В.М.Вержбицкий

ЧИСЛЕННЫЕ МЕТОДЫ

Линейная алгебра и нелинейные уравнения

ВЫСШАЯ ШКОЛА

В.М.Вержбицкий

ЧИСЛЕННЫЕ МЕТОДЫ

Линейная алгебра и нелинейные уравнения

Рекомендовано
Министерством образования
Российской Федерации
в качестве учебного пособия
для студентов математических
и инженерных специальностей
высших учебных заведений

Москва
«Высшая школа» 2000

УДК 519.6
ББК 22.19
В31

Рецензенты:

кафедра математического моделирования систем и процессов Пермского государственного технического университета (зав. кафедрой доктор физ.-мат. наук, профессор П.В. Трусов);

кафедра вычислительной математики Удмуртского государственного университета (зав. кафедрой доктор физ.-мат. наук, профессор Г.Г. Исламов);

доктор физ.-мат. наук, профессор Н.В. Азбелев

ISBN 5-06-003654-5

Оглавление

Предисловие

Предлагаемая вниманию читателя книга явилась результатом многолетнего опыта преподавания автором цикла дисциплин вычислительного характера студентам, обучающимся по специальности «Прикладная математика» и другим специальностям Ижевского государственного технического университета. Она отражает лишь небольшую часть материала основ численных методов, с которой, по мнению автора, должно начинаться их изучение. А именно, в книгу включены численные методы решения основных задач линейной алгебры (к которым, как правило, в итоге сводится решение других задач вычислительной математики, прикладной механики и т.п.), а также методы решения алгебраических и трансцендентных уравнений и их систем. Предполагается, что читатель знаком с линейной алгеброй и математическим (в частности, функциональным) анализом. Тем не менее, чтобы освежить некоторые базовые знания, а также чтобы расширить круг потенциальных пользователей книги, в ее конце приведены два приложения, посвященные нормам в конечномерных векторных пространствах и производным векторных функций векторных аргументов (последнее нужно лишь для облегчения чтения шестой главы); третье приложение содержит постановки заданий для лабораторных работ, выполнение которых автор считает необходимым элементом в освоении численных методов; фигурирующие в этих заданиях точностные константы могут быть заменены другими, в зависимости от конкретных параметров применяемого вычислительного инструментария. Написанию компьютерных программ, реализующих те или иные методы, должно способствовать наличие в книге во многих случаях достаточно детализированных алгоритмов.

Главное, что хотел бы автор от своей читательской аудитории, впервые изучающей численные методы, — это понимания основных идей (которых не так уж много) и умения довести их до числа, содержащего, по возможности, только верные знаки искомого результата. Автор сознательно позволяет себе «не замечать» двух очевидных барьеров на этом пути. Первый барьер связан с тем, что наряду с основными, наиболее важными методами, здесь отводится место и второстепенным, на первый взгляд, методам, т.е. как учебное пособие книга, возможно, несколько перегружена. Второй барьер — более "плотное", чем это принято в учебной литературе, цитирование. Ссылки на другие книги в разных случаях делаются из разных соображений: для указания на источники дополнительной информации и на возможные пути более глубокого изучения той или иной темы, для сопоставления различных терминов, означающих одно и то же, а также для того, чтобы «разделить ответственность» за некоторые приводимые без доказательства

результаты с авторами цитируемых работ. На журнальные статьи ссылки делаются лишь в исключительных случаях (в основном, когда речь идет о приоритетах).

Не вводя специально раздела, в котором были бы сосредоточены все встречающиеся в книге обозначения и сокращения (поскольку, в большинстве своем, они стандартны и должны быть читателю хорошо знакомы), автор обращает внимание на следующие особенности. Во-первых, автор считает удобным использование символа присваивания := и в качестве знака "по определению", что позволяет сразу различать, особенно в цепи преобразований, где равенство фигурирует как утверждение, а где им определяется (вводится, обозначается) некоторая величина. Во-вторых, в некоторых местах трудно было обойтись без использования довольно редко применяемого, но очень удобного в вычислительной математике знака $\lesssim$ означающего неравенство в смысле главных (линейных) частей сравниваемых выражений. В-третьих, учитывая распространенность в тексте стержневых для книги групп слов, таких как "система линейных алгебраических уравнений" и "метод простых итераций", автор часто использует вместо них аббревиатуры СЛАУ и МПИ соответственно. Другие аббревиатуры носят локальный характер и не должны затруднять чтение книги. Наконец, автор старался выдержать общепринятую систему, в которой из написания одной и той же буквы, например, $\mathbf{X}$, x или x, сразу должно быть ясно, матрица это, вектор или скаляр; при этом для обозначения индексов применялись только буквы m, n, l, i, j, k, обычно по умолчанию принимающие только целые значения.

Автор выражает свою признательность коллегам по кафедре прикладной математики и информатики ИжГТУ, в частности, профессору В.Я. Дерру и доценту А.А. Айзиковичу за внимание и поддержку при подготовке рукописи и особенно профессору А.Л. Тептину, сделавшему своевременно ряд ценных замечаний. Автор благодарен А.П. Ананину, А.Н. Борисенкову, Р.М. Валееву, А.В. Дворникову, П.В. Куприянову, А.В. Лосеву, К.В. Перцеву, принимавшим участие в компьютерном наборе рукописи, и, более всего, С.В. Высоцкому, собравшему воедино и приведшему к определенному стандарту набранные в разное время разными исполнителями части рукописи и осуществившему окончательную подготовку оригинал-макета. Особо автор отмечает роль не одного поколения студентов, явно или неявно способствовавших появлению этой книги.

В настоящее время автор работает над продолжением этого издания – книгой «Численные методы (математический анализ и дифференциальные уравнения)».

Автор

ГЛАВА 1 ОБ УЧЕТЕ ПОГРЕШНОСТЕЙ ПРИБЛИЖЕННЫХ ВЫЧИСЛЕНИЙ

Рассматривается круг вопросов, связанных с учетом погрешностей, появление которых неизбежно при численном анализе математических моделей, в частности, при решении линейных алгебраических систем и нелинейных скалярных уравнений. Обращается внимание на заведомо приближенный характер компьютерных операций над действительными числами. Приводятся примеры задач и методов, чрезмерно чувствительных к ошибкам исходных данных и к погрешностям арифметических действий.

1.1. ОБЩАЯ ФОРМУЛА ДЛЯ ОЦЕНКИ ГЛАВНОЙ ЧАСТИ ПОГРЕШНОСТИ

При численном решении математических и прикладных задач почти неизбежно появление на том или ином этапе их решения погрешностей следующих трех типов.

а) ***Погрешность задачи.*** Она связана с приближенным характером исходной содержательной модели (в частности, с невозможностью учесть все факторы в процессе изучения моделируемого явления), а также ее математического описания, параметрами которого служат обычно приближенные числа (например, из-за принципиальной невозможности выполнения абсолютно точных измерений). Для вычислителя погрешность задачи следует считать ***неустранимой*** (безусловной), хотя постановщик задачи иногда может ее изменить.

б) ***Погрешность метода.*** Это погрешность, связанная со способом решения поставленной математической задачи и появляющаяся в результате подмены исходной математической модели другой или конечной последовательностью других, например, линейных моделей. При создании численных методов закладывается возможность отслеживания таких погрешностей и доведения их до сколь угодно малого уровня. Отсюда естественно отношение к погрешности метода как к ***устранимой*** (или условной).

в) ***Погрешность округлений*** (погрешность действий). Этот тип погрешностей обусловлен необходимостью выполнять арифметические операции над числами, усеченными до количества разрядов, зависящего от применяемой вычислительной техники (если, разумеется, не используются специальные программные средства, реализующие, например, арифметику рациональных чисел).

Все три описанных типа погрешностей в сумме дают ***полную погрешность*** результата решения задачи. Поскольку первый тип погрешностей не находится в пределах компетенции вычислителя, для него он служит лишь ориентиром точности, с которой следует рассчитывать математическую модель. Нет смысла решать задачу существенно точнее, чем это диктуется неопределенностью исходных данных. Таким образом, погрешность метода подчиняют погрешности задачи. Наконец, при выводе оценок погрешностей численных методов обычно исходят из предположения, что все операции над числами выполняются точно. Это означает, что погрешность округлений не должна существенно отражаться на результатах реализации методов, т.е. должна подчиняться погрешности метода. Влияние погрешностей округлений не следует упускать из виду ни на стадии отбора и алгоритмизации численных методов, ни при выборе вычислительных и программных средств, ни при выполнении отдельных действий и вычислении значений функций.

Рассмотрим некоторые возможные подходы к учету погрешностей действий ([1, 8, 10, 17, 21, 24, 34] и др.).

Пусть A и a – два "близких" числа; условимся считать A точным, a – приближенным.

Величина $\Delta a := |A - a|$ называется ***абсолютной погрешностью*** приближенного числа a, а $\delta a := \dfrac{\Delta a}{|a|}$ – его ***относительной погрешностью*** *).

Числа Δ_a и δ_a такие, что $\Delta_a \geq \Delta a$ и $\delta_a = \dfrac{\Delta_a}{|a|} \geq \delta a$, называются ***оценками*** или ***границами*** абсолютной и относительной погрешностей соответственно (к Δ_a и δ_a часто применяют также термин ***"предельные погрешности"***). Так как обычно истинные погрешности не известны, то там, где не может возникнуть недоразумений, будем иногда называть Δ_a и δ_a просто абсолютной и относительной погрешностями.

Поставим вопрос о грубом оценивании погрешности результата вычисления значения дифференцируемой функции $u = f(x_1, x_2, \ldots, x_n)$ приближенных аргументов $x_1, x_2, \ldots, x_n$, если известны границы их абсолютных погрешностей $\Delta_{x_1}, \Delta_{x_2}, \ldots, \Delta_{x_n}$ соответственно. В этом случае точные значения аргументов $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$ лежат соответственно на отрезках

---

*) Возможно, чаще относительной погрешностью называют $\dfrac{\Delta a}{|A|}$.

Использование символа Δ в дальнейшем может быть двояким: как для обозначения абсолютной погрешности, так и для обозначения приращения переменной, что будет ясно из контекста либо специально оговорено.

$\left[x_1-\Delta_{x_1}, x_1+\Delta_{x_1}\right]$, $\left[x_2-\Delta_{x_2}, x_2+\Delta_{x_2}\right]$,…, $\left[x_n-\Delta_{x_n}, x_n+\Delta_{x_n}\right]$, а точная абсолютная погрешность результата $u=f(x_1,x_2,\ldots,x_n)$ есть

$$\Delta u=\left|f(\tilde{x}_1,\tilde{x}_2,\ldots,\tilde{x}_n)-f(x_1,x_2,\ldots,x_n)\right|$$

– модуль полного приращения функции. Главной, т.е. линейной частью этого приращения является, как известно, полный дифференциал du. Таким образом, имеем:

$$\Delta u\approx|du|=\left|\sum_{i=1}^{n}\frac{\partial u}{\partial x_i}dx_i\right|\le\sum_{i=1}^{n}\left|\frac{\partial u}{\partial x_i}\right|\cdot|\tilde{x}_i-x_i|\le\sum_{i=1}^{n}\left|\frac{\partial u}{\partial x_i}\right|\cdot\Delta_{x_i},$$

т.е. за границу абсолютной погрешности результата приближенно может быть принята величина

$$\Delta_u=\sum_{i=1}^{n}\left|\frac{\partial u}{\partial x_i}\right|\cdot\Delta_{x_i}. \qquad (1.1)$$

Отсюда легко получается формула приближенной оценки относительной погрешности значения u:

$$\delta_u=\frac{\Delta_u}{|u|}=\sum_{i=1}^{n}\left|\frac{\partial u}{\partial x_i}\right|\cdot\frac{\Delta_{x_i}}{|u|}=\sum_{i=1}^{n}\left|\frac{\partial u}{u\partial x_i}\right|\cdot\Delta_{x_i}=\sum_{i=1}^{n}\left|\frac{\partial\ln u}{\partial x_i}\right|\cdot\Delta_{x_i}. \qquad (1.2)$$

Как частные случаи формул (1.1), (1.2) (точных для функций, линейных относительно x_i или $\ln x_i$ соответственно) можно получить известные правила оценивания погрешностей результатов арифметических действий.

Действительно, пусть $u=\pm x_1\pm x_2\pm\ldots\pm x_n$. Тогда $\left|\frac{\partial u}{\partial x_i}\right|=1$ и $\Delta_{\Sigma(\pm x_i)}=\sum_{i=1}^{n}1\cdot\Delta_{x_i}=\sum_{i=1}^{n}\Delta_{x_i}$, т.е. при сложении и вычитании приближенных чисел их предельные абсолютные погрешности складываются.

Пусть теперь $u=x_1\cdot x_2\cdot\ldots\cdot x_n$, где можно считать все сомножители положительными. Так как $\ln u=\ln x_1+\ln x_2+\ldots+\ln x_n$ и $\frac{\partial\ln u}{\partial x_i}=\frac{1}{x_i}$, то, согласно (1.2),

$$\delta_{\Pi x_i}=\sum_{i=1}^{n}\frac{1}{x_i}\Delta_{x_i}=\sum_{i=1}^{n}\delta_{x_i}. \qquad (1.3)$$

Если же $u=\frac{x_1}{x_2}$, где $x_1x_2>0$, то $\ln u=\ln x_1-\ln x_2$, $\frac{\partial\ln u}{\partial x_i}=\frac{1}{x_i}$ и, значит,

$$\delta_{x_1 x_2} = \frac{\Delta_{x_1}}{x_1} + \frac{\Delta_{x_2}}{x_2} = \delta_{x_1} + \delta_{x_2}.$$

Последнее вместе с (1.3) означает известный результат о сложении предельных относительных погрешностей при умножении и делении приближенных чисел.

Возвращаясь к сложению, рассмотрим относительную погрешность суммы n положительных приближенных чисел $x_1, x_2, \ldots, x_n$, имеющих границы относительных погрешностей $\delta_{x_1}, \delta_{x_2}, \ldots, \delta_{x_n}$ соответственно:

$$\delta(x_1 + x_2 + \ldots + x_n) = \frac{\Delta(x_1 + x_2 + \ldots + x_n)}{x_1 + x_2 + \ldots + x_n} \leq$$

$$\leq \frac{\Delta_{x_1 + x_2 + \ldots + x_n}}{x_1 + x_2 + \ldots + x_n} = \frac{\Delta_{x_1} + \Delta_{x_2} + \ldots + \Delta_{x_n}}{x_1 + x_2 + \ldots + x_n} =$$

$$= \frac{x_1\delta_{x_1} + x_2\delta_{x_2} + \ldots + x_n\delta_{x_n}}{x_1 + x_2 + \ldots + x_n} \leq \frac{x_1\delta^* + x_2\delta^* + \ldots + x_n\delta^*}{x_1 + x_2 + \ldots + x_n} = \delta^*,$$

где $\delta^* = \max_i \delta_{x_i}$. Полученное неравенство говорит о том, что относительная погрешность суммы n положительных приближенных чисел не превосходит максимальной относительной погрешности слагаемых.

С вычитанием приближенных чисел дело обстоит хуже: оценка

$$\delta_{x_1 - x_2} = \frac{\Delta_{x_1 - x_2}}{|x_1 - x_2|} = \frac{\Delta_{x_1} + \Delta_{x_2}}{|x_1 - x_2|}$$

относительной погрешности разности $x_1 - x_2$ двух приближенных положительных чисел указывает на возможность сильного возрастания погрешности при $x_1 - x_2 \to 0$. В этом случае говорят о потере точности при вычитании близких чисел.

Часто возникает ***обратная задача теории погрешностей***: какой точности данные нужно подать на вход, чтобы на выходе получить результат заданной точности? Применительно к поставленной выше прямой задаче оценивания погрешности результата вычисления значения функции при заданных оценках погрешностей аргументов обратная задача заключается в оценивании величин Δx_i (или δx_i) по известной величине Δu. Для случая дифференцируемой функции одной переменной грубое решение обратной задачи тривиально: если $y = f(x)$, то $\Delta y \approx |dy| = |f'(x)|\Delta x$, откуда $\Delta x \approx \frac{\Delta y}{|f'(x)|}$. Для функции большего числа переменных обратная задача,

вообще говоря, некорректна. Нужны дополнительные условия. Например, ***принцип равных влияний*** состоит в предположении, что частные дифференциалы $\left|\frac{\partial u}{\partial x_i}\right|\Delta_{x_i}$ в (1.1) одинаково влияют на погрешность значения функции; тогда

$$\Delta_u = n\left|\frac{\partial u}{\partial x_i}\right|\Delta_{x_i}, \quad \text{откуда} \qquad \Delta_{x_i} = \frac{\Delta_u}{n\left|\frac{\partial u}{\partial x_i}\right|}.$$

В качестве другого довольно естественного допущения можно принять равенство относительных погрешностей всех аргументов, т.е. считать $\delta_{x_i} = \frac{\Delta_{x_i}}{|x_i|} = p$ при всех $i = 1, 2, \ldots, n$. Тогда $\Delta_{x_i} = p|x_i|$ и, значит, $\Delta_u = p\sum_{i=1}^{n}\left|\frac{\partial u}{\partial x_i}x_i\right|$. Из последнего равенства получаем величину $p = \frac{\Delta_u}{\sum_{i=1}^{n}\left|x_i\frac{\partial u}{\partial x_i}\right|}$ (характеризующую относительный уровень точности задания аргументов), на основе которой за границы абсолютных погрешностей аргументов принимаем $\Delta_{x_i} = \frac{|x_i|\Delta_u}{\sum_{i=1}^{n}\left|x_i\frac{\partial u}{\partial x_i}\right|}$. Имеются и другие, более сложные подходы к решению обратной задачи (см., например, [6]).

1.2. СТАТИСТИЧЕСКИЙ И ТЕХНИЧЕСКИЙ ПОДХОДЫ К УЧЕТУ ПОГРЕШНОСТЕЙ ДЕЙСТВИЙ

Рассмотренный выше ***аналитический*** (или ***классический***) ***способ учета погрешностей действий***, предполагающий точное оценивание погрешностей, основанное либо на приведенных в предыдущем параграфе правилах подсчета погрешностей арифметических действий, либо на параллельной работе с верхними и нижними границами исходных данных, имеет два существенных недостатка. Во-первых, этот способ чрезвычайно громоздок и не может быть рекомендован при массовых вычислениях. Во-вторых, он учитывает крайние, наихудшие случаи взаимодействия погрешностей, которые допустимы, но маловероятны. Ясно, что, например, при суммировании нескольких приближенных чисел (полученных в результате измерений, округлений или каким-либо другим путем) среди них почти наверное будут слагаемые как с избытком, так и с недостатком, т.е. произой-

дет частичная компенсация погрешностей. При больших количествах однотипных вычислений вступают в силу уже ***вероятностные*** или ***статистические законы*** формирования погрешностей результатов действий. Например, методами теории вероятностей показывается, что математическое ожидание абсолютной погрешности суммы n слагаемых с одинаковым уровнем абсолютных погрешностей, при достаточно большом n, пропорционально $\sqrt{n}$ ([8, 10, 21, ...]). В частности, если $n > 10$ и все слагаемые округлены до m-го десятичного разряда, то для подсчета абсолютной погрешности суммы S применяют ***правило Чеботарева***

$$\Delta S \approx \sqrt{3n} \cdot 0.5 \cdot 10^{-m}. \qquad (1.4)$$

Различие в результатах классического и статистического подходов к оцениванию погрешности суммы рассмотрим на примере оценки погрешности среднего арифметического нескольких приближенных чисел.

Пусть $x = \frac{1}{n}(x_1 + \ldots + x_n)$ – среднее арифметическое n (>10) приближенных чисел (например, результатов измерений), имеющих одинаковый уровень абсолютных погрешностей $\Delta_{x_i} = 0.5 \cdot 10^{-m}$. Тогда классическая оценка абсолютной погрешности величины x есть

$$\Delta_x = \frac{1}{n}\left(\Delta_{x_1} + \ldots + \Delta_{x_n}\right) = \frac{1}{n} \cdot n \cdot 0.5 \cdot 10^{-m} = 0.5 \cdot 10^{-m} = \Delta_{x_i},$$

т.е. такая же, как и у исходных данных. В то же время по формуле (1.4) имеем

$$\Delta_x \approx \frac{1}{n}\sqrt{3n} \cdot 0.5 \cdot 10^{-m} = \sqrt{\frac{3}{n}} \cdot 0.5 \cdot 10^{-m} = \sqrt{\frac{3}{n}} \cdot \Delta_{x_i} \xrightarrow[n\to\infty]{} 0.$$

Как видим, применение правила Чеботарева приводит к естественному выводу о том, что арифметическое усреднение результатов измерений или наблюдений увеличивает точность, чего нельзя сказать на основе классической теории погрешностей.

Прямое применение вероятностно-статистических оценок погрешностей также является достаточно сложным делом и вряд ли может быть рекомендовано при рядовых массовых вычислениях. Однако именно такие оценки подкрепляют практические правила работы с приближенными числами, составляющие основу так называемого ***технического подхода***. Этот подход связывают с именем известного русского кораблестроителя, математика и механика академика А. Н. Крылова. Согласно ***принципу А. Н. Крылова***, приближенное число должно записываться так, чтобы в нем все значащие цифры, кроме последней, были верными и лишь последняя была бы сомнительна, и притом в среднем*) не более чем на одну единицу. Напомним, что ***значащими цифрами*** числа в его позиционной записи называются все его цифры, начиная с первой ненулевой слева. Знача-

*) "В среднем" здесь понимается в вероятностном смысле.

щую цифру приближенного числа называют ***верной***, если абсолютная погрешность числа не превосходит единицы разряда, в котором стоит эта цифра (или половины единицы; в этом случае иногда применяется термин *верная в узком смысле*).

Чтобы результаты арифметических действий, совершаемых над приближенными числами, записанными в соответствии с принципом А. Н. Крылова, также соответствовали этому принципу, нужно придерживаться следующих нехитрых правил [10, 17, 21].

1. *При сложении и вычитании приближенных чисел в результате следует сохранять столько десятичных знаков, сколько их в приближенном данном с наименьшим количеством десятичных знаков.*

2. *При умножении и делении в результате следует сохранять столько значащих цифр, сколько их имеет приближенное данное с наименьшим числом значащих цифр.*

3. *Результаты промежуточных вычислений должны иметь один–два запасных знака (которые затем должны быть отброшены).*

Таким образом, при техническом подходе к учету погрешностей приближенных вычислений предполагается, что в самой записи приближенного числа содержится информация о его точности. И хотя прямая выгода от применения приведенных правил работы с приближенными числами может быть получена лишь при ручном счете (не нужно оперировать с цифрами, не влияющими на информативную часть приближенного результата), их знание и понимание помогает правильной интерпретации компьютерных расчетов, а иногда и самой организации таковых.

1.3. ПОНЯТИЕ О ПОГРЕШНОСТЯХ МАШИННОЙ АРИФМЕТИКИ

Для представления вещественных чисел в ЭВМ применяют, в основном, два способа: с фиксированной и с плавающей запятой (точкой).

Пусть в основу запоминающего устройства машины положены однотипные физические устройства (базисные элементы), имеющие r устойчивых состояний (как правило, $r = 2,8,16$ и т.п.), причем каждому числу ставится в соответствие одинаковое количество k этих элементов и, кроме того, с помощью таких или более простых элементов может фиксироваться знак. Упорядоченные элементы образуют разрядную сетку машинного слова: в каждом разряде может быть записано одно из базисных чисел $0, 1, \ldots, r-1$ (одна из r "цифр" r-ичной системы счисления) и в специальном разряде отображен знак + или –.

При записи числа с ***фиксированной запятой*** кроме упомянутых r параметров (основания системы счисления) и k (количества разрядов, отводимых под запись цифр числа) указывается еще количество ℓ разря-

дов, выделяемых под дробную часть числа. Таким образом, положительное вещественное число a, представляющее собой в r-ичной системе бесконечную, вообще говоря, непериодическую дробь, здесь будет отображено конечной последовательностью

$$\alpha_1\alpha_2\ldots\alpha_{k-\ell}\alpha_{k-\ell+1}\ldots\alpha_{k-1}\alpha_k,$$

где $\alpha_i \in \{0;1;\ldots;r-1\}$, т.е. реализуется приближенное равенство*)

$$a \approx \mathrm{fix}(a) := \alpha_1 r^{k-\ell-1} + \alpha_2 r^{k-\ell-2} + \cdots + \alpha_{k-\ell} r^0 + \\ + \alpha_{k-\ell+1} r^{-1} + \cdots + \alpha_{k-1} r^{-(\ell-1)} + \alpha_k r^{-\ell}.$$

Диапазон представляемых таким способом чисел определяется числами с наибольшими цифрами во всех разрядах, т.е. наименьшим числом $-(r-1)(r-1)\cdots(r-1)$ и наибольшим $(r-1)(r-1)\cdots(r-1)$, а ***абсолютная точность представления*** есть оценка величины $|a-\mathrm{fix}(a)|$, зависящая от способа округления: это $r^{-\ell}$ при простом отбрасывании "хвоста" $\alpha_{k+1}r^{-(\ell+1)} + \alpha_{k+2}r^{-(\ell+2)} + \ldots$ числа a и половина этой величины при ***правильном округлении*** (т.е. при увеличении α_k на единицу, если $\alpha_{k+1} > \frac{r}{2}$). Заметим, что абсолютная точность представления вещественных чисел с фиксированной запятой одинакова в любой части диапазона. В то же время ***относительная точность***, т.е. оценка величины $\left|\frac{a-\mathrm{fix}(a)}{a}\right|$ (или $\left|\frac{a-\mathrm{fix}(a)}{\mathrm{fix}(a)}\right|$), очевидно, может значительно различаться в зависимости от того, берется a близким к нулю или к границе диапазона. Иными словами, вещественные числа с фиксированной запятой имеют равномерную абсолютную плотность распределения на всем отрезке вещественной оси, определяемом границами диапазона, и неравномерную, возрастающую к границам отрезка, относительную плотность распределения.

В основе значительно чаще употребляемого представления с ***плавающей запятой*** лежит следующая экспоненциальная форма записи вещественного числа:

$$a = M \cdot r^p,$$

где r – ***основание***, p – ***порядок***, а M такое, что $r^{-1} \le |M| < 1 \left(= r^0\right)$ – ***мантисса***. Если под мантиссу выделяется ℓ r-ичных элементов, а под порядок m, то в системе записи с плавающей запятой вещественное число a

---

*) Символ := нами используется в двух близких смыслах:

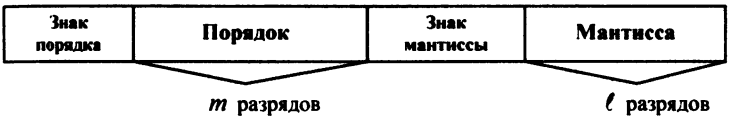
1) положить по определению, задать, выбрать;

2) присвоить, т.е. записать в ячейку памяти компьютера, определяемую идентификатором в левой части, значение, задаваемое правой частью.

представляется конечным числом $\mathrm{fl}(a)$ (от англ. floating – "плавающий") вида

$$a \approx \mathrm{fl}(a) := \pm\left(\beta_1 r^{-1} + \beta_2 r^{-2} + \ldots + \beta_\ell r^{-\ell}\right) \cdot r^{\gamma},$$

где γ – целое число из промежутка $\left[-r^m, r^m - 1\right]$; $\beta_1 \in \{1;\ldots;r-1\}$; $\beta_i \in \{0;1;\ldots;r-1\}$ $(i = 2,\ldots,\ell)$, т.е. ***машинное слово*** условно имеет структуру:



Числа $\pm r^{r^m}$ определяют границы допустимого числового диапазона. Более информативно здесь нужно говорить о диапазоне представимости положительных вещественных чисел, составляющем промежуток $\left[r^{-r^m}, r^{r^m} - 1\right]$. Левую и правую границы этого отрезка называют соответственно ***машинным нулем*** и ***машинной бесконечностью***, так как числа из промежутка $\left[-r^{-r^m}, r^{-r^m}\right]$ машина заменяет нулем, а числа, лежащие за пределами промежутка $\left[-r^{r^m-1}, r^{r^m-1}\right]$, она не воспринимает (без специальных ухищрений).

Важной характеристикой является число ε, называемое ***машинный эпсилон*** и обозначаемое обычно идентификатором *macheps*. Эта характеристика определяется как расстояние между единицей и ближайшим следующим за ней числом системы машинных чисел с плавающей запятой. Так как

$$1 = \left(1 \cdot r^{-1} + 0 \cdot r^{-2} + \ldots + 0 \cdot r^{-\ell} + \ldots\right) \cdot r^{1},$$

а следующее за 1 машинное число есть

$$\left(1 \cdot r^{-1} + 0 \cdot r^{-2} + \ldots + 0 \cdot r^{-(\ell-1)} + 1 \cdot r^{-\ell}\right) \cdot r^{1} = \mathrm{fl}(1+\varepsilon),$$

то за *macheps* можно принять величину

$$\varepsilon = 1 \cdot r^{-\ell} \cdot r^{1} = r^{1-\ell}.$$

Это число непосредственно связано с относительной погрешностью представления чисел в системе с плавающей запятой. Имеем:

$$\left|\frac{a-\mathrm{fl}(a)}{a}\right| = \frac{\beta_{\ell+1} r^{-(\ell+1)} + \beta_{\ell+2} r^{-(\ell+2)} + \cdots}{\beta_1 r^{-1} + \beta_2 r^{-2} + \cdots} \le \frac{1 \cdot r^{-\ell}}{\beta_1 \cdot r^{-1}} \le r^{1-\ell} = \varepsilon . \quad (1.5)$$

Таким образом, *машинный эпсилон**) *служит мерой относительной точности представления вещественных чисел*, причем эта точность одинакова в любой части числового диапазона и зависит лишь от числа r-ичных разрядов, отводимых под мантиссу числа. В то же время оценка абсолютной погрешности

$$|a - \mathrm{fl}(a)| \le |a| \cdot r^{1-\ell}$$

показывает, что расстояние между вещественными числами и конечными приближениями к ним в системе с плавающей запятой неодинаковы в разных частях числового диапазона: *абсолютная плотность машинных чисел больше вблизи нуля при одинаковой относительной плотности их распределения.*

Заметим, что величина *macheps* служит оценкой относительной точности представления вещественного числа a при условии, что $|a| > r^{-r^m}$. Если же $a \in \left[-r^{-r^m}, r^{-r^m}\right]$, то $\mathrm{fl}(a) \equiv 0$ и, значит, относительная погрешность

$$\left|\frac{a-\mathrm{fl}(a)}{a}\right| \equiv 1,$$

т.е. является постоянной достаточно большой величиной, в то время как абсолютная погрешность не превосходит величины r^{-r^m}.

Приведем значения введенных выше теоретических параметров для нескольких типов отечественных ЭВМ 60-80 гг.**) (реальные параметры, естественно, могут незначительно отличаться от приводимых) [1, 2, 47].

---

*) Если трактовать *macheps* как минимальное положительное действительное число, прибавление которого к 1 дает следующее за 1 число с плавающей запятой, то, очевидно, при правильном округлении значение *macheps* будет в два раза меньшим. Действительно, полагая $\varepsilon = \frac{1}{2} r^{1-\ell} = \frac{r}{2} \cdot r^{-(\ell+1)} \cdot r^1$, получаем

$$1 + \varepsilon = \left(1 \cdot r^{-1} + 0 \cdot r^{-2} + \ldots + 0 \cdot r^{-\ell} + \frac{r}{2} \cdot r^{-(\ell+1)}\right) \cdot r^1$$

и, значит, $\mathrm{fl}(1+\varepsilon) = 1 + r^{1-\ell} > 1$. "Мера дискретности" множества машинных чисел, как видим, остается той же: $r^{1-\ell}$.

**) В качестве упражнения читателю предлагается составить несложную программу, которая позволит получить соответствующие параметры используемого им современного компьютера.

Так, для записи числа в 48-разрядном машинном слове БЭСМ-6 40 двоичных разрядов выделяются под мантиссу, 6 – под порядок и 2 – под знаки мантиссы (т.е. числа) и порядка. Отсюда, принимая r=2, ℓ=40, m=6, получаем, что точность представления чисел с плавающей запятой на БЭСМ-6 не хуже $2^{-39}\left(\approx 10^{-12}\right)$, граница машинного нуля $2^{-64}\left(\approx 10^{-19}\right)$, машинной бесконечности $2^{63}\left(\approx 10^{19}\right)$.

Машинное слово СМ ЭВМ имеет 32 двоичных разряда, из которых под мантиссу выделяется 24, а под порядок 7. Зная параметры $r=2,\ \ell=24,\ m=7$, получаем $macheps=2^{-23}\approx 10^{-7}$, машинные нуль и бесконечность $\approx 10^{\mp 38}$.

На ЕС ЭВМ используется представление вещественных чисел по основанию $r=16$. Эти машины имеют относительную точность представления $\approx 10^{-7}$ и диапазон для положительных чисел $\approx 10^{-77} \div 10^{76}$.

Практически любая машина ЕС ЭВМ рассчитана на то, чтобы выделить под запись числа двойное машинное слово, что позволяет более чем вдвое увеличивать точность представления. Двойная точность предусматривается также многими языками программирования.

Обращаясь к арифметическим операциям над машинными числами, прежде всего заметим, что они утрачивают привычные свойства. Особенно это касается свойств ассоциативности и дистрибутивности, нарушаемых при выполнении арифметических операций на любых ЭВМ. (Сохранение или несохранение свойства коммутативности связывают со способом округления чисел.) Так, весьма утрированный пример сравнения

$$\left(r^{D/2}\cdot r^{3D/4}\right)\cdot r^{-D/2} \quad \text{с} \quad r^{D/2}\cdot\left(r^{3D/4}\cdot r^{-D/2}\right),$$

где D таково, что r^D – правая граница числового диапазона, показывает существенность способа расстановок скобок при умножении: в первом случае машина выдаст сообщение о переполнении, в силу

$$r^{D/2}\cdot r^{3D/4} > r^D,$$

а во втором случае будет получен правильный результат. Легко также представить ситуацию с тремя положительными числами a, b, c, когда расставляя по-разному скобки в выражении $a+b-c$, будем получать (или не получать вовсе) разные результаты.

Изучение погрешностей результатов арифметических операций над числами с плавающей запятой производится с помощью представления

$$\mathrm{fl}(a)=a(1+\delta), \quad \text{где} \quad |\delta|\le macheps \qquad (1.6)$$

(чтобы убедиться в справедливости (1.6), достаточно ввести δ равенством $\delta=\dfrac{\mathrm{fl}(a)-a}{a}$, равносильным фигурирующему в (1.6), и воспользоваться доказанным в (1.5) неравенством $|\delta|\le macheps$). Принимая во внимание, что

операции над двумя машинными числами a и b производятся точно (здесь используется двойная длина машинного слова), после чего производится округление, результат любой арифметической операции $\otimes$ также может быть записан в виде

$$\mathrm{fl}(a \otimes b) = (a \otimes b)(1+\delta_1), \qquad (1.7)$$

где $|\delta_1| \le \varepsilon$ (за исключением особых случаев).

Пусть складываются последовательно три положительных числа a_1, a_2, a_3. Тогда, согласно (1.7),

$$\mathrm{fl}(a_1 + a_2) = (a_1 + a_2)(1+\delta_1),$$

где $|\delta_1| \le \varepsilon$;

$$\mathrm{fl}((a_1 + a_2) + a_3) = ((a_1 + a_2)(1+\delta_1) + a_3)(1+\delta_2) =$$
$$= (a_1 + a_2)(1+\delta_1)(1+\delta_2) + a_3(1+\delta_2),$$

где $|\delta_i| \le \varepsilon \ (i = 1,2)$.

Заменяя здесь δ_i бóльшим значением ε, получим оценку абсолютной погрешности суммы трех слагаемых:

$$|\mathrm{fl}(a_1 + a_2 + a_3) - (a_1 + a_2 + a_3)| \le 2(a_1 + a_2)\varepsilon + a_3\varepsilon + (a_1 + a_2)\varepsilon^2.$$

Обращает на себя внимание неравноправность слагаемых в образовании погрешности суммы: меньшую роль в ней играет слагаемое, прибавляемое последним. Природа этого факта очевидна: первые слагаемые неявно (в просуммированном виде) участвуют в процессе каждого последующего сложения.

Если пренебречь степенями ε выше первой, то для суммы n положительных чисел a_i нетрудно получить [2] приближенную оценку абсолютной погрешности вида*)

$$\left| \mathrm{fl}\left(\sum_{i=1}^{n} a_i\right) - \sum_{i=1}^{n} a_i \right| \le |(n-1)(a_1 + a_2) + (n-2)a_3 + \ldots + 2a_{n-1} + a_n| \cdot \varepsilon$$

при последовательном суммировании, начинающемся с a_1. Очевидно, чтобы эта погрешность была минимальной, *последовательность чисел нужно суммировать в порядке возрастания членов*. Только за счет этого можно добиться уменьшения погрешности, как показано в [15], в $n/\log_2 n$ раз.

На основе изучения погрешности произведения нескольких чисел строятся алгоритмы оптимального умножения. Пусть требуется перемножить n чисел a_i таких, что $|a_1| \le |a_2| \le \ldots \le |a_n|$. Погрешность произведения

*) Здесь и далее знак $\le$ используется для обозначения неравенства в смысле главных (линейных) частей.

будет минимальной, если находить его по схеме: умножать a_1 последовательно на a_n, a_{n-1},... до тех пор, пока модуль частичного произведения не станет большим единицы, затем это частичное произведение умножать на a_2, a_3, ... до тех пор, пока новое частичное произведение не станет по модулю меньшим единицы, и так далее до исчерпывания всех сомножителей [15].

Естественно, что расплатой за выигрыш в точности при реализации таких алгоритмов будет проигрыш в скорости счета.

1.4. ПРИМЕРЫ НЕУСТОЙЧИВЫХ ЗАДАЧ И МЕТОДОВ

В силу неизбежного появления погрешностей в исходных данных задачи (в процессе создания математической модели изучаемого объекта или явления), а также погрешностей округления при ее решении, следует иметь представление о том, насколько чувствительными могут оказаться сами задачи и методы их решения к таким погрешностям. Рассмотрим несколько примеров проявления чрезмерной чувствительности.

а) Пусть требуется найти вещественное решение уравнения

$$(x-a)^n = \varepsilon,$$

где ε – очень малое положительное число, а натуральное n достаточно велико. Тогда естественно заменить ε нулем и положить $x \approx a$. Так как точное решение данного уравнения есть $x = a + \sqrt[n]{\varepsilon}$, то абсолютная погрешность при таком подходе составит величину $\sqrt[n]{\varepsilon}$. Много это или мало – судить об этом можно, придавая ε и n численные значения. Например, взяв $\varepsilon = 10^{-10}$, n=10, получим абсолютную погрешность значения корня $x \approx a$, равную 0.1. Относительная погрешность при этом может оказаться сколь угодно большой, если a взять сколь угодно малым.

б) Пример Уилкинсона [1, 2, 26, 53 и др.]. Многочлен

$$P_{20}(x) \equiv (x-1)(x-2)\ldots(x-20) \equiv x^{20} - 210x^{19} + \ldots + 20!$$

имеет 20 хорошо отделимых действительных корней

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

Предположим, что только в одном его коэффициенте, а именно, при x^{19} сделана ошибка порядка *macheps*: вместо -210 в развернутый вид многочлена $P_{20}(x)$ подставлено число $-\left(210 + 2^{-23}\right) \approx -\left(210 + 10^{-7}\right)$. Полученный при этом так называемый ***возмущенный многочлен*** будет иметь следующие корни (ограничимся записью трех цифр после запятой):

$$x_1 \approx 1.000, \qquad x_6 \approx 6.000, \qquad x_{12,13} \approx 11.794 \pm 1.652i,$$

$x_2 \approx 2.000$, $x_7 \approx 7.000$, $x_{14,15} \approx 13.992 \pm 2.519i$,
$x_3 \approx 3.000$, $x_8 \approx 8.007$, $x_{16,17} \approx 16.731 \pm 2.813i$,
$x_4 \approx 4.000$, $x_9 \approx 8.917$, $x_{18,19} \approx 19.502 \pm 1.940i$,
$x_5 \approx 5.000$, $x_{10,11} \approx 10.095 \pm 0.644i$, $x_{20} \approx 20.847$.

