} + \Lambda_{yy} u^{n+2/3} = 0,$$

$$\frac{u^{n+1} - u^{n+2/3}}{\tau} + \Lambda_{zz} u^{n+1} = 0.$$

Для повышения порядка аппроксимации можно использовать схему с весами

$$\frac{u^{n+1/3}-u^n}{\tau}+\Lambda_{xx}\left[(1-\sigma)u^n+\sigma u^{n+1/3}\right]=0,$$

$$\frac{u^{n+2/3}-u^{n+1/3}}{\tau}+\Lambda_{yy}\left[(1-\sigma)u^{n+1/3}+\sigma u^{n+2/3}\right]=0,$$

$$\frac{u^{n+1}-u^{n+2/3}}{\tau}+\Lambda_{zz}[(1-\sigma)u^{n+2/3}+\sigma u^{n+1}]=0.$$

18.3. Методы двуциклического покомпонентного расщепления

Для этих методов отсутствует требование коммутативности операторов Λ_i.

Будем рассматривать численное решение (18.1) не на одном шаге по времени, отрезке $[t^n, t^{n+1}]$, а на двух последовательных шагах $[t^{n-1}, t^{n+1}]$. Пусть теперь разностные локально-одномерные операторы зависят явно от времени, тогда они определены в середине отрезка $\Lambda_i = \Lambda_i(t^n)$. Запишем схему расщепления:

$$\frac{u^{n-1/2}-u^{n-1}}{\tau}+\Lambda_1\frac{u^{n-1/2}+u^{n-1}}{2}=0,$$

$$\frac{u^{n}-u^{n-1/2}}{\tau}+\Lambda_2\frac{u^{n}+u^{n-1/2}}{2}=0,$$

$$\frac{u^{n+1/2}-u^{n}}{\tau}+\Lambda_2\frac{u^{n+1/2}-u^{n}}{2}=0,$$

$$\frac{u^{n+1}-u^{n+1/2}}{\tau}+\Lambda_1\frac{u^{n+1}-u^{n+1/2}}{2}=0. \qquad (18.6)$$

В операторной форме этот метод записывается как $u^{n+1}=\mathbf{T}^n u^{n-1}$, где введено обозначение

$$\mathbf{T}^n=\left(\mathbf{E}+\frac{\tau}{2}\Lambda_1\right)^{-1}\left(\mathbf{E}-\frac{\tau}{2}\Lambda_1\right)\left(\mathbf{E}+\frac{\tau}{2}\Lambda_2\right)^{-1}\left(\mathbf{E}-\frac{\tau}{2}\Lambda_2\right)\times$$

$$\times\left(\mathbf{E}+\frac{\tau}{2}\Lambda_2\right)^{-1}\left(\mathbf{E}-\frac{\tau}{2}\Lambda_2\right)\left(\mathbf{E}+\frac{\tau}{2}\Lambda_1\right)^{-1}\left(\mathbf{E}-\frac{\tau}{2}\Lambda_2\right)=\mathbf{E}-2\tau\Lambda+\frac{(2\tau)^2}{2}(\Lambda)^2+\ldots$$

Если локально-одномерные операторы положительны $\mathbf{A}_i(t) > 0$, то при достаточной гладкости решения и элементов матриц $\mathbf{A}_i(t)$ схема (18.6) абсолютно устойчива и аппроксимирует (18.1) со вторым порядком.

Для неоднородного дифференциального уравнения $\frac{\partial u}{\partial t}+Au=f$ разностная аппроксимация метода расщепления может быть представлена в виде

$$\left(\mathbf{E}+\frac{\tau}{2}\Lambda_1\right)u^{n-1/2}=\left(\mathbf{E}-\frac{\tau}{2}\Lambda_1\right)u^{n-1},$$

$$\left(\mathbf{E}+\frac{\tau}{2}\Lambda_2\right)(u^n-\tau f^n)=\left(\mathbf{E}-\frac{\tau}{2}\Lambda_2\right)u^{n-1/2},$$

$$\left(\mathbf{E}+\frac{\tau}{2}\Lambda_2\right)u^{n+1/2}=\left(\mathbf{E}-\frac{\tau}{2}\Lambda_2\right)(u^n+\tau f^n),$$

$$\left(\mathbf{E}+\frac{\tau}{2}\Lambda_1\right)u^{n+1}=\left(\mathbf{E}-\frac{\tau}{2}\Lambda_1\right)u^{n+1/2},$$

где $f^n=f(t_n)$.

В операторной форме записи решение неоднородной задачи имеет вид: $u^{n+1}=\mathbf{T}^nu^{n-1}+2\tau\mathbf{T}_1^n\mathbf{T}_2^nf^n$, где введены обозначения $\mathbf{T}^n=\mathbf{T}_1^n\mathbf{T}_2^n\mathbf{T}_2^n\mathbf{T}_1^n$ $\mathbf{T}_i^n=\left(\mathbf{E}+\frac{\tau}{2}\Lambda_i^n\right)^{-1}\left(\mathbf{E}-\frac{\tau}{2}\Lambda_i^n\right)$.

Представим разностную аппроксимацию неоднородного дифференциального уравнения с помощью последовательного применения операторов $\Lambda_1,\ldots,\Lambda_N,\Lambda_N,\ldots,\Lambda_1$:

$$\left(\mathbf{E}+\frac{\tau}{2}\Lambda_1\right)u^{n-\frac{N-1}{N}}=\left(\mathbf{E}-\frac{\tau}{2}\Lambda_1\right)u^{n-1},$$

$$\ldots$$

$$\left(\mathbf{E}+\frac{\tau}{2}\Lambda_N\right)(u^n-\tau f^n)=\left(\mathbf{E}-\frac{\tau}{2}\Lambda_N\right)u^{n-1/N},$$

$$\ldots$$

$$\left(\mathbf{E}+\frac{\tau}{2}\Lambda_N\right)u^{n+1/N}=\left(\mathbf{E}-\frac{\tau}{2}\Lambda_N\right)(u^n+\tau f^n),$$

$$\ldots$$

$$\left(\mathbf{E}+\frac{\tau}{2}\Lambda_1\right)u^{n+1}=\left(\mathbf{E}-\frac{\tau}{2}\Lambda_1\right)u^{n+\frac{N-1}{N}}.$$

Рассмотрим примеры использования метода двуциклического расщепления для некоторых задач математической физики.

Пример 1. Трехмерное нестационарное уравнение диффузии, область интегрирования — параллелепипед. Полагаем, что в вертикальном направлении (ось $0z$) коэффициент диффузии в вертикальной плоскости γ зависит от координаты, что характерно для задач геофизики, μ — коэффициент диффузии в горизонтальной плоскости. Задача может быть представлена в виде

$$\frac{\partial u}{\partial t}=\frac{\partial}{\partial z}\gamma\frac{\partial u}{\partial z}+\mu\Delta u+f.$$

Сведем решение рассматриваемой трехмерной задачи к последовательному решению трех одномерных задач. Первая задача имеет вид

$$\frac{\partial u_1}{\partial t} = \frac{\partial}{\partial z}\gamma\frac{\partial u_1}{\partial z} + f,$$

она описывает диффузию в вертикальной плоскости. Вторую и третью задачи запишем

$$\frac{\partial u_2}{\partial t} = \mu\frac{\partial^2 u_2}{\partial x^2}, \frac{\partial u_3}{\partial t} = \mu\frac{\partial^2 u_3}{\partial y^2}.$$

Теперь рассмотрим разностную аппроксимацию исходного дифференциального уравнения

$$\frac{\partial u}{\partial t} + (\Lambda_1 + \Lambda_2 + \Lambda_3)u = f, \text{ где}$$

$$\Lambda_1 u = -\frac{\mu}{h_x^2}(u_{m,j,k+1} - 2u_{mjk} + u_{m,j,k-1}),$$

$$\Lambda_2 u = -\frac{\mu}{h_y^2}(u_{m,j-1,k} - 2u_{mjk} + u_{m,j+1,k}),$$

$$\Lambda_3 u = \frac{1}{h_z}\left[-\frac{\gamma_{m+1/2,jk}}{h_z}(u_{m+1,jk} - u_{mjk}) + \frac{\gamma_{m-1/2}}{h_z}(u_{mjk} - u_{m-1,jk})\right].$$

Разностная схема двуциклического покомпонентного расщепления приобретает вид

$$\frac{u^{n+1/6} - u^n}{\tau} + \Lambda_1\frac{u^{n+1/6} + u^n}{2} = 0,$$

$$\frac{u^{n+2/6} - u^{n+1/6}}{\tau} + \Lambda_2\frac{u^{n+2/6} + u^{n+1/6}}{2} = 0,$$

$$\frac{u^{n+3/6} - u^{n+2/6}}{\tau} + \Lambda_3\frac{u^{n+3/6} + u^{n+2/6}}{2} = \frac{f^{n+1/2}}{2},$$

$$\frac{u^{n+4/6} - u^{n+3/6}}{\tau} + \Lambda_3\frac{u^{n+4/6} + u^{n+3/6}}{2} = \frac{f^{n+1/2}}{2},$$

$$\frac{u^{n+5/6} - u^{n+4/6}}{\tau} + \Lambda_2\frac{u^{n+5/6} + u^{n+4/6}}{2} = 0,$$

$$\frac{u^{n+1} - u^{n+5/6}}{\tau} + \Lambda_1\frac{u^{n+1} + u^{n+5/6}}{2} = 0.$$

Пример 2. Сопряженное нестационарное уравнение переноса и диффузии

$$\frac{\partial u}{\partial t} + (\mathbf{A}_1 + \mathbf{A}_2 + \mathbf{A}_3)u = f,$$

где операторы определены как

$$\mathbf{A}_1 = \frac{\partial(v_1 u)}{\partial x} - \mu\frac{\partial^2 u}{\partial x^2},\quad \mathbf{A}_2 = \frac{\partial(v_2 u)}{\partial y} - \mu\frac{\partial^2 u}{\partial y^2},\quad \mathbf{A}_3 = \frac{\partial(v_3 u)}{\partial z} - \frac{\partial}{\partial z}\gamma\frac{\partial u}{\partial z} + \sigma u.$$

Здесь v_1, v_2, v_3 — компоненты вектора скорости, u — концентрация субстанции, σ — коэффициент поглощения субстанции внешней средой, $\sigma > 0$.

Соответствующую схему расщепления представим в виде

$$\frac{u^{n+1/6} - u^n}{\tau} + \Lambda_1\frac{u^{n+1/6} + u^n}{2} = 0,$$

$$\frac{u^{n+2/6} - u^{n+1/6}}{\tau} + \Lambda_2\frac{u^{n+2/6} + u^{n+1/6}}{2} = 0,$$

$$\frac{u^{n+3/6} - u^{n+2/6}}{\tau} + \Lambda_3\frac{u^{n+3/6} + u^{n+2/6}}{2} = \frac{f^{n+1/2}}{2},$$

$$\frac{u^{n+4/6} - u^{n+3/6}}{\tau} + \Lambda_3\frac{u^{n+4/6} + u^{n+3/6}}{2} = \frac{f^{n+1/2}}{2},$$

$$\frac{u^{n+5/6} - u^{n+4/6}}{\tau} + \Lambda_2\frac{u^{n+5/6} + u^{n+4/6}}{2} = 0,$$

$$\frac{u^{n+1} - u^{n+5/6}}{\tau} + \Lambda_1\frac{u^{n+1} + u^{n+5/6}}{2} = 0,$$

где разностные операторы аппроксимируют соответствующие локально-одномерные дифференциальные операторы. Так для рассматриваемой задачи

$$\Lambda_1 u = -\frac{\mu}{h_x^2}(u_{m+1,jk} - 2u_{mjk} + u_{m-1j,k}) + \frac{(v_1 u)_{m+1,jk} - (v_1 u)_{m-1,jk}}{2h_x},$$

$$\Lambda_2 u = -\frac{\mu}{h_y^2}(u_{m,j+1,k} - 2u_{mjk} + u_{m,j-1,k}) + \frac{(v_2 u)_{m,j+1,k} - (v_2 u)_{m,j-1,k}}{2h_y},$$

$$\Lambda_3 u = \frac{1}{h_z}\left[-\frac{\gamma_{m+1/2,jk}}{h_z}(u_{m+1,jk} - u_{mjk}) + \frac{\gamma_{m-1/2}}{h}(u_{mjk} - u_{m-1,jk})\right] +$$

$$+\frac{(v_3 u)_{m,j,k+1} - (v_3 u)_{m,j,k-1}}{2h_z} + \sigma u_{mjk}.$$

Отметим, что в рассматриваемой задаче на каждом этапе может быть проведено расщепление и по физическим процессам. Для простоты рассмотрим двухмерное уравнение конвекции-диффузии

$$\frac{\partial u}{\partial t} + \frac{\partial(v_1, u)}{\partial x} + \frac{\partial(v_2, u)}{\partial y} - \mu\Delta u - \sigma u = f,$$

в котором компоненты скорости движения среды v_1, v_2 удовлетворяют уравнению неразрывности:

$$\frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y} = 0.$$

Первый этап расщепления задачи по физическим процессам связан с переносом, на нем решается разностный аналог уравнения

$$\frac{\partial u_1}{\partial t} + \frac{\partial (v_1, u)}{\partial x} + \frac{\partial (v_2, u)}{\partial y} = 0.$$

Второй этап расщепления описывает процессы диффузии и поглощения субстанций

$$\frac{\partial u_2}{\partial t} - \mu \Delta u + \sigma u_2 = f.$$

Пример 3. *Расщепление по физическим процессам* системы уравнений газовой динамики (метод крупных частиц).

Система

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho \mathbf{v}) = 0,$$

$$\frac{\partial (\rho u_1)}{\partial t} + \operatorname{div}(\rho \mathbf{v} u_1) + \frac{\partial P}{\partial x} = 0,$$

$$\frac{\partial (\rho u_2)}{\partial t} + \operatorname{div}(\rho \mathbf{v} u_2) + \frac{\partial P}{\partial y} = 0,$$

$$\frac{\partial (\rho e)}{\partial t} + \operatorname{div}(\rho e \mathbf{v}) + \operatorname{div}(P \mathbf{v}) = 0,$$

$$P = P(\rho, \varepsilon), e = \varepsilon + \frac{u_1^2 + u_2^2}{2},$$

u_1, u_2 — компоненты вектора скорости $\mathbf{v}$, P — давление газа, ρ — плотность, ε — внутренняя энергия.

Первый (Эйлеров) этап. Измеряются лишь величины, относящиеся к ячейке в целом, а жидкость внутри каждой ячейки сетки считается моментально замороженной. Расчет производится по формулам, аппроксимирующим уравнения

$$\frac{\partial \rho}{\partial t} = 0,$$

$$\rho \frac{\partial u_1}{\partial t} + \frac{\partial P}{\partial x} = 0, \rho \frac{\partial u_2}{\partial t} + \frac{\partial P}{\partial y} = 0,$$

$$\rho \frac{\partial e}{\partial t} + \operatorname{div}(P \mathrm{v}) = 0.$$

На втором (Лагранжевом) этапе происходит движение газа массы через границы эйлеровых ячеек и перераспределение массы, импульса, энергии по пространству; определяются поля параметров течения газа. Аппроксимируется система уравнений

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho \mathbf{v}) = 0,$$

$$\frac{\partial(\rho u_1)}{\partial t} + \operatorname{div}(\rho u_1 \mathbf{v}) = 0, \frac{\partial(\rho u_2)}{\partial t} + \operatorname{div}(\rho u_2 \mathbf{v}) = 0,$$

$$\frac{\partial(\rho e)}{\partial t} + \operatorname{div}(\rho e \mathbf{v}) = 0.$$

18.4. Методы расщепления с факторизацией оператора

18.4.1. Факторизованная схема расщепления

Пусть для решения дифференциальной задачи

$$\mathbf{B}\frac{\partial u}{\partial t} + \mathbf{A}u = f, u(t_0) = u_0$$

используется разностная схема $\mathbf{B}u^{n+1} = F^n$, где $F^n = (\mathbf{B} - \tau\mathbf{A})u^n + \tau f^n$, $n = 0, 1, \ldots$

Пусть для вычисления F^n затрачивается $O(N)$ действий, число арифметических операций пропорционально числу узлов сетки N. Такие разностные операторы называются *экономичными*.

Пусть $\mathbf{B}_i$ $(i = 1, 2, \ldots, N)$ — экономичные разностные операторы, такие, что $\mathbf{B}_i v = F$.

Назовем схему *разностной схемой с факторизованным оператором* $\mathbf{B}$, если возможно его представление в виде

$$\mathbf{B} = \mathbf{B}_1\mathbf{B}_2 \ldots \mathbf{B}_N.$$

Эта схема будет также экономичной, так как для решения разностного уравнения по-прежнему потребуется $O(N)$ действий. В самом деле, решение уравнения

$$\mathbf{B}_1\mathbf{B}_2 \ldots \mathbf{B}_N u^{n+1} = F^n$$

может быть найдено в результате последовательного решения p уравнений

$$\mathbf{B}_1 u_1 = F^n,$$

$$\mathbf{B}_2 u_2 = u_1,$$

$$\mathbf{B}_3 u_3 = u_2,$$

$$\cdot \quad \ldots$$

$$\mathbf{B}_i u_i = u_{i-1},$$

здесь $i = 2, 3, \ldots, N$. Тогда $u^{n+1} = u_N$. В записи задачи введены обозначения $u_1 = u^{n+1/N}, \ldots, u_i = u^{n+i/N}, \ldots, u_{N-1} = u^{n+\frac{N-1}{N}}$ — промежуточные значения.

Схемы с факторизованным оператором иногда называются также *факторизованными схемами*. Устойчивая схема с факторизованным оператором $\mathbf{B}$, которая представляет собой произведение конечного числа операторов $\mathbf{B}_1, \ldots, \mathbf{B}_N$, является экономичной схемой.

Пример. Метод переменных направлений (продольно-поперечная схема). Приведем запись схемы для решения линейного двумерного уравнения теплопроводности. Расчетные формулы есть

$$\frac{u^{n+1/2} - u^n}{\frac{1}{2}\tau} - (\Lambda_1 u^{n+1/2} + \Lambda_2 u^n) = f^n,$$

$$\frac{u^{n+1} - u^{n+1/2}}{\frac{1}{2}\tau} - (\Lambda_1 u^{n+1/2} + \Lambda_2 u^{n+1}) = f^n.$$

Тогда, исключая $u^{n+1/2}$, получим в операторной форме записи

$$\left(\mathbf{E} - \frac{\tau}{2}\Lambda_1\right)\left(\mathbf{E} - \frac{\tau}{2}\Lambda_2\right) u^{n+1} = \left(\mathbf{E} + \frac{\tau}{2}\Lambda_1\right)\left(\mathbf{E} + \frac{\tau}{2}\Lambda_2\right) u^n,$$

или $\mathbf{B}_1\mathbf{B}_2 u^{n+1} = F^n$, где $\mathbf{B}_1 = \mathbf{E} - \frac{\tau}{2}\Lambda_1$, $\mathbf{B}_2 = \mathbf{E} - \frac{\tau}{2}\Lambda_2$, $F^n = \left(\mathbf{E} + \frac{\tau}{2}\Lambda_1\right)\left(\mathbf{E} + \frac{\tau}{2}\Lambda_2\right) u^n$.

Разностная схема может быть представлена в виде факторизованный схемы расщепления:

$$\mathbf{B}_1 u_1 = F^n, \mathbf{B}_2 u^{n+1} = u_1.$$

18.4.2. Неявная схема расщепления с приближенной факторизацией

Рассмотрим неявную разностную схему

$$\frac{u^{n+1} - u^n}{\tau} + \Lambda u^{n+1} = 0, n = 0, 1, \ldots, \Lambda = \sum_{i=1}^{N} \Lambda_i, \Lambda_i > 0. \tag{18.7}$$

Представим разностную схему (18.7) в виде

$$(\mathbf{E} + \tau\Lambda) u^{n+1} = u^n. \tag{18.8}$$

Факторизуем разностную схему (18.8) приближенно с точностью до членов порядка $O(\tau^2)$. Для этого заменим в (18.8) оператор $\mathbf{E} + \tau\Lambda$ на факторизованный

$$(\mathbf{E} + \tau\Lambda_1)(\mathbf{E} + \tau\Lambda_2)\dots(\mathbf{E} + \tau\Lambda_N) = \mathbf{E} + \tau\Lambda + \tau^2\mathbf{R},$$

где введено обозначение

$$\mathbf{R} = \sum_{i<j} \Lambda_i\Lambda_j + \tau \sum_{i<j<k} \Lambda_i\Lambda_j\Lambda_k + \dots + \tau^{n-2}\Lambda_1\dots\Lambda_n.$$

Врезультате приходим к неявной схеме с приближенной факторизацией

$$\mathbf{B}u^{n+1} = u^n, \mathbf{B} = \prod_{i=1}^{n} \mathbf{B}_i, \mathbf{B}_i = \mathbf{E} + \tau\Lambda_i,$$

или

$$(\mathbf{E} + \tau\Lambda_1)u^{n+1/N} = u^n,$$
$$(\mathbf{E} + \tau\Lambda_2)u^{n+2/N} = u^{n+1/N},$$
$$\dots$$
$$(\mathbf{E} + \tau\Lambda_N)u^{n+1} = u^{n+\frac{N-1}{N}}.$$

Эта схема абсолютно устойчива, имеет первый порядок аппроксимации.

18.4.3. Метод «предиктор-корректор»

Основная идея методов типа «предиктор-корректор» заключается в следующем. На каждом отрезке $[t^n,\ t^{n+1}]$ задача решается в два приема: сначала по схеме первого порядка аппроксимации и со значительным запасом устойчивости находится решение в момент времени $t^{n+1/2} = t^n + \tau/2$ — *предиктор*. После этого на втором этапе расписывается исходное уравнение по схеме более высокого порядка аппроксимации (чаще всего, второго) — корректор. Основная идея семейств таких методов близка к идее построения методов типа Рунге-Кутты для обыкновенных дифференциальных уравнений.

Представим эту схему как следующую схему расщепления:

$$\frac{u^{n+1/4} - u^n}{\tau/2} + \Lambda_1 u^{n+1/4} = 0,$$
$$\frac{u^{n+1/2} - u^{n+1/4}}{\tau/2} + \Lambda_2 u^{n+1/2} = 0,$$
$$\frac{u^{n+1} - u^n}{\tau} + \Lambda u^{n+1/2} = 0.$$

Если в этой схеме расщепления исключить $u^{n+1/4}$, то получим последовательность расчетных формул

$$\left(\mathbf{E}+\frac{\tau}{2}\Lambda_1\right)\left(\mathbf{E}+\frac{\tau}{2}\Lambda_2\right)u^{n+1/2}=\varphi^n,$$

$$\frac{u^{n+1}-u^n}{\tau}+\Lambda u^{n+1/2}=0,$$

далее, исключив, $\Lambda u^{n+1/2}$, получим

$$\frac{u^{n+1}-u^n}{\tau}+\Lambda\left(\mathbf{E}+\frac{\tau}{2}\Lambda_2\right)^{-1}\left(\mathbf{E}+\frac{\tau}{2}\Lambda_1\right)^{-1}u^n=0.$$

Если для разностных операторов выполнены условия $\frac{\tau}{2}\|\Lambda_i\|<1$, $\Lambda_1\geqslant 0$, $\Lambda_2\geqslant 0$, а коэффициенты разностной схемы явно не зависят от времени, то при достаточной гладкости решения дифференциальной задачи разностная схема абсолютно устойчива и аппроксимирует исходную задачу со вторым порядком.