Как видим, весьма малое возмущение (сопоставимое с точностью представления чисел в некоторых ЭВМ) всего лишь в одном коэффициенте даже качественно изменило набор корней данного многочлена: половина из них перестали быть действительными.

в) Линейная система

$$\begin{cases} x + 10y = 11, \\ 100x + 1001y = 1101 \end{cases}$$

имеет единственное решение $x=1, y=1$. Допустив абсолютную погрешность в 0.01 в правой части одного уравнения, получим ***возмущенную систему***

$$\begin{cases} x + 10y = 11.01, \\ 100x + 1001y = 1101 \end{cases}$$

с единственным решением $x=11.01, y=0$. Последнее никак не назовешь близким к решению исходной системы.

г) Для вычисления определенных интегралов вида

$$I_n := \frac{1}{e}\int_0^1 x^n e^x dx,$$

где $n \in \mathbb{N}$, с помощью метода интегрирования "по частям" легко вывести рекуррентную формулу

$$I_n = 1 - n I_{n-1}; \qquad n = 1,2,\ldots; \qquad I_0 = 1 - \frac{1}{e}. \tag{1.8}$$

В [4] приведена сводная таблица результатов подсчета I_n при $n=0,1,2,\ldots,14$ по формуле (1.8), полученных в 60-х годах на разных вычислительных машинах в Чехословакии, Восточной Германии, России. Эта таблица наглядно демонстрирует рост разброса значений интеграла при увеличении значения n. Не приводя здесь полностью этих результатов, отметим лишь, что при $n=14$ спектр значений I_{14} лежит в границах от -148 до 5356. Несмотря на то, что на некоторых ЭВМ вычисления велись с мантиссами, имеющими более десяти десятичных разрядов, ни один результат не имел ни одной верной цифры! Причина подобного явления – в численной неустойчивости схемы (1.8). Безупречная в теоретическом пла-

не, т.е. с точки зрения аналитической математики, она совершенно непригодна с позиций вычислительной математики, поскольку неизбежная погрешность стартового значения I_0 при подсчете I_n увеличивается в $n!$ раз, т.е. катастрофически нарастает.

Если примеры а)–в) указывали прямо на существование ***неустойчивых задач***[*], то в г) мы видим яркий пример ***неустойчивого метода*** вычислений. Последним, как правило, есть альтернативы. Например, для вычисления рассматриваемого в г) интеграла I_n при конкретном значении n можно применять квадратурные формулы, т.е. формулы, специально приспособленные для приближенного вычисления определенных интегралов. А можно воспользоваться и равенством (1.8), только переписать его в виде

$$I_{n-1} = \frac{1}{n}(1 - I_n). \qquad (1.9)$$

Учитывая, что $0 < I_{n+1} < I_n$ и $I_n \xrightarrow[n\to\infty]{} 0$, для подсчета I_n при некотором фиксированном $n=k$ (и для меньших этого k индексов) можно задаться значением $I_N := 0$ и вести счет по формуле (1.9) при $n = N, N-1, \ldots, k+1$. Так как начальная погрешность на каждом шаге теперь уменьшается в n раз, то такой алгоритм будет ***численно устойчивым***. Значение N при этом может быть определено теоретически или экспериментально [2, 4].

Более подробно тема устойчивости и неустойчивости задач и методов развивается при изучении численных процессов решения дифференциальных уравнений. Для решения задач, требующих особого учета возмущений и обобщения самого понятия решения, в последние 10-20 лет рассматриваются особые подходы, приводящие к построению специальных устойчивых численных методов, называемых ***методами регуляризации*** [45, 49].

1.5. ОБУСЛОВЛЕННОСТЬ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ СИСТЕМ

Учитывая распространенность систем линейных алгебраических уравнений (ибо часто именно к ним сводится на определенном этапе процесс математического моделирования), попытаемся количественно охарактеризовать степень неопределенности этих задач. Знание таких характеристик позволяет обоснованно судить о корректности моделей, грамотно подбирать методы и строить алгоритмы, правильно трактовать полученные результаты.

---

[*] См. также пример неустойчивой задачи на собственные значения для $n \times n$-матрицы в п.4 5

Рассмотрим линейную алгебраическую систему, записанную в виде векторно-матричного уравнения

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \qquad (1.10)$$

где $\mathbf{A}$ – невырожденная $n \times n$-матрица коэффициентов данной системы; $\mathbf{b}$ – ненулевой n-мерный вектор свободных членов; $\mathbf{x}$ – n-мерный вектор неизвестных (решение, если трактовать (1.10) как верное равенство).

Пусть правая часть (1.10) получила приращение ("возмущение") $\Delta\mathbf{b}$, т.е. вместо истинного вектора $\mathbf{b}$ используется приближенный вектор $\mathbf{b} + \Delta\mathbf{b}$. Реакцией решения $\mathbf{x}$ на возмущение $\Delta\mathbf{b}$ правой части будет вектор поправок $\Delta\mathbf{x}$, т.е. если $\mathbf{x}$ – решение (1.10), то $\mathbf{x} + \Delta\mathbf{x}$ – решение уравнения

$$\mathbf{A}(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b} + \Delta\mathbf{b}. \qquad (1.11)$$

Понимая под абсолютной погрешностью приближенного вектора норму*) разности между точным и приближенным векторами, а под относительной погрешностью – отношение абсолютной погрешности к норме вектора (точного или приближенного), выясним связь между относительными погрешностями вектора свободных членов и вектора-решения. Иначе, получим оценку вида

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \le (?)\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|},$$

где $\|\bullet\|$ – какая-либо векторная норма, а $(?)$ – неизвестный пока коэффициент связи.

Подставляя (1.10) в (1.11), видим, что поправка $\Delta\mathbf{x}$ связана с возмущением $\Delta\mathbf{b}$ таким же, как и (1.10), равенством

$$\mathbf{A}\,\Delta\mathbf{x} = \Delta\mathbf{b},$$

из которого находим ее явное выражение

$$\Delta\mathbf{x} = \mathbf{A}^{-1}\Delta\mathbf{b}. \qquad (1.12)$$

Нормируя равенства (1.10) и (1.12), будем иметь

$$\|\mathbf{b}\| \le \|\mathbf{A}\| \cdot \|\mathbf{x}\| \qquad \text{и} \qquad \|\Delta\mathbf{x}\| \le \left\|\mathbf{A}^{-1}\right\| \cdot \|\Delta\mathbf{b}\|,$$

где матричная норма должна быть согласованной с выбранной векторной нормой. Эти два числовых неравенства одинакового смысла можно перемножить:

$$\|\mathbf{b}\| \cdot \|\Delta\mathbf{x}\| \le \|\mathbf{A}\| \cdot \|\mathbf{x}\| \cdot \left\|\mathbf{A}^{-1}\right\| \cdot \|\Delta\mathbf{b}\|.$$

Из последнего делением на $\|\mathbf{b}\| \cdot \|\mathbf{x}\|$ получаем искомую связь

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \le \|\mathbf{A}\| \cdot \left\|\mathbf{A}^{-1}\right\| \cdot \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}. \qquad (1.13)$$

*) См. приложение 1

Положительное число $\|\mathbf{A}\|\cdot\|\mathbf{A}^{-1}\|$ – коэффициент этой связи – называют ***числом (мерой) обусловленности*** матрицы $\mathbf{A}$ и обозначают $\operatorname{cond}\mathbf{A}$ (от английского слова conditioned – "обусловленный"). Распространены также обозначения $\nu(\mathbf{A})$ и $\chi(\mathbf{A})$.

Легко показать, что то же самое число $\operatorname{cond}\mathbf{A}=\|\mathbf{A}\|\cdot\|\mathbf{A}^{-1}\|$ служит коэффициентом роста относительных погрешностей при неточном задании элементов матрицы $\mathbf{A}$ в (1.10). А именно, если матрица $\mathbf{A}$ получила возмущение $\Delta\mathbf{A}$ и $\mathbf{x}+\Delta\mathbf{x}$ – решение возмущенной системы

$$(\mathbf{A}+\Delta\mathbf{A})(\mathbf{x}+\Delta\mathbf{x})=\mathbf{b},$$

то справедливы неравенства*)

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|}\le\operatorname{cond}\mathbf{A}\cdot\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}+\Delta\mathbf{A}\|}\qquad\text{и}\qquad\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}+\Delta\mathbf{x}\|}\le\operatorname{cond}\mathbf{A}\cdot\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}\ .\quad(1.14)$$

Итак, неравенства (1.13) и (1.14) показывают, что *чем больше число обусловленности, тем сильнее сказывается на решении линейной системы ошибка в исходных данных*. Грубо говоря, если $\operatorname{cond}\mathbf{A}=O(10^{p})$ и исходные данные имеют погрешность в ℓ-м знаке после запятой, то независимо от способа решения системы (1.10) в результате можно гарантировать не более $\ell-p$ знаков после запятой.

Если число $\operatorname{cond}\mathbf{A}$ велико, то система считается плохо обусловленной. Говорить о том, "что такое хорошо и что такое плохо", в отрыве от контекста решаемой задачи почти бессмысленно, так как здесь может играть роль размерность задачи, точность, с которой должно быть найдено ее решение, точность представления чисел в ЭВМ и т.п. Однако можно дать оценку снизу числа обусловленности. А именно, если используются подчиненные матричные нормы (для которых норма единичной матрицы есть единица), то, очевидно,

---

*) Можно получить оценки, обобщающие в определенном смысле (1.13) и (1.14) в случае одновременного возмущения матрицы $\mathbf{A}$ системы (1.10) и ее правой части $\mathbf{b}$. Более того, величина $\nu(\mathbf{A})=\|\mathbf{A}\|\cdot\|\mathbf{A}^{-1}\|$ является также мерой обусловленности линейного обратимого оператора $\mathbf{A}$ в произвольном нормированном пространстве. Справедливо следующее утверждение [51].

Пусть $\mathbf{A}\mathbf{x}=\mathbf{b}$ – данное, а $\tilde{\mathbf{A}}\tilde{\mathbf{x}}=\tilde{\mathbf{b}}$ – возмущенное линейные операторные уравнения с относительными уровнями возмущений $\delta_{\mathbf{A}}\ge\dfrac{\|\mathbf{A}-\tilde{\mathbf{A}}\|}{\|\mathbf{A}\|}$ и $\delta_{\mathbf{b}}\ge\dfrac{\|\mathbf{b}-\tilde{\mathbf{b}}\|}{\|\mathbf{b}\|}$. Тогда если $\delta_{\mathbf{A}}\nu(\mathbf{A})<1$, то эти уравнения одновременно однозначно разрешимы и справедлива оценка относительной погрешности решения

$$\frac{\|\mathbf{x}-\tilde{\mathbf{x}}\|}{\|\mathbf{x}\|}\le\frac{\nu(\mathbf{A})}{1-\delta_{\mathbf{A}}\cdot\nu(\mathbf{A})}\cdot\delta_{\mathbf{b}}+\frac{\nu^{2}(\mathbf{A})}{1-\delta_{\mathbf{A}}\cdot\nu(\mathbf{A})}\cdot\delta_{\mathbf{A}}.$$

$$\operatorname{cond}\mathbf{A} = \|\mathbf{A}\| \cdot \left\|\mathbf{A}^{-1}\right\| \geq \left\|\mathbf{A} \cdot \mathbf{A}^{-1}\right\| = \|\mathbf{E}\| = 1,$$

т.е. число обусловленности не может быть меньше 1. Можно также указать верхнюю границу для чисел обусловленности, превышение которой при решении линейных систем на конкретной ЭВМ может привести к заведомо ложным результатам. Так, решение считается ненадежным, если $\operatorname{cond}\mathbf{A} \geq (macheps)^{-1}$ или даже $\operatorname{cond}\mathbf{A} \geq (macheps)^{-0.5}$ [22]. При этом заметим, что масштабированием матрицы $\mathbf{A}$ путем умножения на скаляр α ее обусловленность не улучшить, ибо

$$\operatorname{cond}(\alpha\mathbf{A}) = \|\alpha\mathbf{A}\| \cdot \left\|(\alpha\mathbf{A})^{-1}\right\| = \|\mathbf{A}\| \cdot \left\|\mathbf{A}^{-1}\right\| = \operatorname{cond}\mathbf{A}.$$

Классическим примером плохо обусловленной матрицы является так называемая ***матрица Гильберта***

$$\mathbf{H}_n = \left(\frac{1}{i+j-1}\right)_{i,j=1}^{n},$$

демонстрирующая катастрофическое возрастание числа обусловленности с ростом размерности [41, 55]. Так, уже при n=8 $\operatorname{cond}\mathbf{H}_8 > 10^{10}$ и обратная матрица $\mathbf{H}_8^{-1}$, полученная на машине с точностью представления чисел порядка 10^{-8}, может не содержать ни одного верного знака.

Очевидно, число обусловленности зависит от выбора матричной нормы (индуцированной, как правило, той или иной векторной нормой, в терминах которой характеризуется относительная погрешность решения алгебраической системы). Однако нетрудно получить оценку числа обусловленности через собственные числа матрицы. Действительно, пусть собственные числа λ_i матрицы $\mathbf{A}$ упорядочены по модулю:

$$|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|,$$

т.е. ***спектральный радиус*** матрицы $\mathbf{A}$ есть $\rho_{\mathbf{A}} = |\lambda_1|$. Тогда в силу известного неравенства $\rho_{\mathbf{A}} \leq \|\mathbf{A}\|$ и соотношения между собственными числами прямой и обратной матриц, имеем

$$\|\mathbf{A}\| \cdot \left\|\mathbf{A}^{-1}\right\| \geq \rho_{\mathbf{A}} \cdot \rho_{\mathbf{A}^{-1}} = |\lambda_1| \cdot \frac{1}{|\lambda_n|}.$$

Таким образом, оценкой снизу меры обусловленности матрицы $\mathbf{A}$ может служить величина $\left|\frac{\lambda_1}{\lambda_n}\right|$ (называемая иногда ***числом обусловленности Тодда*** [РЖМат, 1983, 10Б937]). Для симметричных матриц эта оценка и на самом деле является числом обусловленности, соответствующим спектральной норме матрицы (индуцированной евклидовой нормой вектора). Учитывая смысл собственных чисел матрицы, можно сказать, что число обусловленности показывает величину отношения наибольшего коэффи-

циента растяжения вектора посредством линейного преобразования $\mathbf{A}$ к наименьшему.

Следует отметить, что непосредственный подсчет чисел обусловленности, особенно при большой размерности матриц, является весьма дорогостоящим делом из-за необходимости обращать матрицы или находить их собственные значения. Поэтому зачастую о приемлемости порядка возможного роста относительной погрешности результата решения какой-либо алгебраической задачи относительно данной матрицы судят либо по каким-то достаточным признакам (например, по доминированию элементов главной диагонали матрицы), либо на основе теоретического изучения матрицы [41, 46], либо путем применения специальных алгоритмов приближенного оценивания cond A [22]. Исследование матриц на обусловленность может быть естественным образом связано со способом решения той или иной алгебраической задачи относительно данной матрицы.

Прокомментируем теперь пример неустойчивой системы, приведенной в примере в) п.1.4.

Матрица коэффициентов системы $\mathbf{A} = \begin{pmatrix} 1 & 10 \\ 100 & 1001 \end{pmatrix}$ имеет обратную $\mathbf{A}^{-1} = \begin{pmatrix} 1001 & -10 \\ -100 & 1 \end{pmatrix}$. Следовательно, число обусловленности в матричной норме, индуцированной векторной нормой-максимум (иначе, нормой l_∞), есть

$$\nu_\infty(\mathbf{A}) = 1101 \cdot 1011 = 1113111 > 10^6.$$

Учитывая, что в данном примере $\mathbf{b} = \begin{pmatrix} 11 \\ 1101 \end{pmatrix}$, $\Delta\mathbf{b} = \begin{pmatrix} 0.01 \\ 0 \end{pmatrix}$, на основе (1.13) получаем оценку относительной погрешности решения в l_∞-нормах

$$\delta_{\mathbf{x}} \le \nu_\infty(\mathbf{A}) \cdot \delta_{\mathbf{b}} = 1113111 \cdot \frac{0.01}{1101} = 10.11.$$

Так как норма-максимум решения $\mathbf{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ равна 1, то оценка абсолютной погрешности $\|\Delta\mathbf{x}\|$ решения равна

$$\|\Delta\mathbf{x}\| = \|\mathbf{x}\| \cdot \delta_{\mathbf{x}} \le 10.11.$$

Как видим, решение $\mathbf{x} + \Delta\mathbf{x} = \begin{pmatrix} 11.01 \\ 0 \end{pmatrix}$ возмущенной системы вписывается в оценку

$$\|\mathbf{x} + \Delta\mathbf{x}\| \le \|\mathbf{x}\| + \|\Delta\mathbf{x}\| \le 1 + 10.11 = 11.11.$$

Аналогичный результат может быть получен через число обусловленности Тодда. Решая характеристическое уравнение

$$\lambda^2 - 1002\lambda + 1 = 0,$$

находим собственные числа матрицы $\mathbf{A}$: $\lambda_1 \approx 1002$ и $\lambda_2 \approx 0.000998$, дающие оценку

$$\operatorname{cond}\mathbf{A} \geq \frac{\lambda_1}{\lambda_2} \approx 1004000 > 10^6.$$

На данном примере также можно наглядно убедиться в том, что *малость* ***невязки*** $\mathbf{r} = \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}$ *плохо обусловленной системы еще не говорит о близости приближенного решения* $\tilde{\mathbf{x}}$ *к точному* $\mathbf{x}$. Действительно, невязки $\mathbf{r}_1$ и $\mathbf{r}_2$ векторов $\tilde{\mathbf{x}}_1 = \begin{pmatrix} 11.01 \\ 0 \end{pmatrix}$ и $\tilde{\mathbf{x}}_2 = \begin{pmatrix} 1 \\ 1.1 \end{pmatrix}$ для основной (невозмущенной) системы из примера в) п.1.4 имеют нормы соответственно $\|\mathbf{r}_1\|_\infty = 0.01$ и $\|\mathbf{r}_2\|_\infty = 100.1$. Вектор $\tilde{\mathbf{x}}_2$, явно более близкий к точному решению $\mathbf{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, чем $\tilde{\mathbf{x}}_1$, имеет существенно большую невязку!

Двумерный случай допускает простую геометрическую трактовку понятия обусловленности. Плохая обусловленность системы двух уравнений с двумя неизвестными означает, что прямые, являющиеся геометрическими образами уравнений, пересекаются на координатной плоскости под очень острым углом. В этом случае небольшое искажение в данных, интерпретируемое как параллельный перенос (при возмущении свободного члена) или поворот прямых (при возмущении матрицы коэффициентов), приводит к значительному перемещению их точки пересечения, т.е. геометрического образа решения.

1.6. ПОГРЕШНОСТИ КОРНЕЙ СКАЛЯРНЫХ УРАВНЕНИЙ С ПРИБЛИЖЕННЫМИ КОЭФФИЦИЕНТАМИ

Пусть решается алгебраическое или трансцендентное уравнение вида

$$f(x) = 0. \qquad (1.15)$$

Это уравнение, являясь нелинейным, может иметь один или несколько корней или не иметь их вовсе. Каждый корень так или иначе зависит от числовых данных уравнения (1.15). Эти числовые данные (будем иногда называть их коэффициентами уравнения) можно считать фиксированными значениями параметров, т.е. уравнение (1.15) целесообразно рассматривать в качестве представителя семейства уравнений

$$f(x, a_1, a_2, \ldots, a_m) = 0 \qquad (1.16)$$

и наличие корней у (1.15) связывать с существованием неявной функции $x = \varphi(a_1, a_2, \ldots, a_m)$ при данном наборе значений параметров (коэффициентов) $a_1, a_2, \ldots, a_m$. Поскольку эти числа, как правило, точно не известны (грубость модели, неточность измерений, усечение чисел при вводе в ЭВМ

и т.п.), встает вопрос о том, как влияет погрешность коэффициентов уравнения (1.15) на погрешности его корней. Иначе, с какой точностью имеет смысл решать данное уравнение, если известно, что его коэффициенты не точны, но имеется информация об уровне их погрешностей?

Реально можно оценить лишь главную часть погрешности корня, понимая под ней, как и в п.1.1, модуль дифференциала. В данном случае речь идет о дифференциале неявной функции нескольких переменных. Подходящую основу для этого находим в математическом анализе: если, например, выполняются условия, при которых уравнение $F(x,y,z)=0$ неявно определяет дифференцируемую функцию $z=z(x,y)$, то выражение дифференциала dz этой функции можно получить из равенства $dF=0$, а именно,

$$dz=-\frac{1}{F'_z}\left(F'_x dx+F'_y dy\right).$$

Аналогично, если $x=\varphi(a_1,a_2,\ldots,a_m)$ – корень уравнения (1.16), то линейная часть его изменения, соответствующая изменениям аргументов – коэффициентов $a_1,a_2,\ldots,a_m$, равна

$$dx=-\frac{1}{\frac{\partial f}{dx}}\left(\frac{\partial f}{\partial a_1}da_1+\frac{\partial f}{\partial a_2}da_2+\ldots+\frac{\partial f}{\partial a_m}da_m\right).$$

Переходя здесь к модулям и заменяя истинные абсолютные погрешности коэффициентов $\Delta a_i\left(=da_i\right)$ их оценками Δ_{a_i}, получим ***формулу для оценки абсолютной погрешности корня:***

$$|\Delta x|\lesssim\frac{1}{\left|\frac{\partial f}{dx}\right|}\left(\left|\frac{\partial f}{\partial a_1}\right|\Delta_{a_1}+\left|\frac{\partial f}{\partial a_2}\right|\Delta_{a_2}+\ldots+\left|\frac{\partial f}{\partial a_m}\right|\Delta_{a_m}\right). \qquad (1.17)$$

Заметим, что при вычислении значений частных производных в (1.17) следует пользоваться фиксированными значениями коэффициентов $a_1,a_2,\ldots,a_m$ (такими, какими они используются при нахождении корня) и приближенным значением того корня, степень неопределенности которого устанавливается. Ясно, что для разных корней одного и того же уравнения значения этой величины могут сильно различаться.

Следуя [23], получаемую с помощью (1.17) оценку будем называть ***безусловной абсолютной погрешностью*** приближенного корня x уравнения (1.16) с приближенными коэффициентами и обозначать б.п.x. Это же x есть точный корень уравнения (1.15), в то время как при численном решении уравнения вместо x будет получено некоторое приближение к нему $\bar{\bar{x}}$. Возникающую при этом остаточную погрешность или погрешность метода назовем ***условной погрешностью*** и обозначим у.п.x.

Итак, в ходе численного решения уравнения (1.15) в предположении, что его коэффициенты точны, находится приближенный корень $\bar{\bar{x}}$ с условной погрешностью у.п.x. Это означает, что точный корень x уравне-

ния (1.15) при данном предположении лежит на интервале ($\bar{\bar{x}}$ –у.п.x, $\bar{\bar{x}}$ + у.п.x). Зная погрешности коэффициентов и приближенный корень $\bar{\bar{x}}$, можно подсчитать безусловную погрешность б.п.x. Истинная величина корня x^* уравнения с приближенными коэффициентами – это потенциально любое число из интервала (x – б.п.x, x + б.п.x). Ясно, что гипотетическое значение x^* – решение рассчитываемой модели – может отличаться от реально полученного значения $\bar{\bar{x}}$ на величину у.п.x + б.п.x, т.е. *полная погрешность корня уравнения с приближенными коэффициентами складывается из погрешностей условной и безусловной*. Так как величиной условной погрешности распоряжается вычислитель, причем ее уменьшение в разумных пределах не вызывает, как правило, больших затруднений, то обычно, чтобы не увеличивать полную погрешность корня приближенного уравнения, условную погрешность задают неравенством

$$\text{у.п.}x \le \text{б.п.}x \tag{1.18}$$

или даже

$$\text{у.п.}x \le 0.1\ \text{б.п.}x .$$

Если воспользоваться ***правилом Ньютона***[*)] оценки близости приближенного корня (нуля) $\bar{\bar{x}}$ дифференцируемой функции $f(x)$ к точному корню x

$$|\bar{\bar{x}} - x| \le \left|\frac{f(\bar{\bar{x}})}{f'(x)}\right| \tag{1.19}$$

и оценкой (1.17), то можно на основе (1.18) получить простой критерий окончания процесса численного решения скалярного уравнения (1.15) с приближенными коэффициентами:

уточнение корня $\bar{\bar{x}} \approx x$ ведется до тех пор, пока не выполнится условие

$$|f(\bar{\bar{x}})| \le \left|\frac{\partial f}{\partial a_1}\right|\Delta_{a_1} + \left|\frac{\partial f}{\partial a_2}\right|\Delta_{a_2} + \ldots + \left|\frac{\partial f}{\partial a_m}\right|\Delta_{a_m} . \tag{1.20}$$

Как видим, большу́ю роль в неравенстве (1.20) играют модули частных производных данной функции по параметрам, называемые ***коэффициентами чувствительности***.

В случае, когда уравнение (1.15) – алгебраическое, т.е.

[*)] По формуле конечных приращений Лагранжа при некотором $\Theta \in (\bar{\bar{x}}; x)$ (или $(x, \bar{\bar{x}})$)

$$f(\bar{\bar{x}}) - f(x) = f'(\Theta)(\bar{\bar{x}} - x).$$

Предполагая, что x – корень, $\bar{\bar{x}} \approx x$, а значит и $\Theta \approx \bar{\bar{x}}$, имеем

$$f(\bar{\bar{x}}) \approx f'(\bar{\bar{x}})(\bar{\bar{x}} - x),$$

откуда следует правило Ньютона (1.19).

$$f(x) \equiv P(x) := a_0 x^n + a_1 x^{n-1} + \ldots + a_{n-1} x + a_n, \qquad (1.21)$$

его корни – функции коэффициентов многочлена $a_0, a_1, \ldots, a_n$, а коэффициенты чувствительности суть числа $|x^n|, |x^{n-1}|, \ldots, |x|, 1$, получающиеся при подстановке сюда вместо x приближенных корней многочлена. Неравенство типа (1.20), закладываемое в процесс уточнения корня $\bar{\bar{x}}$ многочлена (1.21) с приближенными коэффициентами $a_0(\pm\Delta_{a_0})$, $a_1(\pm\Delta_{a_1})$, ... , $a_n(\pm\Delta_{a_n})$, имеет вид

$$|P(\bar{\bar{x}})| \le |\bar{\bar{x}}^n|\Delta_{a_0} + |\bar{\bar{x}}^{n-1}|\Delta_{a_1} + \ldots + |\bar{\bar{x}}|\Delta_{a_{n-1}} + \Delta_{a_n}.$$

Легко заметить, какую роль играют погрешности тех или иных коэффициентов в зависимости от того, малы или велики модули корней.

Получим, наконец, связь между относительными погрешностями коэффициентов и корней многочлена.

На основе оценки (1.17) для многочлена (1.21) имеем

$$\delta x = \left|\frac{\Delta x}{x}\right| \le \frac{\sum_{i=0}^{n}\left|\frac{\partial P}{\partial a_i}\right|\Delta_{a_i}}{|x|\cdot\left|\frac{\partial P}{\partial x}\right|} = \frac{\sum_{i=0}^{n}|x^{n-i}|\cdot|a_i|\cdot\delta_{a_i}}{|x|\cdot\left|\sum_{i=0}^{n-1}(n-i)a_i x^{n-1-i}\right|} = \frac{\sum_{i=0}^{n}|a_i x^{n-i}|\delta_{a_i}}{\left|\sum_{i=0}^{n-1}(n-i)a_i x^{n-i}\right|}.$$

Если сделать допущение, что все коэффициенты многочлена имеют одинаковый уровень δ относительных погрешностей, или положить $\max_{i=0,\ldots,n}\delta_{a_i} = \delta$, то для граничной относительной погрешности δ_x простого ненулевого корня x будем иметь приближенную оценку

$$\delta_x \le \frac{\sum_{i=0}^{n}|a_i x^{n-i}|}{\left|\sum_{i=0}^{n-1}(n-i)a_i x^{n-i}\right|}\cdot\delta. \qquad (1.22)$$

Коэффициент

$$\nu_x = \frac{\sum_{i=0}^{n}|a_i\, x^{n-i}|}{\left|\sum_{i=0}^{n-1}(n-i)\,a_i\, x^{n-i}\right|}$$

связи (1.22) относительных погрешностей корней и коэффициентов (для каждого корня свой) по аналогии с терминологией предыдущего пункта

естественно назвать ***числом*** или ***мерой обусловленности ненулевого простого корня***[*)] ***многочлена с приближенными коэффициентами***.

Обращаясь к примеру Уилкинсона в п.1.4б), с помощью подсчета чисел обусловленности корней можно грубо объяснить картину разного влияния на разные корни внесенного малого возмущения. Если для первого корня число обусловленности $\nu_x \approx 400$, то с увеличением номера корня это число значительно возрастает, достигая, например, на пятнадцатом корне величины большей, чем 10^{10}. Последний (двадцатый) корень имеет меньшее, чем предыдущий, число обусловленности, сравнимое с величиной $1/macheps$. (Значения несколько иначе определенных чисел обусловленности можно найти в [1].)

Отметим, что в числителе выражения для подсчета числа обусловленности ν_x могут отсутствовать слагаемые, соответствующие теоретически точным и при этом точно реализуемым при вводе в ЭВМ значениям коэффициентов многочлена. Это видно из оценки δ_x, где некоторые из Δ_{a_i} могут быть тождественно равными нулю (т.е. соответствующие коэффициенты не считаются параметрами приближенного уравнения (1.16)).

УПРАЖНЕНИЯ

1.1. Найти приближенное значение выражения

$$\frac{3.7894 \cdot 0.29^2}{5.63},$$

если известно, что входящие в него числа (за исключением показателя степени) – приближенные, записанные в соответствии с правилом Крылова.

Сделать теоретическую оценку главной части абсолютной погрешности результата.

1.2. Пусть машинное слово некой ЭВМ состоит из 16 четверичных разрядов.

[*)] В случае нулевого или кратного корня x знаменатель дроби обращается в нуль.

В [26] мерой обусловленности простого корня $x = x(a_1, a_2, \cdots, a_n)$ многочлена (1.21) с $a_0 = 1$ считается норма градиента $\left(\frac{\partial x}{\partial a_1}, \frac{\partial x}{\partial a_2} \cdots \frac{\partial x}{\partial a_n}\right)$. Если все корни $x_1, x_2, \cdots x_n$ – простые, то обусловленность многочлена характеризуется нормой матрицы Якоби $\left(\frac{\partial x_i}{\partial a_j}\right)_{i,j=1}^{n}$ (см приложение 2)

а) Найти приближенно значения машинного нуля M_0, машинной бесконечности M_∞ и машинного эпсилон в системе представления с плавающей запятой, если под мантиссу выделяется 11 разрядов.

б) Оценить абсолютную погрешность и диапазон представимости чисел с фиксированной запятой, если под целую и дробную части числа выделяется поровну разрядов.

(Считается, что для отображения знаков + , – и , используются те же четверичные элементы и что округление производится на основе простого отбрасывания).

1.3. Обращением матрицы коэффициентов решить систему:

$$\begin{cases} x_1 - x_3 = 2, \\ x_1 + x_2 + x_3 = 6, \\ x_2 + 3x_3 = 5. \end{cases}$$

Найдя число обусловленности, оценить возможные относительное и абсолютное отклонение от полученного решения, если допустить относительную ошибку в правой части порядка 0.01 (по некоторой норме).

1.4. Пусть коэффициенты уравнения

$$x^2 - 6.9x - 0.0212 = 0$$

являются приближенными числами, записанными в соответствии с принципом Крылова (коэффициент при x^2 и правую часть считаем числами точными).

а) Найти корни уравнения с максимальной разумной точностью, диктуемой степенью неопределенности коэффициентов.

б) Подсчитать числа обусловленности для найденных корней и на этой основе дать приближенную оценку их относительных и абсолютных погрешностей.

ГЛАВА 2 РЕШЕНИЕ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ СИСТЕМ (ПРЯМЫЕ МЕТОДЫ)

Рассматриваются простые, часто применяемые численные методы решения систем линейных алгебраических уравнений с квадратными матрицами коэффициентов. Попутно решаются задачи обращения матриц и вычисления определителей. Наряду с широко известным методом Гаусса, изучаемым здесь в плане реальных вычислений, вниманию читателя предлагаются также метод вращений (более устойчивый к погрешностям арифметических действий, чем метод Гаусса), метод квадратных корней для решения систем с симметричными матрицами и метод прогонки для систем с трехдиагональными матрицами (т.е. для трехточечных разностных уравнений второго порядка).

2.0. ВВЕДЕНИЕ

Как утверждается в книге известного американского математика Варяха [11], 75% всех расчетных математических задач приходится на решение систем линейных алгебраических уравнений (в дальнейшем, СЛАУ). Это не удивительно, так как математические модели тех или иных явлений или процессов либо сразу строятся как линейные алгебраические, либо сводятся к таковым посредством дискретизации и/или линеаризации. Поэтому трудно переоценить роль, какую играет выбор эффективного (в том или ином смысле) способа решения СЛАУ. Современная вычислительная математика располагает большим арсеналом методов, а математическое обеспечение ЭВМ – многими пакетами прикладных программ. позволяющих решать различные возникающие на практике линейные системы. Чтобы ориентироваться в этом море методов и программ и в нужный момент сделать оптимальный выбор, нужно разбираться в основах построений методов и алгоритмов, учитывающих специфику постановок задач, знать их сильные и слабые стороны и границы применимости.

Все методы решения линейных алгебраических задач (наряду с задачей решения СЛАУ, это и вычисление определителей, и обращение матриц, и задачи на собственные значения) можно разбить на два класса: прямые и итерационные. Как явствует из заглавия, здесь будут рассмотрены только ***прямые методы,*** т.е. такие методы, которые приводят к решению

за конечное число арифметических операций. Если операции реализуются точно, то и решение также будет точным (в связи с чем к классу прямых методов применяют еще название ***точные методы***). ***Итерационным методам,*** т.е. методам, в которых точное решение может быть получено лишь в результате бесконечного повторения единообразных (как правило, простых) действий,будет посвящена следующая глава.

Другое ограничение будет касаться рассматриваемых систем. Условимся говорить о численном решении таких СЛАУ, у которых число уравнений совпадает с числом вещественных неизвестных, причем будем предполагать наличие единственного решения*), если существование и единственность не следует из каких-либо условий.

Такое ограничение здесь довольно естественно, так как решение и недоопределенных, и переопределенных систем, а также систем с комплексными коэффициентами и переменными, в конечном счете, сводится к решению однозначно определенных вещественных систем.

Итак, изучается вопрос о численном решении систем вида

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots & \ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= b_n, \end{cases} \qquad (2.1)$$

или иначе, векторно-матричных уравнений

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \qquad (2.1')$$

где $\mathbf{b} = (b_1, b_2, \ldots, b_n)^T$ – вектор свободных членов и $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ – вектор неизвестных (он же в другой интерпретации может означать и вектор-решение) с вещественными координатами, а $\mathbf{A} = (a_{ij})_{i,j=1}^{n}$ – вещественная $n \times n$-матрица коэффициентов данной системы. Эффективность способов решения системы (2.1) во многом зависит от структуры и свойств матрицы $\mathbf{A}$: размерности, обусловленности (см. п.1.5), симметричности, заполненности (т.е. соотношения между числом ненулевых и нулевых элементов), специфики расположения ненулевых элементов в матрице и др.

Так, размерность системы (т.е. число n) является главным фактором, заставляющим вычислителей отвернуться от весьма привлекательных в

*) Обратим внимание на двоякое употребление в тексте термина "решение":

1) решение – как упорядоченный набор (вектор) значений неизвестных, удовлетворяющих всем уравнениям системы;

2) решение – как процесс нахождения решения в первом смысле.

теоретическом плане и приемлемых на практике при небольших n (2, 3) ***формул Крамера***

$$x_i = \frac{\det \mathbf{A}_i}{\det \mathbf{A}} \qquad (i = 1, 2, \ldots, n),$$

позволяющих находить неизвестные в виде дробей, знаменателем которых является определитель матрицы системы, а числителем – определители матриц $\mathbf{A}_i$, полученные из $\mathbf{A}$ заменой столбца коэффициентов при вычисляемом неизвестном столбцом свободных членов. Если при реализации этих формул определители вычисляются понижением порядка на основе разложения по элементам какой-нибудь строки или столбца матрицы, то на вычисление определителя n-го порядка будет затрачиваться $n!$ операций умножения. Факториальный рост количества арифметических операций (и вообще, очень быстрый рост) с увеличением размерности задачи называют "проклятьем размерности". Что это такое, можно представить, зафиксировав, например, $n = 100$. Оценив величину $100! \approx 10^{158}$ и прикинув потенциальные возможности развития вычислительной техники, приходим к выводу о том, что в обозримом будущем системы сотого порядка в принципе не могут быть решены по формулам Крамера [5, 33]. Заметим при этом, что, во-первых, метод Крамера будет неустойчив, т.е. погрешности округлений будут катастрофически нарастать, во-вторых, размерность $n = 100$ для современных задач не так и велика: довольно часто решаются системы с сотнями и с тысячами неизвестных.

Если осуществлять вычисление обратной матрицы с помощью союзной матрицы, т. е. через алгебраические дополнения, то нахождение решения векторно-матричного уравнения (2.1') по формуле

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

фактически равнозначно применению формул Крамера и также практически непригодно по упомянутым выше причинам для вычислительных целей.

2.1. АЛГОРИТМ РЕШЕНИЯ СЛАУ МЕТОДОМ ГАУССА С ПОСТОЛБЦОВЫМ ВЫБОРОМ ГЛАВНОГО ЭЛЕМЕНТА

Наиболее известным и популярным способом решения линейных систем вида (2.1) является метод Гаусса. Суть его проста – это последовательное исключение неизвестных. В отличие от курсов линейной алгебры, нас будут интересовать вычислительные аспекты метода Гаусса, а именно, технология получения вектора-решения $\mathbf{x}$ из исходных матрицы $\mathbf{A}$ и вектора $\mathbf{b}$, причем, по возможности, минимизирующая влияние неизбежных ошибок округлений. С этой целью, работая с уравнениями системы (2.1), выведем сначала совокупность формул, позволяющих в итоге получить ис-

комые значения неизвестных, а затем на их основе запишем алгоритм решения поставленной задачи.

Будем поэтапно приводить систему (2.1) к треугольному виду, исключая последовательно сначала x_1 из второго, третьего, ... , n-го уравнений, затем x_2 из третьего, четвертого, ... , n-го уравнений преобразованной системы, и т.д.