Далее рассмотрим случай, когда оператор Λ представляется в виде суммы операторов Λ_i: $\Lambda=\sum\limits_i\Lambda_i$. Пусть все эти разностные операторы положительны. Метод «предиктор-корректор» можно записать в виде последовательности расчетных формул

$$\left(\mathbf{E}+\frac{\tau}{2}\Lambda_1\right)u^{n+1/2N}=u^n+\frac{\tau}{2}f^{n+1/2},$$

$$\left(\mathbf{E}+\frac{\tau}{2}\Lambda_2\right)u^{n+2/2N}=u^{n+1/2N},$$

$$\ldots$$

$$\left(\mathbf{E}+\frac{\tau}{2}\Lambda_N\right)u^{n+1/2}=u^{n+\frac{N}{2N}},$$

$$\frac{u^{n+1}-u^n}{\tau}+\Lambda u^{n+1/2}=f^{n+1/2}.$$

Эта последовательность после исключения промежуточных этапов сводится к одному разностному уравнению

$$\frac{u^{n+1}-u^n}{\tau}+\Lambda\prod_{i=N}^{1}\left(\mathbf{E}+\frac{\tau}{2}\Lambda_i\right)^{-1}(u^n+\frac{\tau}{2}f^{n+1/2})=f^{n+1/2}.$$

Приведем пример построения такой схемы. Для нестационарного трехмерного уравнения теплопроводности

$$\frac{\partial u}{\partial t}-a^2\sum_{i=1}^{3}\frac{\partial^2 u}{\partial x_i^2}=0$$

получим следующую разностную схему типа «предиктор-корректор»

$$\frac{u^{n+1/6}-u^n}{\tau/2}+\Lambda_1 u^{n+1/6}=0,$$

$$\frac{u^{n+1/3}-u^{n+1/6}}{\tau/2}+\Lambda_2 u^{n+1/3}=0,$$

$$\frac{u^{n+1/2}-u^{n+1/3}}{\tau/2}+\Lambda_3 u^{n+1/2}=0,$$

$$\frac{u^{n+1}-u^n}{\tau}+\Lambda u^{n+1/2}=0.$$

Далее, исключая промежуточные слои, получим схему, записанную в каноническом виде

$$\frac{u^{n+1}-u^n}{\tau}+\Lambda\frac{u^{n+1}+u^n}{2}+\frac{\tau^2}{4}(\Lambda_1\Lambda_2+\Lambda_1\Lambda_3+\Lambda_2\Lambda_3)\frac{u^{n+1}-u^n}{\tau}+$$
$$+\frac{\tau^3}{8}\Lambda_1\Lambda_2\Lambda_3\frac{u^{n+1}-u^n}{\tau}=0.$$

Схема абсолютно устойчива (для коммутирующих операторов), имеет второй порядок аппроксимации по τ и h_i. Конечно, при практическом решении задач на компьютере используется именно последовательность разностных операторов. Канонический вид схемы удобен для ее теоретического исследования.

Литература

[1] *Белоцерковский О.М.* Численное моделирование в механике сплошных сред. М.: Физико-математическая литература, 1994. 442 с.

[2] *Марчук Г.И.* Методы расщепления. М.: Наука, 1988. 263 с.

[3] *Ковеня В.М., Яненко Н.Н.* Методы расщепления в задачах газовой динамики. Новосибирск: Наука, 1981. 263 с.

Лекция 19. Применение вариационных принципов для построения разностных схем

В необязательной лекции приводятся примеры использования вариационных принципов Лагранжа и Гамильтона для построения разностных схем на основе вариации дискретного аналога лагранжиана (гамильтониана) системы.

Ключевые слова: вариационные принципы Лагранжа и Гамильтона. Разностная схема на неравномерной сетке. Консервативные разностные схемы.

Вариационный принцип Ритца, позволяющий получить МКЭ для уравнений в частных производных эллиптического типа, в том числе и на нерегулярных сетках, рассматривался в лекции 17. Далее для решения нестационарных задач в основном использовался проекционный вариант МКЭ (метод Галеркина). Тем не менее, многие задачи математической физики допускают вариационную постановку. Некоторые величины и законы сохранения могут играть особую роль для задач (пример — закон сохранения гамильтониана для консервативной системы). Необходимы разностные схемы (или численные методы), позволяющие учитывать специфику задачи и вариационные постановки.

Сделать это можно двумя способами. Первый — использовать вариационные принципы для дискретных аналогов соответствующих функционалов. При этом обычно получаются консервативные схемы. Второй способ — построение разностных схем обычными методами (конечных разностей), а затем их модификация, направленная на минимизацию погрешности аппроксимации законов сохранения.

Рассмотрим первый (вариационный) подход.

19.1. Пример использования принципа наименьшего действия (Гамильтона)

Рассматривается задача о движении твердого нерастяжимого стержня длиной 1. Пусть он закреплен в точке 0, а на другой конец стержня действует сила $\mathbf{F}(t)$ (рис. 19.1). Требуется определить движение стержня. Начальная форма стержня считается заданной.

Возможное решение: записать уравнение движения — получится уравнение гиперболического типа; поставить граничные условия; по-

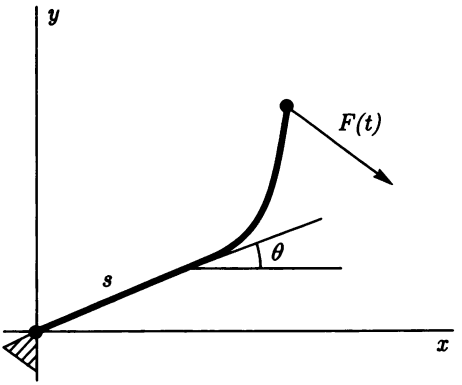


Рис. 19.1

строить разностную схему. Но в задаче допускаются сколь угодно большие колебания стержня.

Другой способ приближенного решения. Введем θ — угол отклонения от оси x — как функцию длины дуги s, времени t. Тогда имеем

$$x = \int_0^s \cos\,\theta(s',\,t)\,ds',$$

$$y = \int_0^s \sin\,\theta(s',\,t)\,ds'.$$

Кинетическая энергия стержня есть

$$T = \int_0^1 \left(\frac{V_x^2}{2} + \frac{V_y^2}{2}\right) ds = \frac{1}{2}\int_0^1 \left(\left[\frac{\partial x}{\partial t}\right]^2 + \left[\frac{\partial y}{\partial t}\right]^2\right) ds,$$

а потенциальная энергия складывается из упругой энергии (изгиба) и работы внешней силы $\mathbf{F}(t)$:

$$U = \int_0^1 \left(\frac{\partial\,\theta(s,\,t)}{\partial\,s}\right)^2 ds - F_x x(1,\,t) - F_y y(1,\,t)$$

(соответствующие коэффициенты полагаются равными 1).

Лагранжиан системы есть $L = T - U$, так что

$$L = \frac{1}{2}\int_0^1\int_0^s \{\dot\theta^2 \sin^2\theta + \dot\theta^2\cos^2\theta\}\, d\xi\, ds - \frac{1}{2}\int_0^1 \left(\frac{\partial\theta}{\partial s}\right)^2 ds -$$

$$-F_x\int_0^1 \cos\theta\, ds - F_y\int_0^1 \sin\theta\, ds.$$

Согласно принципу Гамильтона, функционал действия достигает экстремального значения на истинном движении. Отсюда следует, что

$$\frac{d}{dt}\left(\frac{\partial L}{\partial\dot\theta}\right) - \frac{\partial L}{\partial\theta} = 0$$

и получается уравнение движения для θ:

$$\int_0^1 G\ \cos\left(\theta(s,\ t) - \theta(\sigma,\ t)\right)\ddot\theta(\sigma,\ t)\, d\sigma = \theta''(s,\ t) + F_y\ \sin\ \theta - F_x\ \cos\ \theta -$$

$$-\int_0^1 G\ \sin\left(\theta(s,\ t) - \theta(\sigma,\ t)\right)\dot\theta^2(\sigma,\ t)\, d\sigma.$$

$$\theta(0,\ t) = 0; \qquad \theta'(0,\ t) = 0;$$

$$G(s,\ \sigma) = \begin{cases} 1 - s; & \sigma < s, \quad 0 \leqslant s \leqslant 1; \\ 1 - \sigma; & s \leqslant \sigma, \quad 0 \leqslant \sigma \leqslant 1. \end{cases}$$

Имеем интегро-дифференциальное уравнение для определения $\theta(s,\ t)$, причем как строить его разностную аппроксимацию — непонятно.

Отметим, что $G(s,\ \sigma)$ есть функция Грина для задачи

$$w'' = -g(s), \qquad w'(0) = w(1) = 0.$$

Теперь введем дискретный аналог лагранжиана L_h. Для этого разобьем стержень на n отрезков одинаковой длины Δl (и одинаковой массы). Каждый отрезок характеризуется углом наклона θ_k. Тогда (см. рис. 19.2)

$$x_j = \sum_{k=1}^{j} \Delta l\ \cos\ \theta_k = \Delta l\ \sum_{k=1}^{j} \cos\ \theta_k;$$

$$y_j = \Delta l\ \sum_{k=1}^{j} \sin\ \theta_k;$$

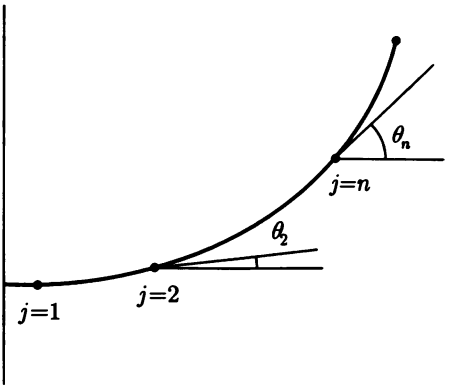


Рис. 19.2

$$T_h = \frac{\Delta l}{2}\sum_{j=1}^{n}\dot{x}+\dot{y} = \frac{(\Delta l)^3}{2}\sum_{j=1}^{n}\frac{\partial z_j}{\partial t}\frac{\partial \bar{z}_j}{\partial t};\ z_j = e^{i\theta_j};$$

$$U_h = \frac{\Delta l}{2}\sum_{j=2}^{n}\left(\frac{\theta_j-\theta_{j-1}}{\Delta l}\right)^2 - F_x\Delta l\sum_{j=1}^{n}\cos\theta_j - F_y\Delta l\sum_{j=1}^{n}\sin\theta_j.$$

Здесь появился аналог конечных элементов или схем с центральными разностями. Интегралы в этом выражении заменены конечными изломами, фактически эти интегралы вычислены методом трапеций, т. е. погрешность в определении лагранжиана L_h есть $O(\Delta l)^2$.

Теперь для построения системы уравнений надо записать

$$\frac{d}{dt}\left(\frac{\partial L_h}{\partial \dot{\theta}_m}\right) - \frac{\partial L_h}{\partial \theta_m} = 0, m = 1,2,\ldots,n$$

т. е. продифференцировать дискретный аналог функционала по всем значениям $\theta_m, \dot{\theta}_m$ на введенной сетке. Для рассматриваемой задачи последнее равенство приводит к соотношению

$$\sum_{j=1}^{n}\frac{\partial^2 L_h}{\partial\theta_k\,\partial\theta_j}\ddot{\theta}_j = \frac{\partial L_h}{\partial\theta_k} - \frac{\partial^2 L_h}{\partial\dot{\theta}_k\,\partial t} - \sum_{j=1}^{n}\frac{\partial^2 L_h}{\partial\theta_k\,\partial\theta_j}\dot{\theta}_j.$$

После подстановки в последнее выражение дискретного аналога ла-

гранжиана и выполнения дифференцирования по всем $\theta_k, \dot{\theta}_k$, получаем:

$$\sum_{k=1}^{n} \alpha_{lk}\ddot{\theta}_k = n^4(\theta_{l-1} - 2\theta_l + \theta_{l+1}) +$$
$$+n^2(F_y \cos\theta_l - F_x \sin\theta_l) - \sum_{k=1}^{n} g_{lk} \sin(\theta_l - \theta_k)\dot{\theta}_k^2.$$

(известно, что $n\,\Delta l = 1$ по построению).

Таким образом, при использовании сеточного аналога вариационного принципа Гамильтона получена дифференциально-разностная система уравнений (дифференциальная по времени, разностная по пространственным переменным).

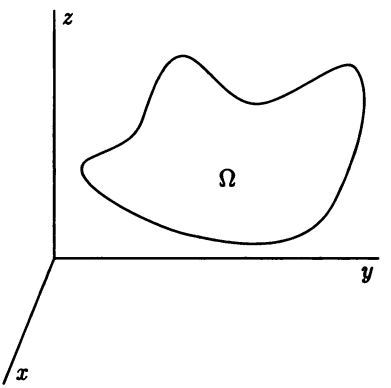


Рис. 19.3

Несколько затруднительно решать эту систему как систему обыкновенных дифференциальных уравнений, так как она не приведена к нормальной форме Коши. Однако с последней дифференциально-алгебраической системой можно работать и решать ее. Алгоритмы решения основаны на том, что g_{lk} — сеточный аналог функции Грина. Обратная матрица — сеточная аппроксимация оператора второй производной. Подробное изложение метода решения в [1].

19.2. Вариационные схемы для решения задач газовой динамики

Более интересные разностные схемы получаются для многомерных задач. Особенно плодотворным вариационный подход оказывается при решении задач механики сплошной среды.