На первом этапе заменим второе, третье, ... , n-е уравнение на уравнения, получающиеся сложением этих уравнений с первым, умноженным соответственно на $-\frac{a_{21}}{a_{11}}, -\frac{a_{31}}{a_{11}}, \ldots, -\frac{a_{n1}}{a_{11}}$. Результатом этого этапа преобразований будет эквивалентная (2.1) система

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1, \\ a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + \ldots + a_{2n}^{(1)}x_n = b_2^{(1)}, \\ a_{32}^{(1)}x_2 + a_{33}^{(1)}x_3 + \ldots + a_{3n}^{(1)}x_n = b_3^{(1)}, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n2}^{(1)}x_2 + a_{n3}^{(1)}x_3 + \ldots + a_{nn}^{(1)}x_n = b_n^{(1)}, \end{cases} \qquad (2.2)$$

коэффициенты которой (с верхним индексом 1) подсчитываются по формулам

$$a_{ij}^{(1)} = a_{ij} - \frac{a_{i1}}{a_{11}}a_{1j}\ , \quad b_i^{(1)} = b_i - \frac{a_{i1}}{a_{11}}b_1\ \ , \text{ где } i, j = 2, 3, \ldots, n.$$

При этом можно считать, что $a_{11} \neq 0$, так как по предположению система (2.1) однозначно разрешима, значит, не могут все коэффициенты при x_1 одновременно равняться нулю, и на первое место всегда можно поставить уравнение с отличным от нуля первым коэффициентом.

На втором этапе проделываем такие же операции, как и на первом, с подсистемой системы (2.2), получающейся исключением первого уравнения. Эквивалентный (2.2) (а значит, и (2.1)) результат второго этапа будет иметь вид

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1, \\ a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + \ldots + a_{2n}^{(1)}x_n = b_2^{(1)}, \\ a_{33}^{(2)}x_3 + \ldots + a_{3n}^{(2)}x_n = b_3^{(2)}, \\ \ldots\ldots\ldots\ldots\ldots \\ a_{n3}^{(2)}x_3 + \ldots + a_{nn}^{(2)}x_n = b_n^{(2)}, \end{cases}$$

где $a_{ij}^{(2)} = a_{ij}^{(1)} - \frac{a_{i2}^{(1)}}{a_{22}^{(1)}}a_{2j}^{(1)}\ , \quad b_i^{(2)} = b_i^{(1)} - \frac{a_{i2}^{(1)}}{a_{22}^{(1)}}b_2^{(1)}\ \ ;\ i, j = 3, \ldots, n.$

Продолжая этот процесс, на $(n-1)$-м этапе так называемого ***прямого хода*** метода Гаусса данную систему (2.1) приведем к треугольному виду:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1, \\ a_{22}^{(1)}x_2 + \ldots + a_{2n}^{(1)}x_n = b_2^{(1)}, \\ \ldots\ldots\ldots\ldots \\ a_{nn}^{(n-1)}x_n = b_n^{(n-1)}. \end{cases} \qquad (2.3)$$

На основе предыдущих рассуждений и формул легко убедиться, что коэффициенты этой системы могут быть получены из коэффициентов данной системы последовательным пересчетом по формулам

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} a_{kj}^{(k-1)}, \quad b_i^{(k)} = b_i^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} b_k^{(k-1)}, \qquad (2.4)$$

где верхний индекс k (номер этапа) должен изменяться от 1 до $n-1$, нижние индексы i и j (в любой очередности) – от $k+1$ до n; по определению полагаем $a_{ij}^{(0)} := a_{ij}$, $b_i^{(0)} := b_i$.

Треугольная, точнее, трапециевидная структура системы (2.3) позволяет последовательно одно за другим вычислять значения неизвестных, начиная с последнего:

$$x_n = \frac{b_n^{(n-1)}}{a_{nn}^{(n-1)}};$$

$$\ldots\ldots\ldots$$

$$x_2 = \frac{b_2^{(1)} - a_{23}^{(1)}x_3 - \ldots - a_{2n}^{(1)}x_n}{a_{22}^{(1)}};$$

$$x_1 = \frac{b_1 - a_{12}x_2 - \ldots - a_{1n}x_n}{a_{11}}.$$

Этот процесс последовательного вычисления значений неизвестных называют ***обратным ходом*** метода Гаусса. Очевидно, он определяется одной формулой

$$x_k = \frac{1}{a_{kk}^{(k-1)}} \left(b_k^{(k-1)} - \sum_{j=k+1}^{n} a_{kj}^{(k-1)} x_j \right), \qquad (2.5)$$

где k полагают равным $n, n-1, \ldots, 2, 1$ и сумма по определению считается равной нулю, если нижний предел суммирования у знака Σ имеет значение больше верхнего.

Итак, решение СЛАУ вида (2.1) методом Гаусса сводится к последовательной реализации вычислений по формулам (2.4) и (2.5).

Учитывая цикличность выполняемых при этом операций, а также нецелесообразность хранения промежуточных результатов (пересчитывае-

мых коэффициентов промежуточного этапа), запишем простой **алгоритм решения линейных систем (2.1) методом Гаусса** [41]:

1) для k=1, 2, ..., n-1,

2) для $i = k+1, \ldots, n$:

3) $t_{ik} := a_{ik} / a_{kk}$,

4) $b_i := b_i - t_{ik} b_k$;

5) для $j=k+1, \ldots, n$:

6) $a_{ij} := a_{ij} - t_{ik} a_{kj}$.

7) $x_n := b_n / a_{nn}$;

8) для $k = n-1, \ldots, 2, 1$:

9) $x_k := \left(b_k - \sum_{j=k+1}^{n} a_{kj} x_j \right) / a_{kk}$.

Подав на его вход квадратную матрицу $\left(a_{ij}\right)_{i,j=1}^{n}$ коэффициентов при неизвестных системы (2.1) и вектор $\left(b_i\right)_{i=1}^{n}$ свободных членов и выполнив три вложенных цикла вычислений прямого хода (строки 1–6) и один цикл вычислений обратного хода (строки 7–9), на выходе алгоритма получим вектор-решение $\left(x_k\right)_{k=1}^{n}$ (в обратном порядке), если, разумеется, ни один из знаменателей не обращается в нуль и все вычисления проводятся точно.

Так как реальные машинные вычисления производятся не с точными, а с усеченными числами, т.е. неизбежны ошибки округления, то анализируя, например, формулы (2.4), можно сделать вывод о том, что выполнение алгоритма может прекратиться или привести к неверным результатам, если знаменатели дробей на каком-то этапе окажутся равными нулю или очень маленькими числами. Чтобы уменьшить влияние ошибок округлений и исключить деление на нуль, на каждом этапе прямого хода уравнения системы (точнее, обрабатываемой подсистемы) обычно переставляют так, чтобы деление производилось на наибольший по модулю в данном столбце (обрабатываемом подстолбце) элемент. Числа, на которые производится деление в методе Гаусса, называются ***ведущими*** или ***главными элементами.*** Отсюда название рассматриваемой модификации метода, исключающей деление на нуль и уменьшающей вычислительные погрешности, ***метод Гаусса с постолбцовым выбором главного элемента*** (или, иначе, ***с частичным упорядочиванием по столбцам***).

Частичное упорядочивание по столбцам требует внесения в алгоритм следующих изменений: между строками 1 и 2 нужно сделать вставку*)

"Найти $m \geq k$ такое, что $|a_{mk}| = \max\limits_{i \geq k}\{|a_{ik}|\}$;

⊗ если $a_{mk} = 0$, остановить работу алгоритма ("однозначного решения нет"),

иначе поменять местами b_k и b_m, a_{kj} и a_{mj} при всех $j = k, \ldots, n$."

Более разумным, наверное, является сравнение $|a_{mk}|$ не с нулем, а с некоторым малым $\varepsilon > 0$, задаваемым вычислителем в зависимости от различных априорных соображений. Счет останавливается или берется под особый контроль, если окажется $|a_{mk}| < \varepsilon$. Заметим, что соответствующая частичному упорядочиванию вставка в алгоритм позволяет фактически в процессе его выполнения проводить алгоритмически исследование системы (2.1) на однозначную разрешимость.

Устойчивость алгоритма к погрешностям исходных данных и результатов промежуточных вычислений можно еще усилить, если выполнять деление на каждом этапе на элемент, наибольший по модулю во всей матрице преобразуемой на данном этапе подсистемы. Такая модификация метода Гаусса, называемая ***методом главных элементов***, применяется довольно редко, поскольку сильно усложняет алгоритм. Усложнение связано как с необходимостью осуществления двумерного поиска главных элементов, так и с необходимостью запоминать номера столбцов, откуда берутся эти элементы (перестановка столбцов означает как бы переобозначение неизвестных, в связи с чем требуется обратная замена).

2.2. ПРИМЕНЕНИЕ МЕТОДА ГАУССА К ВЫЧИСЛЕНИЮ ОПРЕДЕЛИТЕЛЕЙ И К ОБРАЩЕНИЮ МАТРИЦ

Как было упомянуто во введении, решения линейных алгебраических систем можно получать с помощью определителей или обратных матриц. Однако нетрудно увидеть, что более эффективно поступать наоборот: вычислять определители и обращать матрицы в рамках метода Гаусса решения линейных систем.

Действительно, выполняемые в методе Гаусса преобразования прямого хода, приведшие матрицу **A** системы к треугольному виду (см. (2.3)), таковы, что они не изменяют определителя матрицы **A**. Учитывая, что оп-

*) Чтобы частичное упорядочивание было более эффективным, перед этим целесообразно произвести ***масштабирование*** (уравновешивание) системы, например, разделить все числа каждой строки на наибольшее число строки [41]

ределитель треугольной матрицы равен произведению диагональных элементов, имеем

$$\det \mathbf{A} = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ 0 & a_{22}^{(1)} & \dots & a_{2n}^{(1)} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_{nn}^{(n-1)} \end{vmatrix} = a_{11} \cdot a_{22}^{(1)} \dots a_{nn}^{(n-1)} .$$

Таким образом, det **A** *равен произведению всех ведущих элементов метода Гаусса.*

При желании получить det **A** дополнительно к решению СЛАУ **Ax = b** алгоритм предыдущего пункта должен быть пополнен всего лишь одной строкой:

$$10) \quad \det \mathbf{A} = \prod_{k=1}^{n} a_{kk} .$$

Если метод Гаусса используется только для вычисления определителя, из алгоритма его реализации следует изъять строки 4 и 7–9.

Так как перестановка строк матрицы меняет знак определителя, то при постолбцовом выборе главного элемента, т. е. при включении в алгоритм вставки $\otimes$, нужно в результате учесть число p произведенных перестановок, точнее, четность этого числа. Это означает, что при вычислении det **A** алгоритмом Гаусса с частичным упорядочиванием вместо строки 10 должна быть включена строка

$$10') \quad \det \mathbf{A} = (-1)^p \prod_{k=1}^{n} a_{kk} .$$

Для получения матрицы $\mathbf{A}^{-1}$, обратной к матрице $\mathbf{A} = \left(a_{ij}\right)_{i,j=1}^{n}$, будем исходить из того, что она является решением матричного уравнения

$$\mathbf{AX} = \mathbf{E}, \qquad (2.6)$$

где **E** – единичная матрица.

Представляя искомую матрицу $\mathbf{X} = \left(x_{ij}\right)_{i,j=1}^{n}$ как набор (вектор-строку) векторов-столбцов

$$\mathbf{x}_1 = \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{pmatrix}, \quad \dots, \quad \mathbf{x}_n = \begin{pmatrix} x_{1n} \\ x_{2n} \\ \vdots \\ x_{nn} \end{pmatrix},$$

а единичную матрицу $\mathbf{E}$ как набор единичных векторов

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \quad \ldots, \quad \mathbf{e}_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix},$$

матричное уравнение (2.6) в соответствии с правилами умножения матриц подменим эквивалентной системой не связанных между собой векторно-матричных уравнений

$$\mathbf{A}\mathbf{x}_1 = \mathbf{e}_1; \quad \mathbf{A}\mathbf{x}_2 = \mathbf{e}_2; \quad \ldots \quad ; \mathbf{A}\mathbf{x}_n = \mathbf{e}_n. \qquad (2.7)$$

Каждое из последних уравнений имеет вид (2.1') и может быть решено методом Гаусса. При этом специфичным является то обстоятельство, что все СЛАУ (2.7) имеют одну и ту же матрицу коэффициентов, а это означает, что наиболее трудоемкая часть метода Гаусса – приведение матрицы системы к треугольному виду – общая для всех систем (2.7). Так что, если требуется приспособить рассмотренный выше алгоритм решения СЛАУ методом Гаусса к обращению матриц, целесообразно не просто применить его последовательно n раз к системам (2.7), а слегка подкорректировать: "размножить" строки 4 и 9 так, чтобы в роли вектора $\mathbf{b}$ оказались все единичные векторы $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$. Тогда в результате завершения работы алгоритма будут получаться столбец за столбцом (столбцы "перевернуты"!) элементы обратной матрицы $\mathbf{X} = \mathbf{A}^{-1}$. При этом введение в алгоритм частичного упорядочивания, т.е. постолбцовый выбор главного элемента, не требует запоминаний и обратных замен.

2.3. LU - РАЗЛОЖЕНИЕ МАТРИЦ

Пусть $\mathbf{A} = \left(a_{ij}\right)_{i,j=1}^{n}$ – данная $n \times n$-матрица, а $\mathbf{L} = \left(l_{ij}\right)_{i,j=1}^{n}$ и $\mathbf{U} = \left(u_{ij}\right)_{i,j=1}^{n}$ – соответственно нижняя (левая) и верхняя (правая) треугольные матрицы*). Справедлива следующая теорема.

Теорема 2.1 [15, 47, 55] *Если все главные миноры квадратной матрицы* $\mathbf{A}$ *отличны от нуля, то существуют такие нижняя* $\mathbf{L}$ *и верхняя* $\mathbf{U}$ *треугольные матрицы, что* $\mathbf{A} = \mathbf{L}\mathbf{U}$. *Если элементы диагонали одной из матриц* $\mathbf{L}$ *или* $\mathbf{U}$ *фиксированы (ненулевые), то такое разложение единственно.*

*) Общепринятые обозначения L и U связаны с английскими словами lower (нижний) и upper (верхний). Существует другой стандарт обозначения: L и R, определяемый словами left (левый) и right (правый).

Вместо полного доказательства этой теоремы (см., например, [15]) получим формулы для фактического разложения матриц в случае фиксирования диагонали нижней треугольной матрицы $\mathbf{L}$.

Будем находить l_{ij} (при $i > j$) и u_{ij} (при $i \le j$) такие, чтобы

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

Выполнив перемножение матриц, на основе поэлементного приравнивания левых и правых частей приходим к $n \times n$-матрице уравнений

$$\begin{array}{lll} u_{11} = a_{11}, & u_{12} = a_{12}, \dots, & u_{1n} = a_{1n}, \\ l_{21}u_{11} = a_{21}, & l_{21}u_{12} + u_{22} = a_{22}, \dots, & l_{21}u_{1n} + u_{2n} = a_{2n}, \\ \dots\dots\dots & \dots\dots\dots\dots & \dots\dots\dots\dots\dots\dots \\ l_{n1}u_{11} = a_{n1}, & l_{n1}u_{12} + l_{n2}u_{22} = a_{n2}, \dots, & l_{n1}u_{1n} + \dots + l_{n,n-1}u_{n-1,n} + u_{nn} = a_{nn}; \end{array}$$

относительно $n \times n$-матрицы неизвестных

$$\begin{array}{l} u_{11}, \ u_{12}, \dots, \ u_{1n}, \\ l_{21}, \ u_{22}, \dots, \ u_{2n}, \\ \dots\dots\dots\dots \\ l_{21}, \ l_{2n}, \dots, \ u_{nn}. \end{array} \qquad (2.8)$$

Специфика этой системы позволяет находить неизвестные (2.8) одно за другим в следующем порядке.

Из первой строки уравнений имеем

$$u_{1j} = a_{1j} \qquad (j = 1, \dots, n);$$

из оставшейся части первого столбца уравнений

$$l_{i1} = \frac{a_{i1}}{u_{11}} \quad (i = 2, \dots, n);$$

из оставшейся части второй строки

$$u_{2j} = a_{2j} - l_{21}u_{1j} \quad (j = 2, \dots, n);$$

из оставшейся части второго столбца

$$l_{i2} = \frac{a_{i2} - l_{i1}u_{12}}{u_{22}} \quad (i = 3, \ldots, n)$$

и т.д. Последним находится элемент

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

Легко видеть, что все отличные от 0 и 1 элементы матриц **L** и **U** могут быть однозначно вычислены с помощью всего двух формул:

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj} \qquad (i \le j), \tag{2.9}$$

$$l_{ij} = \frac{1}{u_{jj}}\left(a_{ij} - \sum_{k=1}^{j-1} l_{ik}u_{kj}\right) \qquad (i > j). \tag{2.10}$$

При практическом выполнении разложения*) матрицы **A** нужно иметь в виду следующие два обстоятельства.

Во-первых, организация вычислений по формулам (2.9)–(2.10) должна предусматривать переключение счета с одной формулы на другую в соответствии с показанным выше процессом получения неизвестных, приведшим к этим формулам. Это удобно делать, ориентируясь на матрицу неизвестных (2.8) (ее, кстати, можно интерпретировать как n^2–мерный массив для компактного хранения LU-разложения в памяти ЭВМ), а именно, первая строка (2.8) вычисляется по формуле (2.9) при $i = 1$, $j = 1, 2, \ldots, n$; первый столбец (2.8) (без первого элемента) – по формуле (2.10) при $j = 1$, $i = 2, \ldots, n$ и т.д.

Во-вторых, препятствием для осуществимости описанного процесса LU-разложения матрицы **A** может оказаться равенство нулю диагональных элементов матрицы **U**, поскольку на них выполняется деление в формуле (2.10). Отсюда требование теоремы, накладываемое на главные миноры (напомним, что главными минорами матрицы $\mathbf{A} = \left(a_{ij}\right)_{i,j=1}^{n}$ называются определители подматриц $\mathbf{A}_k = \left(a_{ij}\right)_{i,j=1}^{k}$, где $k = 1, 2, \ldots, n-1$). Заметим, что $u_{11} = a_{11}$, т.е. первый диагональный элемент матрицы **U** совпадает с

*) Применяют также термины ***факторизация, декомпозиция.***

первым главным минором **A** и по условию должен быть отличным от нуля. Второй диагональный элемент матрицы **U**

$$u_{22} = a_{22} - l_{21}u_{12} = a_{22} - \frac{a_{21}}{a_{11}}a_{12} = \frac{\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}}{a_{11}}$$

не равен нулю, если отличен от нуля второй главный минор, и т.д. Ясно, что вместо проверки на равенство нулю главных миноров данной матрицы удобнее делать такую проверку для элементов u_{jj} в процессе их вычисления, причем, чтобы уменьшить влияние погрешностей округлений, лучше сравнивать модули u_{jj} с малой положительной константой (допуском).

Для определенных классов матриц требования теоремы о разложении заведомо выполняются. Это относится, например, к ***матрицам с диагональным преобладанием***, т.е. к таким, для которых

$$|a_{ii}| > \sum_{\substack{j=1 \\ (j \neq i)}}^{n} |a_{ij}| \qquad \forall i \in \{1, 2, \ldots, n\}.$$

2.4. РЕШЕНИЕ ЛИНЕЙНЫХ СИСТЕМ И ОБРАЩЕНИЕ МАТРИЦ С ПОМОЩЬЮ LU–РАЗЛОЖЕНИЯ

Если матрица **A** исходной системы (2.1) разложена в произведение треугольных **L** и **U**, то, значит, вместо (2.1') мы можем записать эквивалентное (2.1) уравнение

$$\mathbf{LUx} = \mathbf{b}.$$

Введя вектор вспомогательных переменных $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$, последнее можно переписать в виде системы

$$\begin{cases} \mathbf{Ly} = \mathbf{b}, \\ \mathbf{Ux} = \mathbf{y}. \end{cases}$$

Таким образом, решение данной системы с квадратной матрицей коэффициентов свелось к последовательному решению двух систем с треугольными матрицами коэффициентов.

Получим сначала формулы для вычисления элементов y_i. Для этого запишем уравнение **Ly** = **b** в развернутом виде:

$$\begin{cases} y_1 & & & & & = b_1, \\ l_{21}y_1 & + & y_2 & & & = b_2, \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ l_{n1}y_1 & + & l_{n2}y_2 & + \ldots + & l_{n,n-1}y_{n-1} + y_n & = b_n. \end{cases}$$

Очевидно, все y_i могут быть последовательно найдены при $i = 1, 2, \ldots, n$ по формуле

$$y_i = b_i - \sum_{k=1}^{i-1} l_{ik} y_k . \qquad (2.11)$$

Развернем теперь $\mathbf{Ux} = \mathbf{y}$:

$$\begin{cases} u_{11}x_1 + u_{12}x_2 + \ldots + u_{1n}x_n = y , \\ \qquad u_{22}x_2 + \ldots + u_{2n}x_n = y_2 , \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \qquad\qquad\qquad\qquad u_{nn}x_n = y_n . \end{cases} \qquad (2.12)$$

Отсюда значения неизвестных x_i находятся в обратном порядке, т.е. при $i = n, n-1, \ldots, 2, 1$, по формуле

$$x_i = \frac{1}{u_{ii}} \left(y_i - \sum_{k=i+1}^{n} u_{ik} x_k \right) . \qquad (2.13)$$

Итак, решение СЛАУ посредством LU-факторизации сводится к организации вычислений по четырем формулам: совокупности формул (2.9), (2.10) для получения матрицы $\mathbf{L} + \mathbf{U} - \mathbf{E}$ (2.8) ненулевых и неединичных элементов матриц для $\mathbf{L}$ и $\mathbf{U}$, формулы (2.11) для получения вектора свободных членов треугольной системы (2.12) и формулы (2.13), генерирующей решение исходной системы (2.1).

Обратим внимание на тот факт, что выполнение расчетов по формулам (2.9)–(2.11) можно интерпретировать как преобразование системы (2.1) к треугольной системе (2.12). С системой (2.3) – результатом прямого хода метода Гаусса – последняя имеет не только структурное сходство, но полностью совпадает с ней (совпадение первых уравнений очевидно, коэффициенты при неизвестных и свободные члены вторых уравнений легко выражаются одинаково через исходные данные; можно и дальше проводить сравнение систем (2.3) и (2.12) таким путем, однако обычно идентичность этих систем показывают проще на основе представления прямого хода метода Гаусса как последовательности умножений матрицы $\mathbf{A}$ слева на матрицы очень простой структуры такой, что в итоге получается матрица $\mathbf{U}$, а последовательность обратных преобразований дает $\mathbf{L}$ [1, 25, 47, 55]. Так что решение линейных систем с помощью LU-разложения – это просто другая схема реализации метода Гаусса. В отличие от рассмотренной в п.2.1 так называемой схемы ***единственного деления*** эту называют ***компактной схемой Гаусса*** [16, 54] или ***схемой Холецкого****) [21].

---

*) Андре-Луи Холецкий (1875–1918) – французский военный геодезист.

Чаще схемой Холецкого называют описываемый в следующем пункте основанный на той же идее способ решения симметричных линейных систем (метод квадратных корней).

Вычисление определителя LU-факторизованной матрицы **A** опирается на свойство определителя произведения матриц и сводится к перемножению n чисел:

$$\det \mathbf{A} = \det \mathbf{L} \cdot \det \mathbf{U} = u_{11} \cdot u_{22} \cdot \ldots \cdot u_{nn} .$$

Для обращения LU–факторизацией матрицы **A** можно применить тот же прием, который рассмотрен в п. 2.2, т. е. n-кратно использовать формулы (2.11) и (2.13) для получения столбцов матрицы $\mathbf{A}^{-1}$; при этом в качестве b_i в (2.11) должны фигурировать только 0 или 1: для нахождения первого столбца $\mathbf{A}^{-1}$ полагаем $b_1 = 1$, $b_2 = b_3 = \ldots = b_n = 0$, для второго – $b_2 = 1$, $b_1 = b_3 = \ldots = b_n = 0$, и т.д. Можно однако вывести и специальные формулы для выражения элементов обратной матрицы через элементы матриц **L** и **U**.

Пусть матрицы **A** и **U** обратимы (матрица **L** обратима всегда). Тогда

$$\mathbf{A} = \mathbf{L} \cdot \mathbf{U} \quad \Leftrightarrow \quad \mathbf{A}^{-1} = \mathbf{U}^{-1} \cdot \mathbf{L}^{-1} \quad .$$

Умножая последнее равенство поочередно на **U** слева и на **L** справа, будем иметь

$$\mathbf{U}\mathbf{A}^{-1} = \mathbf{L}^{-1} \quad \text{и} \quad \mathbf{A}^{-1}\mathbf{L} = \mathbf{U}^{-1}. \qquad (2.14)$$

Обозначим, как и ранее, искомые элементы матрицы $\mathbf{A}^{-1}$ через x_{ij}. Учитывая, что треугольные матрицы при обращении сохраняют свою структуру, перепишем равенства (2.14) в виде

$$\begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & u_{nn} \end{pmatrix} \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ * & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ * & * & \cdots & 1 \end{pmatrix},$$

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{pmatrix} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ l_{21} & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ l_{n1} & l_{n2} & \cdots & 1 \end{pmatrix} = \begin{pmatrix} * & * & \cdots & * \\ 0 & * & \cdots & * \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & * \end{pmatrix},$$

где * – некоторые числа. Полученные матричные равенства можно рассматривать как систему $2n^2$ уравнений с n^2 неизвестными x_{ij} $(i, j = 1, 2, \ldots, n)$. Из этих $2n^2$ уравнений ровно n^2 имеют известные пра-

вые части (0 или 1). Выпишем соответствующую им $n\times n$-матрицу уравнений:

$$u_{11}x_{11}+\ldots+u_{1n}x_{n1}=1,\quad u_{11}x_{12}+\ldots+u_{1n}x_{n2}=0,\ \ldots\ ,\quad u_{11}x_{1n}+\ldots+u_{1n}x_{nn}=0,$$
$$x_{21}+x_{22}l_{21}+\ldots+x_{2n}l_{n1}=0,\quad u_{22}x_{22}+\ldots+u_{2n}x_{n2}=1,\ \ldots\ ,\quad u_{22}x_{2n}+\ldots+u_{2n}x_{nn}=0,$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$x_{n1}+x_{n2}l_{21}+\ldots+x_{nn}l_{n1}=0,\qquad x_{n2}+\ldots+x_{nn}l_{n2}=0,\ \ldots\ ,\qquad u_{nn}x_{nn}=1.$$

Короче все эти уравнения могут быть представлены следующими тремя типами связей*):

$$\sum_{k=i}^{n}u_{ik}x_{kj}=0,\quad \text{если } i<j,$$
$$\sum_{k=i}^{n}u_{ik}x_{kj}=1\qquad \text{при } i=j$$

и

$$x_{ij}+\sum_{k=j+1}^{n}x_{ik}l_{kj}=0,\quad \text{если } i>j.$$

Отсюда можно выразить все элементы x_{ij} искомой обратной матрицы $\mathbf{A}^{-1}$:

$$x_{jj}=\frac{1}{u_{jj}}\left(1-\sum_{k=j+1}^{n}u_{ik}x_{kj}\right);\qquad (2.15)$$

$$x_{ij}=-\frac{1}{u_{ii}}\sum_{k=i+1}^{n}u_{ik}x_{kj}\ (i<j);\qquad (2.16)$$

$$x_{ij}=-\sum_{k=j+1}^{n}x_{ik}l_{kj}\qquad (i>j).\qquad (2.17)$$

Формулы (2.15)–(2.17) позволяют эффективно обращать LU-факторизованную матрицу, если соблюдать определенную технологию их использования. А именно, как видно из записанной выше матрицы уравнений, следует сначала из последнего с т о л б ц а уравнений найти x_{nn}, $x_{n-1,n},\ldots,x_{2n},x_{1n}$, затем из оставшейся части последней с т р о к и уравнений найти $x_{n,n-1},\ldots,x_{n2},x_{n1}$, потом переключиться на предпоследний

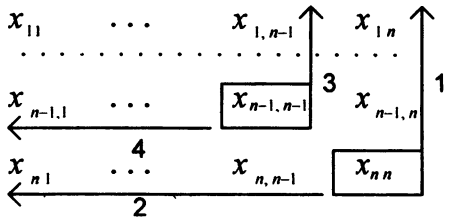
---

*) Используя символ Кронекера δ_{ij}, первые две формулы можно совместить:

$$\sum_{k=i}^{n}u_{ik}x_{kj}=\delta_{ij}$$

Это относится и к формулам (2.15), (2.16)

столбец и т.д. Схематично последовательность вычислений элементов обратной матрицы можно изобразить пронумерованными стрелками следующим образом:



При этом стрелка 1 означает, что фиксируем $j = n$ и ведем счет по формулам (2.15), (2.16) при $i = n,\ n-1,\ \ldots,1$; стрелка 2 – счет по формуле (2.17) при $i = n$ и $j = n-1,\ n-2,\ \ldots,1$ и т.д.

Возвращаясь к началу п.2.3, заметим, что так же употребительно фиксирование единичной диагонали у правой треугольной матрицы, т.е. представление

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

В этом случае l_{ij} и u_{ij} находятся по формулам

$$l_{ij} = a_{ij} - \sum_{k=1}^{j-1} l_{ik} u_{kj} \qquad (i \geq j),$$

$$u_{ij} = \frac{1}{l_{ii}} \left(a_{ij} - \sum_{k=1}^{i-1} l_{ik} u_{kj} \right) \qquad (i < j),$$

где индексы фиксируются так, чтобы вычислялись поочередно столбец $(l_{i1})_{i=1}^{n}$, затем строка $(u_{1j})_{j=2}^{n}$ и т.д.

Решение системы (2.1) с таким образом факторизованной матрицей коэффициентов получают по формулам

$$y_i = \frac{1}{l_{ii}} \left(b_i - \sum_{k=1}^{i-1} l_{ik} y_k \right), \qquad i = 1, 2, \ldots, n,$$

$$x_i = y_i - \sum_{k=i+1}^{n} u_{ik} x_k, \qquad i = n, n-1, \ldots, 1.$$

Детерминант матрицы A равен произведению $l_{11}l_{22}\ldots l_{nn}$, а для подсчета элементов обратной матрицы используют совокупность формул

$$x_{ii} = \frac{1}{l_{ii}}\left(1 - \sum_{k=i+1}^{n} x_{ik}l_{ki}\right),$$

$$x_{ij} = -\frac{1}{l_{jj}} \sum_{k=i+1}^{n} x_{ik}l_{kj} \qquad (i > j),$$

$$x_{ij} = -\sum_{k=i+1}^{n} u_{ik}x_{kj} \qquad (i < j)$$

с такой организацией вычислений, при которой сначала вычисляется последняя строка (x_{nj}) при $j = n, n-1, \ldots, 1$, затем последний столбец (x_{in}) при $i = n-1, \ldots, 1$, потом предпоследняя строка $(x_{n-1,j})$ при $j = n-1, \ldots, 1$, и т.д.

В отличие от рассмотренной в п.2.1 схемы единственного деления схема Холецкого менее удобна для усовершенствования с целью уменьшения влияния вычислительных погрешностей путем выбора подходящих ведущих элементов. Достоинством же ее можно считать то, что LU-разложение матрицы A играет роль обратной матрицы, может помещаться в память ЭВМ на место матрицы A и использоваться, например, при решении нескольких систем, имеющих одну и ту же матрицу коэффициентов и разные правые части.

2.5. РАЗЛОЖЕНИЕ СИММЕТРИЧНЫХ МАТРИЦ. МЕТОД КВАДРАТНЫХ КОРНЕЙ

Объем вычислений, требующихся для решения линейных алгебраических задач с симметрическими матрицами, можно сократить почти вдвое, если учитывать симметрию при треугольной факторизации матриц.

Пусть $\mathbf{A} = (a_{ij})_{i,j=1}^{n}$ – данная симметрическая матрица, т.е. $a_{ij} = a_{ji}$.

Будем строить ее представление в виде $\mathbf{A} = \mathbf{U}^T\mathbf{U}$, где

$$\mathbf{U} = \begin{pmatrix} u_{11} & u_{12} & \ldots & u_{1n} \\ 0 & u_{22} & \ldots & u_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & u_{nn} \end{pmatrix}, \quad \mathbf{U}^T = \begin{pmatrix} u_{11} & 0 & \ldots & 0 \\ u_{12} & u_{22} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ u_{1n} & u_{2n} & \ldots & u_{nn} \end{pmatrix}.$$

Аналогично тому, как это делалось в п.2.3, составим систему $\frac{n(n+1)}{2}$ уравнений относительно такого же количества неизвестных (элементов матрицы $\mathbf{U}$):

$$\begin{aligned} u_{11}^2 = a_{11},\ u_{12}u_{11} &= a_{12},\ \ldots,\ u_{11}u_{1n} &&= a_{1n},\\ u_{12}^2 + u_{22}^2 &= a_{22},\ \ldots,\ u_{12}u_{1n} + u_{22}u_{2n} &&= a_{2n},\\ &\cdots\cdots\cdots\cdots\cdots\cdots &&\cdots\\ & &u_{1n}^2 + u_{2n}^2 + \ldots + u_{nn}^2 &= a_{nn}. \end{aligned}$$

Из первой строки уравнений находим сначала $u_{11} = \sqrt{a_{11}}$, затем $u_{1j} = \frac{a_{1j}}{u_{11}}$ при $j = 2, \ldots, n$. Из второй – $u_{22} = \sqrt{a_{22} - u_{12}^2}$, затем $u_{2j} = \frac{a_{2j} - u_{12}u_{1j}}{u_{22}}$ при $j = 3, \ldots, n$, и т.д. Завершается процесс вычислением

$$u_{nn} = \sqrt{a_{nn} - \sum_{k=1}^{n-1} u_{kn}^2}.$$

Таким образом, матрица $\mathbf{U}$ может быть определена совокупностью формул

$$\begin{aligned} u_{ii} &= \sqrt{a_{ii} - \sum_{k=1}^{i-1} u_{ki}^2} \qquad \text{при} \quad i = 1, 2, \ldots, n;\\ u_{ij} &= \frac{a_{ij} - \sum_{k=1}^{i-1} u_{ki}u_{kj}}{u_{ii}} \qquad \text{при} \quad j = 2, \ldots, n; \quad j > i \\ &\left(u_{ij} = 0 \quad \text{при} \quad j < i\right). \end{aligned} \tag{2.18}$$

Осуществимости вещественного $\mathbf{U}^T\mathbf{U}$-разложения вещественной симметрической матрицы $\mathbf{A}$ по этим формулам могут помешать два обстоятельства: обращение в нуль элемента u_{ii} при каком-либо $i \in \{1, 2, \ldots, n\}$ и отрицательность подкоренного выражения. Известно, что для важного в приложениях класса симметрических положительно определенных матриц разложение по формулам (2.18) выполнимо [15, 54 и др.]*).

*) Более универсальным, чем $\mathbf{U}^T\mathbf{U}$-разложение Холецкого, является $\mathbf{U}^*\mathbf{D}\mathbf{U}$-разложение, пригодное для эрмитовых матриц, частным случаем которых являются симметрические (см., например, [5, 16, 47]).

При наличии $\mathbf{U}^T\mathbf{U}$-разложения решение симметричной системы $\mathbf{Ax} = \mathbf{b}$ сводится к последовательному решению двух треугольных систем

$$\mathbf{U}^T\mathbf{y} = \mathbf{b} \quad \text{и} \quad \mathbf{Ux} = \mathbf{y}.$$

Первая из них имеет вид

$$\begin{cases} u_{11}y_1 & = b_1, \\ u_{12}y_1 + u_{22}y_2 & = b_2, \\ \dots\dots\dots\dots\dots & \dots \\ u_{1n}y_1 + u_{2n}y_2 + \dots + u_{nn}y_n & = b_n, \end{cases}$$

откуда получаем вспомогательные неизвестные $y_1, y_2, \dots, y_n$ по формуле

$$y_i = \frac{b_i - \sum_{k=1}^{i-1} u_{ki}y_k}{u_{ii}}, \tag{2.19}$$

полагая в ней $i = 1, 2, \dots, n$. Из второй системы

$$\begin{cases} u_{11}x_1 + u_{12}x_2 + \dots + u_{1n}x_n = y_1, \\ u_{22}x_2 + \dots + u_{2n}x_n = y_2, \\ \dots\dots\dots \\ u_{nn}x_n = y_n \end{cases}$$

находим искомые значения x_i в обратном порядке, т.е. при $i = n, n-1, \dots, 1$ по формуле

$$x_i = \frac{y_i - \sum_{k=i+1}^{n} u_{ik}x_k}{u_{ii}}. \tag{2.20}$$

Решение симметричных СЛАУ по формулам (2.18)–(2.20) называют ***методом квадратных корней*** или ***схемой Холецкого.*** В случае систем с положительно определенными матрицами можно ожидать хороших результатов применения такого метода (особенно, если в процессе решения делать проверку на немалость $|u_{ii}|$ дабы избежать большого роста погрешностей). В противном случае нет, например, гарантий, что в процессе разложения не появятся чисто мнимые числа, что кстати может не отразиться на результатах, если в алгоритме реализации метода квадратных корней предусмотреть возможность появления мнимых чисел [54].

2.6. МЕТОД ПРОГОНКИ РЕШЕНИЯ СИСТЕМ С ТРЕХДИАГОНАЛЬНЫМИ МАТРИЦАМИ КОЭФФИЦИЕНТОВ

Часто возникает необходимость в решении линейных алгебраических систем, матрицы которых, являясь слабо заполненными, т.е. содержащими немного ненулевых элементов, имеют определенную структуру. Среди таких систем выделим системы с матрицами ленточной структуры, в которых ненулевые элементы располагаются на главной диагонали и на нескольких побочных диагоналях. Для решения систем с ленточными матрицами коэффициентов метод Гаусса можно трансформировать в более эффективные методы.

Рассмотрим наиболее простой случай ***ленточных систем***, к которым, как увидим впоследствии, сводится решение задач сплайн-интерполяции функций, дискретизации краевых задач для дифференциальных уравнений методами конечных разностей, конечных элементов и др. А именно, будем искать решение такой системы, каждое уравнение которой связывает три "соседних" неизвестных:

$$b_i x_{i-1} + c_i x_i + d_i x_{i+1} = r_i \,, \qquad (2.21)$$

где $i = 1, 2, \ldots, n$; $b_1 = 0, d_n = 0$. Такие уравнения называют ***трехточечными разностными уравнениями второго порядка***[*]. Система (2.21) имеет трехдиагональную структуру, что хорошо видно из следующего, эквивалентного (2.21), векторно-матричного представления:

$$\begin{bmatrix} c_1 & d_1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ b_2 & c_2 & d_2 & 0 & \ldots & 0 & 0 & 0 \\ 0 & b_3 & c_3 & d_3 & \ldots & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \ldots & b_{n-1} & c_{n-1} & d_{n-1} \\ 0 & 0 & 0 & 0 & \ldots & 0 & b_n & c_n \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ \vdots \\ r_{n-1} \\ r_n \end{bmatrix}.$$

Преследуя, как и в п.2.1, цель избавиться от ненулевых элементов в поддиагональной части матрицы системы, предположим, что существуют такие наборы чисел δ_i и λ_i $(i = 1, 2, \ldots, n)$, при которых

$$x_i = \delta_i x_{i+1} + \lambda_i \,, \qquad (2.22)$$

---

[*] Чаще вместо принятой здесь записи системы (2.21), предполагающей как бы наличие фиктивных неизвестных x_0 и x_{n+1} с нулевыми коэффициентами, считают в (2.21) i изменяющимся от 2 до $n-1$, выделяя первое и последнее уравнения системы соответственно

$$c_1 x_1 + d_1 x_2 = r_1 \qquad \text{и} \qquad b_n x_{n-1} + c_n x_n = r_n$$

в отдельные строки (так называемые ***краевые условия*** разностного уравнения).

т.е. трехточечное уравнение второго порядка (2.21) преобразуется в двухточечное уравнение первого порядка (2.22). Уменьшим в связи (2.22) индекс на единицу и полученное выражение $x_{i-1} = \delta_{i-1} x_i + \lambda_{i-1}$ подставим в данное уравнение (2.21):

$$b_i \delta_{i-1} x_i + b_i \lambda_{i-1} + c_i x_i + d_i x_{i+1} = r_i ,$$

откуда

$$x_i = -\frac{d_i}{c_i + b_i \delta_{i-1}} x_{i+1} + \frac{r_i - b_i \lambda_{i-1}}{c_i + b_i \delta_{i-1}} .$$

Последнее равенство имеет вид (2.22) и будет точно с ним совпадать, иначе говоря, представление (2.22) будет иметь место, если при всех $i = 1, 2, \ldots, n$ выполняются рекуррентные соотношения

$$\delta_i = -\frac{d_i}{c_i + b_i \delta_{i-1}} , \quad \lambda_i = \frac{r_i - b_i \lambda_{i-1}}{c_i + b_i \delta_{i-1}} . \qquad (2.23)$$

Легко видеть, что, в силу условия $b_1 = 0$, процесс вычисления δ_i, λ_i может быть начат со значений

$$\delta_1 = -\frac{d_1}{c_1} , \quad \lambda_1 = \frac{r_1}{c_1}$$

и продолжен далее по формулам (2.23) последовательно при $i = 2, 3, \ldots, n$, причем при $i = n$, в силу $d_n = 0$, получим $\delta_n = 0$. Следовательно, полагая в (2.22) $i = n$, будем иметь

$$x_n = \lambda_n = \frac{r_n - b_n \lambda_{n-1}}{c_n + b_n \delta_{n-1}}$$

(где λ_{n-1}, δ_{n-1} – уже известные с предыдущего шага числа). Далее по формулам (2.22) последовательно находятся $x_{n-1}, x_{n-2}, \ldots, x_1$ при $i = n-1, n-2, \ldots, 1$ соответственно.

Таким образом, решение уравнений вида (2.21) описываемым способом, называемым ***методом прогонки***[*], сводится к вычислениям по трем простым формулам: нахождение так называемых ***прогоночных коэффициентов*** δ_i, λ_i по формулам (2.23) при $i = 1, 2, \ldots, n$ (***прямая прогонка***) и затем получение неизвестных x_i по формуле (2.22) при $i = n, n-1, \ldots, 1$ (***обратная прогонка***).

Для успешного применения метода прогонки нужно, чтобы в процессе вычислений не возникало ситуаций с делением на нуль, а при больших размерностях систем не должно быть быстрого роста погрешностей округлений.

Будем называть прогонку ***корректной****, если знаменатели прогоночных коэффициентов* (2.23) *не обращаются в нуль, и* ***устойчивой****, если* $|\delta_i| < 1$ *при всех* $i \in \{1, 2, \ldots, n\}$.

---

[*] Термин, характерный, в основном, для отечественной литературы по вычислительной математике, введён в 50-х годах (см., например, [1]).

Приведем простые достаточные условия корректности и устойчивости прогонки, которые во многих приложениях метода автоматически выполняются.

Теорема 2.2. *Пусть коэффициенты* b_i *и* d_i *уравнения* (2.21) *при* $i = 2, 3, \ldots, n-1$ *отличны от нуля и пусть*

$$|c_i| > |b_i| + |d_i| \qquad \forall\ i = 1, 2, \ldots, n. \tag{2.24}$$

Тогда прогонка (2.23), (2.22) *корректна и устойчива* (*т.е.* $c_i + b_i\delta_{i-1} \neq 0$, $|\delta_i| < 1$).

Доказательство. Воспользуемся методом математической индукции для установления обоих нужных неравенств одновременно.

При $i = 1$, в силу (2.24), имеем:

$$|c_1| > |d_1| \geq 0$$

– неравенство нулю знаменателя первой пары прогоночных коэффициентов, а также

$$|\delta_1| = \left|-\frac{d_1}{c_1}\right| < 1.$$

Предположим, что знаменатель $(i-1)$–х прогоночных коэффициентов не равен нулю и что $|\delta_{i-1}| < 1$. Тогда, используя свойства модулей, условия теоремы и индукционные предположения, получаем:

$$\begin{aligned} |c_i + b_i\delta_{i-1}| &\geq |c_i| - |b_i\delta_{i-1}| > |b_i| + |d_i| - |b_i| \cdot |\delta_{i-1}| = \\ &= |d_i| + |b_i|\left(1 - |\delta_{i-1}|\right) > |d_i| > 0, \end{aligned}$$

а с учетом этого

$$|\delta_i| = \left|-\frac{d_i}{c_i + b_i\delta_{i-1}}\right| = \frac{|d_i|}{|c_i + b_i\delta_{i-1}|} < \frac{|d_i|}{|d_i|} = 1.$$

Следовательно, $c_i + b_i\delta_{i-1} \neq 0$ и $|\delta_i| < 1$ при всех $i \in \{1, 2, \ldots, n\}$, т.е. имеет место утверждаемая в данных условиях корректность и устойчивость прогонки. Теорема доказана.

Пусть A – матрица коэффициентов данной системы (2.21), удовлетворяющих условиям теоремы 2.2, и пусть

$$\delta_1 = -\frac{d_1}{c_1},\ \delta_i = -\frac{d_i}{c_i + b_i\delta_{i-1}}\ \ (i = 2, 3, \ldots, n-1),\ \delta_n = 0$$

– прогоночные коэффициенты, определяемые первой из формул (2.23), а

$$\Delta_i = c_i + b_i\delta_{i-1} \quad (i = 2\ ,\ 3\ ,\ \ldots\ ,\ n)$$

– знаменатели этих коэффициентов (отличные от нуля согласно утверждению теоремы 2.2). Непосредственной проверкой легко убедиться, что имеет место представление $\mathbf{A} = \mathbf{LU}$, где

$$\mathbf{L} = \begin{bmatrix} c_1 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ b_2 & \Delta_2 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & b_3 & \Delta_3 & 0 & \ldots & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \ldots & b_{n-1} & \Delta_{n-1} & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & b_n & \Delta_n \end{bmatrix},$$

$$\mathbf{U} = \begin{bmatrix} 1 & -\delta_1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 1 & -\delta_2 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & 1 & -\delta_3 & \ldots & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \ldots & 0 & 1 & -\delta_{n-1} \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 1 \end{bmatrix},$$

единственное в силу утверждения теоремы 2.1. Как видим, LU-разложение трехдиагональной матрицы $\mathbf{A}$ может быть выполнено очень простым алгоритмом, вычисляющим Δ_i и δ_i при возрастающих значениях i. При необходимости попутно может быть вычислен

$$\det\ \mathbf{A} = c_1 \prod_{i=2}^{n} \Delta_i .$$

В заключение этого пункта заметим, что, во-первых, имеются более слабые условия корректности и устойчивости прогонки, чем требуемое в теореме 2.2 условие строгого диагонального преобладания в матрице $\mathbf{A}$ (см., например, [17, 47, 48]). Во-вторых, применяется ряд других, отличных от рассмотренной нами правой прогонки, методов подобного типа, решающих как поставленную здесь задачу (2.21) для систем с трехдиагональными матрицами (левая прогонка, встречная прогонка, немонотонная, циклическая, ортогональная прогонки и т.д.), так и для более сложных систем с матрицами ленточной структуры или блочно-матричной структуры (например, матричная прогонка). Выводы и исследование различных вариантов метода прогонки можно найти, например, в [27], [48].

2.7. МЕТОД ВРАЩЕНИЙ РЕШЕНИЯ ЛИНЕЙНЫХ СИСТЕМ

Вернемся к рассмотрению линейных систем общего вида (2.1).

Предположим, что методом Гаусса решается система $\mathbf{Ax} = \mathbf{b}$ с матрицей коэффициентов

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

Приведение такой системы*) к треугольному виду прямым ходом метода Гаусса равносильно следующей последовательности эквивалентных преобразований матрицы $\mathbf{A}$:

$$\mathbf{A} \sim \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & -1 & 1 & 2 \\ 0 & -1 & -1 & 2 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 4 \\ 0 & 0 & -1 & 4 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 8 \end{pmatrix}.$$

Очевидно, в случае $n \times n$-матрицы такого типа прямой ход метода Гаусса допускает рост элементов матрицы до 2^{n-1}. При больших n это может привести если не к переполнению разрядной сетки ЭВМ, то к сильному влиянию погрешностей округлений, причем в данной ситуации не даст эффекта и постолбцовый выбор главного элемента.

Рассмотренный пример оправдывает поиск других подходов к построению прямых методов решения линейных систем (2.1), возможно, более сложных, чем метод Гаусса, но не допускающих большого роста элементов в процессе преобразований и как следствие численно более устойчивых.

Как и в методе Гаусса, цель прямого хода преобразований в новом методе – приведение системы к треугольному виду последовательным обнулением поддиагональных элементов сначала первого столбца, затем второго и т.д. Делается это следующим образом.

Пусть c_1 и s_1 – некоторые отличные от нуля числа. Умножим первое уравнение системы (2.1) на c_1, второе – на s_1 и сложим их; полученным уравнением заменим первое уравнение системы. Затем первое уравнение исходной системы умножаем на $-s_1$, второе – на c_1 и результатом их сло-

*) Пример заимствован из [25].

жения заменяем второе уравнение. Таким образом, первые два уравнения системы (2.1) заменяются уравнениями

$$(c_1a_{11}+s_1a_{21})x_1+\ (c_1a_{12}+s_1a_{22})x_2+\ldots+\ (c_1a_{1n}+s_1a_{2n})x_n = c_1b_1+s_1b_2\ ,$$
$$(-s_1a_{11}+c_1a_{21})x_1+(-s_1a_{12}+c_1a_{22})x_2+\ldots+(-s_1a_{1n}+c_1a_{2n})x_n = -s_1b_1+c_1b_2\ .$$

На введенные два параметра c_1 и s_1 наложим два условия:

$$-s_1a_{11}+c_1a_{21}=0$$

– условие обнуления (т.е. исключение x_1 из второго уравнения) и

$$c_1^2+s_1^2=1$$

– условие нормировки. Легко проверить, что за c_1 и s_1, удовлетворяющие этим условиям, можно принять

$$c_1=\frac{a_{11}}{\sqrt{a_{11}^2+a_{21}^2}},\quad s_1=\frac{a_{21}}{\sqrt{a_{11}^2+a_{21}^2}}. \tag{2.25}$$

Эти числа можно интерпретировать как косинус и синус некоторого угла α_1 (отсюда название ***метод вращений***, так как один промежуточный шаг прямого хода такого метода может рассматриваться как преобразование вращения на угол α_1 расширенной матрицы системы в плоскости, определяемой индексами обнуляемого элемента*)).

После фиксирования c_1 и s_1 способом (2.25) система (2.1) принимает вид

$$\begin{cases} a_{11}^{(1)}x_1+a_{12}^{(1)}x_2+\ldots+a_{1n}^{(1)}x_n = b_1^{(1)}, \\ a_{22}^{(1)}x_2+\ldots+a_{2n}^{(1)}x_n = b_2^{(1)}, \\ a_{31}x_1+a_{32}x_2+\ldots+a_{3n}x_n = b_3, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1+a_{n2}x_2+\ldots+a_{nn}x_n = b_n, \end{cases} \tag{2.26}$$

где

$$a_{1j}^{(1)}=c_1a_{1j}+s_1a_{2j}\quad (j=1,2,\ldots,n),\quad b_1^{(1)}=c_1b_1+s_1b_2\ ;$$
$$a_{2j}^{(1)}=-s_1a_{1j}+c_1a_{2j}\quad (j=2,3,\ldots,n),\quad b_2^{(1)}=-s_1b_1+c_1b_2\ .$$

*) Более подробно эта интерпретация, как и вообще идея ортогональных преобразований, будет рассмотрена позже применительно к решению полной алгебраической проблемы собственных значений (см. пп.4.4, 4.6). Там же, в замечании 4.10 описан другой способ численно устойчивого решения СЛАУ – метод отражений.

Далее первое уравнение системы (2.26) заменяется новым, полученным сложением результатов умножения первого и третьего уравнений (2.26) соответственно на

$$c_2 = \frac{a_{11}^{(1)}}{\sqrt{\left(a_{11}^{(1)}\right)^2 + a_{31}^2}} \quad \text{и} \quad s_2 = \frac{a_{31}}{\sqrt{\left(a_{11}^{(1)}\right)^2 + a_{31}^2}},$$

а третье – уравнением, полученным сложением результатов умножения тех же уравнений соответственно на $-s_2$ и c_2. Получаем систему

$$\begin{cases} a_{11}^{(2)}x_1 + a_{12}^{(2)}x_2 + \ldots + a_{1n}^{(2)}x_n = b_1^{(2)}, \\ a_{22}^{(1)}x_2 + \ldots + a_{2n}^{(1)}x_n = b_2^{(1)}, \\ a_{32}^{(1)}x_2 + \ldots + a_{3n}^{(1)}x_n = b_3^{(1)}, \\ a_{41}x_1 + a_{42}x_2 + \ldots + a_{4n}x_n = b_4, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n, \end{cases}$$

где

$$a_{1j}^{(1)} = c_2 a_{1j}^{(1)} + s_2 a_{3j} \quad (j = 1, 2, \ldots, n), \quad b_1^{(1)} = c_2 b_1^{(1)} + s_2 b_3;$$

$$a_{3j}^{(1)} = -s_2 a_{1j}^{(1)} + c_2 a_{3j} \quad (j = 2, 3, \ldots, n), \quad b_3^{(1)} = -s_2 b_1^{(1)} + c_2 b_3.$$

Проделав такие преобразования $n-1$ раз, придем к системе

$$\begin{cases} a_{11}^{(n-1)}x_1 + a_{12}^{(n-1)}x_2 + \ldots + a_{1n}^{(n-1)}x_n = b_1^{(n-1)}, \\ a_{22}^{(1)}x_2 + \ldots + a_{2n}^{(1)}x_n = b_2^{(1)}, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n2}^{(1)}x_2 + \ldots + a_{nn}^{(1)}x_n = b_n^{(1)} \end{cases} \qquad (2.27)$$

такого же вида, какой приняла система (2.1) после первого этапа преобразований прямого хода метода Гаусса. Однако в отличие от (2.2) система (2.27) обладает замечательным свойством: длина любого вектора-столбца (иначе, евклидова норма) расширенной матрицы системы (2.27) остается такой же, как у соответствующего столбца исходной системы (2.1)*). Чтобы убедиться в этом, достаточно посмотреть на результат следующих преобразований, учитывающих условия нормировки:

$$\left(a_{1j}^{(1)}\right)^2 + \left(a_{2j}^{(1)}\right)^2 = c_1^2 a_{1j}^2 + 2c_1 s_1 a_{1j} a_{2j} + s_1^2 a_{2j}^2 + s_1^2 a_{1j}^2 - 2c_1 s_1 a_{1j} a_{2j} + c_1^2 a_{2j}^2 =$$

$$= (c_1^2 + s_1^2) a_{1j}^2 + (c_1^2 + s_1^2) a_{2j}^2 = a_{1j}^2 + a_{2j}^2.$$

Оно показывает сохранение величины суммы квадратов изменяемых на данном промежуточном шаге преобразований пары элементов любого столбца; так как остальные элементы столбцов при этом остаются неиз-

*) Разумеется, в предположении точной реализации метода, в частности, формул типа (2.25).

менными, то, значит, на любом этапе преобразований длина столбца будет одной и той же, т.е. не будет наблюдаться роста элементов (но какой ценой это достигается! коэффициенты первого уравнения пересчитываются $n-1$-раз!).

Дальше точно так же за $n-2$ промежуточных шага преобразуем подсистему

$$\begin{cases} a_{22}^{(1)}x_2 + \ldots + a_{2n}^{(1)}x_n = b_2^{(1)}, \\ \ldots\ldots\ldots\ldots\ldots \\ a_{n2}^{(1)}x_2 + \ldots + a_{nn}^{(1)}x_n = b_n^{(1)} \end{cases}$$

системы (2.27), создавая нули под элементом $a_{22}^{(1)}$, и т.д.

В результате $n-1$ таких этапов прямого хода исходная система (2.1) будет приведена к треугольному виду (ср. с (2.3) в п.2.1):

$$\begin{cases} a_{11}^{(n-1)}x_1 + a_{12}^{(n-1)}x_2 + \ldots + a_{1n}^{(n-1)}x_n = b_1^{(n-1)}, \\ a_{22}^{(n-1)}x_2 + \ldots + a_{2n}^{(n-1)}x_n = b_2^{(n-1)}, \\ \ldots\ldots\ldots\ldots\ldots \\ a_{nn}^{(n-1)}x_n = b_n^{(n-1)}. \end{cases}$$

Нахождение отсюда неизвестных $x_n, x_{n-1}, \ldots, x_1$ не отличается от рассмотренного ранее обратного хода метода Гаусса.

2.8. ДВА ЗАМЕЧАНИЯ К ПРИМЕНЕНИЮ ПРЯМЫХ МЕТОДОВ

а) О контроле точности и уточнении приближенного решения в рамках прямого метода.

Как отмечалось во введении, прямые методы приводят к точному решению СЛАУ при точном выполнении предусматриваемых соответствующими алгоритмами арифметических операций (без округлений). Реальные же вычисления базируются на арифметике машинных (т.е. усеченных до определенного количества разрядов) чисел. Как отражается на результате решения системы подмена арифметики действительных чисел машинной арифметикой, зависит от самой решаемой системы, параметров применяемой ЭВМ, способов реализации алгоритмов (см. по этому поводу, в частности, гл. 1). В любом случае, практически вместо точного решения СЛАУ (2.1) прямой метод дает п р и б л и ж е н н о е р е ш е н и е (обозначим его $\mathbf{x}^{(0)}$). Подставив $\mathbf{x}^{(0)}$ в выражение

$$\boldsymbol{\xi} = \mathbf{b} - \mathbf{A}\mathbf{x},$$

называемое ***невязкой***, по малости полученного вектора-значения $\xi^{(0)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}$ можно с осторожностью судить о близости найденного решения $\mathbf{x}^{(0)}$ к точному решению $\mathbf{x}$. Если, например, $\left\|\xi^{(0)}\right\|$ – недостаточно малая величина, то следует искать вектор-***поправку*** $\mathbf{p}$ такой, что $\mathbf{x}^{(0)} + \mathbf{p} = \mathbf{x}$ есть точное решение системы (2.1), т.е.

$$\mathbf{A}\left(\mathbf{x}^{(0)} + \mathbf{p}\right) = \mathbf{b}.$$

Последнее равносильно векторно-матричному уравнению

$$\mathbf{A}\mathbf{p} = \xi^{(0)}.$$

Как видим, нахождение поправки сводится к решению такой же системы, как и (2.1), где в качестве вектора свободных членов должен быть взят вектор невязок. Поскольку матрица коэффициентов осталась той же, что и у исходной системы, нет надобности в выполнении прямого хода преобразований коэффициентов при неизвестных (иначе, LU-разложения); достаточно выполнить только действия, касающиеся новых свободных членов (решить две треугольные системы: $\mathbf{L}\mathbf{z} = \xi^{(0)}$ и $\mathbf{U}\mathbf{p} = \mathbf{z}$). Прибавив найденную поправку $\mathbf{p} = \mathbf{p}^{(0)}$ к $\mathbf{x}^{(0)}$, получаем уточненное приближенное решение $\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \mathbf{p}^{(0)}$. В случае, если величина $\left\|\mathbf{p}^{(0)}\right\|$ (или $\left\|\mathbf{p}^{(0)}\right\| / \left\|\mathbf{x}^{(1)}\right\|$, если контролируется относительная, а не абсолютная погрешность) окажется недостаточно малой, процесс уточнения может быть повторен: ищется поправка $\mathbf{p}^{(1)}$ как приближенное решение уравнения $\mathbf{A}\mathbf{p} = \xi^{(1)}$, где $\xi^{(1)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(1)}$; тогда более точным должно быть решение $\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \mathbf{p}^{(1)}$. Как аргументировано утверждается в [15], сходимость к нулю невязок в таком процессе уточнения решения может не наблюдаться*), т.е. следить нужно за установлением знаков самого решения. Обычно делают не более двух-трех шагов уточнения, причем рекомендуется производить вычисление невязок в режиме накопления. Если в этом процессе не происходит сближения $\mathbf{x}^{(k)}$ при $k = 2, 3$, то это говорит скорее всего о том, что данная система плохо обусловлена и ее решение не может быть найдено с требуемой точностью без привлечения дополнительной информации об исходной задаче. В таких случаях закономерно ставить вопрос о том, что понимать под точным решением системы и, возможно, обращаться к методам нахождения ее псевдорешений. (Достаточно глубокие исследования затронутых здесь вопросов, основанные на изучении поведения методов при введении ошибок в исходные данные, а также ряд других све-

*) См. также пример поведения невязок у плохо обусловленной системы на приближённых решениях в гл. 1 (пп.1.4, 1.5).

дений, в частности, точностные характеристики прямых методов, можно найти в [15]).

Хотя описанный здесь контроль точности по невязкам и уточнение решений не требует больших вычислительных затрат, требуемая память ЭВМ должна быть увеличена вдвое, так как при этом нужно удерживать в памяти исходные данные.

б) О вычислительных затратах.

Одним из факторов, предопределяющих выбор того или иного метода при решении конкретной задачи, является вычислительная эффективность метода. Особенностью прямых методов является то, что здесь можно точно подсчитать требуемое количество арифметических операций. Приведем пример такого подсчета для метода прогонки решения n-мерной системы с трехдиагональной матрицей коэффициентов (см. п.2.6). Необходимые операции и их число наглядно видны из следующей таблицы (вычитание отождествляется со сложением):

Таблица 2.1

Расчетные формулы метода прогонки	Умножений	Делений	Сложений
$\Delta_i = c_i + b_i\delta_{i-1} \quad (i = 2, \ldots, n)$	$n-1$		$n-1$
$\delta_1 = -\frac{d_1}{c_1};\ \delta_i = -\frac{d_i}{\Delta_i} \quad (i = 2, \ldots, n-1)$		$n-1$	
$\lambda_1 = \frac{r_1}{c_1};\ \lambda_i = \frac{r_i - b_i\lambda_{i-1}}{\Delta_i} \quad (i = 2, \ldots, n)$	$n-1$	n	$n-1$
$x_n = \lambda_n;\ \ x_i = \delta_i x_{i+1} + \lambda_i \quad (i = n-1, \ldots, 1)$	$n-1$		$n-1$
Итого арифметических действий	$3(n-1)$	$2n-1$	$3(n-1)$

Учитывая, что часто операции сложения выполняются намного быстрее, чем умножения и деления, обычно ограничиваются подсчетом последних. Так, аналогично показанному нетрудно проверить, что для решения n-мерной СЛАУ методом Гаусса (без выбора главного элемента) требуется $\frac{n^3}{3} + n^2 - \frac{n}{3}$ умножений и делений*), а методом квадратных корней – $\frac{n^3}{6} + \frac{3}{2}n^2 + \frac{n}{3}$ плюс n операций извлечения корня (см. [8, 47, 54] и др.). Метод вращений предполагает вчетверо больше операций умножения, чем метод Гаусса [25]. При больших значениях размерности n существенным является старший член выражения для подсчета числа арифметических операций. Можно сказать, что вычислительные затраты на операции ум-

*) Несложно подсчитать и общее число операций, включая сложение [16, 17, 21], а также время решения n-мерной системы, если известна скорость выполнения различных операций [11].

ножения и деления в методе Гаусса составляют величину $O\left(\frac{n^3}{3}\right)$, в методе квадратных корней $O\left(\frac{n^3}{6}\right)$, в методе вращений $O\left(\frac{4}{3}n^3\right)$, в то время как прогонка требует всего $O(5n)$ таких операций.

УПРАЖНЕНИЯ

2.1. а) Решить систему

$$\begin{cases} 0.1x_1 + 2x_2 - 10x_3 = 0.6, \\ 0.3x_1 + 6.01x_2 - 25x_3 = 1.852, \\ 0.4x_1 + 8.06x_2 + 10.001x_3 = 2.91201 \end{cases}$$

методом Гаусса, пошагово выполняя предписания алгоритма из п.2.1.

б) Перемножением ведущих элементов метода Гаусса найти детерминант матрицы коэффициентов данной системы.

в) Выполнить задания а), б), имитируя работу модельного компьютера, в котором под запись мантиссы числа в режиме с плавающей запятой выделяется три десятичных разряда.

Проделать то же, проводя частичное упорядочивание по столбцам.

Сравнить результаты в) с результатами а) и б).

г) Подсчитав невязки приближенных решений данной системы, полученных в в), произвести итерационное уточнение этих решений в той же вычислительной среде (см. п.2.8а).

2.2. а) Выполняя LU-разложение матрицы

$$\mathbf{A} = \begin{pmatrix} 4 & -3 & 2 \\ 8 & -8 & 7 \\ 12 & -5 & 5 \end{pmatrix}$$

по формулам (2.9)–(2.10), найти решение системы $\mathbf{Ax} = \mathbf{b}$, где $\mathbf{b} = (0; -12; 4)^T$.

б) Используя полученное в а) LU-разложение, найти матрицу $\mathbf{A}^{-1}$ двумя способами: решая подсистемы $\mathbf{AX}_i = \mathbf{e}_i$, где $\mathbf{X}_i$ и $\mathbf{e}_i$ – столбцы со-

ответственно искомой и единичной матриц, и реализуя вычисления по формулам (2.15)–(2.17).

2.3. а) Методом квадратных корней решить систему

$$\begin{cases} 16x_1 - 8x_2 - 4x_3 = -8, \\ -8x_1 + 13x_2 - 4x_3 - 3x_4 = 7, \\ -4x_1 - 4x_2 + 9x_3 = 6, \\ -3x_2 + 3x_4 = -3. \end{cases}$$

б) С помощью полученного в а) $\mathbf{U}^T\mathbf{U}$-разложения Холецкого найти детерминант матрицы коэффициентов данной системы и обратную ей матрицу.

в) Подсчитав число обусловленности, выяснить, какую относительную погрешность может иметь результат а), если в одной компоненте правой части данной системы допустить абсолютную ошибку 0.01.

2.4. При каких $n \in N$ можно гарантировать корректность и устойчивость метода прогонки для решения системы (2.21), где:

$$\begin{aligned} b_i &= 1+i \quad \text{при } i = 2, 3, \ldots, n; \\ c_i &= 15+i \quad \text{при } i = 1, 2, \ldots, n; \\ d_i &= -i \quad \text{при } i = 1, 2, \ldots, n-1? \end{aligned}$$

2.5. Вывести формулы левой прогонки для решения системы (2.21) (т.е. такие, при которых неизвестные x_i вычислялись бы в порядке возрастания индексов).

Решить систему предыдущего упражнения при $n = 7$, $r_i = i^2 + 14i - 1$ (где $i = 1, \ldots, 6$), $r_7 = 202$ по формулам левой и правой прогонок.

2.6. Пусть в (2.21) $b_{i+1} = d_i$ при всех $i = 1, 2, \ldots, n-1$. Записать для этого случая расчетные формулы метода квадратных корней. Подсчитать число требуемых арифметических операций и сравнить его с аналогичным результатом для метода прогонки (см. п.2.8б).

2.7. Решить систему из упражнения 2.1 методом вращений:

а) используя всю разрядную сетку калькулятора или компьютера;

б) работая, как и в упражнении 2.1в), с тремя значащими цифрами.

Проанализировать результаты, сравнивая их с результатами упражнений 2.1а) и 2.1в).

ГЛАВА 3 ИТЕРАЦИОННЫЕ МЕТОДЫ РЕШЕНИЯ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ СИСТЕМ И ОБРАЩЕНИЯ МАТРИЦ

Рассматриваются итерационные способы решения систем линейных алгебраических уравнений и обращения матриц, служащие серьезной альтернативой прямым методам решения таких задач, по крайней мере в случаях, когда их размерность велика. Показывается логика построения нескольких наиболее важных итерационных процессов, таких как методы простых итераций, Якоби, Зейделя, релаксации, Шульца, и изучаются условия сходимости последовательностей приближений, получаемых этими методами, к искомым решениям; дается первое представление о методах установления.

3.1. РЕШЕНИЕ СЛАУ МЕТОДОМ ПРОСТЫХ ИТЕРАЦИЙ

Система

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \tag{3.1}$$

где $\mathbf{A} = \left(a_{ij}\right)_{i,j=1}^{n}$ – $n \times n$–матрица, а $\mathbf{x} = (x_1, \ldots, x_n)^T$ и $\mathbf{b} = (b_1, \ldots, b_n)^T$ – n–мерные векторы-столбцы, тем или иным способом (таких способов существует бесконечное множество; некоторые из них будут рассмотрены ниже) может быть преобразована к эквивалентной ей системе вида

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{c}, \tag{3.2}$$

где $\mathbf{x}$ – тот же вектор неизвестных, а $\mathbf{B}$ и $\mathbf{c}$ – некоторые новые матрица и вектор соответственно. Систему (3.2) можно трактовать как задачу о неподвижной точке линейного отображения $\mathbf{B}$ в пространстве R_n и по аналогии со скалярным случаем (более подробно изучаемым в гл.5) определить последовательность приближений $\mathbf{x}^{(k)}$ к неподвижной точке $\mathbf{x}^*$ рекуррентным равенством

$$\mathbf{x}^{(k+1)} = \mathbf{B}\mathbf{x}^{(k)} + \mathbf{c}, \qquad k = 0, 1, 2, \ldots . \tag{3.3}$$

Итерационный процесс (3.3), начинающийся с некоторого вектора $\mathbf{x}^{(0)} = \left(x_1^{(0)}, \ldots, x_n^{(0)}\right)^T$, будем называть ***методом простых итераций*** (коротко МПИ).

Изучим комплекс вопросов о сходимости этого процесса. А именно: 1) какие нужно предъявить требования к $\mathbf{B}$, $\mathbf{c}$ и $\mathbf{x}^{(0)}$, чтобы последовательность $\left(\mathbf{x}^{(k)}\right)$ при $k \to \infty$ имела пределом $\mathbf{x}^*$ – неподвижную точку задачи (3.2) (и значит, решение эквивалентной (3.2) исходной системы (3.1))? 2) с какой скоростью сходится этот процесс, т.е. каков закон убывания абсолютных погрешностей получаемых по формуле (3.3) приближений? 3) сколько нужно сделать итераций по формуле (3.3), чтобы при заданном начальном приближении найти решение задачи (3.2) с заданной точностью?

Ответы на подобные вопросы теории итерационных методов в R_n часто опираются на следующие два утверждения о сходимости степенных матричных рядов, точнее "матричной геометрической прогрессии". Во втором из них, а также всюду далее под нормой матрицы понимается мультипликативная норма такая, что $\|\mathbf{E}\| = 1$ ($\mathbf{E}$ – единичная матрица).

Лемма 3.1*). *Условие, что все собственные числа матрицы* $\mathbf{B}$ *по модулю меньше* 1, *является необходимым и достаточным для того, чтобы*:

1) $\mathbf{B}^k \to \mathbf{0}$ *при* $k \to \infty$ $(k \in N)$;

2) *матрица* $\mathbf{E} - \mathbf{B}$ *имела обратную и*

$$(\mathbf{E} - \mathbf{B})^{-1} = \mathbf{E} + \mathbf{B} + \mathbf{B}^2 + \ldots + \mathbf{B}^k + \ldots .$$

Лемма 3.2**) . *Если* $\|\mathbf{B}\| \le q < 1$, *то матрица* $\mathbf{E} - \mathbf{B}$ *имеет обратную матрицу* $(\mathbf{E} - \mathbf{B})^{-1} = \sum_{k=0}^{\infty} \mathbf{B}^k$ *и при этом* $\left\|(\mathbf{E} - \mathbf{B})^{-1}\right\| \le \dfrac{1}{1-q}$.

Доказательство. Рассмотрим матричный ряд

$$\mathbf{E} + \mathbf{B} + \mathbf{B}^2 + \ldots + \mathbf{B}^k + \ldots \quad . \tag{3.4}$$

---

*) В некоторых литературных источниках лемма 3.1 называется ***леммой Неймана*** (см., например, [42]).

) В функциональном анализе для более общего случая, когда $\mathbf{B}$ – линейный оператор, действующий в полных нормированных пространствах, эту лемму называют ***теоремой Банаха* ([29] и др.).

В силу условия леммы и вытекающего из мультипликативного свойства нормы неравенства $\|\mathbf{B}^k\| \le \|\mathbf{B}\|^k$, этот ряд можно промажорировать сходящимся числовым рядом:

$$\|\mathbf{E}\| + \|\mathbf{B}\| + \|\mathbf{B}^2\| + \ldots + \|\mathbf{B}^k\| + \ldots \le 1 + q + q^2 + \ldots + q^k + \ldots = \frac{1}{1-q}.$$

Следовательно, ряд (3.4) сходится, т.е. существует матрица

$$\mathbf{V} = \mathbf{E} + \mathbf{B} + \mathbf{B}^2 + \ldots + \mathbf{B}^k + \ldots$$

такая, что $\|\mathbf{V}\| \le \frac{1}{1-q}$. Так как

$$(\mathbf{E} - \mathbf{B})\mathbf{V} = (\mathbf{E} - \mathbf{B})\left(\mathbf{E} + \mathbf{B} + \mathbf{B}^2 + \ldots + \mathbf{B}^k + \ldots\right) =$$
$$= \mathbf{E} + \mathbf{B} + \mathbf{B}^2 + \ldots + \mathbf{B}^k + \ldots - \mathbf{B} - \mathbf{B}^2 - \mathbf{B}^3 - \ldots - \mathbf{B}^{k+1} - \ldots = \mathbf{E},$$

то $\mathbf{V} = (\mathbf{E} - \mathbf{B})^{-1}$. Лемма доказана.

Доказательство леммы 1 более сложно. Его можно найти во многих учебных пособиях по вычислительной математике и функциональному анализу (см., например, [8, 21, 29, 34, 42]).

Теорема 3.1. *Необходимым и достаточным условием сходимости метода простых итераций* (3.3) *при любом начальном векторе* $\mathbf{x}^{(0)}$ *к решению* $\mathbf{x}^*$ *системы* (3.2) *является требование, чтобы все собственные числа матрицы* $\mathbf{B}$ *были по модулю меньше* 1.

Доказательство. *Достаточность.* Пусть $\max \lambda_B < 1$, тогда по лемме 3.1 общий член $\mathbf{B}^k$ ряда (3.4) стремится к нуль-матрице и существует матрица $(\mathbf{E} - \mathbf{B})^{-1}$, являющаяся пределом частичных сумм $\left(\mathbf{E} + \mathbf{B} + \mathbf{B}^2 + \ldots + \mathbf{B}^k\right)$ при $k \to \infty$. Применяя рекурсию в равенстве (3.3), определяющем МПИ, получим:

$$\mathbf{x}^{(k+1)} = \mathbf{B}\mathbf{x}^{(k)} + \mathbf{c} = \mathbf{B}^2\mathbf{x}^{(k-1)} + (\mathbf{B} + \mathbf{E})\mathbf{c} = \ldots =$$
$$= \mathbf{B}^{k+1}\mathbf{x}^{(0)} + \left(\mathbf{E} + \mathbf{B} + \mathbf{B}^2 + \ldots + \mathbf{B}^k\right)\mathbf{c}. \qquad (3.5)$$

В силу сказанного выше, предел последнего выражения существует при любом фиксированном $\mathbf{x}^{(0)}$ и равен $(\mathbf{E} - \mathbf{B})^{-1}\mathbf{c}$. Следовательно, итерационный процесс (3.3) сходится и

$$\mathbf{x}^* := \lim_{k \to \infty} \mathbf{x}^{(k)} = (\mathbf{E} - \mathbf{B})^{-1}\mathbf{c}.$$

Подставляя $\mathbf{x}^*$ в уравнение (3.2), преобразованное к виду $(\mathbf{E}-\mathbf{B})\mathbf{x}=\mathbf{c}$, имеем:

$$(\mathbf{E}-\mathbf{B})(\mathbf{E}-\mathbf{B})^{-1}\mathbf{c}=\mathbf{c},$$

т.е. вектор $\mathbf{x}^*=\lim\limits_{k\to\infty}\mathbf{x}^{(k)}$ удовлетворяет системе (3.2). (Заметим, что это $\mathbf{x}^*$ – единственное решение (3.2). Действительно, допустив, что наряду с $\mathbf{x}^*$ таким, что $\mathbf{x}^*=\mathbf{B}\mathbf{x}^*+\mathbf{c}$, имеется $\mathbf{x}^{**}$, удовлетворяющее такому же равенству $\mathbf{x}^{**}=\mathbf{B}\mathbf{x}^{**}+\mathbf{c}$, получаем $\mathbf{x}^*-\mathbf{x}^{**}=\mathbf{B}\left(\mathbf{x}^*-\mathbf{x}^{**}\right)$. Последнее означает, что $\lambda=1$ по определению является собственным числом матрицы $\mathbf{B}$, что противоречит условию).

Необходимость. Как видно из представления общего члена итерационной последовательности $\left(\mathbf{x}^{(k)}\right)$ в форме (3.5), существование $\lim\limits_{k\to\infty}\mathbf{x}^{(k+1)}$ при любых векторах $\mathbf{x}^{(0)}$ и $\mathbf{c}$ (в том числе и нулевых, что гарантирует существование предела каждого слагаемого в правой части (3.5)) влечет сходимость матриц $\mathbf{B}^{k+1}$ к нуль-матрице и сходимость ряда $\sum\limits_{k=0}^{\infty}\mathbf{B}^k$ к $(\mathbf{E}-\mathbf{B})^{-1}$. Согласно лемме 3.1, это равносильно условию $\lambda_B<1$ для каждого собственного числа матрицы $\mathbf{B}$.

Теорема доказана.

Теорема 3.2. *Пусть* $\|\mathbf{B}\|\le q<1$. *Тогда при любом начальном векторе* $\mathbf{x}^{(0)}$ *МПИ* (3.3) *сходится к единственному решению* $\mathbf{x}^*$ *задачи* (3.2) *и при всех* $k\in N$ *справедливы оценки погрешности*:

1) $\left\|\mathbf{x}^*-\mathbf{x}^{(k)}\right\|\le\dfrac{q}{1-q}\left\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\right\|$ (***апостериорная***);

2) $\left\|\mathbf{x}^*-\mathbf{x}^{(k)}\right\|\le\dfrac{q^k}{1-q}\left\|\mathbf{x}^{(1)}-\mathbf{x}^{(0)}\right\|$ (***априорная***).*)

(Одно и то же обозначение $\|\cdot\|$ здесь используется для матричных и векторных норм, согласованных между собой, т.е. таких, что $\|\mathbf{A}\mathbf{x}\|\le\|\mathbf{A}\|\cdot\|\mathbf{x}\|$).

*) Лат. a priori и a posteriori означают соответственно "до опыта" и "из опыта", т.е. априорной оценкой можно воспользоваться до начала счета, а апостериорной – лишь после проведения k-й итерации.