Рассмотрим следующий пример. В бесконечном объеме (вакууме) находится область Ω, занятая газом (см. рис. 19.3). Масса газа $m = \int\limits_{\Omega} \rho\, dV$, $\rho(x,\ y,\ z)$ — плотность среды.

Движение системы подчиняется следующим уравнениям:

$$\rho \frac{d\mathbf{W}}{dt} + \operatorname{grad} P = 0 \quad \text{(движение)};$$

$$\frac{d\rho}{dt} + \rho \operatorname{div} \mathbf{W} = 0 \quad \text{(неразрывность)};$$

$$\rho \frac{d\varepsilon}{dt} + P \operatorname{div} \mathbf{W} = 0 \quad \text{(закон сохранения энергии)};$$

$$F(\rho,\ P,\ \varepsilon) = 0 \quad \text{(уравнение состояния)}.$$

Эти уравнения приведены в лагранжевых переменных. Напомним, что полная производная определена как $\frac{d}{dt} = \frac{\partial}{\partial t} + (\mathbf{W}, \nabla)$. Здесь все обозначения традиционные, через $\mathbf{W}$ обозначена скорость движения среды, а через V — элемент объема. Рассматривается случай невязкого нетеплопроводного газа. Запишем лагранжиан для движения среды, представив ее как систему частиц, а затем устремим число частиц в бесконечность.

$$L = T - U = \int\limits_{\Omega} \frac{(\mathbf{W},\ \mathbf{W})}{2} \rho\, dV - \int\limits_{\Omega} \rho\varepsilon\, dV = \int\limits_{\Omega} \left(\frac{(\mathbf{W},\ \mathbf{W})}{2} - \varepsilon \right) dm.$$

Движение среды должно доставлять минимум функционалу действия

$$S = \int\limits_{0}^{t} L(\tau)\, d\tau.$$

По закону сохранения массы $\delta(\rho dV) = \delta(dm) = 0$. Найдем δS — вариацию S:

$$\delta S = \int\limits_{0}^{t} \delta L(\tau)\, d\tau = \int\limits_{0}^{t} d\tau \int\limits_{\Omega} ((\mathbf{W},\ \delta\mathbf{W}) - \delta\varepsilon)\, dm.$$

Согласно первому началу термодинамики в предположении адиабатичности процесса,

$$\delta\varepsilon = \frac{P}{\rho^2}\, \delta\rho.$$

Используя кинематическое соотношение

$$\delta\mathbf{W} = \delta \left(\frac{d\mathbf{r}}{dt} \right) = \frac{d}{dt} (\delta\mathbf{r})$$

получим, что

$$\frac{\delta\rho}{\rho^2} = \frac{\delta\left(\frac{1}{V}\right)}{\rho\left(\frac{1}{V}\right)} = -\frac{\delta(dV)}{\rho\, dV} = -\frac{1}{\rho}\mathrm{div}(\delta\mathbf{r}).$$

Тогда

$$\delta S = -\int_0^t d\tau \int_\Omega \left(\rho\frac{d\mathbf{W}}{dt} + \mathrm{grad}\ P\right)\delta\mathbf{r}\, dV +$$

$$+\left[\int_\Omega (\mathbf{W},\ \delta\mathbf{r})\,\rho\, dV\right]\Bigg|_0^t + \int_0^t d\tau \oiint_{\partial\Omega} (\delta\mathbf{r},\ \mathbf{n})\, P\, d\sigma = 0.$$

Независимые вариации траекторий $\delta\mathbf{r}$ полагают равными нулю при $\tau = 0$ и $\tau = t$. Считая также $(\delta\mathbf{r},\ \mathbf{n})|_{\partial\Omega} = 0$, получим, что вариация лагранжиана равна нулю на решении системы уравнений Эйлера газовой динамики.

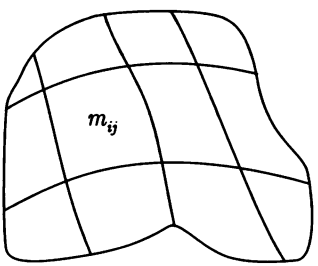


Рис. 19.4

Рассмотрим дискретный аналог функционала действия. Для этого введем в Ω сетку с ячейками, пронумерованными по левому нижнему углу. При отображении Ω на квадрат сетка отображается на равномерную квадратную сетку (см. рис. 4). Тогда

$$L_h = \sum_{i,\,j} m_{ij}\left(\frac{1}{2} < u^2 + v^2 >_{ij} - \varepsilon_{ij}\right),$$

где $m_{ij} = \rho_{ij}V_{ij}$ — масса ячейки (ρ_{ij} — ее плотность) $< u^2 + v^2 >$ — усредненная по ячейке плотность кинетической энергии

$$< u^2 + v^2 >_{ij} = \frac{1}{4}\sum_{l,\,k=0}^{1} (u^2_{i+l,\,j+k} + v^2_{i+l,\,j+k}).$$

Условие адиабатичности течения имеет следствием для дискретной системы выражение $m_{ij}\, d\varepsilon_{ij} = -P_{ij}\, dV_{ij}$.

Кинематические соотношения (уравнения движения для узлов сетки) есть

$$\frac{dx_{ij}}{dt} = u_{ij}, \frac{dy_{ij}}{dt} = v_{ij}.$$

Объем каждой ячейки вычисляется в предположении, что границы ячейки — отрезки прямых:

$$V_{ij} = \frac{1}{2}\,[(x_{i+1j} - x_{ij+1})(y_{i+1j+1} - y_{ij}) - (x_{i+1j+1} - x_{ij})(y_{i+1j} - y_{ij+1})].$$

Теперь необходимо получить соответствующую систему дифференциально-разностных соотношений, аппроксимирующих уравнения Эйлера.

Запишем функционал действия

$$S_h = \int_0^t L_h(\tau)\, d\tau$$

и найдем его вариацию:

$$0 = \delta S_h = \int_0^t \delta \left(\sum_{\omega_h} m_{ij} \left(\frac{< u^2 + v^2 >_{ij}}{2} - \varepsilon_{ij} \right) \right) d\tau =$$

$$= \sum_{\omega_h} \left\{ \delta \left(\frac{m_{ij}}{8} \sum_{l,k=0}^{1} (u^2_{i+l,\, j+k} + v^2_{i+l,\, j+k}) \right) - \delta(m_{ij}\varepsilon_{ij}) \right\} =$$

$$= \sum_{\omega_h} \frac{1}{8}\, \delta(m_{ij} \sum_{l,k=0}^{1} u^2_{i+l,\, j+k}) + \sum_{\omega_h} \frac{1}{8}\, \delta(m_{ij} \sum_{l,k=0}^{1} v^2_{i+l,\, j+k}) + \sum_{\omega_h} P_{ij}\, \delta V_{ij}.$$

Первое слагаемое в последнем равенстве можно переписать как

$$\sum_{\omega_h} \delta(u^2_{ij} \cdot \frac{1}{8} \sum_{l,k=0}^{1} m_{i-k,\, j-l}) = M_{ij} u_{ij} \delta u_{ij} = M_{ij}\, \frac{du_{ij}}{dt}\, \delta x_{ij},$$

где

$$M_{ij} = \frac{1}{4}\,(m_{i-1j} + m_{i-1j-1} + m_{ij} + m_{ij-1}).$$

После преобразований получаются следующие соотношения:

$$M_{ij}\,\frac{du_{ij}}{dt} = \sum_{l,\,k=0}^{1} P_{i-l,\,j-k}\,\frac{\partial\, V_{i-l,\,j-k}}{\partial\, x_{ij}};$$

$$M_{ij}\,\frac{dv_{ij}}{dt} = \sum_{l,\,k=0}^{1} P_{i-l,\,j-k}\,\frac{\partial\, V_{i-l,\,j-k}}{\partial\, y_{ij}};$$

т. е. возникает система дифференциально-разностных уравнений движения.

Уравнение адиабатичности трактуется, как уравнение для определения энергии:

$$M_{ij}\,\frac{d\varepsilon_{ij}}{dt} = -P_{ij}\,\frac{dV_{ij}}{dt} =$$

$$= -P_{ij}\sum_{l,\,k=0}^{1}\left(\frac{\partial\, V_{ij}}{\partial\, x_{i+l,\,j+k}}\,u_{i+l,\,j+k} + \frac{\partial\, V_{ij}}{\partial\, y_{i+l,\,j+k}}\,y_{i+l,\,j+k}\right);$$

Система дифференциально-разностных соотношений замыкается уравнением состояния.

Теперь, заменяя производные по времени на конечные разности, получим разностную схему для решения уравнений газовой динамики на нерегулярной подвижной сетке. Обычно для решения таких уравнений используются явные разностные схемы. Подробнее о вариационных методах получения схем для уравнений газовой динамики в [2].

19.3. Вариационная схема для уравнения теплопроводности на криволинейной сетке

Рассмотрим линейное уравнение теплопроводности

$$\frac{\partial\, u}{\partial\, t} + \operatorname{div}\,(-k\,\operatorname{grad}\,u) = 0 \tag{19.1}$$

с условиями

$$(-k\,\operatorname{grad}\,u,\ \mathbf{n}) = 0$$

в ограниченной области Ω с криволинейной границей. При этом нигде в $\bar{\Omega}$ уравнение не вырождено, т. е. $k(x,\ y) > 0$ во всех точках области, включая граничные.

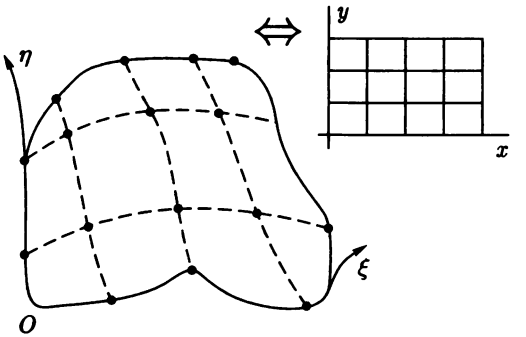


Рис. 19.5

В области Ω каким-либо образом введена сетка с четырехугольными ячейками. Сетка считается связной, т. е. для любых двух вершин ячеек существует ломаная, их соединяющая и состоящая из ребер ячеек (рис. 19.5).

Пусть сетка построена так, что существует преобразование, переводящее область Ω в параллелограмм (прямоугольник) с равномерной сеткой внутри. Тогда координатные линии x, y переходят в координатные кривые криволинейного базиса ξ, η.

Перепишем уравнение (19.1) в виде системы

$$\frac{\partial u}{\partial t} + \operatorname{div}\mathbf{W} = 0, \tag{19.2}$$

$$\mathbf{W} + k \operatorname{grad} u = 0.$$

Рассмотрим функционал

$$F[u] = \int\limits_{\Omega} \left(\frac{(\mathbf{W},\ \mathbf{W})}{k} - \frac{\partial}{\partial t} u^2 \right) dx\, dy. \tag{19.3}$$

Найдем $\delta F[u]$:

$$\delta F[u] = \int\limits_{\Omega} \delta \left(\frac{(\mathbf{W},\ \mathbf{W})}{k} - \frac{\partial}{\partial t} u^2 \right) dx\, dy =$$

$$= \int\limits_{\Omega} \left(-2\operatorname{div}\mathbf{W} \cdot \delta u - 2\delta u \frac{\partial u}{\partial t} - 2u \frac{\partial}{\partial t} \delta u \right) dx\, dy,$$

$$\delta \frac{(\mathbf{W},\ \mathbf{W})}{k} = 2 \left(\frac{\mathbf{W}}{k},\ \delta \mathbf{W} \right) = 2 \left(\frac{\mathbf{W}}{k},\ -k \operatorname{grad} \delta u \right) = -2\,(\delta u, \operatorname{div}\mathbf{W}).$$

Отсюда при $\frac{\partial}{\partial t}\,\delta u = 0$ минимум функционала достигается на решении уравнения теплопроводности.

Для построения разностной схемы введем дискретный аналог функционала $F_h(\mathbf{W}^h)$, т. е. в дискретном аналоге основной расчетной величиной будет поток тепла.

Прежде чем построить функционал, рассмотрим ячейку разностной сетки (рис. 19.6). Температуру u_{ij} и коэффициент теплопроводности (или температуропроводности) k_{ij} отнесем к центру ячейки (точке пересечения диагоналей). В дальнейшем считаем, что термодинамические величины постоянны во всей ячейке. Векторы теплового потока отнесем к углам ячейки (рис. 19.6), а к центрам соответствующих ребер — проекции потоков на координатные оси. Считаем, что i увеличивается по мере увеличения координаты ξ; j — по мере увеличения η; проекции векторов потока направлены вдоль соответствующих координатных линий. Заметим, что проекции потоков для двух ячеек сонаправлены с векторами внешней нормали, а для двух — противонаправлены.

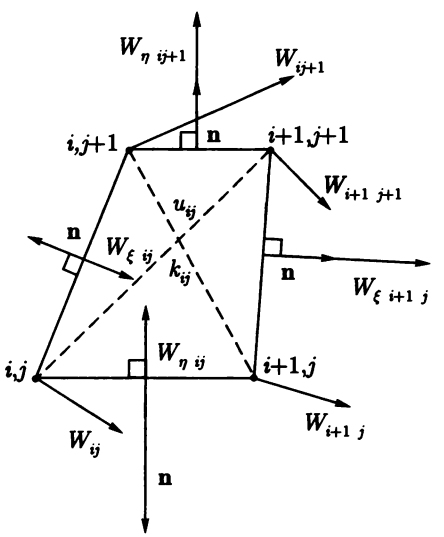


Рис. 19.6

Проинтегрируем уравнение

$$\frac{\partial u}{\partial t} + \mathrm{div}\mathbf{W} = 0$$

по элементарной ячейке разностной сетки. Имеем:

$$S_{ij}\frac{du_{ij}}{dt}+(W_{\eta\,ij}\Gamma_{\xi\,ij}-W_{\xi\,i+1j}\Gamma_{\eta\,i+1j}-W_{\eta\,ij+1}\Gamma_{\xi\,ij+1}+W_{\xi\,ij}\Gamma_{\eta\,ij})=0, \tag{19.4}$$

где Γ — длины соответствующих ребер, S_{ij} — площадь элементарной ячейки.

Так как координаты всех вершин выпуклого четырехугольника известны, то поиск длин, площадей и углов — элементарная геометрическая задача.

Уравнение (19.4) — дискретный аналог уравнения (19.2). Если возможно определить все потоки в моменты времени t^n, t^{n+1}, а после применить аппроксимацию (19.2) по времени с какими-либо весами, то будет построена разностная схема для расчета температуры.

Учтем, что

$$2u\frac{\partial u}{\partial t}=-2u\mathrm{div}\mathbf{W}.$$

Построим дискретный аналог (19.3):

$$F_h(\mathbf{W}^h)=\sum_{i,j\in\omega_h}\left[S_{ij}\cdot\left(\sum_{p,l=0}^{1}\frac{W^2_{i+p,j+l}}{k_{ij}}\right)+\right.$$

$$\left.2u^{(\sigma)}_{ij}(W_{\eta ij}\Gamma_{\xi ij}-W_{\xi i+1j}\Gamma_{\eta i+1j}-W_{\eta ij+1}\Gamma_{\xi ij+1}+W_{\xi ij}\Gamma_{\eta ij})\right]. \tag{19.5}$$

Скалярные квадраты, входящие в первое слагаемое дискретного аналога функционала, выражаются через контравариантные проекции следующим образом (рис. 19.7):

$$W_{ij}=\frac{1}{\sin^2\varphi_1}(W^2_{\eta ij}+W^2_{\xi ij}+2W_{\eta ij}W_{\xi ij}\cos\varphi_1),$$

$$W_{i+1j}=\frac{1}{\sin^2\varphi_2}(W^2_{\eta ij}+W^2_{\xi i+1j}-2W_{\eta ij}W_{\xi i+1j}\cos\varphi_2)$$

и т. д.

Знак «+» или «−» определяется по правилу: компоненту потока приписывается знак «+», если проекция потока сонаправлена с внешней нормалью, а знак «−» — если противонаправлена. Таким образом, для углов ячейки φ_1 и φ_3 получим знак «+» в последнем слагаемом (проекции одинаковых знаков), а для углов φ_2 и φ_4 — знак «−» (проекции теплового потока в произведении разных знаков).

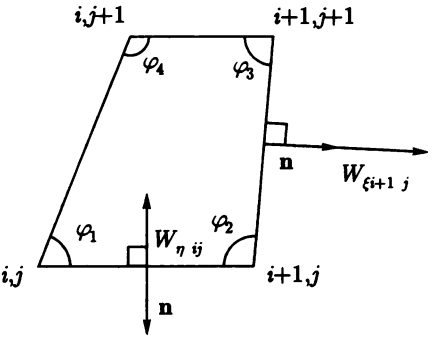


Рис. 19.7

Для получения явной схемы положим в (19.5) вес верхнего слоя по времени $\sigma = 0$ и дифференцируем (19.5) по всем $W_{\xi\, ij}, W_{\eta\, ij}$. Приравнивая производные нулю, получим схему для определения потоков, затем из (19.4) ищем все u_{ij}^{n+1}.

Для построения неявной схемы в (19.4) считаем $\sigma = 1$, а вместо (19.5) пишем следующую дискретизацию:

$$S_{ij}\,\frac{u_{ij}^{n+1}-u_{ij}^{n}}{\tau} = -\sum_{l,p=0}^{1}(-1)^{l}(W_{\xi})_{i+l,j}^{n+1}\Gamma_{\eta\, i+l,j} + (-1)^{p}(W_{\eta})_{i,j+p}^{n+1}\Gamma_{\xi\, i,j+p}.$$

Выражая отсюда неизвестное пока значение u_{ij}^{n+1} в (19.5), получим выражение, зависящее от $u_{ij}^{n}, \{W_{\xi}^{n+1}\}, \{W_{\eta}^{n+1}\}$, причем $F_h(\mathbf{W}^h)$ есть сумма квадратов контравариантных проекций.

Дифференцируя, получим линейную систему уравнений для определения потоков. Можно показать, что матрица системы будет обладать следующими свойствами:

1. $\mathbf{A} = \mathbf{A}^* > 0$;
2. $\mathbf{A}$ имеет ленточную структуру;
3. $\mathbf{A}$ является разреженной.

Можно применить эффективные итерационные методы решения системы.

Доказано, что неявная схема будет безусловно устойчивой, а явная — условно устойчивой.

Метод легко обобщается на случай $k = k(x,\ y,\ u)$, если уравнение не вырождается. Кроме того, метод может быть обобщен и на случай других граничных условий (не обязательно отсутствия потоков). В этом случае в функционал (19.3) добавляются соответствующие интегралы по границам, а в (19.5) — суммы по поверхностям.

Подробнее об этих схемах можно прочитать в [2].

19.4. Задачи для самостоятельного решения

1. Уравнение Кортевега-Де Фриза

   Одно из самых замечательных уравнений математической физики — уравнение Кортевега-Де Фриза (сокращенно КДФ) часто записывают в виде

   $$u_t - 6uu_x + u_{xxx} = 0$$

   или

   $$\tilde{u}_t + \tilde{u}\tilde{u}_x = \tilde{u}_{xxx}.$$

   (a) Найти преобразование, переводящее эти формы записи друг в друга.

   (b) Рассматриваем задачу для уравнения $u_t - 6uu_x + u_{xxx} = 0$ в области $x \in [-10;\ 10]$ с условием периодичности.

   Для решения иногда используют трехслойную разностную схему на шаблоне рис. 19.8 (третья производная расписывается по пяти точкам симметричным образом, некоторые коэффициенты могут обратиться в нуль). Исследовать ее на аппроксимацию и устойчивость. Какое условие устойчивости получено? Построить разностную схему на шаблоне, напоминающем шаблон схемы Саульева для решения уравнения теплопроводности (рис. 19.9а, б).

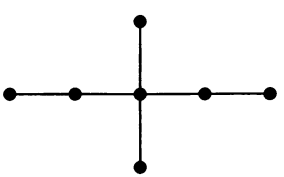


Рис. 19.8

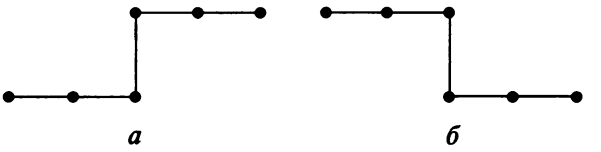


Рис. 19.9

Исследовать получившиеся схемы на аппроксимацию и устойчивость. Можно ли использовать прогонку для вычислений на верхнем слое?

(c) Уравнение $u_t - 6uu_x + u_{xxx} = 0$ имеет бесконечное число законов сохранения. Укажем несколько из них:

$$\int u dx = \text{const}_1,$$

$$\int u^2 dx = \text{const}_2$$

$$\int \left(\frac{(u'_x)^2}{2} + u^3 \right) dx = \text{const}_3 = I_1.$$

Как построить *консервативную* разностную схему, чтобы на сеточном уровне выполнялись законы сохранения $\int u dx = \text{const}$?

(d) Особую роль играет третий из приведенных выше законов сохранения. Он является гамильтонианом для уравнения $u_t - 6uu_x + u_{xxx} = 0$, т. е.

$$\frac{\partial u}{\partial t} = \frac{d}{dx} \left(\frac{\delta I_1}{\delta u} \right).$$

Здесь $\frac{\delta I_1}{\delta u}$ — вариационная производная функционала I_1, способ получения разностной схемы, сохраняющей гамильтониан системы.

Решение. Запишем сеточный аналог гамильтониана I_1:

$$I_1^h = \sum_{m=-\infty}^{+\infty} \left[\frac{1}{2} \left(\frac{u_{m+1} - u_m}{h} \right)^2 + (u_m)^3 \right] h = I_1 + O(h^2).$$

Контрольный вопрос: почему I_1^h аппроксимирует I_1 с точностью $O(h^2)$?

На сеточном уровне взятие вариационной производной означает дифференцирование по всем u_m^n и деление на h. Тогда получаем сеточную запись вариационной производной:

$$\frac{\delta I_1^h}{\delta u} = \frac{1}{h}\frac{\partial I_1^h}{\partial u_m} = 3u_m^2 + \frac{1}{h^2}((u_m - u_{m-1}) - (u_{m+1} - u_m)) =$$

$$= 3u_m^2 - \frac{u_{m+1} - 2u_m + u_{m-1}}{h^2} + O(h^2).$$

Аппроксимируя дискретный аналог (4.4.3) с естественным для этого вторым порядком по h, получаем

$$\frac{\partial u_m}{\partial t} = \frac{3(u_{m+1})^2 - 3(u_{m-1})^2}{2h} - \frac{u_{m+2} - 2u_{m+1} + 2u_{m-1} - u_{m-2}}{2h^3}.$$

Заменяя производную по времени разностью $\frac{u_m^{n+1} - u_m^{n-1}}{2\tau}$ и вычисляя правую часть на n слое, получаем одну из схем пункта 2.

Конечно, возможны и другие аппроксимации гамильтониана, варьирование которых приводит к другим разностным схемам. Все они будут записываться на симметричных шаблонах, на сеточном уровне для этих схем также будет выполняться закон сохранения $\int u dx = \text{const}$.

Удастся ли получить вариационную схему на несимметричном шаблоне типа Саульева?

Литература

[1] *Хайрер Э., Ваннер Г.* Решение обыкновенных дифференциальных уравнений. Жесткие и дифференциально-алгебраические системы. М.: Мир, 1999. 685 с.

[2] *Самарский А.А., Колдоба А.В., Повещенко Ю.А., Тишкин В.Ф., Фаворский А.П.* Разностные схемы на нерегулярных сетках. Минск, ЗАО "Критерий", 1996. 196 с.

Приложение. Параллельные вычисления на кластерах из персональных компьютеров в математической физике

В. Е. Карпов, А. И. Лобанов

В Приложении на примере решения конкретной задачи по проектированию установки рассмотрены основные схемы распараллеливания численных методов.

Ключевые слова: параллельные вычисления, кластер, метод Якоби, статическая модель, динамическая модель, потоковая модель, внутренний параллелизм.

1. Введение

Одним из источников сложных задач, описываемых уравнениями в частных производных, является современная физика плазмы. Для исследований в области управляемого термоядерного синтеза создаются установки, способные генерировать мощные энергетические импульсы. Актуальными становятся вопросы моделирования динамики плазмы в таких установках.