Доказательство. Вычитая из (3.3) равенство $\mathbf{x}^{(k)} = \mathbf{B}\mathbf{x}^{(k-1)} + \mathbf{c}$, имеем $\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} = \mathbf{B}\left(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right)$. Переходя в последнем к нормам, получаем неравенство

$$\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| \le q\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\|, \qquad (3.6)$$

из которого видно, в силу условия $q < 1$, что элементы итерационной последовательности $\left(\mathbf{x}^{(k)}\right)$ сближаются с ростом номера k. С помощью (3.6) оценим разность между $(k+m)$-м и k-м членами этой последовательности при некотором $m \in N$:

$$\begin{aligned}
\left\|\mathbf{x}^{(k+m)} - \mathbf{x}^{(k)}\right\| &= \left\|\mathbf{x}^{(k+m)} - \mathbf{x}^{(k+m-1)} + \mathbf{x}^{(k+m-1)} - \mathbf{x}^{(k+m-2)} + \right. \\
&\left. + \mathbf{x}^{(k+m-2)} - \ldots - \mathbf{x}^{(k+1)} + \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| \le \\
&\le \left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| + \left\|\mathbf{x}^{(k+2)} - \mathbf{x}^{(k+1)}\right\| + \ldots + \left\|\mathbf{x}^{(k+m)} - \mathbf{x}^{(k+m-1)}\right\| \le \\
&\le q\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| + q^2\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| + \ldots + q^m\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| = \\
&= \frac{q\left(1-q^m\right)}{1-q}\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| \le \frac{q^k}{1-q}\left(1-q^m\right)\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\|.
\end{aligned}$$

Рассматривая итоговое неравенство при $k \to \infty$ и фиксированном m, видим, что $\left(\mathbf{x}^{(k)}\right)$ является фундаментальной последовательностью и, в силу полноты пространства R_n, имеет предел. Обозначим его $\mathbf{x}^*$. Переходя к пределу в равенстве (3.3), получаем $\mathbf{x}^* = \mathbf{B}\mathbf{x}^* + \mathbf{c}$, т.е. $\mathbf{x}^* = \lim\limits_{k\to\infty} \mathbf{x}^{(k)}$ – решение уравнения (3.2). При этом $\mathbf{x}^*$ – единственное решение (3.2), так как предположив существование другого решения $\mathbf{x}^{**} \ne \mathbf{x}^*$ и нормируя равенство $\mathbf{x}^* - \mathbf{x}^{**} = \mathbf{B}\left(\mathbf{x}^* - \mathbf{x}^{**}\right)$, приходим к противоречащему условию теоремы неравенству $\|\mathbf{B}\| \ge 1$.

Справедливость утверждаемых в теореме оценок погрешности видна из неравенств

$$\left\|\mathbf{x}^{(k+m)} - \mathbf{x}^{(k)}\right\| \le \frac{q\left(1-q^m\right)}{1-q}\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| \le \frac{q^k}{1-q}\left(1-q^m\right)\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\|,$$

если в них теперь зафиксировать k и перейти к пределу при $m \to \infty$. Теорема доказана.

Замечание 3.1. Последние неравенства говорят еще о том, что априорная оценка, как правило, грубее апостериорной.

Замечание 3.2. Теорема 3.2 могла быть доказана на основе леммы 3.2 и теоремы 3.1. В частности, сходимость $\left(\mathbf{x}^{(k)}\right)$ к решению $\mathbf{x}^*$ системы (3.2) сразу следует из теоремы 3.1, в силу соотношений $|\lambda_{\mathbf{B}}| \le \|\mathbf{B}\| < 1$. Из леммы 3.2 также легко вывести другую априорную оценку погрешности k-го приближения: вычитая из равенства

$$\mathbf{x}^* = (\mathbf{E}-\mathbf{B})^{-1}\mathbf{c} = \left(\mathbf{E}+\mathbf{B}+\ldots+\mathbf{B}^k+\ldots\right)\mathbf{c}$$

равенство

$$\mathbf{x}^{(k)} = \mathbf{B}^k\mathbf{x}^{(0)} + \left(\mathbf{E}+\mathbf{B}+\ldots+\mathbf{B}^{k-1}\right)\mathbf{c}$$

(см. (3.5)), имеем:

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| = \left\|\left(\mathbf{B}^k + \mathbf{B}^{k+1} + \ldots\right)\mathbf{c} - \mathbf{B}^k\mathbf{x}^{(0)}\right\| \le$$

$$\le \left\|\mathbf{B}^k\right\| \cdot \left\|(\mathbf{E}-\mathbf{B})^{-1}\mathbf{c} - \mathbf{x}^{(0)}\right\| \le q^k\left(\left\|\mathbf{x}^{(0)}\right\| + \frac{\|\mathbf{c}\|}{1-q}\right).$$

Замечание 3.3. Априорная оценка позволяет подсчитывать заранее число итераций k, достаточное для получения решения $\mathbf{x}^*$ с заданной точностью ε (в смысле допустимого уровня абсолютных погрешностей) при выбранном начальном векторе $\mathbf{x}^{(0)}$. Для этого нужно найти наименьшее целое решение неравенства

$$\frac{q^k}{1-q}\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\| \le \varepsilon$$

относительно переменной k (или неравенства $q^k\left(\left\|\mathbf{x}^{(0)}\right\| + \frac{\|\mathbf{c}\|}{1-q}\right) \le \varepsilon$ в соответствии с результатом предыдущего замечания). Апостериорной же оценкой удобно пользоваться непосредственно в процессе вычислений и останавливать этот процесс, как только выполнится неравенство

$$\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| \le \frac{1-q}{q}\varepsilon.$$

Отметим, что неравенство $\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| \le \varepsilon$ будет гарантией того, что и $\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \le \varepsilon$, только в том случае, когда $q \le \frac{1}{2}$.

Замечание 3.4. По поводу выбора начального приближения.

Как установлено выше, сходимость МПИ (3.3) гарантируется при любом начальном векторе $\mathbf{x}^{(0)}$. Очевидно, итераций потребуется тем меньше, чем ближе $\mathbf{x}^{(0)}$ к $\mathbf{x}^*$. Если нет никакой дополнительной информации о решении задачи (3.2) (например, может быть известным решение близкой задачи или грубое решение данной задачи), за $\mathbf{x}^{(0)}$ обычно принимают вектор $\mathbf{c}$ свободных членов системы (3.2). Мотивация этого может

быть такой: матрица **B** "мала", значит вектор **Bx** "мал", следовательно, и вектор $\mathbf{x}^*$ не должен сильно отличаться от вектора **c**. При выборе $\mathbf{x}^{(0)} = \mathbf{c}$ фигурирующая в теореме 3.2 априорная оценка принимает вид

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \le \frac{\|\mathbf{c}\|}{1-q} q^{k+1} \quad \forall\, k \in N.$$

3.2. МЕТОД ЯКОБИ

Вернемся к рассмотрению задачи (3.1). После выяснения условия, которому должна удовлетворять матрица коэффициентов приведенной системы (3.2) для сходимости МПИ (3.3), следует осуществить приведение системы (3.1) к виду (3.2) так, чтобы это условие выполнялось. Рассмотрим один из способов такого приведения, достаточно эффективный в определенных случаях.

Представим матрицу **A** системы (3.1) в виде

$$\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{R},$$

где **D** – диагональная, а **L** и **R** – соответственно левая и правая строго треугольные (т.е. с нулевой диагональю) матрицы. Тогда система (3.1) может быть записана в виде

$$\mathbf{Lx} + \mathbf{Dx} + \mathbf{Rx} = \mathbf{b}, \tag{3.7}$$

и если на диагонали исходной матрицы нет нулей, то эквивалентной (3.1) задачей вида (3.2) будет

$$\mathbf{x} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{R})\mathbf{x} + \mathbf{D}^{-1}\mathbf{b}, \tag{3.8}$$

т.е. в (3.2) и (3.3) следует положить

$$\mathbf{B} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{R}), \qquad \mathbf{c} = \mathbf{D}^{-1}\mathbf{b}.$$

Основанный на таком приведении системы (3.1) к виду (3.2) метод простых итераций (3.3) называют ***методом Якоби***[*]. В векторно-матричных обозначениях он определяется формулой

$$\mathbf{x}^{(k+1)} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{R})\mathbf{x}^{(k)} + \mathbf{D}^{-1}\mathbf{b}, \qquad k = 0, 1, 2, \ldots . \tag{3.9}$$

Чтобы записать метод Якоби (3.9) решения системы (3.1) в развернутом виде, достаточно заметить, что обратной матрицей к матрице $\mathbf{D} = \left(a_{ii}\right)_{i=1}^{n}$ служит диагональная матрица $\mathbf{D}^{-1}$ с элементами диагонали

[*] Карл Густав Якоб Якоби (1804–1851) – немецкий математик.

$d_{ii} = 1/a_{ii}$. Поэтому представление (3.8) системы (3.1), записанной в виде (3.7), равнозначно выражению "диагональных неизвестных" через остальные:

$$\begin{cases} x_1 = -\left(a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n - b_1\right)/a_{11}, \\ x_2 = -\left(a_{21}x_1 + a_{23}x_3 + \ldots + a_{2n}x_n - b_2\right)/a_{22}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n = -\left(a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{n,n-1}x_{n-1} - b_n\right)/a_{nn}. \end{cases} \tag{3.10}$$

Теперь для записи итерационного процесса (3.9) осталось в равенствах системы (3.10) только "навесить" индексы, соответствующие номерам приближений (т.е. $k = 0, 1, 2, \ldots$):

$$\begin{cases} x_1^{(k+1)} = -\left(a_{12}x_2^{(k)} + a_{13}x_3^{(k)} + \ldots + a_{1n}x_n^{(k)} - b_1\right)/a_{11}, \\ x_2^{(k+1)} = -\left(a_{21}x_1^{(k)} + a_{23}x_3^{(k)} + \ldots + a_{2n}x_n^{(k)} - b_2\right)/a_{22}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n^{(k+1)} = -\left(a_{n1}x_1^{(k)} + a_{n2}x_2^{(k)} + \ldots + a_{n,n-1}x_{n-1}^{(k)} - b_n\right)/a_{nn}. \end{cases} \tag{3.9'}$$

Установим простой достаточный признак сходимости метода Якоби к решению системы (3.1).

Теорема 3.3. *В случае диагонального преобладания в матрице* **A** *системы* (3.1) *метод Якоби* (3.9) *сходится.*

Доказательство. ***Диагональное преобладание*** в матрице **A** означает, что

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

Следовательно, в любой строке ***матрицы итерирования*** (иначе, ***матрицы перехода***)

$$\mathbf{B} = \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} & -\frac{a_{13}}{a_{11}} & \ldots & -\frac{a_{1,n-1}}{a_{11}} & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & -\frac{a_{23}}{a_{22}} & \ldots & -\frac{a_{2,n-1}}{a_{22}} & -\frac{a_{2n}}{a_{22}} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots & \ldots\ldots & \ldots\ldots \\ -\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & -\frac{a_{n3}}{a_{nn}} & \ldots & -\frac{a_{n,n-1}}{a_{nn}} & 0 \end{pmatrix} \tag{3.11}$$

метода Якоби сумма модулей элементов меньше единицы. Значит, по крайней мере, одна из норм матрицы **B** (согласованная, в частности, с век-

торной нормой-максимум) меньше единицы. Таким образом, существуют нормы, в которых к методу Якоби, рассматриваемому как метод простых итераций, применима теорема 3.2, т.е. метод Якоби сходится. Так как сходимость по одной норме в пространстве R_n означает сходимость по любой другой, тем самым теорема 3.3 доказана.

Замечание 3.5. Обратим внимание на то, что к методу Якоби при условии диагонального преобладания в матрице **A** относится полностью заключение теоремы 3.2, а также предыдущие замечания; нужно лишь учесть в них, что матрица **B** определяется с помощью (3.11), а вектор **c** – равенством

$$\mathbf{c} = \left(\frac{b_1}{a_{11}}; \frac{b_2}{a_{22}}; \ldots; \frac{b_n}{a_{nn}} \right)^T .$$

При этом матрица **D** заведомо обратима.

Следствием теоремы 1, устанавливающим необходимые и достаточные условия сходимости метода Якоби, является следующая теорема .

Теорема 3.4. *Метод Якоби* (3.9) *сходится к решению системы* (3.1) *в том и только в том случае, когда все корни уравнения*

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

по модулю меньше единицы.

Действительно, чтобы все собственные числа матрицы $\mathbf{B} = -\mathbf{D}^{-1}(\mathbf{L}+\mathbf{R})$ были по модулю меньше единицы, как этого требует теорема 3.1 для данного случая, нужно, чтобы меньше единицы были модули всех корней характеристического уравнения

$$\det\left(-\mathbf{D}^{-1}(\mathbf{L}+\mathbf{R}) - \lambda\mathbf{E}\right) = 0.$$

Последнее же эквивалентно уравнению

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

которое в записи через элементы исходной матрицы **A** и фигурирует в формулировке теоремы.

3.3. МЕТОД ЗЕЙДЕЛЯ

Под ***методом Зейделя***[*)] обычно понимается такое видоизменение метода простых итераций (3.3) решения СЛАУ, приведенных к виду (3.2), при котором для подсчета i-й компоненты $(k+1)$-го приближения к искомому вектору $\mathbf{x}^*$ используются уже найденные на этом, т.е. $(k+1)$-м шаге, новые значения первых $i-1$ компонент. Это означает, что если система (3.1) тем или иным способом сведена к системе (3.2) с матрицей коэффициентов $\mathbf{B}=\left(b_{ij}\right)_{i,j=1}^{n}$ и вектором свободных членов $\mathbf{c}=\left(c_i\right)_{i=1}^{n}$, то приближения к ее решению по методу Зейделя определяются системой равенств

$$\begin{cases} x_1^{(k+1)} = b_{11}x_1^{(k)} + b_{12}x_2^{(k)} + \ldots + b_{1n}x_n^{(k)} + c_1, \\ x_2^{(k+1)} = b_{21}x_1^{(k+1)} + b_{22}x_2^{(k)} + \ldots + b_{2n}x_n^{(k)} + c_2, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x_n^{(k+1)} = b_{n1}x_1^{(k+1)} + b_{n2}x_2^{(k+1)} + \ldots + b_{nn}x_n^{(k+1)} + c_n, \end{cases} \qquad (3.12)$$

где $k=0,1,2,\ldots$, а $x_i^{(0)}$ – компоненты заданного (выбранного) начального вектора $\mathbf{x}^{(0)}$.

Применительно к компьютерным расчетам один полный, т.е. векторный итерационный шаг метода Зейделя может интерпретироваться как реализация формулы (3.2), где под знаком равенства следует понимать знак присваивания, а под $\mathbf{x}$ – один и тот же линейный массив из n элементов, на нулевом шаге заполненный компонентами заданного начального вектора $\mathbf{x}^{(0)}$. В таком случае на обычной однопроцессорной вычислительной машине элементы массива $\mathbf{x}$ будут постепенно замещаться новыми элементами.

Аналогичный взгляд на МПИ (3.3) показывает, что для компьютерной реализации одного его шага нужно целиком сохранять n-элементный массив $\mathbf{x}$, подставляемый в правую часть, до тех пор, пока не сформируется полностью другой n-элементный массив – результат данного итерационного шага. В связи с такой интерпретацией метод Зейделя называют ***методом последовательных смещений***, а метод простых итераций – ***одновременных смещений***.

Остановимся подробнее на случае, когда приведение системы (3.1) к виду (3.2) основано на представлении (3.7), т.е. когда метод Зейделя есть

---

*) Людвиг Зейдель (1821–1896) – немецкий астроном и математик.

модификация метода Якоби. Запись соответствующих расчетных формул здесь сводится к верхней индексации системы (3.10) по типу (3.12):

$$\begin{cases} x_1^{(k+1)} = -\left(a_{12}x_2^{(k)} + a_{13}x_3^{(k)} + \ldots + a_{1n}x_n^{(k)} - b_1\right)/a_{11}, \\ x_2^{(k+1)} = -\left(a_{21}x_1^{(k+1)} + a_{23}x_3^{(k)} + \ldots + a_{2n}x_n^{(k)} - b_2\right)/a_{22}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n^{(k+1)} = -\left(a_{n1}x_1^{(k+1)} + a_{n2}x_2^{(k+1)} + \ldots + a_{n,n-1}x_{n-1}^{(k+1)} - b_n\right)/a_{nn}, \end{cases} \tag{3.13}$$

где $k = 0, 1, 2, \ldots$; $\mathbf{x}^{(0)} = \left(x_1^{(0)}, \ldots, x_n^{(0)}\right)^T$ задается.

Для анализа сходимости метода Зейделя (3.13) обратимся к его векторно-матричной форме. Легко видеть, что если неявный вид метода Якоби, вытекающий из представления (3.7) системы (3.1), есть

$$\mathbf{L}\mathbf{x}^{(k)} + \mathbf{D}\mathbf{x}^{(k+1)} + \mathbf{R}\mathbf{x}^{(k)} = \mathbf{b} \quad \text{(ср. с (3.9))},$$

то равнозначный (3.13) неявный вид метода Зейделя в векторно-матричных обозначениях суть

$$\mathbf{L}\mathbf{x}^{(k+1)} + \mathbf{D}\mathbf{x}^{(k+1)} + \mathbf{R}\mathbf{x}^{(k)} = \mathbf{b}.$$

Следовательно, тот же вектор $\mathbf{x}^{(k+1)}$, который фигурирует в левой части совокупности равенств (3.13), может быть получен по формуле

$$\mathbf{x}^{(k+1)} = -(\mathbf{L}+\mathbf{D})^{-1}\mathbf{R}\mathbf{x}^{(k)} + (\mathbf{L}+\mathbf{D})^{-1}\mathbf{b}. \tag{3.14}$$

Последнее выражение определяет не что иное, как МПИ (3.3) для системы вида (3.2), где

$$\mathbf{B} = -(\mathbf{L}+\mathbf{D})^{-1}\mathbf{R}, \qquad \mathbf{c} = (\mathbf{L}+\mathbf{D})^{-1},$$

т.е. результат применения одного шага метода Зейделя (3.13), полученного на основе $(\mathbf{L}+\mathbf{D}+\mathbf{R})$ – разложения матрицы $\mathbf{A}$, можно расценивать как шаг МПИ для эквивалентной (3.1) задачи о неподвижной точке

$$\mathbf{x} = -(\mathbf{L}+\mathbf{D})^{-1}\mathbf{R}\mathbf{x} + (\mathbf{L}+\mathbf{D})^{-1}\mathbf{b} \tag{3.15}$$

(разумеется, если треугольная матрица $\mathbf{L}+\mathbf{D}$ обратима). Эта связь между методом Зейделя и методом простых итераций позволяет легко переформулировать некоторые утверждения о сходимости МПИ применительно к методу Зейделя (3.13).

Теорема 3.5. *Для сходимости метода Зейделя необходимо и достаточно, чтобы все корни уравнения*

$$\begin{vmatrix} a_{11}\lambda & a_{12} & \ldots & a_{1n} \\ a_{21}\lambda & a_{22}\lambda & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1}\lambda & a_{n2}\lambda & \ldots & a_{nn}\lambda \end{vmatrix} = 0 \tag{3.16}$$

были по модулю меньше единицы.

Доказательство. Применяя теорему 3.1 к МПИ (3.14), составляем характеристическое уравнение, определяющее собственные числа λ матрицы $\mathbf{B} = -(\mathbf{L}+\mathbf{D})^{-1}\mathbf{R}$:

$$\det\left(-(\mathbf{L}+\mathbf{D})^{-1}\mathbf{R} - \lambda\mathbf{E}\right) = 0.$$

Это уравнение равносильно уравнению

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

которое с учетом $\mathbf{A} = \mathbf{L}+\mathbf{D}+\mathbf{R}$ совпадает с (3.16).

Замечание 3.6. Уравнение (3.16), а также метод (3.13), являющийся частным случаем более общей формы метода Зейделя (3.12), называют иногда соответственно ***уравнением*** и ***методом Некрасова*** [54]. Метод (3.12) называют еще и ***методом Гаусса-Зейделя*** [41, 42].

Прямым следствием теоремы 2 для метода Зейделя (3.13) является следующая теорема.

Теорема 3.6. *Пусть* $\left\|(\mathbf{L}+\mathbf{D})^{-1}\mathbf{R}\right\| \le t < 1$. *Тогда при любом начальном векторе* $\mathbf{x}^{(0)}$ *метод Зейделя* (3.13) *сходится к решению* $\mathbf{x}^*$ *системы* (3.1) *и справедливы оценки погрешности*

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \le \frac{t}{1-t}\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| \le \frac{t^k}{1-t}\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\|. \qquad (3.17)$$

Ясно, что для непосредственного использования оценок (3.17) нужно предварительно выполнить обращение треугольной матрицы $\mathbf{L}+\mathbf{D}$ и перемножить матрицы $(\mathbf{L}+\mathbf{D})^{-1}$ и $\mathbf{R}$. В таком случае частично теряется смысл в поэлементной реализации метода Зейделя (3.13); вместо этого можно проводить итерации по формуле (3.14) до тех пор, пока не выполнится условие $\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| \le \frac{1-t}{t}\varepsilon$, где $\varepsilon > 0$ – требуемая точность. В частности, такой подход может быть рекомендован при решении СЛАУ методом Зейделя на ЭВМ с векторной обработкой информации.

Как видим, условия сходимости методов Зейделя и простых итераций, вообще говоря, различаются. Но некоторые достаточные условия можно применять к обоим методам одновременно.

Теорема 3.7. *Если в матрице* $\mathbf{A}$ *системы* (3.1) *имеет место диагональное преобладание, то метод Зейделя* (3.13) *сходится, причем быстрее, чем метод Якоби* (3.9′).

Доказательство. Вычитая тождественное (3.13) равенство (3.14) из равенства (3.15), рассматриваемого как верное равенство при подстановке в него решения $\mathbf{x}^*$, получаем

$$\mathbf{x}^* - \mathbf{x}^{(k+1)} = -(\mathbf{L}+\mathbf{D})^{-1}\mathbf{R}\left(\mathbf{x}^* - \mathbf{x}^{(k)}\right). \qquad (3.18)$$

Введем в рассмотрение вектор ошибок

$$\Delta^{(k)} = \mathbf{x}^* - \mathbf{x}^{(k)}$$

с компонентами $\delta_i^{(k)} = x_i^* - x_i^{(k)}$. Через элементы матрицы $\mathbf{A}$ исходной системы (3.1) равенство (3.18) теперь можно записать так (см. соответствие между (3.14) и (3.13)):

$$\delta_i^{(k+1)} = -\frac{1}{a_{ii}}\sum_{j=1}^{i-1} a_{ij}\delta_j^{(k+1)} - \frac{1}{a_{ii}}\sum_{j=i+1}^{n} a_{ij}\delta_j^{(k)},$$

где $i = 1, 2, \ldots, n$; $k = 0, 1, 2, \ldots$. Переходя к модулям, отсюда имеем:

$$\left|\delta_i^{(k+1)}\right| \le \frac{1}{|a_{ii}|}\sum_{j=1}^{i-1}|a_{ij}|\cdot\left|\delta_j^{(k+1)}\right| + \frac{1}{|a_{ii}|}\sum_{j=i+1}^{n}|a_{ij}|\cdot\left|\delta_j^{(k)}\right| \le$$

$$\le \left(\max_j\left|\delta_j^{(k+1)}\right|\right)\cdot\frac{1}{|a_{ii}|}\sum_{j=1}^{i-1}|a_{ij}| + \left(\max_j\left|\delta_j^{(k)}\right|\right)\cdot\frac{1}{|a_{ii}|}\sum_{j=i+1}^{n}|a_{ij}|.$$

Обозначим $\alpha_i = \frac{1}{|a_{ii}|}\sum_{j=1}^{i-1}|a_{ij}|$, $\beta_i = \frac{1}{|a_{ii}|}\sum_{j=i+1}^{n}|a_{ij}|$ и $\left\|\Delta^{(k)}\right\|_\infty = \max_i\left|\delta_i^{(k)}\right|$ (где $\left\|\Delta^{(k)}\right\|_\infty$ может трактоваться как абсолютная погрешность k-го приближения по методу Зейделя*)). В этих обозначениях последнее неравенство имеет вид

$$\left|\delta_i^{(k+1)}\right| \le \alpha_i\left\|\Delta^{(k+1)}\right\|_\infty + \beta_i\left\|\Delta^{(k)}\right\|_\infty. \qquad (3.19)$$

Пусть $m \in \{1,2,\ldots,n\}$ – значение индекса i, при котором реализуется равенство $\left|\delta_m^{(k+1)}\right| = \max_i\left|\delta_i^{(k+1)}\right| = \left\|\Delta^{(k+1)}\right\|_\infty$. Тогда из (3.19) следует

$$\left\|\Delta^{(k+1)}\right\|_\infty \le \alpha_m\left\|\Delta^{(k+1)}\right\|_\infty + \beta_m\left\|\Delta^{(k)}\right\|_\infty,$$

*) Индекс ∞ у знака нормы использован согласно обозначению соответствующего частного случая l_p-нормы (нормы Гёльдера [16], см. приложение 1).

т.е. при этом фиксированном $i = m$ выполняется неравенство

$$\left\|\Delta^{(k+1)}\right\|_\infty \le \frac{\beta_m}{1-\alpha_m}\left\|\Delta^{(k)}\right\|_\infty . \qquad (3.20)$$

Так как в условиях диагонального преобладания справедливо неравенство $\alpha_i + \beta_i < 1$, а это неравенство, в свою очередь, влечет неравенство $\frac{\beta_i}{1-\alpha_i} \le \alpha_i + \beta_i$ (проверьте!), причем равенство в последнем случае имеет место лишь при $i = 1$, то абсолютная погрешность приближений по методу Зейделя (3.13), согласно (3.20), убывает со скоростью геометрической прогрессии, знаменатель которой, вообще говоря, меньше, чем для соответствующего этому случаю метода Якоби (3.9′). Такое мажорирование последовательности величин $\left\|\mathbf{x}^* - \mathbf{x}^{(k+1)}\right\|$ позволяет сделать заключение о справедливости доказываемой теоремы.

Замечание 3.7. В соответствии с последней теоремой в методе Зейделя (3.13) вместо оценок (3.17), требующих дополнительных затрат на обращение треугольной матрицы, можно использовать оценки погрешности метода Якоби. Естественно, они заведомо грубее.

Остановимся еще на одном важном для приложений классе систем вида (3.1), для которых имеет место сходимость метода Зейделя (3.13).

Определение 3.1 [21]. *Система* $\mathbf{Ax} = \mathbf{b}$ ***называется нормальной,*** *если матрица* $\mathbf{A}$ *– симметричная положительно определенная.*

Теорема 3.8. *Если система* (3.1) *– нормальная, то метод Зейделя* (3.13) *сходится.*

Доказательство этой теоремы заключается в проверке того, что положительная определенность матрицы $\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{L}^T$ влечет выполнимость условия теоремы 3.5 (т.е. собственные числа матрицы $-(\mathbf{L}+\mathbf{D})^{-1}\mathbf{L}^T$ по модулю меньше единицы). Это доказательство можно найти; например, в [8, 21].

Любая линейная система $\mathbf{Ax} = \mathbf{b}$ легко может быть симметризована умножением на матрицу $\mathbf{A}^T$. Более того, справедлива следующая теорема.

Теорема 3.9 [21]. *Пусть* $\det \mathbf{A} \neq 0$. *Тогда система* $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$ – *нормальная.*[*)]

Таким образом, если, например, известно, что система (3.1) однозначно разрешима, но в ее матрице коэффициентов нет диагонального преобладания, метод Зейделя типа (3.13) можно применять к системе $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$. Правда, при этом возникают трудности со своевременным окончанием процесса итерирования, обеспечивающим заданную точность приближенного решения, так как приведенные ранее оценки погрешности (см. теорему 3.6 и замечание 3.7) в этом случае "не работают". Да и сходимость в этом случае может оказаться весьма медленной.

Наряду с рассмотренными, применяют и другие способы приведения систем (3.1) к виду (3.2) для их решения методами простых итераций и Зейделя. Достаточно общий подход к этой процедуре заключается в том, что эквивалентное (3.1) уравнение $\mathbf{0} = \mathbf{b} - \mathbf{A}\mathbf{x}$ умножается на некоторую неособенную матрицу $\mathbf{H}$ (матричный параметр) и к обеим частям прибавляется вектор $\mathbf{x}$. Полученное уравнение

$$\mathbf{x} = \mathbf{x} + \mathbf{H}(\mathbf{b} - \mathbf{A}\mathbf{x}),$$

переписанное в виде

$$\mathbf{x} = (\mathbf{E} - \mathbf{H}\mathbf{A})\mathbf{x} + \mathbf{H}\mathbf{b},$$

имеет структуру (3.2). Проблема теперь заключается в подборе матрицы $\mathbf{H}$ такой, чтобы матрица $\mathbf{B} = \mathbf{E} - \mathbf{H}\mathbf{A}$ обладала нужными свойствами для сходимости применяемых методов; для некоторых классов матриц $\mathbf{A}$ имеются определенные рекомендации [21, 34]. Заметим, что матрица $\mathbf{H}$ может быть как постоянной (в этом случае говорят о ***стационарном*** итерационном процессе), так и изменяющейся от шага к шагу. В последнем случае данное уравнение $\mathbf{A}\mathbf{x} = \mathbf{b}$ подменяется последовательностью эквивалентных ему задач $\mathbf{x} = \mathbf{B}_k\mathbf{x} + \mathbf{c}_k$, и соответствующий итерационный процесс называется ***нестационарным***.

3.4. ПОНЯТИЕ О МЕТОДЕ РЕЛАКСАЦИИ

В случаях, когда применение оценок погрешностей в методах простых итераций и Зейделя невозможно из-за отсутствия констант $q < 1$ или $t < 1$, ограничивающих сверху какие-либо нормы матрицы итерирования соответствующего метода (см. теоремы 3.2 и 3.6), эти методы неэффективны и, более того, как будет показано в п.3.7, малонадежны ввиду медлен-

---

[*)] Переход от системы $\mathbf{A}\mathbf{x} = \mathbf{b}$ к системе $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$ (или в более общем случае к $\mathbf{A}^*\mathbf{A}\mathbf{x} = \mathbf{A}^*\mathbf{b}$) называют ***симметризацией Гаусса***.

ной сходимости. Рассмотрим одно обобщение метода Зейделя, позволяющее иногда в несколько раз ускорить сходимость итерационной последовательности.

Пусть $z_i^{(k)}$ – обозначение i-й компоненты k-го приближения к решению системы (3.1) по методу Зейделя, а обозначение $x_i^{(k)}$ будем использовать для i-й компоненты k-го приближения новым методом. Этот метод определим равенством

$$x_i^{(k+1)} = x_i^{(k)} + \omega\left(z_i^{(k+1)} - x_i^{(k)}\right), \qquad (3.21)$$

где $i = 1, 2, \ldots, n$; $k = 0, 1, 2, \ldots$; $x_i^{(0)}$ – задаваемые начальные значения; ω – числовой параметр, называемый ***параметром релаксации***. Очевидно, при $\omega = 1$ метод (3.21), называемый ***методом релаксации (ослабления)***, совпадает с методом Зейделя*).

Конкретизируем метод релаксации для случая, когда исходная система (3.1) представляется в виде (3.7) и, следовательно, метод Зейделя имеет вид (3.13).

Пользуясь введенными здесь обозначениями, запишем на основе (3.13) дополнительное к (3.21) равенство для выражения компонент векторов $\mathbf{z}^{(k)} = \left(z_i^{(k)}\right)_{i=1}^{n}$ через компоненты векторов $\mathbf{x}^{(k)} = \left(x_i^{(k)}\right)_{i=1}^{n}$:

$$z_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)}\right). \qquad (3.22)$$

Таким образом, метод релаксации можно понимать как поочередное применение формул (3.22) и (3.21) при каждом $k = 0, 1, 2, \ldots$. Действительно, задав начальные значения $x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}$ и параметр ω, при $k = 0$, полагая $i = 1, 2, \ldots, n$, вычислим

$$z_1^{(1)}, x_1^{(1)}; z_2^{(1)}, x_2^{(1)}; \ldots; z_n^{(1)}, x_n^{(1)},$$

при $k = 1$, так же полагая $i = 1, 2, \ldots, n$, находим

$$z_1^{(2)}, x_1^{(2)};\ z_2^{(2)}, x_2^{(2)};\ \ldots\ ;\ z_n^{(2)}, x_n^{(2)},$$

и т.д. Но можно избавиться от вспомогательной последовательности $\left(\mathbf{z}^{(k)}\right)$, подставив (3.22) в (3.21). Для $i = 1,2,\ldots,n$ будем иметь:

$$x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \frac{\omega}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)}\right). \qquad (3.23)$$

---

*) Метод Зейделя в качестве представителя семейства релаксационных методов называют иногда ***методом полной релаксации***.

От формулы (3.23), объединяющей формулы (3.22) и (3.21) и пригодной для проведения покоординатных вычислений, мало отличающихся от вычислений по методу Зейделя, легко перейти к векторно-матричной записи процесса релаксации. С этой целью перепишем (3.23) в виде

$$a_{ii}x_i^{(k+1)} + \omega\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} = (1-\omega)a_{ii}x_i^{(k)} - \omega\sum_{j=i+1}^{n} a_{ij}x_j^{(k)} + \omega b_i$$

и далее, учитывая аддитивное представление матрицы $\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{R}$, получаем векторно-матричный итерационный процесс в неявной форме

$$(\mathbf{D} + \omega\mathbf{L})\mathbf{x}^{(k+1)} = (1-\omega)\mathbf{D}\mathbf{x}^{(k)} - \omega\mathbf{R}\mathbf{x}^{(k)} + \omega\mathbf{b}.$$

Умножив последнее равенство слева на матрицу $(\mathbf{D} + \omega\mathbf{L})^{-1}$, приходим к эквивалентному (3.23) методу простых итераций

$$\mathbf{x}^{(k+1)} = (\mathbf{D} + \omega\mathbf{L})^{-1}\big((1-\omega)\mathbf{D} - \omega\mathbf{R}\big)\mathbf{x}^{(k)} + \omega(\mathbf{D} + \omega\mathbf{L})^{-1}\mathbf{b} \qquad (3.24)$$

(подстановка сюда значения $\omega = 1$ превращает (3.24) в МПИ (3.14), эквивалентный методу Зейделя (3.13)).

Представление метода релаксации (3.23) в виде (3.24) позволяет сделать для него некоторые утверждения о сходимости на основании соответствующих теорем о сходимости МПИ. Например, можно применить теоремы 3.1 и 3.2, полагая в них $\mathbf{B} = (\mathbf{D} + \omega\mathbf{L})^{-1}\big((1-\omega)\mathbf{D} - \omega\mathbf{R}\big)$, правда, получаемые при этом результаты вряд ли будут вызывать интерес. Более глубокие результаты на этом пути получают, изучая спектральные свойства таких матриц $\mathbf{B}$. Так, установлено, что *для сходимости процесса* (3.23) *необходимо, чтобы* $\omega \in (0;2)$. Для некоторых классов СЛАУ (3.1) это требование к параметру релаксации является и достаточным. Справедлива следующая теорема, обобщающая теорему 3.8.

Теорема 3.10. (***Островского-Рейча*** [41, 48]). *Для нормальной системы* $\mathbf{A}\mathbf{x} = \mathbf{b}$ *метод релаксации* (3.23) *сходится при любом* $\mathbf{x}^{(0)}$ *и любом* $\omega \in (0;2)$.

Поскольку итерационный процесс (3.23) содержит параметр, естественно распорядиться им так, чтобы сходимость последовательности $\left(\mathbf{x}^{(k)}\right)$ была наиболее быстрой. Очевидно, это достигается в том случае, когда спектральный радиус матрицы $\mathbf{B} = (\mathbf{D} + \omega\mathbf{L})^{-1}\big((1-\omega)\mathbf{D} - \omega\mathbf{R}\big)$ будет минимальным. В общем случае задача нахождения оптимального значения $\omega = \omega_0$ не решена, и в практических расчетах применяют метод проб и ошибок. Однако для отдельных важных классов задач такие значения удается выразить через собственные числа матрицы $\mathbf{D}^{-1}(\mathbf{L} + \mathbf{R})$ (т.е. корни уравнения, фигурирующего в теореме 3.4) и даже оценить ускорение, дос-

тигаемое введением в процесс Зейделя оптимального параметра релаксации. Существенно отметить, что это оптимальное значение $\omega_0 \in (1;2)$. При значениях $\omega \in (1;2)$ метод (3.23) называют ***методом последовательной верхней релаксации*** (сокращенно ПВР- или SOR-методом*)). Ввиду неэффективности метода (3.23) при $\omega \in (0;1)$, называемого в этом случае ***методом нижней релаксации***, название метод ПВР в последнее время относят ко всему семейству методов (3.23), т.е. для любых $\omega \in (0;2)$. При этом случай $\omega \in (1;2)$ называют ***сверхрелаксацией***.

3.5. О ДРУГИХ ИТЕРАЦИОННЫХ МЕТОДАХ РЕШЕНИЯ СЛАУ

В основе построения и изучения или, по крайней мере, понимания многих итерационных методов лежит связь между системами алгебраических уравнений и методами дискретизации дифференциальных уравнений, их порождающими.

В простейшем абстрактном, но далеко не самом общем случае, легко установить такую связь между СЛАУ (3.1) и абстрактным дифференциальным уравнением

$$\frac{d\mathbf{y}}{dt} + \mathbf{A}\mathbf{y}(t) = \mathbf{b} \qquad (3.25)$$

с начальным условием $\mathbf{y}(0) = \mathbf{x}^{(0)}$, где t – абстрактная скалярная переменная, изменяющаяся на промежутке $[0;+\infty)$, а матрица $\mathbf{A}$ и вектор $\mathbf{b}$ те же, что и в уравнении (3.1).

Пусть постоянный вектор $\mathbf{x}$ и переменный вектор $\mathbf{y} = \mathbf{y}(t)$ – решения задач (3.1) и (3.25) соответственно. Введем вектор

$$\mathbf{z}(t) = \mathbf{x} - \mathbf{y}(t).$$

Учитывая равенство $\frac{d\mathbf{z}}{dt} = -\frac{d\mathbf{y}}{dt}$, из совместного рассмотрения (3.1) и (3.25) выясняем, что $\mathbf{z}(t)$ удовлетворяет однородному дифференциальному уравнению

$$\frac{d\mathbf{z}}{dt} = -\mathbf{A}\mathbf{z}(t)$$

с начальным условием $\mathbf{z}(0) = \mathbf{x} - \mathbf{x}^{(0)}$. Решением этой начальной задачи служит вектор

$$\mathbf{z}(t) = e^{-\mathbf{A}t} \cdot \mathbf{z}(0),$$

*) От англ. Successive over relaxation

и если спектр A лежит в правой полуплоскости (в частности, если, например, матрица A положительно определена), то $z(t) \xrightarrow[t\to\infty]{} 0$ при любых $z(0)$. Таким образом, решение x системы (3.1) (стационарной задачи) может быть получено как предел решения $y(t)$ задачи Коши (3.25) (эволюционной задачи) при $t \to \infty$ с произвольным начальным вектором $\mathbf{x}^{(0)}$.

Методы приближенного решения стационарных задач, основанные на нахождении решений нестационарных задач, асимптотически эквивалентных данным задачам, для достаточно больших значений искусственной скалярной переменной, называются ***методами установления***[*].

Будем далее считать параметром скалярную величину τ_k, которую применительно к задаче (3.25) можно интерпретировать как шаг (вообще говоря, переменный), с которым на полуоси $[0;+\infty)$ фиксируются точки

$$t_0(=0), t_1, t_2, \ldots,$$

т.е. $t_{k+1} = t_k + \tau_k$, где $k = 0, 1, 2, \ldots$.

При "замораживании" $t = t_k$ уравнение (3.25) принимает вид

$$\left.\frac{d\mathbf{y}}{dt}\right|_{t=t_k} = -\mathbf{A}\mathbf{y}(t_k) + \mathbf{b}. \qquad (3.26)$$

Для производной в его левой части при малых τ_k на основе определения можно записать приближенное равенство

$$\left.\frac{d\mathbf{y}}{dt}\right|_{t=t_k} = \lim_{\tau_k \to 0} \frac{\mathbf{y}(t_k + \tau_k) - \mathbf{y}(t_k)}{\tau_k} \approx \frac{\mathbf{y}(t_{k+1}) - \mathbf{y}(t_k)}{\tau_k}.$$

Теперь ясно, что полагая $\mathbf{x}^{(k)} := \mathbf{y}(t_k)$ (заметим, что $\mathbf{y}(t_0) = \mathbf{y}(0) = \mathbf{x}^{(0)}$), равенство (3.26) можно приближенно заменить равенством

$$\frac{\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}}{\tau_k} = -\mathbf{A}\mathbf{x}^{(k)} + \mathbf{b}, \qquad (3.27)$$

---

[*] Иногда к дифференциальным уравнениям переходят не от исходной стационарной задачи, а от какого-то конкретного итерационного метода ее решения. Получающуюся при этом асимптотически эквивалентную дифференциальную задачу называют ***непрерывным аналогом*** соответствующего итерационного метода (см., например, [14, 18]).