Соответствие с экспериментами дает численное моделирование на основе сложных математических моделей. При этом резко возрастает объем выполняемых расчетов. Лет 10–12 назад использовать полные модели не позволяли возможности вычислительной техники. Сейчас требуется применение параллельных вычислительных систем. Для проведения параллельных расчетов разрабатываются специальные методы. Однако существует ряд апробированных методов и комплексов программ для обычных компьютеров. Анализ их внутреннего параллелизма и адаптация для выполнения на параллельных ЭВМ становятся необходимыми.

Один из способов повышения производительности ЭВМ — развитие элементной базы, ведущее к уменьшению времени выполнения процессором элементарных операций. Согласно эмпирическому закону Мура (Moore) [1], производительность процессоров удваивается каждые полтора года, но конечному пользователю требуется решение задачи уже сейчас. Неудовлетворенные запросы заставляют разработчиков компьютерных систем искать другие способы повышения производительности, например, за счет совмещения по времени выполнения операций несколькими физическими устройствами.

Идея одновременного использования двух и более устройств в вычислительных машинах стала внедряться с конца 1950-х годов. Первоначальные изменения в конструкции не затрагивали алгоритмических основ решения задач. Они допускали совместную работу нескольких программ, требуя более совершенной организации процесса выполнения заданий, но практически не позволяли ускорить решение задачи.

Существенного повышения производительности удалось добиться после внесения в архитектуру ЭВМ принципов параллельной обработки данных. Известные приемы конвейеризации этапов выполнения различных команд процессора, применение в одном процессоре нескольких функциональных устройств позволяют ускорить решение задачи на уровне машинных команд. Такой параллелизм используется при работе процессоров с очень большим командным словом — VLIW-процессоров (Very Large Instruction Word). На современном этапе этот параллелизм скрыт от прикладного программиста. Ответственность за его эффективное использование возлагается на разработчиков процессоров или компиляторов.

Наибольшего увеличения производительности удалось достичь с появлением параллельных вычислений. Термин «параллельные вычисления» относится к ускорению решения задач за счет одновременного использования нескольких процессоров. Параллельные вычисления стали особенно популярными в конце 80-х–начале 90-х годов XX века.

К применению параллельной техники оказались не готовы специалисты, освоившие программирование на последовательных машинах. Потребовалось создание алгоритмов, пригодных для реализации на параллельных ЭВМ. Научная мысль одновременно продвигалась в двух направлениях. Первым стало исследование на параллельность уже существующих алгоритмов для последовательных компьютеров [2]. Вторым — разработка новых алгоритмов, специально предназначенных для параллельных комплексов [3, 4].Необходимое условие для параллельной реализации алгоритма — наличие у него *внутреннего параллелизма*, т. е. существование ярусно-параллельной формы с большой шириной ярусов. В концепции неограниченного параллелизма [5] лежит предположение о том, что вычислительная система не накладывает никаких ограничений на алгоритм. Тогда это же условие полагается достаточным. Опыт реализации показал ошибочность предположения о достаточности. На машинах с одной архитектурой алгоритмы могли показывать высокую эффективность, в то время как на машинах с другой архитектурой эти же алгоритмы работали медленнее своих последовательных версий. Но анализ алгоритмов на наличие внутреннего параллелизма необходим для любой параллельной программы, а концепция неограниченного параллелизма позволяет получить оценку максимально возможного ускорения.

Для сложных задач такой анализ — весьма трудоемкий процесс. При его проведении используются различные инструментальные программы. По большей части это компиляторы, которые не только автоматически выявляют наличие внутреннего параллелизма, но и готовят пригодный к параллельному исполнению машинный код.

Для вычислительных комплексов с распределенной памятью наличие у программы значительного внутреннего параллелизма еще не означает, что она будут эффективно работать. Свою роль здесь начинает играть соотношение времен выполнения параллельных частей программы и обменов информацией между процессорами. Если у пользователя есть возможность снять временной профиль последовательной программы на своей вычислительной системе, то эти соотношения могут быть учтены.

В качестве примера использования параллельных технологий рассмотрим задачу модернизации экспериментальной установки.

2. Расчет электрического поля установки РС-20 с использованием кластера из персональных компьютеров

Установка РС-20 была создана в Курчатовском институте и первоначально предназначалась для проведения радиационно-биологических исследований. Необходимость проведения экспериментов по созданию сверхмощных импульсов тока потребовала модернизации для увеличения напряжений и плотности заряда, пропускаемого через плазменный прерыватель тока.

Первоначально эскиз конструкции установки выглядел как спайка двух коаксиальных конденсаторов различного диаметра, внешние обкладки которых соединены металлической поверхностью (рис. 1). В больший конденсатор помещена диэлектрическая вставка, представляющая собой полый усеченный конус. В полиэтиленовую вставку врезаны пары компенсаторных колец. Каждая пара соединена между собой плоскими металлическими кольцами, проходящими через диэлектрик. Внутри всей установки поддерживается вакуум. Внешние обкладки конденсаторов заземлены. На внутренний цилиндрический электрод подается постоянное напряжение. К компенсаторным кольцам подведены постоянные напряжения, увеличивающиеся от нижнего кольца к верхнему. Требуется определить потенциалы и напряженность электрического поля внутри установки, а также оптимальную форму и расположение компенсаторных колец.

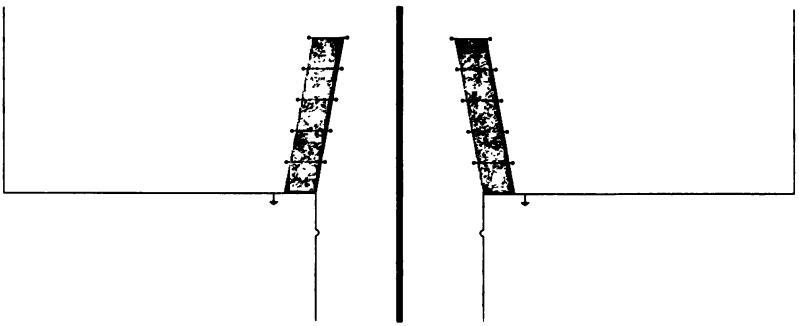


Рис. 1. Первоначальный эскиз установки RS-20

3. Математическая модель и выбор численного метода

Ввиду осевой симметрии целесообразно перейти в цилиндрическую систему координат $(r,\ z,\ \theta)$. Распределение потенциала электрического поля ϕ описывается уравнением Лапласа:

$$\frac{1}{r}\frac{\partial}{\partial r}(\varepsilon r\frac{\partial \varphi}{\partial r}) + \frac{\partial}{\partial z}\varepsilon\frac{\partial \varphi}{\partial z} = 0. \tag{6}$$

На металлических поверхностях значения потенциалов заданы эскизом конструкции. Так как корректно поставить граничные условия на верхнем и нижнем срезе конденсаторов не представляется возможным, будем считать, что конденсаторы продолжаются в обе стороны на бесконечное расстояние. В этом случае на входе и выходе установки получаем граничные условия вида

$$\varphi = \frac{\ln(r) - \ln(r_1)}{\ln(r_2) - \ln(r_1)}\varphi_2 + \frac{\ln(r) - \ln(r_2)}{\ln(r_1) - \ln(r_2)}\varphi_1. \tag{7}$$

Рассматриваемая область не является односвязной (имеются внутренние границы на компенсаторных кольцах). Кроме того, известно, что в плоском случае в окрестности угловых точек решение ведет себя как $\text{const} + r^{\frac{\pi}{\beta}}$, где β — величина угла. Для входящих углов, у которых $\beta > \pi$ (в данном случае это точка стыка коаксиальных конденсаторов), производные от решения стремятся к бесконечности как $r^{\frac{\pi}{\beta}-1}$.

Наличие угловой точки предполагает высокие напряженности электрического поля и требует подробных сеток для расчета потенциала. Коэффициент диэлектрической проницаемости среды ϵ терпит разрыв на границе раздела сред вакуум — диэлектрик.

Целесообразно использовать для решения поставленной задачи один из явных методов, характеризующихся высокой степенью внутренней параллельности. Применялась пятиточечная разностная аппроксимация, при этом система линейных уравнений решалась итерационным методом Якоби:

$$\frac{1}{h_{r-}+h_{r+}}\left(\varepsilon_{n+\frac{1}{2},m}\frac{\varphi^{i}_{n+1,m}-\varphi^{i+1}_{n,m}}{h_{r+}}-\varepsilon_{n-\frac{1}{2},m}\frac{\varphi^{i+1}_{n,m}-\varphi^{i}_{n-1,m}}{h_{r-}}\right)+$$
$$+\frac{1}{4r_{n,m}}\left(\varepsilon_{n+\frac{1}{2},m}\frac{\varphi^{i}_{n+1,m}-\varphi^{i+1}_{n,m}}{h_{r+}}+\varepsilon_{n-\frac{1}{2},m}\frac{\varphi^{i+1}_{n,m}-\varphi^{i}_{n-1,m}}{h_{r-}}\right)+ \tag{8}$$
$$+\frac{1}{h_{z-}+h_{z+}}\left(\varepsilon_{n,m+\frac{1}{2}}\frac{\varphi^{i}_{n,m+1}-\varphi^{i+1}_{n,m}}{h_{z+}}-\varepsilon_{n,m-\frac{1}{2}}\frac{\varphi^{i+1}_{n,m}-\varphi^{i}_{n,m-1}}{h_{z-}}\right)=0,$$

где верхний индекс указывает на номер итерации. Индексы + и − относятся к величинам шагов сетки справа и сверху от рассматриваемой точки и слева и снизу, соответственно. Для учета разрыва коэффициента диэлектрической проницаемости в качестве значения ϵ в полуцелых точках вблизи границы раздела сред выбирались средневзвешенные значения. Использовался следующий критерий выхода из итераций:

$$\max_{n,m}\frac{\left|\varphi^{k+1}_{n,m}-\varphi^{k}_{n,m}\right|}{\varphi^{k+1}_{n,m}}\leqslant\delta. \tag{9}$$

4. Модели организации параллельных вычислений для комплексов с распределенной памятью

Эффективность работы параллельных программ на вычислительных комплексах с распределенной памятью определяется тремя составляющими:

- степенью внутреннего параллелизма алгоритма;
- количеством данных, передаваемых между процессорами;
- степенью равномерности загрузки процессоров.

Существуют другие факторы, влияющие на производительность. Но их учет относится к компетенции производителей компиляторов, в то время как ответственность за три выше перечисленных фактора ложится на плечи прикладного программиста. Степень внутреннего параллелизма является свойством выбранного алгоритма и никак не связана с организацией совместной работы процессоров. Количество передаваемых данных и равномерность загрузки вычислительных мощностей, напротив, напрямую связаны с моделью организации параллельных расчетов. При

анализе эффективности параллельной версии алгоритма используется не количество данных, а время их передачи. Скорость передачи одного и того же количества данных в рамках одного вычислительного комплекса может полагаться постоянной, что позволяет сравнивать между собой разные модели организации вычислений. Переход к другому вычислительному комплексу требует повторного сравнения эффективностей моделей.

Для многопроцессорных компьютерных систем с распределенной памятью существуют три модели организации параллельных вычислений: потоковая, динамическая и статическая. Рассмотрим особенности каждой из них на примере.

Пусть задан следующий фрагмент программного кода:

```
do k = 1, maxiter
  do i = 1, n
    a(i) = f(a(i),b(i))
  enddo
enddo
print *,a
```

Для внутреннего цикла массивы a и b являются входными (они требуются для вычислений на любом процессоре). Массив a относится и к выходным, так как «полный» массив требуется, по крайней мере, на одном процессоре после завершения параллельных вычислений.

Потоковая модель представляет собой модель вычислений по классической схеме master-worker и требует наличия двух программ с различными кодами. Все вычисления, за исключением параллельных частей алгоритма, выполняются только на главном процессоре (master). Параллельные части выполняются на рабочих процессорах (workers). Для приведенного выше фрагмента кода организация вычислений в потоковой модели выглядит следующим образом:

```
master (фрагмент программы 1)
do k = 1, maxiter
```

(балансировка загрузки рабочих процессов и передача им соответствующих частей ранее вычисленных массивов a и b)
(сбор вычисленных частей массива a от рабочих процессоров для последующего использования)

```
enddo
print *,a
```

```
worker (фрагмент программы 2)
(прием своей части массивов a и b и границ цикла)
do i = imin, imax
  a(i) = f(a(i),b(i))
enddo
```

(отправка вычисленной части массива a главному процессору)

В потоковой модели входные данные рассылаются в необходимых объемах всем рабочим процессорам, что требует значительного времени на передачу. Выходные данные собираются только главным процессором, на что требуется меньшее время.