которое можно рассматривать как некий ***явный*** итерационный процесс. Его называют ***двухслойным***[*] ***итерационным методом*** [48] или ***методом Ричардсона*** [47].

Более общий вид семейство двухслойных итерационных методов примет, если ввести в (3.27) невырожденный матричный параметр $\mathbf{B}_k$:

$$\mathbf{B}_k \frac{\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}}{\tau_k} = -\mathbf{A}\mathbf{x}^{(k)} + \mathbf{b}. \qquad (3.28)$$

Различные конкретные итерационные процессы решения СЛАУ (3.1) (в том числе и все рассмотренные выше) получаются из (3.28) фиксированием матриц $\mathbf{B}_k$ и скаляров τ_k. При этом, если $\mathbf{B}_k$ и τ_k не зависят от k, т.е. одни и те же на каждой итерации, то (3.28) определяет ***стационарный метод***, в противном случае – ***нестационарный***. В общем случае, за исключением $\mathbf{B}_k \equiv \mathbf{E}$, (3.28) – ***неявный метод***.

Выбор параметров $\mathbf{B}_k$, τ_k в (3.28) осуществляют, добиваясь удовлетворения каких-то отдельных или совокупности нескольких, возможно в чем-то противоречивых требований таких, как простота, хорошая структура и легкая обращаемость матриц $\mathbf{B}_k$, и в то же время, как можно более быстрая сходимость последовательности $\left(\mathbf{x}^{(k)}\right)$ к решению $\mathbf{x}^*$ системы (3.1). Разумеется, оптимальность или, скорее, квазиоптимальность некоторых методов рассматриваемого семейства удается установить лишь при очень жестких ограничениях на решаемую систему (3.1).

Так, например, доказано [47, 48], что если система (3.1) – нормальная с известными границами $\lambda_{\min} > 0$, $\lambda_{\max} > 0$ спектра ее матрицы коэффициентов, то при заранее зафиксированном (максимальном в реализуемом процессе) числе итераций K метод (3.27) будет обеспечивать наименьшую погрешность, иначе, минимизировать $\left\|\mathbf{x}^* - \mathbf{x}^{(K)}\right\|$ в том случае, когда параметры τ_k вычисляются по формуле

$$\tau_k = \frac{2}{(\lambda_{\max} + \lambda_{\min}) + (\lambda_{\max} - \lambda_{\min})t_{k+1}}, \qquad (3.29)$$

где $k = 0, 1, \ldots, K-1$, а $t_k = \cos\dfrac{(2k-1)\pi}{2K}$ – корни полинома Чебышева K-й степени. Совокупность формул (3.27), (3.29) называют ***явным итерационным методом с чебышевским набором параметров***. Имеется обобщение приведенного утверждения и на неявный случай.

---

[*] Смысл термина "двухслойный" становится понятным при изучении численных процессов решения уравнений математической физики. Изучая же численное интегрирование систем ОДУ, обнаруживаем, что (3.27) есть не что иное, как явный метод Эйлера (с переменным шагом) для задачи (3.25).

Дальнейшее формальное развитие методы установления получают как сугубо неявные методы вида (3.28) с матрицами $\mathbf{B}_k$, представляемыми в виде произведения простых легко обращаемых (например, ленточных) матриц, в связи с чем такие методы называются ***методами расщепления***. Из методов расщепления наиболее известными являются ***методы переменных направлений***[*] и ***попеременно-треугольный метод***. Неформальное изучение этих методов более целесообразно по месту их применения: при численном решении многомерных задач математической физики.

Рассматриваются также ***трехслойные итерационные методы*** (в частности, с чебышевскими параметрами), связывающие уже не два, а три соседних приближения: $\mathbf{x}^{(k+1)}$, $\mathbf{x}^{(k)}$ и $\mathbf{x}^{(k-1)}$. В отличие от предыдущих, такие методы являются ***двухшаговыми***.

Другой большой класс методов итерационного решения СЛАУ (3.1) – это так называемые ***методы вариационного типа***. К ним относятся методы минимальных невязок, минимальных поправок, минимальных итераций, наискорейшего спуска, сопряженных градиентов и т.п. Хорошего понимания и обоснования таких методов можно достигнуть лишь с привлечением теории оптимизации, ибо решение линейной алгебраической системы здесь подменяется решением эквивалентной экстремальной задачи.

А именно, пусть $\mathbf{Ax} = \mathbf{b}$ – нормальная n-мерная система, т.е. $\mathbf{A}$ – положительно определенная симметричная матрица, и пусть $(\cdot,\cdot)$ – скалярное произведение в пространстве R_n. Образуем квадратичный функционал

$$\Phi(\mathbf{x}) = (\mathbf{Ax},\mathbf{x}) - 2(\mathbf{b},\mathbf{x}) + c, \qquad (3.30)$$

где $c \in R_1$ – произвольная постоянная. Задача решения нормальной системы (3.1) и задача минимизации функционала (3.30) эквивалентны ([46] и др.). Действительно, нормальная система имеет и притом единственное решение; обозначим его $\mathbf{x}^*$. Тогда при любом векторе $\mathbf{x} = \mathbf{x}^* + \Delta$

$$\Phi(\mathbf{x}) = \Phi(\mathbf{x}^* + \Delta) = (\mathbf{A}(\mathbf{x}^* + \Delta), \mathbf{x}^* + \Delta) - 2(\mathbf{b}, \mathbf{x}^* + \Delta) + c =$$

$$= (\mathbf{Ax}^*, \mathbf{x}^*) + (\mathbf{A}\Delta, \mathbf{x}^*) + (\mathbf{Ax}^*, \Delta) + (\mathbf{A}\Delta, \Delta) - 2(\mathbf{b}, \mathbf{x}^*) - 2(\mathbf{b}, \Delta) + c =$$

$$= \Phi(\mathbf{x}^*) + (\mathbf{A}\Delta, \mathbf{x}^*) + (\mathbf{Ax}^*, \Delta) - 2(\mathbf{Ax}^*, \Delta) + (\mathbf{A}\Delta, \Delta) =$$

$$= \Phi(\mathbf{x}^*) + (\mathbf{A}\Delta, \Delta) \quad > \quad \Phi(\mathbf{x}^*),$$

в силу самосопряженности и положительности $\mathbf{A}$, значит,

$$\Phi(\mathbf{x}^*) = \min_{\mathbf{x} \in R_n} \Phi(\mathbf{x}).$$

---

[*] В зарубежной литературе используется аббревиатура ADI – Alternating Direction Implicite [16].

Теперь можно применять различные методы численной минимизации функционала $\Phi(\mathbf{x})$ (функции n переменных $x_1, x_2, \ldots, x_n$).

Одним из наиболее популярных и хорошо разработанных методов подобного типа является ***метод сопряженных градиентов***. Приведем без вывода алгоритм, быть может, недостаточно подробный, но вполне определенный, чтобы с его помощью можно было решать нормальные СЛАУ (3.1) таким способом [41].

1.1. Задать $\mathbf{x}^{(0)}$ (начальный вектор) и число $\varepsilon > 0$ (уровень допустимых погрешностей).

1.2. Вычислить вектор $\xi^{(0)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}$ (невязка начального приближения).

1.3. Положить $\mathbf{p}^{(0)} = \xi^{(0)}$, $k = 0$ (номер итерации).

2.1. Вычислить скаляр $\alpha_k = \left(\xi^{(k)}, \mathbf{p}^{(k)}\right) \Big/ \left(\xi^{(k)}, \mathbf{A}\mathbf{p}^{(k)}\right)$.

2.2. Вычислить вектор $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{p}^{(k)}$ (очередное приближение).

2.3. Вычислить $\xi^{(k+1)} = \xi^{(k)} - \alpha_k \mathbf{A}\mathbf{p}^{(k)}$ (невязка $(k+1)$-го приближения*).

2.4. Проверить выполнение неравенства $\left\|\xi^{(k+1)}\right\| \le \varepsilon$; если "да", остановить работу алгоритма и вывести результаты.

3.1. Вычислить скаляр $\beta_k = \left(\xi^{(k+1)}, \mathbf{A}\mathbf{p}^{(k)}\right) \Big/ \left(\mathbf{p}^{(k)}, \mathbf{A}\mathbf{p}^{(k)}\right)$.

3.2. Вычислить вектор $\mathbf{p}^{(k+1)} = \xi^{(k+1)} - \beta_k \mathbf{p}^{(k)}$ (новое направление минимизации).

3.3. Положить $k := k + 1$ и вернуться к шагу 2.1.

Интересно определить место, которое занимает этот метод в общей классификации методов решения линейных алгебраических систем. Дело в том, что метод сопряженных градиентов, являясь по форме итерационным, фактически должен быть отнесен к прямым методам, ибо доказано, что с его помощью минимум квадратичной функции (3.30) от n переменных, иначе, решение n-мерной линейной системы (3.1), достигается ровно за n шагов при любом начальном векторе $\mathbf{x}^{(0)}$. Применяют же метод сопряженных градиентов именно как итерационный метод (что видно и из при-

---

*) Полагая $\xi^{(k)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k)}$, видим, что в силу 2.2,

$$\xi^{(k+1)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k+1)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k)} - \mathbf{A}\alpha_k \mathbf{p}^{(k)} = \xi^{(k)} - \alpha_k \mathbf{A}\mathbf{p}^{(k)}.$$

Такое выражение невязки $\xi^{(k+1)}$ позволяет обходиться без вычисления вектора $\mathbf{A}\mathbf{x}^{(k+1)}$.

веденного алгоритма), имея в виду два обстоятельства. Во-первых, реальный вычислительный процесс может быть довольно далек от идеального и, вследствие неизбежных ошибок округления, на n-м шаге может быть не достигнута нужная точность. Во-вторых, если размерность n решаемой задачи велика, то число шагов, достаточное для получения решения системы с нужной точностью (т.е. выход по критерию 2.4), может оказаться значительно меньшим этой (n) теоретической величины.

Простейший вариант ***метода минимальных невязок*** определяется совокупностью формул

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \tau_k \xi^{(k)}, \qquad \xi^{(k)} = \mathbf{A}\mathbf{x}^{(k)} - \mathbf{b}, \qquad \tau_k = \frac{\left(\mathbf{A}\xi^{(k)}, \xi^{(k)}\right)}{\left(\mathbf{A}\xi^{(k)}, \mathbf{A}\xi^{(k)}\right)}.$$

Его можно рассматривать как явный двухслойный итерационный процесс (3.27), в котором параметр τ_k на каждом итерационном шаге $k = 0, 1, 2, \ldots$ выбирается таким, чтобы минимизировалась евклидова норма невязки $\xi^{(k+1)}$ получаемого приближения $\mathbf{x}^{(k+1)}$.

Действительно, вычтем из вектора $\xi^{(k+1)} = \mathbf{A}\mathbf{x}^{(k+1)} - \mathbf{b}$ вектор $\xi^{(k)}$. Имеем

$$\xi^{(k+1)} - \xi^{(k)} = \mathbf{A}\mathbf{x}^{(k+1)} - \mathbf{A}\mathbf{x}^{(k)} = \mathbf{A}\mathbf{x}^{(k)} - \mathbf{A}\tau_k\xi^{(k)} - \mathbf{A}\mathbf{x}^{(k)},$$

т.е.

$$\xi^{(k+1)} = \xi^{(k)} - \tau_k \mathbf{A}\xi^{(k)}.$$

Возводя последнее равенство в квадрат (в смысле скалярного умножения векторов), получаем

$$\left(\xi^{(k+1)}, \xi^{(k+1)}\right) = \left(\xi^{(k)}, \xi^{(k)}\right) - 2\tau_k\left(\mathbf{A}\xi^{(k)}, \xi^{(k)}\right) + \tau_k^2\left(\mathbf{A}\xi^{(k)}, \mathbf{A}\xi^{(k)}\right)$$

или, что то же,

$$\left\|\xi^{(k+1)}\right\|^2 = \left\|\xi^{(k)}\right\|^2 - 2\tau_k\left(\mathbf{A}\xi^{(k)}, \xi^{(k)}\right) + \tau_k^2 \cdot \left\|\mathbf{A}\xi^{(k)}\right\|^2.$$

Легко видеть, что минимум этой положительной квадратичной функции (значит, и величины $\left\|\xi^{(k+1)}\right\|$) достигается именно при указанном в записи метода значении τ_k.

В случае нормальной системы для метода минимальных невязок можно получить ту же оценку скорости сходимости, что и для метода простой итерации

$$\mathbf{x}^{(k+1)} = \left(\mathbf{E} - \tau\mathbf{A}\right)\mathbf{x}^{(k)} + \tau\mathbf{b}$$

при оптимальном значении параметра $\tau = \dfrac{2}{\lambda_{\min} + \lambda_{\max}}$ (в предположении, что известны границы $\lambda_{\min}$ и $\lambda_{\max}$ спектра матрицы $\mathbf{A}$) [47].

Рассмотренные здесь методы далеко не исчерпывают все многообразие итерационных способов решения СЛАУ. В частности, нами совсем не затрагивалась проблема решения больших разреженных систем, где на первый план выходят блочные методы, максимально сохраняющие исходную разреженность матриц.

3.6. БЫСТРОСХОДЯЩИЙСЯ ИТЕРАЦИОННЫЙ СПОСОБ ОБРАЩЕНИЯ МАТРИЦ

Согласно леммам 3.1, 3.2 (см. п.3.1), если матрица $\mathbf{B} = \mathbf{E} - \mathbf{A}$ мала (в смысле ее нормы или собственных значений), то обратная к $\mathbf{A}$ матрица

$$\mathbf{A}^{-1} = (\mathbf{E} - \mathbf{B})^{-1} = \mathbf{E} + \mathbf{B} + \mathbf{B}^2 + \ldots,$$

в принципе, может быть найдена сколь угодно точно приближенным суммированием данного матричного ряда. Однако такой непосредственный подход к вычислению имеет два очевидных недостатка: во-первых, его можно применить лишь для обращения матриц, близких к единичной, во-вторых, сходимость последовательностей частичных сумм этого ряда будет медленной даже при достаточно малых нормах матриц $\mathbf{B}$. Поэтому, пользуясь отмеченным фактом лишь как теоретической основой, построим итерационный процесс, определяющий существенно более быстро сходящуюся последовательность приближений к обратной для $\mathbf{A}$ матрице $\mathbf{A}^{-1}$. Будем далее обозначать эти приближения, получаемые на k-м шаге, через $\mathbf{U}_k$, а их ***невязки*** $\mathbf{E} - \mathbf{A}\mathbf{U}_k$ – через Ψ_k.

Лемма 3.3. *Если для матрицы* $\mathbf{A}$ *найдется такая обратимая матрица* $\mathbf{U}_0$, *что модули всех собственных чисел матрицы* $\Psi_0 = \mathbf{E} - \mathbf{A}\mathbf{U}_0$ *меньше единицы, то матрица* $\mathbf{A}$ *обратима и для обратной матрицы справедливо представление*

$$\mathbf{A}^{-1} = \mathbf{U}_0(\mathbf{E} - \Psi_0)^{-1} = \mathbf{U}_0\left(\mathbf{E} + \Psi_0 + \Psi_0^2 + \ldots\right). \qquad (3.31)$$

Доказательство. Из равенства

$$\mathbf{A}\mathbf{U}_0 = \mathbf{E} - \Psi_0, \qquad (3.32)$$

в силу обратимости $\mathbf{U}_0$ и $\mathbf{E} - \Psi_0$ (последнее по лемме 3.1), имеем

$$\mathbf{A} = (\mathbf{E} - \Psi_0)\mathbf{U}_0^{-1} = \left(\left((\mathbf{E} - \Psi_0)^{-1}\right)^{-1}\mathbf{U}_0^{-1}\right) = \left(\mathbf{U}_0(\mathbf{E} - \Psi_0)^{-1}\right)^{-1},$$

т.е. матрица $\mathbf{A}$ обратима и

$$\mathbf{A}^{-1} = \mathbf{U}_0(\mathbf{E} - \Psi_0)^{-1}.$$

Доказательство завершается разложением $(\mathbf{E} - \Psi_0)^{-1}$ в матричный ряд (лемма 3.1).

Очевидным следствием лемм 3.2 и 3.3 является следующая лемма.

Лемма 3.4. *Пусть матрица* $\mathbf{U}_0$ *обратима и* $\|\Psi_0\| < 1$. *Тогда*:

1) *существует матрица* $\mathbf{A}^{-1}$;

2) *справедливо представление* $\mathbf{A}^{-1}$ *по формуле* (3.31);

3) *имеет место оценка* $\left\|\mathbf{A}^{-1}\right\| \le \dfrac{\|\mathbf{U}_0\|}{1-\|\Psi_0\|}$.

Для построения итерационного процесса зафиксируем в разложении (3.31) $m+1$ первых слагаемых и будем считать первым приближением к $\mathbf{A}^{-1}$ матрицу

$$\mathbf{U}_1 = \mathbf{U}_0\left(\mathbf{E} + \Psi_0 + \ldots + \Psi_0^m\right).$$

Найдем выражение невязки Ψ_1 этого приближения через невязку Ψ_0 предыдущего (в данном случае начального) приближения $\mathbf{U}_0$:

$$\begin{aligned}\Psi_1 &= \mathbf{E} - \mathbf{A}\mathbf{U}_1 = \mathbf{E} - \mathbf{A}\mathbf{U}_0\left(\mathbf{E} + \Psi_0 + \ldots \Psi_0^m\right) = \\ &= \mathbf{E} - \left(\mathbf{E} - \Psi_0\right)\left(\mathbf{E} + \Psi_0 + \cdots + \Psi_0^m\right) = \mathbf{E} - \left(\mathbf{E} - \Psi_0^{m+1}\right) = \Psi_0^{m+1}.\end{aligned} \tag{3.33}$$

Благодаря полученной связи между невязками, можно утверждать, что если выполняются условия лемм 3.3 или 3.4 по отношению к матрицам $\mathbf{U}_0$, Ψ_0, то для матриц $\mathbf{U}_1$, Ψ_1 они будут выполнены и подавно. Следовательно, к матрицам $\mathbf{U}_1$, Ψ_1 можно применить все рассуждения, проведенные выше для $\mathbf{U}_0$, Ψ_0. Таким образом, приходим к итерационному процессу

$$\begin{cases}\Psi_k = \mathbf{E} - \mathbf{A}\mathbf{U}_k, \\ \mathbf{U}_{k+1} = \mathbf{U}_k\left(\mathbf{E} + \Psi_k + \ldots + \Psi_k^m\right).\end{cases} \tag{3.34}$$

где $k = 0, 1, 2, \ldots$ – номер итерации; $\mathbf{U}_0$ – задаваемая начальная матрица, близкая к $\mathbf{A}^{-1}$ в указанном выше смысле, а $m \in \mathbf{N}$ – параметр метода.

Изучим сходимость этого процесса.

Теорема 3.11. *Пусть квадратные матрицы* $\mathbf{A}$ *и* $\mathbf{U}_0$ *таковы, что матрица* $\mathbf{U}_0$ *обратима и* $\|\Psi_0\| < 1$. *Тогда существует обратная к* $\mathbf{A}$ *матрица* $\mathbf{A}^{-1}$ *и к ней сходится последовательность матриц* $\mathbf{U}_k$, *определяемая итерационным процессом* (3.34). *При этом имеет место точное равенство*

$$\mathbf{A}^{-1} - \mathbf{U}_k = \left(\mathbf{A}^{-1} - \mathbf{U}_0\right)\Psi_0^{(m+1)^k - 1} \tag{3.35}$$

и справедливы оценки погрешности:

1) $\left\|\mathbf{A}^{-1}-\mathbf{U}_k\right\| \le \dfrac{\left\|\mathbf{U}_k\Psi_k\right\|}{1-\left\|\Psi_k\right\|}$;

2) $\left\|\mathbf{A}^{-1}-\mathbf{U}_k\right\| \le \dfrac{\left\|\mathbf{U}_0\right\|}{1-\left\|\Psi_0\right\|}\cdot\left\|\Psi_0\right\|^{(m+1)^k}$.

Доказательство. Существование $\mathbf{A}^{-1}$ следует из леммы 3.4. Упомянутая повторяемость рассуждений и выкладок, проведенных на первом итерационном шаге, позволяет считать очевидными равенства типа (3.31), (3.33) для k-й итерации:

$$\mathbf{A}^{-1} = \mathbf{U}_k\left(\mathbf{E}-\Psi_k\right)^{-1} = \mathbf{U}_k\left(\mathbf{E}+\Psi_k+\Psi_k^2+\ldots\right), \tag{3.36}$$

$$\Psi_k = \mathbf{E}-\mathbf{A}\mathbf{U}_k = \Psi_{k-1}^{m+1} = \Psi_{k-2}^{(m+1)^2} = \ldots = \Psi_0^{(m+1)^k}. \tag{3.37}$$

Из (3.31) имеем

$$\begin{aligned}\mathbf{A}^{-1}-\mathbf{U}_0 &= \mathbf{U}_0\left(\mathbf{E}+\Psi_0+\Psi_0^2+\ldots\right)-\mathbf{U}_0 = \\ &= \mathbf{U}_0\left(\mathbf{E}+\Psi_0+\Psi_0^2+\ldots\right)\Psi_0 = \mathbf{A}^{-1}\Psi_0,\end{aligned} \tag{3.38}$$

а из (3.36) аналогично (с учетом (3.37)) получаем

$$\mathbf{A}^{-1}-\mathbf{U}_k = \mathbf{A}^{-1}\Psi_k = \mathbf{A}^{-1}\Psi_0^{(m+1)^k}. \tag{3.39}$$

Заменяя здесь в правой части $\mathbf{A}^{-1}\Psi_0$ на $\mathbf{A}^{-1}-\mathbf{U}_0$ (см. (3.38)), получаем утверждаемое в теореме равенство (3.35). Переходя в нем к нормам, в соответствии с условием заключаем, что

$$\left\|\mathbf{A}^{-1}-\mathbf{U}_k\right\| \le \left\|\mathbf{A}^{-1}-\mathbf{U}_0\right\|\cdot\left\|\Psi_0\right\|^{(m+1)^k-1} \xrightarrow[k\to\infty]{} 0,$$

т.е. имеет место сходимость $\left(\mathbf{U}_k\right)_{k=1}^{\infty}$ к $\mathbf{A}^{-1}$ по норме, а значит, и поэлементная сходимость.

Для доказательства первой оценки (апостериорной) вычтем $\mathbf{U}_k$ из (3.36):

$$\mathbf{A}^{-1}-\mathbf{U}_k = \mathbf{U}_k\left(\mathbf{E}+\Psi_k+\Psi_k^2+\ldots\right)-\mathbf{U}_k = \mathbf{U}_k\Psi_k\left(\mathbf{E}-\Psi_k\right)^{-1}.$$

Отсюда по лемме 3.2 с учетом (3.37) получаем требуемую оценку 1).

Вторая оценка (априорная) может быть найдена в результате загрубления первой. Но можно вывести ее непосредственно из равенства (3.39), подставив в его правую часть вместо $\mathbf{A}^{-1}$ выражение $\mathbf{U}_0\left(\mathbf{E}-\Psi_0\right)^{-1}$ (см. (3.31)):

$$\mathbf{A}^{-1}-\mathbf{U}_k = \mathbf{U}_0\left(\mathbf{E}-\Psi_0\right)^{-1}\Psi_0^{(m+1)^k}.$$

Переход к нормам в последнем равенстве и привлечение леммы 3.2 завершает доказательство теоремы.

Равенства (3.34) определяют фактически не один, а целое семейство итерационных методов обращения. Фиксированием параметра $m = 1, 2, \ldots$ можно получать конкретные процессы $(m+1)$-го порядка скорости сходимости*). Этот порядок может быть сколь угодно большим, однако обычно ограничиваются процессами второго $(m = 1)$ и третьего $(m = 2)$ порядков. Приоритет процесса второго порядка связан с его простотой и более ранним появлением: первая публикация об этом методе относится к 1933 г. и принадлежит Г. Шульцу [61], в связи с чем и все семейство (3.34) естественно называть ***методом Шульца***** . Метод третьего порядка целесообразно использовать из тех соображений, что он, как показал М. Альтман [59], обладает свойством минимальности вычислительных затрат, требующихся для обращения матриц с заданной точностью методами семейства (3.34).

Отметим, что как сам быстросходящийся итерационный процесс (3.34), так и представленные теоремой 3.11 результаты можно без каких-либо особых дополнительных условий отнести к более общей задаче обращения линейных ограниченных операторов в полных нормированных пространствах.

Процесс (3.34) построения приближений к обратной матрице легко видоизменить подобно тому, как это было сделано с методом простых итераций решения СЛАУ, когда для более оперативного учета получаемой на текущей итерации информации перешли от него к методу Зейделя (см. п.3.3). Например, ***зейделева модификация метода Шульца второго порядка*** может быть определена равенствами

$$\begin{cases} \Psi_k = \mathbf{E} - \mathbf{A}\mathbf{U}_k, \\ \mathbf{U}_{k+1} = \mathbf{U}_k + \mathbf{U}_k \underline{\Psi_k} + \mathbf{U}_{k+1}\overline{\Psi_k}, \end{cases} \tag{3.40}$$

где $k = 0, 1, 2, \ldots$; $\Psi_k = \underline{\Psi_k} + \overline{\Psi_k}$, а $\underline{\Psi_k}$ и $\overline{\Psi_k}$ – соответственно нижняя треугольная и строго верхняя треугольная матрицы [14]. При реализации этой модификации нужно либо расписывать формулы (3.40) поэлементно (чтобы не работать с заведомо нулевыми элементами), либо формировать матрицу $\mathbf{U}_{k+1}$ постепенным замещением старых элементов новыми, осуществляя на k-й итерации цикл присвоений

$$\mathbf{U} := \mathbf{U} + \mathbf{U}\Psi,$$

*) Определение порядка итерационного процесса см далее в п 5.3

**) В разных литературных источниках можно встретить и другие названия этого метода: *Хотеллинга, Бодевига* (Бодвига), а также *Hobo*.

где до начала цикла в правой части в двумерном массиве $\mathbf{U}$ должна содержаться матрица $\mathbf{U}_k$, а в двумерном массиве Ψ – матрица Ψ_k (заполнение массивов новыми элементами производится по строкам). Процесс (3.40) при том же шаговом объеме вычислений и такой же простоте, что и в методе Шульца второго порядка, может дать определенный выигрыш в скорости сходимости. Это можно показать сравнением невязок первых приближений при тех или иных предположениях относительно начальной невязки, иначе, при тех или иных требованиях к начальной матрице $\mathbf{U}_0$ [14].

Вообще, проблема выбора начального приближения $\mathbf{U}_0$ в рассматриваемых здесь процессах итерационного обращения матриц не позволяет относиться к ним как к самостоятельным универсальным методам, конкурирующим с прямыми методами обращения, основанными, например, на LU-разложении матриц. Имеются некоторые рекомендации по выбору $\mathbf{U}_0$ (см. [8, 59] и др.), обеспечивающие выполнение условия*) $\rho(\Psi_0) < 1$, являющегося необходимым и достаточным для сходимости процесса (3.34). Однако при этом, во-первых, требуется знать оценку сверху спектра обращаемой матрицы $\mathbf{A}$ либо матрицы $\mathbf{A}\mathbf{A}^T$ (а именно, если $\mathbf{A}$ – симметричная положительно определенная и $\rho(\mathbf{A}) \le \beta$, то можно взять $\mathbf{U}_0 = \alpha\mathbf{E}$, где $\alpha \in \left(0, \frac{2}{\beta}\right)$; если же $\mathbf{A}$ – произвольная невырожденная матрица и $\rho\left(\mathbf{A}\mathbf{A}^T\right) \le \beta$, то полагают $\mathbf{U}_0 = \alpha\mathbf{A}^T$, где также $\alpha \in \left(0; \frac{2}{\beta}\right)$; можно, конечно, упростить ситуацию и, воспользовавшись тем, что $\rho\left(\mathbf{A}\mathbf{A}^T\right) \le \left\|\mathbf{A}\mathbf{A}^T\right\|$, положить $\mathbf{U}_0 = \frac{\mathbf{A}^T}{\left\|\mathbf{A}\mathbf{A}^T\right\|}$). Во-вторых, при таком задании начальной матрицы нет гарантии, что $\|\Psi_0\|$ будет малой (возможно, даже окажется $\|\Psi_0\| > 1$), и высокий порядок скорости сходимости обнаружится далеко не сразу.

Все сказанное выше не означает, что подобные методы обращения матриц (и операторов) не имеют своей сферы применения. В частности, ниже (в п.6.2) будет рассматриваться способ решения систем нелинейных уравнений, базирующийся на методе Ньютона с приближенным обращением матриц Якоби по методу Шульца.

---

*) Через $\rho(\cdot)$ здесь обозначается спектральный радиус указанной в скобках матрицы.

3.7. О РОЛИ ОШИБОК ОКРУГЛЕНИЯ В ИТЕРАЦИОННЫХ МЕТОДАХ

Обратимся, наконец, к вопросам практической реализации итерационных методов решения линейных алгебраических задач.

Многие утверждения о сходимости итерационных процессов говорят о том, что решение поставленной задачи при определенных условиях может быть найдено этим процессом сколь угодно точно, причем погрешность каждого приближения может быть эффективно проконтролирована (см. теоремы 3.2, 3.6, 3.11, а также теорему 3.3 с замечанием 3.5 и теорему 3.7 с замечанием 3.7). Нетрудно понять, что все это справедливо на самом деле лишь до тех пор, пока на погрешность метода (остаточную погрешность) не наложится вычислительная погрешность (погрешность округлений), неизбежная при любых реальных компьютерных расчетах. Особенно существенное и даже пагубное влияние на результат решения задачи итерационным методом могут оказать ошибки округления в тех случаях, когда утверждения о сходимости метода не содержат эффективных оценок погрешности (теоремы 3.1, 3.4, 3.5, 3.8, 3.10).

Рассмотрим различие между реальным и идеальным итерационными процессами на простейшем объекте – на методе простой итерации.

Пусть на k-м итерационном шаге вычислений по методу (3.3) ошибки округлений составляют вектор $\boldsymbol{\gamma}^{(k)}$. Тогда в отличие от идеального МПИ (3.3), генерирующего последовательность приближений $\mathbf{x}^{(k)}$ к решению $\mathbf{x}^*$ системы (3.1) такому, что

$$\mathbf{x}^* = \mathbf{B}\mathbf{x}^* + \mathbf{c}, \tag{3.41}$$

реальный МПИ будет иметь вид

$$\tilde{\mathbf{x}}^{(k+1)} = \mathbf{B}\tilde{\mathbf{x}}^{(k)} + \mathbf{c} + \boldsymbol{\gamma}^{(k)}. \tag{3.42}$$

Изучим поведение векторов

$$\boldsymbol{\mu}_k := \tilde{\mathbf{x}}^{(k)} - \mathbf{x}^*$$

– ошибок приближений $\tilde{\mathbf{x}}^{(k)}$, получаемых реальным МПИ (3.42).

Вычитая (3.41) из (3.42), имеем

$$\tilde{\mathbf{x}}^{(k+1)} - \mathbf{x}^* = \mathbf{B}\left(\tilde{\mathbf{x}}^{(k)} - \mathbf{x}^*\right) + \boldsymbol{\gamma}^{(k)},$$

т.е.

$$\begin{aligned}\boldsymbol{\mu}_{k+1} &= \mathbf{B}\boldsymbol{\mu}_k + \boldsymbol{\gamma}^{(k)} = \mathbf{B}\left(\mathbf{B}\boldsymbol{\mu}_{k-1} + \boldsymbol{\gamma}^{(k-1)}\right) + \boldsymbol{\gamma}^{(k)} = \\ &= \mathbf{B}^2\left(\mathbf{B}\boldsymbol{\mu}_{k-2} + \boldsymbol{\gamma}^{(k-2)}\right) + \mathbf{B}\boldsymbol{\gamma}^{(k-1)} + \boldsymbol{\gamma}^{(k)} = \cdots = \\ &= \mathbf{B}^{k+1}\boldsymbol{\mu}_0 + \left(\mathbf{B}^k\boldsymbol{\gamma}^{(0)} + \mathbf{B}^{k-1}\boldsymbol{\gamma}^{(1)} + \cdots + \mathbf{B}\boldsymbol{\gamma}^{(k-1)} + \boldsymbol{\gamma}^{(k)}\right).\end{aligned} \tag{3.43}$$

Первое слагаемое в последнем выражении отвечает за погрешность идеального МПИ и может быть сделано сколь угодно малым в процессе итерирования при условии $\rho(\mathbf{B}) < 1$ (см. лемму 3.1). Чтобы оценить второе слагаемое, предположим, что порог абсолютных погрешностей округлений, допускаемых на каждой итерации, есть γ, т.е.

$$\left\|\boldsymbol{\gamma}^{(k)}\right\| \le \gamma \qquad \forall k \in N_0.$$

Тогда

$$\left\|\mathbf{B}^k\boldsymbol{\gamma}^{(0)} + \mathbf{B}^{k-1}\boldsymbol{\gamma}^{(1)} + \cdots + \mathbf{B}\boldsymbol{\gamma}^{(k-1)} + \boldsymbol{\gamma}^{(k)}\right\| \le$$
$$\le \gamma\left\|\mathbf{E} + \mathbf{B} + \cdots + \mathbf{B}^k\right\|,$$

и, если $\|\mathbf{B}\| \le q < 1$, то второе слагаемое в (3.43), хотя и не стремится к нулю, но ограничено по норме величиной

$$\gamma\frac{1-q^k}{1-q} < \frac{\gamma}{1-q}.$$

При условии же $\rho(\mathbf{B}) < 1$, теоретически обеспечивающем сходимость идеального МПИ (3.3), малость этого второго слагаемого отнюдь не гарантируется, что означает допустимость ситуаций, когда в ходе реальных итераций погрешность округлений будет накапливаться вплоть до переполнения множества чисел, представляемых используемой ЭВМ.

Более детальный анализ влияния ошибок округления на итерационный процесс с попыткой пролить свет на природу этого влияния можно найти, например, в [6]. Здесь же ограничимся напоминанием о том, что необходимо с осторожностью применять процессы, когда для них нет эффективных оценок погрешности, и по возможности, учитывать влияние ошибок округления, если такие оценки есть. Например, применительно к МПИ решения СЛАУ выше фактически доказана

Теорема 3.12.[*] *Пусть* $\|\mathbf{B}\| \le q < 1$ *и приближения* $\tilde{\mathbf{x}}^{(k)}$ *к решению* $\mathbf{x}^*$ *системы* (3.2) *получаются посредством равенства* (3.42), *где* $\boldsymbol{\gamma}^{(k)}$ *– вектор ошибок округлений таких, что* $\left\|\boldsymbol{\gamma}^{(k)}\right\| \le \gamma$. *Тогда погрешность* k*-го приближения при любом* $k \in N$ *можно оценить неравенством*

$$\left\|\mathbf{x}^* - \tilde{\mathbf{x}}^{(k)}\right\| \le \frac{q}{1-q}\left\|\tilde{\mathbf{x}}^{(k)} - \tilde{\mathbf{x}}^{(k-1)}\right\| + \frac{\gamma}{1-q}. \qquad (3.44)$$

[*] См. также [1].

Действительно, для последовательности $\left(\mathbf{x}^{(k)}\right)$, получаемой МПИ (3.3), справедливо равенство

$$\mathbf{x}^* - \mathbf{x}^{(k+1)} = \mathbf{B}^{k+1}\left(\mathbf{x}^* - \mathbf{x}^{(0)}\right).$$

Следовательно, считая, что процессы (3.3) и (3.42) начинаются с одного начального приближения $\mathbf{x}^{(0)} = \tilde{\mathbf{x}}^{(0)}$, в идентичном (3.43) равенстве

$$\mathbf{x}^* - \tilde{\mathbf{x}}^{(k+1)} = \mathbf{B}^{k+1}\left(\mathbf{x}^* - \tilde{\mathbf{x}}^{(0)}\right) - \left(\mathbf{B}^k\boldsymbol{\gamma}^{(0)} + \mathbf{B}^{k-1}\boldsymbol{\gamma}^{(1)} + \cdots + \mathbf{B}\boldsymbol{\gamma}^{(k-1)} + \boldsymbol{\gamma}^{(k)}\right)$$

можно заменить $\mathbf{B}^{k+1}\left(\mathbf{x}^* - \tilde{\mathbf{x}}^{(0)}\right)$ на $\mathbf{x}^* - \mathbf{x}^{(k+1)}$. Таким образом, погрешности $(k+1)$-х приближений реального (3.42) и идеального (3.3) методов различаются лишь слагаемым, оцененным выше по норме величиной $\frac{\gamma}{1-q}$, т.е. и для процесса (3.42) можно воспользоваться оценкой, выведенной в теореме 3.2.

Отметим, что как непосредственно видно из оценки (3.44) (при q, приближающихся к единице), роль ошибок округлений в образовании общей погрешности тем сильнее, чем медленнее сходимость итерационного процесса.

УПРАЖНЕНИЯ

3.1. Записать итерационный процесс Якоби нахождения решения системы

$$\begin{cases} 5x_1 + 2x_2 - x_3 + x_4 = 9, \\ x_1 - 4x_2 + 2x_4 = 10, \\ 2x_1 + 3x_2 - 9x_3 - x_4 = -10, \\ 3x_1 + x_3 - 6x_4 = -5. \end{cases}$$

Каким должен быть критерий окончания процесса итерирования, чтобы максимальная из абсолютных погрешностей компонент приближенного решения не превышала заданного малого $\varepsilon > 0$?

3.2. Сделать по пять итераций методов Якоби и Зейделя для системы

$$\begin{cases} 10x_1 + x_2 - 2x_3 = 10, \\ x_1 - 5x_2 + x_3 = 10, \\ 3x_1 - x_2 + 10x_3 = -5. \end{cases}$$

Сколько верных знаков можно гарантировать в приближенных решениях, полученных тем и другим способами?

3.3. Доказать, что при любом начальном векторе $\left(x^{(0)}, y^{(0)}, z^{(0)}\right)^T$ последовательности векторов $\left(x_1^{(k)}, y_1^{(k)}, z_1^{(k)}\right)^T$ и $\left(x_2^{(k)}, y_2^{(k)}, z_2^{(k)}\right)^T$, определяемые при $k = 0, 1, 2, \ldots$ равенствами

$$\begin{cases} x_1^{(k+1)} = 0.1x_1^{(k)} + 0.2y_1^{(k)} - 3, \\ y_1^{(k+1)} = 0.2x_1^{(k)} - 0.1y_1^{(k)} + 0.1z_1^{(k)} + 2, \\ z_1^{(k+1)} = -0.3x_1^{(k)} + 0.2z_1^{(k)} - 1 \end{cases}$$

и

$$\begin{cases} x_2^{(k+1)} = \left(2y_2^{(k)} - 30\right)/9, \\ y_2^{(k+1)} = \left(2x_2^{(k)} + z_2^{(k)} + 20\right)/11, \\ z_2^{(k+1)} = -\left(3x_2^{(k)} + 10\right)/8, \end{cases}$$

сходятся, причем к одному и тому же предельному вектору $\left(x^*, y^*, z^*\right)^T$. Записать линейную систему, решением которой служит этот предельный вектор. За сколько шагов итераций по данным формулам можно получить предельный вектор с точностью $\varepsilon = 10^{-6}$ (по норме-максимум), если начать счет с нулевого вектора ?