Динамическая модель является видоизменением потоковой модели. Для ее работы и на главном, и на рабочих процессорах запускается один и тот же исходный код. Вычисления выполняются на всех процессорах (за исключением операций вывода, выполнение которых может быть возложено только на главный процессор). Фрагмент кода при динамической организации вычислений может иметь следующий вид:

```
do k = 1, maxiter
  (балансировка загрузки процессов)
do i = imin, imax
  a(i) = f(a(i),b(i))
enddo
```

(рассылка вычисленной части массива a на все другие процессоры и сбор информации от них)

```
enddo
print *,a (возможно, только на главном процессоре)
```

В динамической модели входные данные никогда не передаются между процессорами. Вместо этого производится широковещательная рассылка выходных данных по окончании параллельного яруса. Сравнительная эффективность моделей организации вычислений определяется количеством входных и выходных данных для каждой параллельной части программы. Обе модели (и потоковая, и динамическая) могут применять динамическую балансировку загрузки процессоров непосредственно перед выполнением параллельного яруса.

Статическая модель организации вычислений похожа на динамическую, но все передачи данных определяются их фиксированным распределением по процессорам. Оно не может изменяться во время работы программы. Все процессоры выполняют идентичные вычисления.

После завершения параллельного яруса у каждого процессора есть свой срез общего объема информации. Передачи данных определяются не необходимостью выполнения параллельных частей программы, а способом их дальнейшего использования. Модельный фрагмент кода для статической модели может быть записан:

```
  do k = 1, maxiter
    do i = imin, imax
      a(i) = f(a(i),b(i))
    enddo
  enddo
(рассылка и сбор вычисленной части массива a)
print *,a (возможно, только на одном процессоре)
```

Статическая модель позволяет уменьшить объемы передачи информации между процессорами. Она позволяет также решать задачи, требующие большого объема данных, которые принципиально не могут быть решены в последовательном режиме из-за технических ограничений. Но в статической модели невозможна динамическая балансировка загрузки процессоров, что может ухудшить производительность программы при неправильном распределении данных.

5. Выбор модели организации параллельных вычислений

Для решения уравнения (8) был выбран достаточно простой алгоритм, внутренний параллелизм которого практически очевиден. Тем не менее, для определения эффективной модели распараллеливания целесообразно использование системы анализа программ на параллельность BERT 77 (http://www.plogic.com/bert-des.html). BERT 77 опирается на полученные оценки выполнения элементарных конструкций языка программирования FORTRAN 77 без предварительного исполнения анализируемой программы. Эти оценки строятся для каждой тройки: вычислительная система, компилятор, система коммуникации. Были проведены исследования для разных комбинаций компонентов тройки. Они показали, что времена выполнения программных конструкций (включая вычисление выражений, организацию циклов, доступ к элементам многомерных массивов, организацию вызовов функций и подпрограмм) могут быть оценены при использовании небольшого количества (порядка 200) предварительно вычисленных времен выполнения элементарных операций для конкретной пары «вычислительная система–компилятор». Эксперименты по измерению скоростей передачи информации для различных коммуникационных систем приводят к зависимости времени пере-

дачи данных от их объема, подобной изображенной на рис. 2. Такие зависимости в системе BERT 77 приближаются непрерывными кусочно-линейными функциями вида $t = t_i^0 + k_i \times nbytes$ при $n_i \leqslant nbytes \leqslant n_{i+1}$, где n_i пробегает значения 0, 1024, 2048, 4096 и т. д., t — полное время передачи данных, — латентности, k_i — скорости передачи, а *nbytes* — количество передаваемых данных.

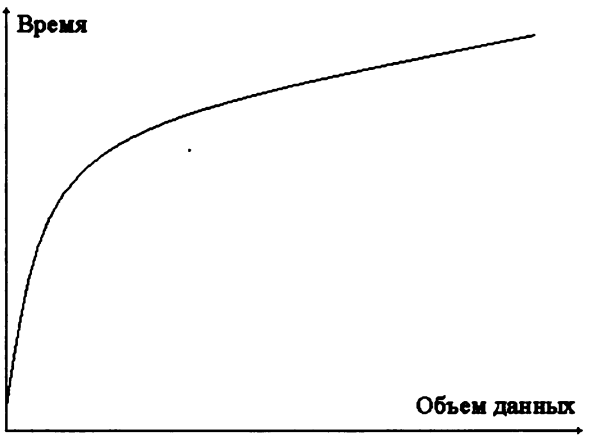


Рис. 2. Типовой график зависимости времени передачи информации от ее объема

Проанализируем эффективность различных моделей организации вычислений.

5.1. Потоковая модель

Входными переменными для параллельного яруса являются значения потенциала ϕ во всех узлах сетки на предыдущей итерации, а выходными — значения потенциала на текущей итерации и значение относительной погрешности (5). Относительная погрешность является редукционной переменной по отношению к операции min, т. е. ее частичные значения могут быть вычислены на рабочих процессорах, а окончательное значение — применением редукционной операции при приеме частичных значений. Если задействовано N рабочих процессоров, то при идеальной балансировке загрузки время выполнения вычислений будет составлять τ/N.

Количество узлов сетки на один процессор при нарезке рабочих областей по вертикали и правильной балансировке вряд ли будет различаться на двоичный порядок. Поэтому объемы передаваемых данных попа-

дут на один участок кусочно-линейной аппроксимации, скажем, на участок j. Скорость передачи информации на нем составляет k_j байтов в единицу времени. Если главный процессор будет отправлять рабочему процессору с номером i $nbytes_i$ байт, то время передачи составит $t^i = t_j^0 + k_j \times nbytes^i$. Полное время передачи входных данных главным процессором при использовании 8-байтовых данных будет $t = N \times t_j^0 + k_j \times nnodes \times 8$, где $nnodes$ — полное число узлов расчетной сетки. Независимо от значений N и j это время на порядок превышает время выполнения одной итерации. Даже без анализа передачи выходной информации видна неэффективность модели на имеющейся конфигурации вычислительной системы.

5.2. Динамическая модель

В динамической модели передача входных данных от главного процессора к рабочим отсутствует. Но вычисленные значения потенциала и относительной погрешности должны быть широковещательно разосланы всем процессорам. Аналогично получаем ту же оценку времени вычислений на одном процессоре — τ/N и оценку снизу для времени передачи данных $t = N \times t_j^0 + k_j \times (N-1) \times ((nnodes/N)+1) \times 8$, которые доказывают неэффективность применения и этой модели.

5.3. Статическая модель

Как правило, статическая модель характеризуется наименьшим количеством обменов информацией между процессами, но требует тщательной предварительной балансировки загрузки процессоров.

Удобно обеспечить баланс загрузки для случая 6 процессоров, распределяя между ними области ответственности, как показано на рис. 3. Процессор, отвечающий за верхний и нижний срезы, будет загружен меньше, чем остальные, но можно поручить ему отображение текущего состояния решения задачи. Незначительное ухудшение балансировки получается при разделении средних областей ответственности пополам, а верхнего и нижнего срезов между разными компьютерами для 12 процессоров.

При идеальной балансировке для N процессоров время выполнения каждой итерации по-прежнему будет τ/N. Для следующей итерации процессоры должны обменяться с соседями значениями потенциалов на граничных слоях, передав два слоя, и приняв два слоя от них. Это займет не более чем

$$t = 4 \times (t_j^0 + k_j \times nnodes_r \times 8)$$

секунд, где $nnodes_r$ — количество узлов в одном радиальном слое, а константы t_j^0 и k_j выбираются для объема передаваемой информации $nnodes_r \times 8$ байт. Оценка ускорения может быть получена как $\tau/(\tau/N + t)$.

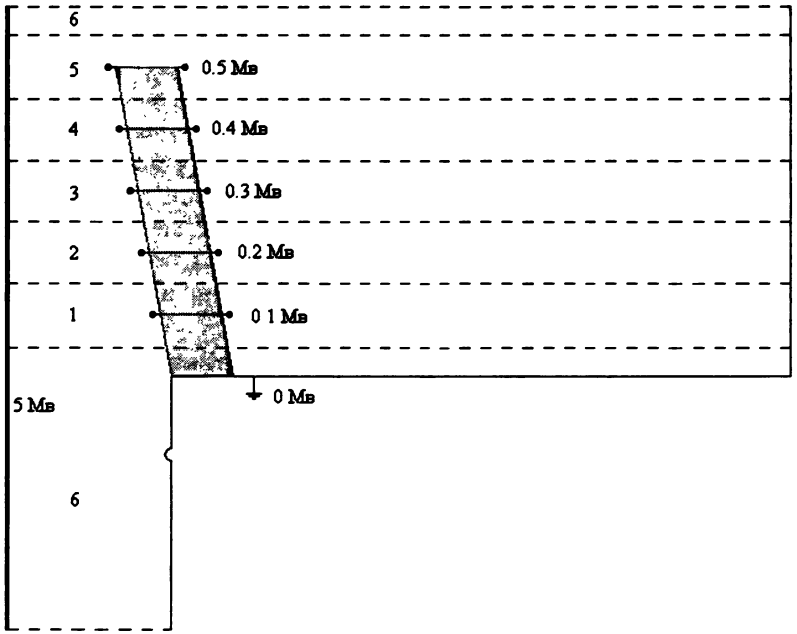


Рис. 3. Разделение на области ответственности для 6 процессоров при использовании статической модели вычислений

Если проводить проверку условия выхода из итераций (9) на каждой итерации, то дополнительно потребуется рассылка значения частичной погрешности по всем процессорам. Это приведет к дополнительным затратам на пересылку, в лучшем случае, $\log(N-1) \times (t_1^0 + k_1 \times 8)$ секунд.

Исправить ситуацию возможно, вычисляя значение погрешности не на каждой итерации, а, скажем, после каждой 100-й итерации. Рискуем выполнить лишние 100 итераций, что составляет доли процента от их общего количества, но при этом дополнительные коммуникационные затраты не влияют на эффективность работы программы.

Предварительная оценка ускорения вычислений на имеющемся комплексе с использованием пакета PVM составляет 5,7 раза для 6 процессоров и 10,1 раза для 12 процессоров. Реальное ускорение, естественно, будет несколько меньше из-за невозможности точной предварительной балансировки загрузки процессоров в статической модели. Проведенный расчет с использованием коммуникационного пакета PVM показал реальное ускорение по отношению к последовательному варианту 5,03 раза для 6 процессоров и 8,7 раза — для 12.

На основе расчетов вариантов модернизации установки было принято решение изменить форму компенсаторных колец и распределение потенциала на кольцах. Окончательный вариант приведен на рис. 4. Вычисления показали значительное снижение напряженности внутри диэлектрика в области нижних компенсаторных колец (рис. 5) и позволили сделать вывод об отсутствии пробоя. Данный вариант был принят за основу модернизации установки РС-20.

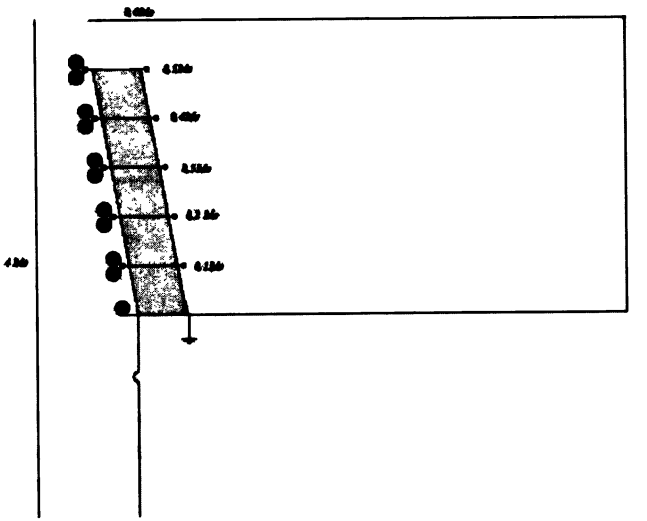


Рис. 4. Окончательный вариант расположения компенсаторных колец

6. Заключение

Рассмотренный пример относится к решению сравнительно простой задачи математической физики. Динамика плазмы (в частности, в РС-20) описывается гораздо более сложной системой уравнений в частных производных. Однако эта задача позволяет проанализировать методы организации параллельных вычислений и отладить технологию применения кластеров для сложных расчетов. В задачах динамики плазмы, например, наиболее эффективной оказывается динамическая модель организации вычислений.