3.4. Пусть методом Якоби решение системы

$$b_i x_{i-1} + c_i x_i + d_i x_{i+1} = r_i$$
$$\left(i = 1, 2, \ldots, n; \qquad b_1 = d_n = 0\right)$$

с нужной точностью достигается за k шагов. Существуют ли такие k и n, при которых метод Якоби эффективнее метода прогонки по числу арифметических операций ?

3.5. Для линейной системы

$$\begin{cases} x_1 + 2x_2 + 3x_3 = 5, \\ 2x_1 - x_2 + 2x_4 = 8, \\ 3x_1 + x_2 - 2x_3 - x_4 = -1, \\ 4x_1 - x_3 + 2x_4 = 8 \end{cases}$$

записать метод Зейделя и обосновать его сходимость. Каковы расчетные формулы метода ПВР в этом случае ?

3.6. Дана система

$$\begin{cases} 7x_1 + 5x_2 + x_3 = 2.2, \\ 5x_1 + 8x_2 + 2x_3 = 2.4, \\ x_1 + 2x_2 + 4x_3 = 1.6. \end{cases}$$

Найти четвертое приближение к ее решению по методу минимальных невязок, начиная итерационный процесс с нулевого вектора. За сколько итераций по методу Якоби достигается примерно такая же величина евклидовой нормы невязки? Сравнить вычислительные затраты, требующиеся для реализации одного шага каждого из этих методов.

3.7. Предположим, что некоторая $n \times n$-система вида $\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{c}$ с $\|\mathbf{B}\| \approx 0.5$ решается методом простых итераций с уровнем абсолютных погрешностей арифметических операций порядка 10^{-6}. Допустим, что при этом $\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\| \approx 1$. Каким числом следует ограничить количество итераций, чтобы вычислительная погрешность не стала существенно превышать погрешность метода?

3.8. Даны матрицы

$$\mathbf{A} = \begin{pmatrix} 1 & -2 & 3 \\ -1 & 1 & 2 \\ 2 & -1 & -1 \end{pmatrix} \quad \text{и} \quad \mathbf{U}_0 = \begin{pmatrix} -0.1 & 0.6 & 0.9 \\ -0.4 & 0.9 & 0.6 \\ 0.1 & 0.4 & 0.1 \end{pmatrix}.$$

а) Подсчитав невязку $\Psi_0 = \mathbf{E} - \mathbf{A}\mathbf{U}_0$, убедиться в существовании матрицы $\mathbf{A}^{-1}$ и оценить какую-либо ее норму.

б) Сделать по два приближения к $\mathbf{A}^{-1}$ методом Шульца второго и третьего порядков и оценить близость полученных приближений к $\mathbf{A}^{-1}$.

в) Сравнить оценки погрешностей приближений с истинными ошибками, найдя $\mathbf{A}^{-1}$ каким-нибудь прямым методом.

ГЛАВА 4 МЕТОДЫ РЕШЕНИЯ АЛГЕБРАИЧЕСКИХ ПРОБЛЕМ СОБСТВЕННЫХ ЗНАЧЕНИЙ

Затрагивается наиболее сложная задача вычислительной линейной алгебры – нахождение собственных чисел и собственных векторов матриц. Рассматриваются современные подходы к решению спектральных задач для вещественных матриц умеренной размерности, базирующиеся на прямой и обратной итерациях (в том числе, со сдвигами), а также на приведении матриц к диагональной или треугольной формам ортогональными преобразованиями подобия. Показываются идеи методов, выводятся расчетные формулы, даются конкретные алгоритмы, позволяющие в оговоренных ситуациях решать частичные и полные проблемы собственных значений до конца.

4.1. СОБСТВЕННЫЕ ПАРЫ МАТРИЦЫ И ИХ ПРОСТЕЙШИЕ СВОЙСТВА

Пусть $\mathbf{A}$ – вещественная $n \times n$-матрица, $\mathbf{y} = \mathbf{y}(t)$ – n-мерная векторная функция скалярного аргумента t, и пусть ищутся нетривиальные решения системы дифференциальных уравнений

$$\frac{d\mathbf{y}}{dt} = \mathbf{A}\mathbf{y} \qquad (4.1)$$

в виде $\mathbf{y} = e^{\lambda t}\mathbf{x}$, где $\mathbf{x} \in C_n$, $\lambda \in C$. Подставляя $\mathbf{y}$ и $\frac{d\mathbf{y}}{dt}$ в (4.1), получаем

$$\lambda e^{\lambda t}\mathbf{x} = \mathbf{A}e^{\lambda t}\mathbf{x},$$

т.е. система (4.1) действительно будет иметь решения заданного вида в том и только том случае, если найдутся такие пары чисел λ и ненулевых векторов $\mathbf{x}$, что

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}. \qquad (4.2)$$

Имеется ряд других примеров из областей, лежащих за пределами линейной алгебры, в которых также приходят к необходимости решать подобные (4.2) алгебраические задачи, называемые ***задачами на собственные значения*** (см., например, [30]). При этом различают ***полную*** (***алгебраическую*** или, иначе, ***матричную***) ***проблему собственных значений***, предполагающую нахождение всех ***собственных пар*** $\{\lambda, \mathbf{x}\}$ матрицы $\mathbf{A}$, и ***частичные проблемы собственных значений***, состоящие, как правило, в нахождении одного или нескольких ***собственных чисел*** λ и, возможно,

соответствующих им ***собственных векторов***[*] $\mathbf{x}$. Чаще всего, в последнем случае речь идет о нахождении наибольшего и наименьшего по модулю собственных чисел; знание таких характеристик матрицы позволяет, например, делать заключения о сходимости тех или иных итерационных методов, оптимизировать параметры итерационных методов, учитывать влияние на результаты решения алгебраических задач погрешностей исходных данных и вычислительных погрешностей (потребность в таких числах неоднократно возникала в гл.3). Имеются и несколько иные постановки частичных проблем [5, 6, 17, 30].

Трактуя $\mathbf{A}$ в равенстве (4.2) как матрицу линейного преобразования в R_n, задачу на собственные значения можно сформулировать так: для каких ненулевых векторов $\mathbf{x}$ и чисел λ линейное преобразование вектора с помощью матрицы $\mathbf{A}$ не изменяет направления этого вектора в R_n, т.е. сводится к "растяжению" этого вектора в λ раз? Эта задача, очевидно, эквивалентна задаче исследования однородной СЛАУ[**] с параметром: при каких λ система

$$(\mathbf{A} - \lambda \mathbf{E})\mathbf{x} = 0 \qquad (4.3)$$

имеет нетривиальные решения? Найти эти решения.

Теоретически эта задача легко решаема: нужно найти корни так называемого ***характеристического*** или иначе ***"векового"*** уравнения

$$\det(\mathbf{A} - \lambda \mathbf{E}) = 0 \qquad (4.4)$$

и, подставляя их поочередно в (4.3), получать из соответствующих переопределенных систем собственные векторы. Практическая реализация этого в сущности простого подхода сопряжена с рядом трудностей, возрастающих с ростом размерности решаемой задачи. Трудности эти обусловлены развертыванием ***"векового" определителя*** $\det(\mathbf{A} - \lambda \mathbf{E})$ и вычислением корней получающегося при этом многочлена n-й степени, а также поиском линейно независимых решений вырожденных СЛАУ. В связи с этим такой непосредственный подход к решению алгебраической проблемы собственных значений обычно применяют лишь при очень малых размерностях матриц $\mathbf{A}$ $(n=2,3)$; уже при $n \geq 4$ на первый план выходят специальные численные методы решения таких задач. Ниже будут рассмотрены некоторые из этих методов так, чтобы можно было понять идеи, лежащие в их основе, и в то же время получить возможность решать поставленные задачи до конца для некоторых классов матриц (более полное и глубокое изложение этой темы см. в монографиях [26, 44, 53, 54] и в учебных пособиях [1, 5, 6, 8, 15, 21, 28, 34]).

---

[*] Собственное число и собственное значение в данном контексте – синонимы. В более общем случае, когда $\mathbf{A}$ в (4.2) – некоторый оператор, ***собственные элементы*** $\mathbf{x}$ могут иметь другую природу.

[**] Здесь, как и ранее, аббревиатура СЛАУ означает "система линейных алгебраических уравнений", а обозначение $\mathbf{E}$ зарезервировано за единичной матрицей.

Следует заметить, что если в недалеком прошлом численные методы решения задач на собственные значения опирались, как правило, на классический подход, т.е. на развертывание вековых определителей, в частности, в простейшем случае с помощью приведения матрицы **A** подходящим преобразованием к так называемой ***сопровождающей матрице*** [26, 53]

$$\mathbf{C} = \begin{pmatrix} c_1 & c_2 & c_3 & \dots & c_{n-1} & c_n \\ 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 1 & \dots & 1 & 0 \end{pmatrix},$$

где в первой строке стоят коэффициенты уравнения (4.4), записанного в виде

$$(-1)^n\left(\lambda^n - c_1\lambda^{n-1} - c_2\lambda^{n-2} - \dots - c_{n-1}\lambda - c_n\right) = 0,$$

то современные методы решения полной проблемы ориентированы на алгоритмическое построение из матрицы **A** такой матрицы, определенные элементы которой являлись бы приближенными значениями собственных чисел **A**, причем параллельно формировались бы и ее собственные векторы.

Прежде чем приступить к изучению методов нахождения собственных чисел и векторов, вспомним некоторые простые их **свойства**, требующиеся в дальнейшем.

Свойство 1. *Если* $\{\lambda, \mathbf{x}\}$ *– собственная пара матрицы* **A**, *а* $\alpha\,(\neq 0)$ *– некоторое число, то* $\{\lambda, \alpha\mathbf{x}\}$ *также является собственной парой для* **A**.

Действительно, умножив верное для данных λ и $\mathbf{x}$ равенство (4.2) на число α, получаем верное равенство

$$\mathbf{A}(\alpha\mathbf{x}) = \lambda(\alpha\mathbf{x}).$$

Оно означает, что каждому собственному числу λ соответствует бесчисленное множество собственных векторов, различающихся лишь скалярным множителем. Такие векторы задают одно направление в n-мерном пространстве; в соответствие этому направлению можно поставить нормированный вектор или орт. (Вообще говоря, одному собственному числу может соответствовать и несколько линейно независимых собственных векторов.)

Свойство 2. *Пусть* $\{\mu, \mathbf{x}\}$ *– собственная пара матрицы* $\mathbf{A} - p\mathbf{E}$ *при некотором* $p \in R$. *Тогда* $\{\lambda := \mu + p, \mathbf{x}\}$ *собственная пара матрицы* **A**.

Чтобы убедиться в этом, заметим, что по условию

$$(\mathbf{A} - p\mathbf{E})\mathbf{x} = \mu\,\mathbf{x} \qquad (4.5)$$

при данных μ и $\mathbf{x}$ – верное равенство. Рассмотрим равенство $\mathbf{A}\mathbf{x} = \lambda\,\mathbf{x}$ при $\lambda = \mu + p$:

$$\mathbf{A}\mathbf{x} = (\mu + p)\mathbf{x}.$$

Оно равносильно (4.5), и значит, справедливо, с другой стороны, говорит о том, что $\{\lambda, \mathbf{x}\}$ – собственная пара $\mathbf{A}$.

Как видим, *прибавление к данной матрице* $\mathbf{A}$ *скалярной матрицы* $p\mathbf{E}$ *не изменяет ее собственных векторов и смещает спектр*[*] *исходной матрицы на число* p (влево при $p>0$).

Свойство 3. *Если* $\{\lambda, \mathbf{x}\}$ *– собственная пара обратимой матрицы* $\mathbf{A}$, *то* $\left\{\frac{1}{\lambda}, \mathbf{x}\right\}$ *– собственная пара матрицы* $\mathbf{A}^{-1}$.

Справедливость этого свойства очевидна: умножив верное для данных λ и $\mathbf{x}$ равенство $\mathbf{A}\mathbf{x} = \lambda\,\mathbf{x}$ слева на матрицу $\frac{1}{\lambda}\mathbf{A}^{-1}$, получаем

$$\frac{1}{\lambda}\mathbf{x} = \mathbf{A}^{-1}\mathbf{x},$$

что и означает утверждаемое.

Свойство 4. *Собственными числами диагональных и треугольных матриц являются их диагональные элементы.*

Этот факт легко усматривается из очевидного представления характеристических уравнений (4.4) для таких матриц в виде

$$\prod_{i=1}^{n}(\lambda - a_{ii}) = 0.$$

Последнее равенство свидетельствует, что *диагональные* и *треугольные вещественные матрицы имеют только вещественные собственные значения* (ровно n с учетом возможной их кратности). *Вещественность собственных чисел присуща и очень важному в приложениях классу симметричных матриц* [1, 16].

---

[*] Напомним, что ***спектром матрицы*** называется множество всех ее собственных значений.

Определение 4.1. ***Отношением Рэлея****[1] для $n \times n$-матрицы* $\mathbf{A}$ *называется функционал* $\rho(\mathbf{x}) = \dfrac{(\mathbf{Ax}, \mathbf{x})}{(\mathbf{x}, \mathbf{x})}$, *определенный на множестве ненулевых n-мерных векторов* $\mathbf{x}$.

Свойство 5. *Пусть* $\mathbf{x}^*$ *– собственный вектор матрицы* $\mathbf{A}$, *тогда* $\rho(\mathbf{x}^*)$ *– ее собственное число.*

Для доказательства этого утверждения обозначим через λ^* собственное число матрицы $\mathbf{A}$, соответствующее вектору $\mathbf{x}^*$. Подставляя $\mathbf{Ax}^* = \lambda^* \mathbf{x}^*$ в вытекающее из определения 4.1 равенство

$$(\mathbf{Ax}^*, \mathbf{x}^*) = \rho(\mathbf{x}^*)(\mathbf{x}^*, \mathbf{x}^*),$$

имеем

$$\lambda^*(\mathbf{x}^*, \mathbf{x}^*) = \rho(\mathbf{x}^*)(\mathbf{x}^*, \mathbf{x}^*),$$

откуда после деления на $(\mathbf{x}^*, \mathbf{x}^*) \neq 0$ получаем утверждаемое: $\lambda^* = \rho(\mathbf{x}^*)$.

Отношение Рэлея обладает рядом других ценных свойств. Например, если матрица $\mathbf{A}$ – симметричная положительно определенная со спектром $\lambda_1 > \lambda_2 > \ldots > \lambda_n$, то $\max \rho(\mathbf{x}) = \lambda_1$, $\min \rho(\mathbf{x}) = \lambda_n$, $\rho(\mathbf{x}) \in [\lambda_n, \lambda_1]$ при любых n-мерных $\mathbf{x} \neq \mathbf{0}$ и, кроме того, $\operatorname{grad} \rho(\mathbf{x}) = \mathbf{0}$ тогда и только тогда, когда $\mathbf{x}$ – собственный вектор матрицы $\mathbf{A}$ (см. [1, 44]). Эти свойства служат основой для некоторых способов локализации собственных значений и построения градиентных методов их вычисления.

В дальнейшем (п.4.2, п.4.3) будет полезно следующее экстремальное свойство отношения Рэлея.

Свойство 6. *Минимум евклидовой нормы вектора* $\xi(\lambda) := \mathbf{Ax} - \lambda \mathbf{x}$ *для любого фиксированного ненулевого вектора* $\mathbf{x}$ *достигается при* $\lambda = \rho(\mathbf{x})$.

Смысл этого факта в следующем: *если некоторый вектор* $\mathbf{x}$ *считать приближением к собственному вектору матрицы* $\mathbf{A}$ (а значит, ξ – его невязкой), *то отношение Рэлея* $\rho(\mathbf{x})$ *будет наилучшим приближением*

---

[1] Лорд Рэлей (до получения титула лорда – Стретт Джон Уильям, 1842–1919) – английский физик, один из основоположников теории колебаний.

к соответствующему этому вектору собственному числу в смысле евклидовой метрики.

Доказательство свойства 4.6 для симметричных **A** и вещественных **x** весьма просто. Действительно, рассмотрим квадрат евклидовой нормы невязки

$$\|\xi(\lambda)\|_2^2 = (\mathbf{Ax} - \lambda\mathbf{x}, \mathbf{Ax} - \lambda\mathbf{x}) =$$

$$= (\mathbf{Ax}, \mathbf{Ax}) - 2\lambda(\mathbf{Ax}, \mathbf{x}) + \lambda^2(\mathbf{x}, \mathbf{x}) = q(\lambda)(\mathbf{x}, \mathbf{x}),$$

где $q(\lambda) := \lambda^2 - 2\lambda\rho(\mathbf{x}) + \dfrac{(\mathbf{Ax}, \mathbf{Ax})}{(\mathbf{x}, \mathbf{x})}$. Очевидно, квадратный трехчлен $q(\lambda)$ всегда имеет минимум при $\lambda = \rho(\mathbf{x})$, а поскольку $(\mathbf{x}, \mathbf{x}) > 0$, это значение λ доставляет минимум величине $\|\xi(\lambda)\|_2^2$ и, следовательно, величине $\|\xi(\lambda)\|_2$.

4.2. СТЕПЕННОЙ МЕТОД

Рассмотрим простейший метод решения частичных проблем собственных значений, который вряд ли может быть отнесен к широко применяемым методам решения таких задач, но многозначащий для понимания и построения других, более эффективных методов.

Пусть о вещественной $n \times n$-матрице **A** известно, что она является ***матрицей простой структуры***, т.е. имеет ровно n линейно независимых собственных векторов (базис):

$$\mathbf{x}_1 = \begin{pmatrix} x_{11} \\ x_{12} \\ \dots \\ x_{1n} \end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix} x_{21} \\ x_{22} \\ \dots \\ x_{2n} \end{pmatrix}, \dots, \mathbf{x}_n = \begin{pmatrix} x_{n1} \\ x_{n2} \\ \dots \\ x_{nn} \end{pmatrix}. \qquad (4.6)$$

Пусть нумерация этих векторов отвечает упорядочению соответствующих им собственных чисел по убыванию модулей (где первое из неравенств – строгое):

$$|\lambda_1| > |\lambda_2| \geq \dots \geq |\lambda_n|. \qquad (4.7)$$

Ставим **задачу приближенного вычисления наибольшего по модулю собственного числа** λ_1 (вещественного, в силу предположения о строгом доминировании его модуля) **и соответствующего ему собственного вектора** $\mathbf{x}_1$ данной матрицы **A**.

Возьмем произвольный ненулевой вектор $\mathbf{y}^{(0)}$ и запишем его разложение по базису из собственных векторов $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$:

$$\mathbf{y}^{(0)} = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \dots + c_n\mathbf{x}_n. \qquad (4.8)$$

При этом без ограничения общности можно считать, что $c_1 \neq 0$, так как в противном (маловероятном) случае можно взять другой начальный вектор $\mathbf{y}^{(0)}$.

Выполним первую итерацию вектора $\mathbf{y}^{(0)}$ умножением (4.8) слева на матрицу $\mathbf{A}$:

$$\mathbf{y}^{(1)} = \mathbf{A}\mathbf{y}^{(0)} = c_1\mathbf{A}\mathbf{x}_1 + c_2\mathbf{A}\mathbf{x}_2 + \ldots + c_n\mathbf{A}\mathbf{x}_n.$$

Так как $\{\lambda_i, \mathbf{x}_i\}$ при всех $i \in \{1,2,\ldots,n\}$ по предположению являются собственными парами матрицы $\mathbf{A}$, то, в силу (4.2), последнее можно переписать в виде

$$\mathbf{y}^{(1)} = c_1\lambda_1\mathbf{x}_1 + c_2\lambda_2\mathbf{x}_2 + \ldots + c_n\lambda_n\mathbf{x}_n.$$

Для второй итерации по тому же принципу получаем

$$\begin{aligned}\mathbf{y}^{(2)} &= \mathbf{A}\mathbf{y}^{(1)} = \mathbf{A}^2\mathbf{y}^{(0)} = \\ &= c_1\lambda_1\mathbf{A}\mathbf{x}_1 + c_2\lambda_2\mathbf{A}\mathbf{x}_2 + \ldots + c_n\lambda_n\mathbf{A}\mathbf{x}_n = \\ &= c_1\lambda_1^2\mathbf{x}_1 + c_2\lambda_2^2\mathbf{x}_2 + \ldots + c_n\lambda_n^2\mathbf{x}_n.\end{aligned}$$

Очевидно, k-я итерация вектора $\mathbf{y}^{(0)}$ с помощью матрицы $\mathbf{A}$ дает вектор

$$\mathbf{y}^{(k)} = \mathbf{A}\mathbf{y}^{(k-1)} = \mathbf{A}^k\mathbf{y}^{(0)} = c_1\lambda_1^k\mathbf{x}_1 + c_2\lambda_2^k\mathbf{x}_2 + \ldots + c_n\lambda_n^k\mathbf{x}_n \qquad (4.9)$$

или, с учетом представления $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ в исходном базисе (см. (4.6)),

$$\mathbf{y}^{(k)} = \begin{pmatrix} y_1^{(k)} \\ y_2^{(k)} \\ \ldots \\ y_n^{(k)} \end{pmatrix} = c_1\lambda_1^k\begin{pmatrix} x_{11} \\ x_{12} \\ \ldots \\ x_{1n} \end{pmatrix} + c_2\lambda_2^k\begin{pmatrix} x_{21} \\ x_{22} \\ \ldots \\ x_{2n} \end{pmatrix} + \ldots + c_n\lambda_n^k\begin{pmatrix} x_{n1} \\ x_{n2} \\ \ldots \\ x_{nn} \end{pmatrix}.$$

Беря отношения компонент ***итерированного вектора****) $\mathbf{y}^{(k)}$ к соответствующим компонентам предыдущего вектора $\mathbf{y}^{(k-1)}$, будем иметь:

$$\begin{aligned}\frac{y_i^{(k)}}{y_i^{(k-1)}} &= \frac{c_1\lambda_1^k x_{1i} + c_2\lambda_2^k x_{2i} + \ldots + c_n\lambda_n^k x_{ni}}{c_1\lambda_1^{k-1}x_{1i} + c_2\lambda_2^{k-1}x_{2i} + \ldots + c_n\lambda_n^{k-1}x_{ni}} = \\ &= \lambda_1 \cdot \frac{1 + \frac{c_2}{c_1}\cdot\frac{x_{2i}}{x_{1i}}\left(\frac{\lambda_2}{\lambda_1}\right)^k + \ldots + \frac{c_n}{c_1}\cdot\frac{x_{ni}}{x_{1i}}\left(\frac{\lambda_n}{\lambda_1}\right)^k}{1 + \frac{c_2}{c_1}\cdot\frac{x_{2i}}{x_{1i}}\left(\frac{\lambda_2}{\lambda_1}\right)^{k-1} + \ldots + \frac{c_n}{c_1}\cdot\frac{x_{ni}}{x_{1i}}\left(\frac{\lambda_n}{\lambda_1}\right)^{k-1}}.\end{aligned} \qquad (4.10)$$

*) Термин взят из [31].

Предел дроби в последнем равенстве при сделанных допущениях равен 1 в процессе $k \to \infty$, и значит, $y_i^{(k)} \Big/ y_i^{(k-1)} \xrightarrow[k\to\infty]{} \lambda_1$ для каждого $i \in \{1,2,...,n\}$, при котором $x_{1i} \neq 0$ (заметим, что числа $x_{11}, x_{12}, \ldots, x_{1n}$ не могут быть одновременно нулями, так как $\mathbf{x}_1$ – базисный вектор и поэтому не может быть нулевым).

Представляя вектор $\mathbf{y}^{(k)}$ на основе (4.9) в виде

$$\mathbf{y}^{(k)} = c_1\lambda_1^k\left[\mathbf{x}_1 + \frac{c_2}{c_1}\left(\frac{\lambda_2}{\lambda_1}\right)^k \mathbf{x}_2 + \ldots + \frac{c_n}{c_1}\left(\frac{\lambda_n}{\lambda_1}\right)^k \mathbf{x}_n\right], \qquad (4.11)$$

можно сделать вывод, что при тех же исходных допущениях, в силу $\left|\frac{\lambda_i}{\lambda_1}\right|^k \xrightarrow[(i\neq 1)]{k\to\infty} 0$, в фигурирующей в скобках выражения (4.11) линейной комбинации векторов $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ с ростом k начнет доминировать первое слагаемое. Это означает, что вектор $\mathbf{y}^{(k)}$ от итерации к итерации будет давать все более хорошие приближения к собственному вектору $\mathbf{x}_1$ по направлению, т.е. с точностью до скалярного множителя $c_1\lambda_1^k$ (см. п.4.1, свойство 1).

Таким образом, как показывают приведенные рассуждения, метод нахождения "старшей" собственной пары матрицы простой структуры, называемый ***степенным методом****), в своей основе весьма примитивен и состоит в следующем: берется произвольный вектор $\mathbf{y}^{(0)}$ $(\neq \mathbf{0})$, простыми итерациями $\mathbf{y}^{(k)} = \mathbf{A}\mathbf{y}^{(k-1)}$ строится последовательность векторов $\mathbf{y}^{(k)}$ и параллельно рассматриваются последовательности отношений соответствующих компонент векторов k-й и $(k-1)$-й итераций (отношения с чрезвычайно малыми по модулю знаменателями следует игнорировать). Как только установятся несколько первых цифр во всех этих отношениях (что выясняется проверкой выполнения приближенных равенств $\frac{y_i^{(k)}}{y_i^{(k-1)}} \approx \frac{y_i^{(k-1)}}{y_i^{(k-2)}}$), так можно считать, что найдено наибольшее по модулю собственное число с точностью, определяемой последним установившимся в отношениях знаком, и соответствующий ему собственный вектор, за который принимается последний итерированный вектор $\mathbf{y}^{(k)}$.

Для практической реализации такая схема нахождения старшей собственной пары мало пригодна по многим причинам и требует определен-

---

*) Этимология данного термина совершенно ясна. Можно встретить и другие названия: ***счет на установление*** [28], ***итерационный метод фон Мизеса*** [32]. Иногда применяют латинскую аббревиатуру PM (от англ. Power method) [44].

ной доводки. Рассмотрим некоторые из этих причин и соответственно пути модификации вышеописанного простейшего алгоритма.

Анализируя выражение $\mathbf{y}^{(k)}$ в форме (4.11), видим, что при достаточно большом числе итераций k за счет множителя λ_1^k в процессе счета может произойти либо превышение допустимых для ЭВМ чисел, если $|\lambda_1|>1$, либо пропадание значащих цифр итерированных векторов, если $|\lambda_1|<1$. Устранить это явление можно достаточно легко, введя в итерационный процесс нормировку итерированных векторов (т.е. приведение к единичной длине по той или иной метрике) на каждой итерации или через некоторое фиксированное число итерационных шагов.

Так, пошаговая нормировка векторов порождает следующий

РМ-алгоритм.

Шаг 1. Ввести $n\times n$-матрицу $\mathbf{A}$, задать n-мерный вектор $\mathbf{y}^{(0)}$, вычислить $\left\|\mathbf{y}^{(0)}\right\|$ и вектор $\mathbf{x}^{(0)}:=\mathbf{y}^{(0)}\Big/\left\|\mathbf{y}^{(0)}\right\|$; положить k=1.

Шаг 2. Вычислить вектор $\mathbf{y}^{(k)}=\mathbf{A}\mathbf{x}^{(k-1)}$.

Шаг 3. Вычислить $\left\|\mathbf{y}^{(k)}\right\|$ и $\mathbf{x}^{(k)}:=\mathbf{y}^{(k)}\Big/\left\|\mathbf{y}^{(k)}\right\|$.

Шаг 4. Вычислить отношения $\lambda_i^{(k)}=y_i^{(k)}\Big/x_i^{(k-1)}$ (координат векторов $\mathbf{y}^{(k)}$ и $\mathbf{x}^{(k-1)}$) при i=1,2,...,n таких, что $\left|x_i^{(k-1)}\right|>\delta$, где $\delta>0$ – некоторое задаваемое малое число (допуск).

Шаг 5. Подвергнуть числа $\lambda_i^{(k)}$ тесту на сходимость.

Если обнаруживается совпадение требуемого числа знаков в $\lambda_i^{(k)}$ и $\lambda_i^{(k-1)}$ ($\lambda^{(0)}$ можно задавать произвольно), то работу алгоритма прекратить и за старшее собственное число λ_1 принять усредненное (по i) значение $\lambda_i^{(k)}$, а за нормированный старший собственный вектор $\mathbf{x}_1$ – вектор $\mathbf{x}^{(k)}$.

В противном случае – вернуться к шагу 2.

Слабым местом данного алгоритма, очевидно, является последний шаг, т.е. решение проблемы своевременного останова работы алгоритма. Этот шаг описан из рациональных соображений и не может гарантировать во всех случаях (даже при сделанных допущениях) получения собственной пары $\{\lambda_1,\mathbf{x}_1\}$ с наперед заданной точностью, поскольку при разработке метода не было получено никаких оценок погрешности.

Относительно характера сходимости степенного метода можно утверждать (см. формулы (4.10) и (4.11)), что в указанных условиях итераци-

онный процесс является линейным*), т.е. сходится со скоростью геометрической прогрессии, знаменатель которой определяется в основном величиной отношения $\left|\frac{\lambda_2}{\lambda_1}\right|$. Это означает, что сходимость будет тем лучше и, как следствие, критерий останова в шаге 5 тем надежнее, чем сильнее доминирует в спектре матрицы $\mathbf{A}$ собственное число λ_1. Подмеченный факт вкупе со свойством 2 п.4.1 позволяет существенно ускорить нахождение наибольшего по модулю собственного числа матрицы $\mathbf{A}$ путем удачного смещения ее спектра, чему могут способствовать какие-либо априорные сведения об исходной задаче**).

То же свойство 2 собственных пар позволяет применять непосредственно степенной метод **для нахождения наименьшего по модулю собственного числа** λ_n знакоопределенной матрицы $\mathbf{A}$ в случае, когда наибольшее λ_1 уже найдено. Для этого достаточно найти наибольшее по модулю собственное число Λ матрицы $\mathbf{A}-\lambda_1\mathbf{E}$; соответствующий ему собственный вектор этой матрицы и число $\lambda_n=\Lambda+\lambda_1$ будут образовывать искомую собственную пару.

Действительно, вычитая из верного для собственной пары $\{\lambda_n,\mathbf{x}_n\}$ равенства $\mathbf{A}\mathbf{x}_n=\lambda_n\mathbf{x}_n$ тождество $\lambda_1\mathbf{x}_n=\lambda_1\mathbf{x}_n$, получаем верное равенство

$$(\mathbf{A}-\lambda_1\mathbf{E})\mathbf{x}_n=(\lambda_n-\lambda_1)\mathbf{x}_n,$$

означающее, что $\Lambda:=\lambda_n-\lambda_1$ и $\mathbf{x}_n$ служат собственной парой матрицы $\mathbf{A}-\lambda_1\mathbf{E}$. Так как для знакоопределенной матрицы справедливо неравенство $|\lambda_n-\lambda_1|\geq|\lambda_i-\lambda_1|$ при любом $i\in\{1,2,\ldots,n\}$, то Λ – наибольшее по модулю собственное число матрицы $\mathbf{A}-\lambda_1\mathbf{E}$ и может быть найдено степенным методом.

Знание старшего собственного числа λ_1 матрицы $\mathbf{A}$ простой структуры, получаемого в процессе прямых итераций по формулам (4.9), (4.10), в предположении, что

$$|\lambda_1|>|\lambda_2|>|\lambda_3|\geq\ldots\geq|\lambda_n|,$$

---

*) См. определение 5.1 в п.5.3.

) См. по этому поводу [53, 54]. В [53] приведен **пример, наглядно показывающий эффективность подходящего сдвига : если некая матрица $\mathbf{A}$ шестого порядка имеет собственные числа $\lambda_i=21-i$, то непосредственное применение степенного метода к вычислению λ_1 порождает итерационный процесс, сходящийся со скоростью порядка $\left(\frac{19}{20}\right)^k$, в то время как сдвиг на величину $p=17$, оставляющий соответствующее λ_1 число $\mu_1:=\lambda_1-p=3$ старшим в спектре матрицы $\mathbf{A}-p\mathbf{E}$, позволяет найти его степенным методом, сходящимся уже со скоростью порядка $\left(\frac{2}{3}\right)^k$.

позволяет без больших дополнительных затрат найти **приближенное значение второго по модулю собственного числа** λ_2. Это можно сделать по формуле

$$\lambda_2 \approx \frac{y_i^{(k+1)} - \lambda_1 y_i^{(k)}}{y_i^{(k)} - \lambda_1 y_i^{(k-1)}}, \qquad (4.12)$$

вычисляя фигурирующие в правой части отношения для достаточно больших k и для всех i=1,2,....,n, при которых абсолютная величина знаменателя не меньше некоторого порогового значения, и затем усредняя результат. Понятно, что при этом неизбежна потеря точности.

Для обоснования*) приближенного равенства (4.12) подставим в его правую часть выражения компонент $(k+1)$-го, k-го и $(k-1)$-го итерированных векторов в соответствии с представлением (4.9) в исходном базисе. После взаимного уничтожения по паре первых членов в числителе и в знаменателе будем иметь:

$$\frac{y_i^{(k+1)} - \lambda_1 y_i^{(k)}}{y_i^{(k)} - \lambda_1 y_i^{(k-1)}} = \frac{c_2\lambda_2^{k+1}x_{2i} - c_2\lambda_1\lambda_2^k x_{2i} + ... + c_n\lambda_n^{k+1}x_{ni} - c_n\lambda_1\lambda_n^k x_{ni}}{c_2\lambda_2^k x_{2i} - c_2\lambda_1\lambda_2^{k-1}x_{2i} + ... + c_n\lambda_n^k x_{ni} - c_n\lambda_1\lambda_n^{k-1}x_{ni}} =$$

$$= \frac{c_2\lambda_2^{k+1}x_{2i}\left(1 - \frac{\lambda_1}{\lambda_2} + \sum_{j=3}^{n}\frac{\lambda_j^{k+1} - \lambda_1\lambda_j^k}{\lambda_2^{k+1}}\cdot\frac{c_j}{c_2}\cdot\frac{x_{ji}}{x_{2i}}\right)}{c_2\lambda_2^k x_{2i}\left(1 - \frac{\lambda_1}{\lambda_2} + \sum_{j=3}^{n}\frac{\lambda_j^k - \lambda_1\lambda_j^{k-1}}{\lambda_2^k}\cdot\frac{c_j}{c_2}\cdot\frac{x_{ji}}{x_{2i}}\right)} \xrightarrow[k\to\infty]{} \lambda_2,$$

так как $\frac{\lambda_j^{k+1} - \lambda_1\lambda_j^k}{\lambda_2^{k+1}} = \left(\frac{\lambda_j}{\lambda_2}\right)^{k+1} - \frac{\lambda_1}{\lambda_2}\cdot\left(\frac{\lambda_j}{\lambda_2}\right)^k \xrightarrow[k\to\infty]{} 0$ при всех j=3,...,n.

Вернемся к вопросу о недостатках степенного метода нахождения наибольшего по модулю собственного числа и путях их устранения. При этом **далее ограничимся рассмотрением класса симметричных положительно определенных матриц**. Известно, что такие матрицы имеют положительный вещественный спектр $\lambda_1, \lambda_2, ..., \lambda_n$, ортонормированный базис из собственных векторов $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ и, естественно, являются матрицами простой структуры.

Обсудим шаг 4 предложенного выше РМ-алгоритма.

Вычисление на каждом итерационном шаге отношений **всех** пар соответствующих компонент векторов $\mathbf{x}$ и $\mathbf{y}=\mathbf{Ax}$, да еще с определенными проверками, при больших размерностях n требует значительных вычисли-

*) Другое обоснование см., например, в [21]. Там же показано, что за соответствующий λ_2 собственный вектор $\mathbf{x}_2$ можно принять нормированный вектор $\mathbf{y}^{(k+1)} - \lambda_1\mathbf{y}^{(k)}$.

тельных затрат, хотя и дает о старшем собственном числе λ_1 дополнительную информацию: как утверждается в [32], значение λ_1 заключено между наименьшим и наибольшим из этих отношений, т.е. имеются двусторонние оценки λ_1 на каждой итерации.

Чтобы упростить соответствующую шагу 4 РМ-алгоритма процедуру, проведем следующие рассуждения.

Пусть R_n – евклидово пространство, $\mathbf{A}$ – симметричная положительно определенная матрица и последовательность итерированных векторов $\mathbf{y}^{(k)}$ строится, как и ранее, по формулам (4.9).

Рассмотрим скалярные произведения $\left(\mathbf{y}^{(k)},\mathbf{y}^{(k)}\right)$ и $\left(\mathbf{y}^{(k)},\mathbf{y}^{(k-1)}\right)$. Выполняя умножение правых частей (4.9) по правилам умножения многочленов и учитывая ортонормированность собственных векторов, т.е. условие $\left(\mathbf{x}_i,\mathbf{x}_j\right)=\delta_{ij}$ при $i, j \in \{1, 2, \ldots, n\}$, имеем:

$$\left(\mathbf{y}^{(k)},\mathbf{y}^{(k)}\right)=c_1^2\lambda_1^{2k}+c_2^2\lambda_2^{2k}+\ldots+c_n^2\lambda_n^{2k},$$

$$\left(\mathbf{y}^{(k)},\mathbf{y}^{(k-1)}\right)=c_1^2\lambda_1^{2k-1}+c_2^2\lambda_2^{2k-1}+\ldots+c_n^2\lambda_n^{2k-1}.$$

Отношение этих чисел

$$\frac{\left(\mathbf{y}^{(k)},\mathbf{y}^{(k)}\right)}{\left(\mathbf{y}^{(k)},\mathbf{y}^{(k-1)}\right)}=\lambda_1\cdot\frac{1+\left(\frac{c_2}{c_1}\right)^2\left(\frac{\lambda_2}{\lambda_1}\right)^{2k}+\ldots+\left(\frac{c_n}{c_1}\right)^2\left(\frac{\lambda_n}{\lambda_1}\right)^{2k}}{1+\left(\frac{c_2}{c_1}\right)^2\left(\frac{\lambda_2}{\lambda_1}\right)^{2k-1}+\ldots+\left(\frac{c_n}{c_1}\right)^2\left(\frac{\lambda_n}{\lambda_1}\right)^{2k-1}} \qquad (4.13)$$

в оговоренных выше условиях при $k\to\infty$ имеет пределом наибольшее собственное число λ_1, причем скорость сходимости к пределу будет больше, чем в степенном методе, опирающемся на отношения (4.10) $\left(O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{2k}\right)\right.$ против $\left.O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{k}\right)\right)$.

Базирующаяся на таком подходе модификация степенного метода называется ***методом скалярных произведений***. Реализовать ее можно, например, в виде следующего ***SP-алгоритма***[*].

1. Ввести: данную симметричную $n\times n$-матрицу $\mathbf{A}$, произвольный n-мерный начальный вектор $\mathbf{y}^{(0)}$ $(\neq\mathbf{0})$, малое число $\varepsilon>0$ (определяющее допустимую абсолютную погрешность искомого собственного числа λ_1), число $\lambda^{(0)}$ для начального сравнения (например, 0). Положить k=1 (счетчик итераций).

[*] SP от англ. Scalar product

2. Вычислить скаляры $s^{(0)} = \left(\mathbf{y}^{(0)}, \mathbf{y}^{(0)}\right)$, $\left\|\mathbf{y}^{(0)}\right\| = \sqrt{s^{(0)}}$ и вектор $\mathbf{x}^{(0)} = \mathbf{y}^{(0)} \Big/ \left\|\mathbf{y}^{(0)}\right\|$.