Но и приведенная задача представляла практический интерес. Во-первых, для физиков. Модернизированная установка активно эксплуатируется. Во-вторых, для математиков-прикладников. На примере зада-

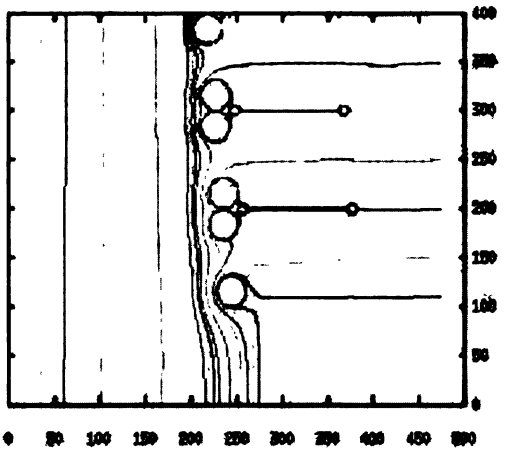


Рис. 5. Распределение потенциала для этого варианта расположения компенсаторных колец (приведен фрагмент расчетной области)

чи проведен анализ алгоритмов на наличие внутреннего параллелизма и отработана технология оценок эффективности распараллеливания без предварительного исполнения последовательных вариантов.

Литература

[1] *Воеводин В.В., Воеводин Вл.В.* Параллельные вычисления. — СПб.: БХВ-Петербург, 2002. — 608 с.

[2] *Воеводин В.В.* Параллельные структуры алгоритмов и программ. — М.: ОВМ АН СССР, 1987. — 148 с.

[3] *Фаддеева В.Н., Фаддеев Д.К.* Параллельные вычисления в линейной алгебре // Кибернетика. 1982. № 3. С. 18-31, 44.

[4] *Самарский А.А., Вабищевич П.Н.* Аддитивные схемы для задач математической физики. М.: Наука, 1999. 319 с.

[5] *Воеводин В.В.* Математические модели и методы в параллельных процессах. М.: Наука, 1986. 296 с.

Учебное издание

Петров Игорь Борисович
Лобанов Алексей Иванович
ЛЕКЦИИ ПО ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКЕ
Учебное пособие

Литературный редактор *Н. Черкесова*
Корректор *Ю. Голомазова*
Компьютерная верстка *А. Широбоков*
Обложка *М. Автономова*

Подписано в печать 31.07.2006. Формат 60x90 $^{1}/_{16}$.
Гарнитура Таймс. Бумага офсетная. Печать офсетная.
Усл. печ. л. 33,0. Тираж 2000 экз. Заказ № 3339.

ООО «ИНТУИТ.ру»
Интернет-Университет Информационных Технологий, www.intuit.ru
Москва, Электрический пер., 8, стр.3.
E-mail: admin@intuit.ru, http://www.intuit.ru

ООО «БИНОМ. Лаборатория знаний»
Москва, проезд Аэропорта, д. 3
Телефон: (495) 157-1902, 157-5272
E-mail: Lbz@aha.ru, http://www.Lbz.ru

Список книг Интернет-Университета Информационных Технологий

Алгоритмы, структуры данных, вычисления

1. **Лекции по вычислительной математике**, А.И. Лобанов и др., 2006, 528 с.
2. **Графы и алгоритмы. Структуры данных. Модели вычислений**, В.Е. Алексеев, В.А. Таланов, 2006, 320 с.
3. **Нейрокомпьютерные системы**, М.С. Тарков, 2006, 144 с.
4. **Нечеткие множества и нейронные сети**, Г.Э. Яхъяева, 2006, 320 с.

Архитектура ЭВМ

5. **Архитектура и технологии IBM eServer zSeries**, Э.К. Лецкий и др., 2005, 640 с.
6. **Архитектуры и топологии многопроцессорных вычислительных систем**, А.В. Богданов и др., 2004, 176 с.
7. **Основы микропроцессорной техники**, 3-е изд., Ю.В. Новиков и др., 2006, 360 с.
8. **Основы теории и организации ЭВМ**, В.В. Гуров и др., 2006, 272 с.

Безопасность информационных технологий

9. **Основы информационной безопасности**, 3-е изд., В.А. Галатенко, 2006, 208 с.
10. **Основы сетевой безопасности: криптографические алгоритмы и протоколы взаимодействия**, О.Р. Лапонина, 2005, 608 с.
11. **Стандарты информационной безонасности**, В.А. Галатенко, 2006, 264 с.

Интернет-технологии

12. **Flash MX для профессиональных программистов**, М.А. Капустин и др., 2006, 512 с.
13. **Основы web-технологий**, П.Б. Храмцов и др., 2003, 512 с.

История и социальные вопросы

14. **Основы права интеллектуальной собственности**, А.Г. Серго и др. 2005, 344 с.

Операционные системы

15. **Операционная система Linux**, Г.В. Курячий и др., 2005, 392 с.
16. **Операционная система Solaris**, Ф.И. Торчинский, 2005, 472 с.
17. **Операционная система Unix**, Г.В. Курячий, 2004, 320 с.
18. **Основы онерационных систем**, 2-е изд., В.Е. Карпов и др. 2006, 536 с.

Разработка приложений

19. **Введение в анализ, синтез и моделирование систем**, В.М. Казиев, 2006, 248 с.
20. **Введение в теорию программирования**, С.В. Зыков, 2004, 400 с.
21. **Интеграция приложений на основе WebSphere MQ**, В.А. Макушкин и др., 2005, 336 с.
22. **Компонентный подход в программировании**, В.В. Кулямин, 2006, 464 с.
23. **Объектно-орнентированный анализ и проектирование с использованием UML и IBM Rational Rose**, А.В. Леоненков, 2006, 320 с.
24. **Основы менеджмента программных проектов**, И.Н. Скопин, 2004, 336 с.
25. **Основы тестирования программного обеспечения**, В.П. Котляров, 2006, 360 с.
26. **Программирование в стандарте POSIX**, В.А. Галатенко, 2004, 560 с.
27. **Проектирование информационных систем**, В.И. Грекул и др., 2005, 296 с.
28. **Стили и методы программирования**, Н.Н. Непейвода, 2005, 320 с.

Сетевые технологии

29. **Основы локальных сетей**, Ю.В. Новиков и др., 2005, 360 с.
30. **Основы сетей передачи данных**, 2-е изд., В.Г. Олифер и др., 2005, 176 с

Системы и языки программирования

31. **Основы программирования на C#**, В.А. Биллиг, 2006, 488 с.
32. **Основы программирования на PHP**, Н.В. Савельева, 2005, 264 с.
33. **Основы программирования на языке Пролог**, П.А. Шрайнер, 2005, 176 с
34. **Основы функционального программирования**, Л.В. Городняя, 2004, 280 с.
35. **Программирование на Java**, Н.А. Вязовик, 2003, 592 с.
36. **Программирование на языке Pascal**, Т.А. Андреева, 2006, 240 с.
37. **Сборник заданий по основаниям программирования**, В.В. Пупышев и др., 2006, 352 с.
38. **Язык программирования Python**, Р.А. Сузи, 2006, 328 с.
39. **Язык программирования Си++**, 2-е изд.,А.Л. Фридман, 2004, 264 с.
40. **Язык Си и особенности работы с ним**, Н.И. Костюкова и др., 2006, 208с.

Технологии баз данных

41. **Data Mining**, И.А. Чубукова, 2006, 384 с.
42. **Основы SQL**, Л.Н. Полякова, 2004, 368 с.
43. **Основы баз данных**, С.Д. Кузнецов, 2005, 488 с.
44. **Основы проектирования приложений баз данных**, И.Ю. Баженова, 2006, 328 с.

Человеко-машинное взаимодействие

45. **Интеллектуальные робототехнические системы**, В.Л. Афонини др., 2005, 208 с.

Основы информатики и математики

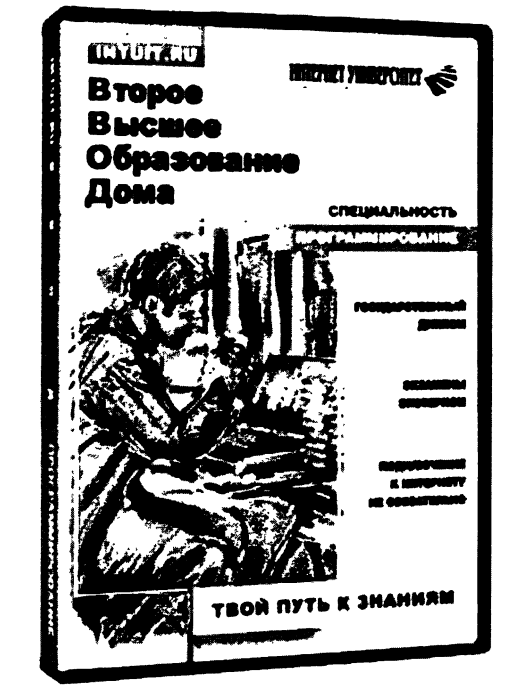
46. **Математическая теория формальных языков**, А.Е. Пентус и др., 2006, 248 с.
47. **Начала алгебры. Часть 1**, А.В. Михалев, А.А. Михалев, 2005, 272 с.
48. **Преподавание информатики и мятематических основ информатики**, под. ред. А.В. Михалева, 2005, 144 с.
49. **Common Intermediate Language и системное программирование в Microsoft .NET**, А.М. Чеповский и др., 2006, 328 с.
50. **Основы программирования**, В.В. Борисенко, 2005, 328 с.
51. **Работа с текстовой информацией. Microsoft Office Word 2003**, О.Б. Калугина, В.С. Люцарев, 2005, 152 с.
52. **Работа с электронными таблицами. Microsoft Office Excel 2003**, О.Б. Калугина, В.С. Люцарев, 2006, 240 с.

Информационные технологии от первого лица

53. **Технология программирования**, А.Н. Терехов, 2006, 152 с.

Архитектура информационных технологий

54. **Архитектура и стратегия. "Инь" и "янь" информационных технологий**, А. Данилин, А. Слюсаренко, 2005, 504 с.
55. **Готовы ли Вы к войне за клиента? Стратегия управления взаимоотношениями с клиентами (CRM)**, П. Черкашин, 2004, 384 с.
56. **OpenView. Network Node Manager. Разработка и реализация корпоративного решения**, Джон Бломмерс, 2005, 264 с.
57. **Объектно-ориентированное конструирование программных систем + CD**, Бертран Мейер, 2005, 1232 с.
58. **Я могу работать в современном офисе + CD**, А. Прохоров, 2005, 264 с.



INTUIT.RU
ЛОКАЛЬНАЯ ВЕРСИЯ 1.01

ТВОЙ ПУТЬ К ЗНАНИЯМ

дистанционное обучение без подключения к Интернету

Автономная система обучения учебным программам Интернет-Университета Информационных Технологий. Подключение к интернету не обязательно. В системе имеется более **70** учебных курсов, предусмотрены возможности сдачи тестов и экзаменов. Система поддержиавает многопользовательскую работу на одном компьютере.

Рекомендуемые системные требования

Операционная система: **Windows 98/ME/NT/2000/XP/2003**

Минимальные требования к аппаратуре: **Pentium III, HDD 300MB, RAM 128Mb, видеорежим 800x600 true color, CD-ROM 12x, мышь**

ЛЕКЦИИ ПО ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКЕ
УЧЕБНОЕ ПОСОБИЕ

В книге рассматриваются основные понятия и методы вычислительной математики. Курс содержит главы, посвященные классическим численным методам анализа и линейной алгебры, решению систем обыкновенных дифференциальных уравнений и уравнений математической физики. Рассматриваются методы решения жестких систем ОДУ. При рассмотрении методов решения дифференциальных уравнений в частных производных обсуждаются методы решения нелинейных задач. Особое внимание уделяется решению систем уравнений в частных производных гиперболического типа. В качестве примеров рассматриваются численные методы решения задач газовой динамики. Дается представление о современных методах решения уравнений математической физики, как конечно-разностных методов, так и вариационных и проекционных методах. Большинство лекций снабжено задачами для рассмотрения на семинарских занятиях и для самостоятельного решения.

Игорь Петров
Алексей Лобанов

ISBN 5-94774-542-9