3. Вычислить $\mathbf{y}^{(k)} = \mathbf{A}\mathbf{x}^{(k-1)}$ (итерация нормированного вектора).

4. Вычислить: $s^{(k)} = \left(\mathbf{y}^{(k)}, \mathbf{y}^{(k)}\right)$ и $t^{(k)} = \left(\mathbf{y}^{(k)}, \mathbf{x}^{(k-1)}\right)$ (скалярные произведения), $\left\|\mathbf{y}^{(k)}\right\| = \sqrt{s^{(k)}}$, $\mathbf{x}^{(k)} = \mathbf{y}^{(k)} \Big/ \left\|\mathbf{y}^{(k)}\right\|$ (приближение к нормированному собственному вектору), $\lambda^{(k)} = s^{(k)} / t^{(k)}$ (приближение к собственному числу λ_1).

5. Если $\left|\lambda^{(k)} - \lambda^{(k-1)}\right| > \varepsilon$, положить $k{:=}k{+}1$ и вернуться к шагу 3, иначе завершить работу алгоритма, считая $\lambda_1 {:\approx} \lambda^{(k)}$, $\mathbf{x}_1 {:\approx} \mathbf{x}^{(k)}$.

Замечание 4.1. Данный алгоритм позволяет более быстро (т.е. за меньшее число итераций), чем РМ-алгоритм, найти с нужной точностью наибольшее собственное число симметричной матрицы, но при этом точность приближенного равенства $\mathbf{x}_1 {:\approx} \mathbf{x}^{(k)}$ для соответствующего собственного вектора может оказаться недостаточной (почему?).

Замечание 4.2. Очевидно, в методе скалярных произведений вместо отношения (4.13), стремящегося к λ_1 при $k \to \infty$, можно с тем же успехом взять отношение $\left(\mathbf{y}^{(k+1)}, \mathbf{y}^{(k)}\right)$ к $\left(\mathbf{y}^{(k)}, \mathbf{y}^{(k)}\right)$, имеющее тот же предел. Последнее же есть не что иное, как отношение Рэлея:

$$\frac{\left(\mathbf{y}^{(k+1)}, \mathbf{y}^{(k)}\right)}{\left(\mathbf{y}^{(k)}, \mathbf{y}^{(k)}\right)} = \frac{\left(\mathbf{A}\mathbf{y}^{(k)}, \mathbf{y}^{(k)}\right)}{\left(\mathbf{y}^{(k)}, \mathbf{y}^{(k)}\right)} = \rho\left(\mathbf{y}^{(k)}\right);$$

отсюда другое название метода скалярных произведений – ***метод частных Рэлея***. В соответствии со свойствами 5, 6 предыдущего пункта можно сказать, что этим методом на каждом итерационном шаге ищется наилучшее для вычисленного вектора $\mathbf{y}^{(k)}$ приближение к собственному числу λ_1 в смысле евклидовой нормы невязки.

Наличие ортонормированного базиса из собственных векторов $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ матрицы $\mathbf{A}$ позволяет применять степенной метод (метод скалярных произведений) для последовательного вычисления собственных пар $\{\lambda_i, \mathbf{x}_i\}$ при $i \geq 2$ более совершенными, чем определяемый формулой (4.12), способами. Рассмотрим один из них.

Пусть первая (старшая) собственная пара $\{\lambda_1, \mathbf{x}_1\}$ уже найдена, причем $\|\mathbf{x}_1\| = \sqrt{(\mathbf{x}_1, \mathbf{x}_1)} = 1$. Возьмем произвольный ненулевой вектор $\mathbf{z}^{(0)}$ и образуем вектор

$$\mathbf{y}^{(0)} = \mathbf{z}^{(0)} - \left(\mathbf{z}^{(0)}, \mathbf{x}_1\right)\mathbf{x}_1. \qquad (4.14)$$

Так как

$$\left(\mathbf{y}^{(0)}, \mathbf{x}_1\right) = \left(\mathbf{z}^{(0)}, \mathbf{x}_1\right) - \left(\mathbf{z}^{(0)}, \mathbf{x}_1\right)(\mathbf{x}_1, \mathbf{x}_1) = 0,$$

то вектор $\mathbf{y}^{(0)}$ ортогонален $\mathbf{x}_1$, т.е. его проекция на первый базисный вектор системы $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ равна нулю. Значит, разложение (4.8) вектора $\mathbf{y}^{(0)}$ по этому базису имеет вид

$$\mathbf{y}^{(0)} = c_2\mathbf{x}_2 + c_3\mathbf{x}_3 + \ldots + c_n\mathbf{x}_n$$

и, соответственно, степенные итерации этого вектора типа (4.9) порождают векторы

$$\mathbf{y}^{(k)} = c_2\lambda_2^k\mathbf{x}_2 + c_3\lambda_3^k\mathbf{x}_3 + \ldots + c_n\lambda_n^k\mathbf{x}_n. \qquad (4.15)$$

Легко видеть (ср. с (4.13)), что если $|\lambda_2| > |\lambda_i|$ при всех $i \in \{3, \ldots, n\}$, то $\dfrac{\left(\mathbf{y}^{(k)}, \mathbf{y}^{(k)}\right)}{\left(\mathbf{y}^{(k)}, \mathbf{y}^{(k-1)}\right)} \xrightarrow[k\to\infty]{} \lambda_2$ со скоростью $O\left(\left|\dfrac{\lambda_3}{\lambda_2}\right|^{2k}\right)$ и $\mathbf{x}^{(k)} = \dfrac{\mathbf{y}^{(k)}}{\left\|\mathbf{y}^{(k)}\right\|} \xrightarrow[k\to\infty]{} \mathbf{x}_2$ со скоростью $O\left(\left|\dfrac{\lambda_3}{\lambda_2}\right|^{k}\right)$.

Следующая собственная пара $\{\lambda_3, \mathbf{x}_3\}$ может быть найдена приближенно тем же методом, если за начальный вектор последовательности $\left(\mathbf{y}^{(k)}\right)$ принять вектор

$$\mathbf{y}^{(0)} = \mathbf{z}^{(0)} - \left(\mathbf{z}^{(0)}, \mathbf{x}_1\right)\mathbf{x}_1 - \left(\mathbf{z}^{(0)}, \mathbf{x}_2\right)\mathbf{x}_2,$$

ортогональный одновременно $\mathbf{x}_1$ и $\mathbf{x}_2$ при любом $\mathbf{z}^{(0)}$, и т.д.

Известны и другие способы последовательного нахождения собственных пар, опирающиеся на непосредственное применение степенного метода. При этом имеются возможности понижения размерности при нахождении каждой последующей собственной пары.

Замечание 4.3. В реальных расчетах, в силу неизбежных ошибок округлений, в представлении (4.15) итерированного вектора $\mathbf{y}^{(k)}$ при вычислении второй собственной пары появится малое, но растущее с увеличением номера k слагаемое, соответствующее проекции $\mathbf{y}^{(k)}$ на первый собст-

венный вектор $\mathbf{x}_1$. Поэтому реальный алгоритм должен предусматривать возврат к началу процесса итерирования, т.е. проведение операции ортогонализации по формуле (4.14) с $\mathbf{z}^{(0)} := \mathbf{y}^{(m)}$ через некоторое число итераций $k=m$.

Замечание 4.4. Не всегда бывает известным, выполняются ли оговоренные выше условия, при которых изучался степенной метод. В таких ситуациях при его применении нужно принимать особые меры предосторожности. Целесообразно, например, контролировать, сближаются ли члены последовательности $\left(\lambda^{(k)}\right)$ посредством проверки неравенств

$$\left|\lambda^{(k+1)} - \lambda^{(k)}\right| < \left|\lambda^{(k)} - \lambda^{(k-1)}\right|$$

(***прием Гарвика*** [1, 28]), а также осуществлять итерационный процесс с разных начальных векторов (в случае кратности находимого собственного числа это просто необходимо для вычисления степенным методом всех соответствующих ему собственных векторов).

Замечание 4.5. Как отмечалось выше, степенной метод сходится линейно, точнее, имеет лишь асимптотическую скорость сходимости геометрической прогрессии. При слабом доминировании модуля вычисляемого собственного числа эта сходимость может оказаться чрезвычайно медленной. Ускорения итерационного процесса можно достигнуть за счет быстрого накопления степеней матриц по схеме

$$\mathbf{A} \cdot \mathbf{A} = \mathbf{A}^2, \quad \mathbf{A}^2 \cdot \mathbf{A}^2 = \mathbf{A}^4$$

и т.д., что позволяет производить не последовательное, пошаговое, а скачкообразное построение последовательности $\left(\mathbf{y}^{(k)}\right)$ с помощью равенств вида

$$\mathbf{y}^{(k)} = \mathbf{A}^{k-m}\mathbf{y}^{(m)}$$

при фиксированных $m \in \{0, 1, \ldots, k-1\}$ (таких, при которых $k-m$ является некоторой целой степенью двойки) и нормированием сразу после очередного скачка. Здесь, правда, нужно особенно внимательно относиться к риску выхода за границы диапазона компьютерных чисел в процессе счета.

Если не требуется находить собственный вектор $\mathbf{x}_1$, то более быстро вычислять максимальное по модулю собственное число λ_1 можно на основе соотношения [21]

$$\lambda_1^k + \lambda_2^k + \ldots + \lambda_n^k = \mathrm{Sp}\,\mathbf{A}^k \qquad \forall k \in N.$$

Вычислив $\mathbf{A}^k$ по закону удвоения степеней, а затем $\mathbf{A}^{k+1} = \mathbf{A}^k \cdot \mathbf{A}$, находим отношение следов (сумм диагональных элементов) этих матриц:

$$\frac{\operatorname{Sp}\mathbf{A}^{k+1}}{\operatorname{Sp}\mathbf{A}^{k}}=\frac{\lambda_1^{k+1}\left(1+\left(\frac{\lambda_2}{\lambda_1}\right)^{k+1}+\ldots+\left(\frac{\lambda_n}{\lambda_1}\right)^{k+1}\right)}{\lambda_1^{k}\left(1+\left(\frac{\lambda_2}{\lambda_1}\right)^{k}+\ldots+\left(\frac{\lambda_n}{\lambda_1}\right)^{k}\right)}\xrightarrow[k\to\infty]{}\lambda_1.$$

Замечание 4.6. Более популярный способ улучшения сходимости степенного метода – это применение Δ^2-процесса Эйткена*). Считается, что если $\lambda^{(k-1)}$, $\lambda^{(k)}$, $\lambda^{(k+1)}$ являются тремя последовательными приближениями к собственному числу, полученными степенным методом, то число

$$\tilde{\lambda}:=\lambda^{(k-1)}-\frac{\left(\lambda^{(k)}-\lambda^{(k-1)}\right)^2}{\lambda^{(k+1)}-2\lambda^{(k)}+\lambda^{(k-1)}}$$

ближе к пределу этой последовательности, чем каждое из них. Этот факт может быть использован в реальных алгоритмах либо через несколько итерационных шагов (например, через два на третий), либо на завершающем этапе вычислений. Для искомого собственного вектора такое ускорение может производиться п о к о о р д и н а т н о .

4.3. ОБРАТНЫЕ ИТЕРАЦИИ

В предыдущем пункте было показано, что при определенных условиях наименьшее по модулю собственное число λ_n может быть найдено степенным методом, когда уже известно наибольшее λ_1. Если же проблема состоит лишь в нахождении младшей собственной пары матрицы $\mathbf{A}$, то можно обойтись и без вычисления λ_1, применяя степенной метод к матрице $\mathbf{A}^{-1}$.

В самом деле, если данная матрица $\mathbf{A}$ имеет собственные пары

$$\{\lambda_1,\mathbf{x}_1\},\{\lambda_2,\mathbf{x}_2\},\ldots,\{\lambda_{n-1},\mathbf{x}_{n-1}\},\{\lambda_n,\mathbf{x}_n\},$$

то по свойству 3 п.4.1 собственными парами матрицы $\mathbf{A}^{-1}$ будут

$$\left\{\frac{1}{\lambda_1},\mathbf{x}_1\right\},\left\{\frac{1}{\lambda_2},\mathbf{x}_2\right\},\ldots,\left\{\frac{1}{\lambda_{n-1}},\mathbf{x}_{n-1}\right\},\left\{\frac{1}{\lambda_n},\mathbf{x}_n\right\}.$$

При этом упорядочиванию спектра $\mathbf{A}$

$$|\lambda_1|\geq|\lambda_2|\geq\ldots\geq|\lambda_{n-1}|>|\lambda_n|$$

*) Более подробно об этом процессе см. в п.5.8а).

соответствует цепочка неравенств

$$\left|\frac{1}{\lambda_n}\right| > \left|\frac{1}{\lambda_{n-1}}\right| \geq \ldots \geq \left|\frac{1}{\lambda_2}\right| \geq \left|\frac{1}{\lambda_1}\right|$$

для собственных чисел $\gamma_1 := \frac{1}{\lambda_n}$, $\gamma_2 := \frac{1}{\lambda_{n-1}}$, ..., $\gamma_{n-1} := \frac{1}{\lambda_2}$, $\gamma_n := \frac{1}{\lambda_1}$ матрицы $\mathbf{A}^{-1}$. Это значит, что наименьшим по модулю собственным числом данной матрицы **A** является величина, обратная наибольшему по модулю собственному числу матрицы $\mathbf{A}^{-1}$. Последнее же может быть получено прямыми итерациями произвольного начального вектора $\mathbf{y}^{(0)}$ посредством матрицы $\mathbf{A}^{-1}$ по аналогичной (4.9) формуле

$$\mathbf{y}^{(k)} = \mathbf{A}^{-1}\mathbf{y}^{(k-1)}, \qquad k=1,2,\ldots \tag{4.16}$$

При достаточно больших $k \in N$ последовательность отношений одноименных координат векторов $\mathbf{y}^{(k)}$ и $\mathbf{y}^{(k-1)}$ должна давать приближенное значение $\frac{1}{\lambda_n}$, а вектор $\mathbf{y}^{(k)}$ (желательно его нормирование) можно принять за собственный вектор $\mathbf{x}_n$.

Вместо прямых итераций (4.16), требующих предварительного обращения исходной матрицы **A**, обычно предпочитают строить ту же последовательность векторов $\left(\mathbf{y}^{(k)}\right)$, решая при $k=1,2,3,\ldots$ линейные системы

$$\mathbf{A}\mathbf{y}^{(k)} = \mathbf{y}^{(k-1)}. \tag{4.17}$$

Так как все эти системы имеют одну и ту же матрицу коэффициентов, то самая трудоемкая часть метода Гаусса для их решения – LU-факторизация матрицы **A** – может быть выполнена лишь один раз.

Построение последовательности векторов, приближающих собственный вектор $\mathbf{x}_n$ по неявной формуле (4.17), называют ***обратными итерациями***, а процесс решения частичных проблем собственных значений на этой основе – ***методом обратных итераций***[*].

Применение обратных итераций к нахождению младшей собственной пары матрицы **A** не требует написания специального алгоритма, достаточно лишь заменить один шаг в алгоритмах предыдущего пункта. А именно, наполнение шага 2 в **PM**-алгоритме для матриц простой структуры и шага 3 в **SP**-алгоритме для симметричных матриц должно быть следующим:

решить уравненне $\mathbf{A}\mathbf{y}^{(k)} = \mathbf{y}^{(k-1)}$.

Полученный алгоритм называют ***INVIT-алгоритмом***[**] [44].

---

[*] Другое название ***обратный степенной метод*** [11].

[**] От англ. Inverse iteration.

Метод обратных итераций, а точнее, ***обратные итерации со сдвигами*** часто применяют в тех случаях, когда нужно с большой точностью найти собственный вектор, отвечающий какому-либо собственному числу из спектра заданной матрицы при условии, что известно приближенное значение этого числа. При этом, очевидно, прямое решение однородной системы (4.3) заведомо неприменимо, так как подстановка в нее значения λ, хоть сколько-нибудь отличного от собственного, сделает систему однозначно разрешимой, т.е. допускающей только тривиальное решение.

Пусть для собственного числа λ_j матрицы простой структуры A известно его приближение σ такое, что

$$|\lambda_j - \sigma| < |\lambda_i - \sigma| \qquad \forall\, i \neq j. \tag{4.18}$$

Начиная с вектора $\mathbf{x}^{(0)}$ такого, что $\left\|\mathbf{x}^{(0)}\right\| = 1$, образуем последовательность нормированных векторов $\left(\mathbf{x}^{(k)}\right)$ по формулам

$$(\mathbf{A} - \sigma\mathbf{E})\mathbf{y}^{(k)} = \mathbf{x}^{(k-1)}; \tag{4.19}$$

$$\mathbf{x}^{(k)} = \mathbf{y}^{(k)} \Big/ \left\|\mathbf{y}^{(k)}\right\|, \qquad k = 1, 2, \ldots . \tag{4.20}$$

Изучим поведение этой последовательности, для чего запишем разложение векторов $\mathbf{y}^{(k)}$ и $\mathbf{x}^{(k-1)}$ по базису из собственных векторов $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ с некоторыми коэффициентами $c_i^{(k)}$ и $b_i^{(k-1)}$ соответственно:

$$\begin{aligned} \mathbf{y}^{(k)} &= c_1^{(k)}\mathbf{x}_1 + c_2^{(k)}\mathbf{x}_2 + \ldots + c_n^{(k)}\mathbf{x}_n, \\ \mathbf{x}^{(k-1)} &= b_1^{(k-1)}\mathbf{x}_1 + b_2^{(k-1)}\mathbf{x}_2 + \ldots + b_n^{(k-1)}\mathbf{x}_n . \end{aligned} \tag{4.21}$$

Подставляя это в (4.19) и учитывая, что по определению

$$\mathbf{A}\mathbf{x}_i = \lambda_i \mathbf{x}_i \qquad \forall\, i \in \{1, 2, \ldots, n\},$$

имеем

$$\begin{aligned} &(\lambda_1 - \sigma)c_1^{(k)}\mathbf{x}_1 + (\lambda_2 - \sigma)c_2^{(k)}\mathbf{x}_2 + \ldots + (\lambda_n - \sigma)c_n^{(k)}\mathbf{x}_n = \\ &= b_1^{(k-1)}\mathbf{x}_1 + b_2^{(k-1)}\mathbf{x}_2 + \ldots + b_n^{(k-1)}\mathbf{x}_n, \end{aligned}$$

откуда, в силу единственности разложения вектора по базису, следует

$$(\lambda_i - \sigma)c_i^{(k)} = b_i^{(k-1)} \qquad \forall\, i \in \{1, 2, \ldots, n\}.$$

Анализируя получающиеся отсюда выражения

$$c_i^{(k)} = \frac{b_i^{(k-1)}}{\lambda_i - \sigma} \tag{4.22}$$

коэффициентов разложения вектора $\mathbf{y}^{(k)}$ по базису из собственных векторов, видим, что вследствие малости модуля знаменателя $\lambda_j - \sigma$ по сравнению с другими знаменателями $\lambda_i - \sigma$ (см. (4.18)), можно рассчитывать на преимущественное возрастание коэффициентов $c_j^{(k)}$ именно при собственном векторе $\mathbf{x}_j$ с ростом k. Значит, чем сильнее неравенство в (4.18), тем сильнее (быстрее) будет доминировать составляющая собственного вектора $\mathbf{x}_j$ в представлении (4.21) вектора $\mathbf{y}^{(k)}$, а значит, и вектора $\mathbf{x}^{(k)}$, получаемого из $\mathbf{y}^{(k)}$ нормированием (4.20). Последнее же говорит о том, что каков бы ни был начальный вектор $\mathbf{x}^{(0)}$ $(\neq \mathbf{0})$*), быстрое доминирование $c_j^{(k)}$ среди остальных коэффициентов $c_i^{(k)}$ происходит еще и за счет числителей дробей (4.22).

Следует заметить, что обратные итерации со сдвигами (4.19), (4.20) позволяют не только найти собственный вектор $\mathbf{x}_j$, но и служат основой для уточнения приближенного равенства $\lambda_j \approx \sigma$.

Действительно, формулы (4.19), (4.20) определяют не что иное, как метод обратных итераций для нахождения наименьшего по модулю собственного числа матрицы $\mathbf{A} - \sigma \mathbf{E}$, и, если σ существенно ближе к λ_j, чем к любому другому собственному числу λ_i матрицы $\mathbf{A}$, то уточняющие λ_j значения, согласно свойству 2 п.4.1, можно получать при k=1,2,... по формуле

$$\lambda_j^{(k)} = \sigma + \left\langle \frac{x_i^{(k-1)}}{y_i^{(k)}} \right\rangle, \qquad (4.23)$$

где $x_i^{(k-1)}$ и $y_i^{(k)}$ – координаты векторов $\mathbf{x}^{(k-1)}$ и $\mathbf{y}^{(k)}$ соответственно, а $\langle \cdot \rangle$ – знак усреднения по всем i, при которых $y_i^{(k)} \neq 0$ [28].

Как показывает практика вычислений, сходимость процесса обратных итераций со сдвигами характеризуется высокой скоростью (по сравнению с обычным степенным методом). Но еще более высокая скорость сходимости может быть получена введением переменных сдвигов, определяемых какой-нибудь последовательностью чисел $\sigma_0, \sigma_1, \sigma_2, \ldots$, сходящейся к находимому собственному числу. Не вызывает сомнений целесообразность

*) Лишь бы при его выборе не попасть на ортогональный $\mathbf{x}_j$ вектор; зачастую берут вектор $\mathbf{x}^{(0)}$ с равными координатами.

использования в роли таких чисел приближений $\lambda_j^{(k)}$ к собственному числу λ_j, получаемых по формуле (4.23).

Таким образом, ***обратные итерации с переменными сдвигами*** можно определить совокупностью равенств

$$\begin{gathered}\left(\mathbf{A}-\lambda_j^{(k-1)}\mathbf{E}\right)\mathbf{y}^{(k)}=\mathbf{x}^{(k-1)},\\ \mathbf{x}^{(k)}=\mathbf{y}^{(k)}\Big/\left\|\mathbf{y}^{(k)}\right\|,\\ \lambda_j^{(k)}=\lambda_j^{(k-1)}+\left\langle\frac{x_i^{(k-1)}}{y_i^{(k)}}\right\rangle,\end{gathered}\qquad(4.24)$$

где k=1,2,..., а число $\lambda_j^{(0)}\approx\lambda_j$ и нормированный вектор $\mathbf{x}^{(0)}$ задаются.

Скорость сходимости такого процесса – квадратичная [16, 28], в то время как в случае постоянных сдвигов – лишь линейная, хотя и с малыми, как правило, знаменателями геометрической прогрессии. Зачастую бывает достаточно сделать 2-3 итерации, чтобы получить заданную собственную пару с реально возможной точностью. Нужно только видеть разницу в цене реализации обратных итераций с постоянными и с переменными сдвигами: в первом случае при каждом k решаются линейные системы с одной и той же матрицей коэффициентов (как это было и при обратных итерациях (4.17) без сдвигов), во втором случае на разных шагах приходится решать совершенно различные системы.

В методе (4.24) так же, как и в предыдущем, неясно, как подбирать начальный сдвиг $\sigma=\lambda_j^{(0)}$, за исключением случаев, когда решается частичная проблема заведомо в такой постановке, при которой требуется найти собственное число, ближайшее к заданному значению, и соответствующий ему собственный вектор.

Более определенной в этом смысле, к тому же более быстро сходящейся является следующая модификация метода (4.24) – ***обратные итерации с отношениями Рэлея***, применяемые для решения симметричных задач на собственные значения.

Ее основу составляет **RQI-*алгоритм***[*]:

0. Задать вектор $\mathbf{x}^{(0)}$ такой, что $\left\|\mathbf{x}^{(0)}\right\|=1$.

Для k=1,2,...:

1. Вычислить $\rho_{k-1}=\left(\mathbf{A}\mathbf{x}^{(k-1)},\mathbf{x}^{(k-1)}\right)\Big/\left(\mathbf{x}^{(k-1)},\mathbf{x}^{(k-1)}\right)$.

---

[*] RQI – Rayleigh quotient iteration (англ.).

2. Найти $\mathbf{y}^{(k)}$ из уравнения $(\mathbf{A}-\rho_{k-1}\mathbf{E})\mathbf{y}^{(k)}=\mathbf{x}^{(k-1)}$.

3. Нормировать $\mathbf{y}^{(k)}$, т.е. положить $\mathbf{x}^{(k)}=\mathbf{y}^{(k)}\Big/\left\|\mathbf{y}^{(k)}\right\|$.

4. Проверить ρ_k, $\mathbf{x}^{(k)}$ на сходимость.

После "штатного" останова работы алгоритма*) при некотором $k=k_0$ в качестве собственной для данной матрицы $\mathbf{A}$ объявляется пара $\left\{\rho_{k_0},\mathbf{x}^{(k_0)}\right\}$.

Сдвиги на отношения Рэлея при наличии ортонормированного базиса из собственных векторов $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n$ обеспечивают асимптотически кубическую скорость сходимости ***последовательности Рэлея*** $\mathbf{x}^{(0)},\mathbf{x}^{(1)},\mathbf{x}^{(2)},\ldots$ к некоторому из векторов этого базиса [16, 44]. К какому именно, зависит от выбора начального вектора этой последовательности; беря различные линейно независимые векторы $\mathbf{x}^{(0)}$, можно получать разные собственные пары данной симметричной матрицы $\mathbf{A}$. При этом, правда, без дополнительных условий (типа (4.18) применительно к ρ_0 в роли σ) нельзя гарантировать, что найденное как предел последовательности $\rho_0,\rho_1,\rho_2,\ldots$ собственное число будет ближайшим к числу ρ_0.

Чтобы если не обосновать, то хотя бы осмыслить RQI-алгоритм, нужно вспомнить свойство 6 п.4.1, согласно которому, при выбранном векторе $\mathbf{x}^{(0)}$ вычисленное на первом шаге при $k=1$ отношение Рэлея $\rho_0=\left(\mathbf{A}\mathbf{x}^{(0)},\mathbf{x}^{(0)}\right)\Big/\left(\mathbf{x}^{(0)},\mathbf{x}^{(0)}\right)$ можно считать некоторым приближением к собственному числу, связанному с заданным в R_n направлением $\mathbf{x}^{(0)}$. С этим начальным приближением к какому-то собственному числу λ_j далее выполняются обратные итерации с переменными сдвигами, как и в (4.24), только приближения к λ_j находятся не через отношения координат векторов $\mathbf{y}^{(k)}$ и $\mathbf{x}^{(k-1)}$, а через отношения Рэлея (как в методе скалярных произведений, см. замечание 4.2 в п.4.2), причем поскольку здесь ρ_k приближает собственное число данной матрицы $\mathbf{A}$, а не "сдвинутой", то нет необходимости корректировать получаемое значение на величину смещения спектра, что имело место в формуле (4.23) и в последней из формул (4.24).

Замечание 4.7. RQI-алгоритм допускает использование любых векторных норм. Более естественно здесь применение евклидовой нормы; в

*) Один из вариантов останова: $\left\|\mathbf{y}^{(k)}\right\|>C$, где $C>0$ – большая константа [44].

таком случае, в силу $\left\|\mathbf{x}^{(k-1)}\right\|_2 = \sqrt{\left(\mathbf{x}^{(k-1)}, \mathbf{x}^{(k-1)}\right)} = 1$, вычисление ρ_{k-1} можно производить по формуле

$$\rho_{k-1} = \left(\mathbf{A}\mathbf{x}^{(k-1)}, \mathbf{x}^{(k-1)}\right).$$

Замечание 4.8. Ясно, что применение переменных сдвигов в методе обратных итераций сильно ухудшает от шага к шагу обусловленность матриц решаемых там СЛАУ (они быстро приближаются к вырожденным). Однако, как показали проведенные Уилкинсоном [53] исследования, это не сказывается на достижимой точности получаемых таким методом результатов. Более того, Парлетт [44] аргументировано утверждает полезность плохой обусловленности (редкий случай!) матриц линейных систем в методах обратных итераций с хорошими сдвигами; объяснением этому парадоксальному явлению служит сосредоточение ошибок округлений именно в направлении искомого собственного вектора, что только ускоряет доминирование нужной составляющей.

4.4. МЕТОД ВРАЩЕНИЙ ЯКОБИ РЕШЕНИЯ СИММЕТРИЧНОЙ ПОЛНОЙ ПРОБЛЕМЫ СОБСТВЕННЫХ ЗНАЧЕНИЙ

Дальнейшее изучение методов решения алгебраических проблем собственных значений существенно опирается на матричное ***преобразование подобия***. Напомним, что *подобными называются матрицы* $\mathbf{A}$ *и* $\mathbf{B} = \mathbf{C}^{-1}\mathbf{A}\mathbf{C}$, *где* $\mathbf{C}$ *– произвольная невырожденная матрица.*

Пополним перечень простейших свойств собственных пар (см. п.4.1) еще двумя свойствами.

Свойство 7. *Пусть* $\{\lambda, \mathbf{x}\}$ *– собственная пара матрицы* $\mathbf{B} = \mathbf{C}^{-1}\mathbf{A}\mathbf{C}$. *Тогда* $\{\lambda, \mathbf{C}\mathbf{x}\}$ *– собственная пара матрицы* $\mathbf{A}$.

Чтобы убедиться в справедливости этого свойства, достаточно подставить выражение $\mathbf{B} = \mathbf{C}^{-1}\mathbf{A}\mathbf{C}$ в верное для пары $\{\lambda, \mathbf{x}\}$ равенство $\mathbf{B}\mathbf{x} = \lambda\mathbf{x}$: имеем $\mathbf{C}^{-1}\mathbf{A}\mathbf{C}\mathbf{x} = \lambda\mathbf{x}$, откуда после умножения слева на матрицу $\mathbf{C}$ получаем равенство $\mathbf{A}\mathbf{C}\mathbf{x} = \lambda\mathbf{C}\mathbf{x}$, означающее истинность утверждения.

Как видим, *преобразование подобия сохраняет неизменным спектр любой матрицы.*

Свойство 8. *Пусть* $\mathbf{A}$ *– n-мерная матрица простой структуры* (см. п.4.2), *а матрицы* $\mathbf{\Lambda} = \mathrm{diag}(\lambda_i)$ *и* $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$ *образованы из ее собственных чисел и собственных векторов соответственно. Тогда* $\mathbf{\Lambda} = \mathbf{X}^{-1}\mathbf{A}\mathbf{X}$.

Действительно, то, что $\{\lambda_i, \mathbf{x}_i\}$ являются собственными парами матрицы $\mathbf{A}$, означает, что

$$\mathbf{A}\mathbf{x}_i = \lambda_i \mathbf{x}_i \quad \forall i \in \{1,2,...,n\}.$$

Эти n равенств могут быть записаны в виде одного матричного равенства

$$\mathbf{A}\mathbf{X} = \mathbf{X}\mathbf{\Lambda}. \qquad (4.25)$$

В силу простой структуры $\mathbf{A}$, все ее собственные векторы, т.е. столбцы матрицы $\mathbf{X}$, линейно независимы, поэтому матрица $\mathbf{X}$ обратима. Умножив равенство (4.25) слева на матрицу $\mathbf{X}^{-1}$, получим нужное представление $\mathbf{\Lambda} = \mathbf{X}^{-1}\mathbf{A}\mathbf{X}$.

Так как для диагональной матрицы $\mathbf{\Lambda}$, образованной из собственных чисел, собственными векторами могут служить единичные векторы исходного базиса ($\mathbf{\Lambda}\mathbf{e}_i = \lambda_i \mathbf{e}_i \ \forall i \in \{1,2,...,n\}$), то применяя к последнему случаю свойство 7 с $\mathbf{C}=\mathbf{X}$ и с $\mathbf{x} = \mathbf{e}_i$ (т.е. с $\mathbf{C}\mathbf{x} = \mathbf{X}\mathbf{e}_i = \mathbf{x}_i$), приходим к другой формулировке свойства 8:

если $\{\lambda_i, \mathbf{e}_i\}$ *– собственные пары матрицы* $\mathbf{\Lambda} = \mathrm{diag}(\lambda_i) = \mathbf{X}^{-1}\mathbf{A}\mathbf{X}$,
то $\{\lambda_i, \mathbf{x}_i\}$ *– собственные пары матрицы* $\mathbf{A}$

(обозначения те же, что и в свойстве 8).

Далее (в пределах данного пункта) будем рассматривать только симметричные вещественные матрицы. Пользуясь известным фактом о наличии у таких матриц полной ортонормированной системы собственных векторов, т.е. тем, что заявленная выше матрица $\mathbf{X}$ из собственных векторов в этом случае является ортогональной ($\mathbf{X}^{-1} = \mathbf{X}^T$), запишем как следствие свойства 8 равенство

$$\mathbf{\Lambda} = \mathbf{X}^T\mathbf{A}\mathbf{X}. \qquad (4.26)$$

Значит, для всякой симметричной матрицы $\mathbf{A}$ найдется диагональная матрица $\mathbf{\Lambda}$, ей ортогонально подобная. Вопрос теперь состоит в том, как реализовать хотя бы приближенно равенство (4.26), которое позволило бы найти сразу все собственные числа матрицы $\mathbf{A}$ (элементы диагонали матрицы $\mathbf{\Lambda}$) и все соответствующие им собственные векторы (столбцы матрицы $\mathbf{X}$)? Один из возможных ответов на этот вопрос состоит в применении к $\mathbf{A}$ последовательности однотипных преобразований, сохраняющих спектр и приводящих в пределе данную матрицу к диагональному виду.

Для этих целей будем использовать преобразования с помощью так

называемой ***матрицы плоских вращений***

$$\mathbf{T}_{ij} = \begin{array}{c} \begin{matrix} \quad & i & \quad & j & \quad \end{matrix} \\ \begin{pmatrix} 1 & \vdots & & \vdots & 0 \\ \cdots & c & \cdots & -s & \cdots \\ & \vdots & & \vdots & \\ \cdots & s & \cdots & c & \cdots \\ 0 & \vdots & & \vdots & 1 \end{pmatrix} \end{array} \begin{matrix} \\ i \\ \\ j \\ \\ \end{matrix}. \qquad (4.27)$$

Она получается из единичной матрицы заменой двух единиц и двух нулей на пересечениях i-х и j-х строк и столбцов числами c и $(\pm s)$, как показано в (4.27), такими, что

$$c^2 + s^2 = 1. \qquad (4.28)$$

Условие нормировки (4.28) позволяет интерпретировать числа c и s как косинус и синус некоторого угла α, и, так как умножение любой матрицы на матрицу $\mathbf{T}_{ij}$ изменяет у нее только две строки и два столбца по формулам поворота на угол α в плоскости, определяемой выбранной парой индексов i и j, то это полностью оправдывает название матрицы $\mathbf{T}_{ij}$.

Матрица $\mathbf{T}_{ij}$ ортогональна при любых $i, j \in \{1, 2, \ldots, n\}$ (проверьте!), и значит, матрица

$$\mathbf{B} = \mathbf{T}_{ij}^T \mathbf{A} \mathbf{T}_{ij} \qquad (4.29)$$

подобна $\mathbf{A}$, т.е. имеет тот же набор собственных чисел, что и матрица $\mathbf{A}$.

Классический итерационный метод вращений, предложенный Якоби (1846 г.), предполагает построение последовательности матриц

$$\mathbf{B}_0 (:= \mathbf{A}), \mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_k, \ldots$$

с помощью преобразований типа (4.29)

$$\mathbf{B}_k = \mathbf{T}_{ij}^T \mathbf{B}_{k-1} \mathbf{T}_{ij} \qquad (4.30)$$

такой, что на k-м шаге *обнуляется максимальный по модулю элемент* матрицы $\mathbf{B}_{k-1}$ предыдущего шага (а значит, и симметричный ему элемент). Эта стратегия определяет способ фиксирования пары индексов i, j, задающих позиции $(i,i), (j,j), (i,j), (j,i)$ существенных элементов в матрице вращения $\mathbf{T}_{ij}$, и угол поворота α, конкретизирующий значения этих элементов $c = \cos\alpha$ и $\pm s = \pm \sin\alpha$. На каждом шаге таких преобразований пересчитываются только две строки (или два столбца, что неважно в силу симметрии) матрицы предыдущего шага. Хотя, к сожалению, нельзя рассчитывать, что таким путем за конечное число шагов можно точно найти диагональную матрицу Λ, ибо полученные на некотором этапе преобразований нулевые элементы на следующем этапе станут, вообще говоря, ненулевыми, но нужное предельное поведение

$$\mathbf{B}_k \xrightarrow[k \to \infty]{} \Lambda,$$

как будет показано ниже, есть.

Определив идею метода вращений, рассмотрим теперь его несколько подробней.

Пусть $\mathbf{A} = (a_{ml})_{m,l=1}^{n}$ – исходная симметричная матрица, а $\mathbf{B} = (b_{ml})_{m,l=1}^{n}$ – матрица, получающаяся после одного шага преобразований по формуле (4.29). Обозначим соответственно через $\widetilde{\mathbf{A}}$ и $\widetilde{\mathbf{B}}$ двумерные подматрицы этих матриц*), определяемые фиксированием позиции (i,j) некоторого элемента a_{ij} матрицы $\mathbf{A}$:

$$\widetilde{\mathbf{A}} = \begin{pmatrix} a_{ii} & a_{ij} \\ a_{ij} & a_{jj} \end{pmatrix}, \quad \widetilde{\mathbf{B}} = \begin{pmatrix} b_{ii} & b_{ij} \\ b_{ji} & b_{jj} \end{pmatrix},$$

а через $\widetilde{\mathbf{T}}$ – такую же подматрицу матрицы $\mathbf{T}_{ij}$:

$$\widetilde{\mathbf{T}} = \begin{pmatrix} c & -s \\ s & c \end{pmatrix}.$$

Очевидно, что равенство (4.29), записанное для матриц $\mathbf{A}, \mathbf{B}, \mathbf{T}_{ij}$, будет верным и для их подматриц $\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}, \widetilde{\mathbf{T}}$. Пользуясь этим, подсчитаем элементы матрицы $\widetilde{\mathbf{B}}$, выполняя умножение в правой части двумерного аналога (4.29):

$$\widetilde{\mathbf{B}} = \widetilde{\mathbf{T}}^T \widetilde{\mathbf{A}} \widetilde{\mathbf{T}} = \begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} ca_{ii} + sa_{ij} & ca_{ij} - sa_{ii} \\ ca_{ij} + sa_{jj} & ca_{jj} - sa_{ij} \end{pmatrix} =$$

$$= \begin{pmatrix} c^2 a_{ii} + 2csa_{ij} + s^2 a_{jj} & c^2 a_{ij} - csa_{ii} + csa_{jj} - s^2 a_{ij} \\ c^2 a_{ij} - csa_{ii} + csa_{jj} - s^2 a_{ij} & c^2 a_{jj} - 2csa_{ij} + s^2 a_{ii} \end{pmatrix}.$$

Отсюда видим, что $b_{ij} = b_{ji} = 0$, если

$$\left(c^2 - s^2\right) a_{ij} - cs\left(a_{ii} - a_{jj}\right) = 0,$$

т.е. если

$$\frac{cs}{c^2 - s^2} = \frac{a_{ij}}{a_{ii} - a_{jj}}.$$

Учитывая тригонометрическую интерпретацию чисел $c = \cos\alpha$ и $s = \sin\alpha$, в соответствии с чем можно считать

$$cs = \frac{\sin 2\alpha}{2}, \quad c^2 - s^2 = \cos 2\alpha,$$

приходим к выводу, что матрица $\mathbf{B}$ будет иметь нулевые внедиагональные элементы $b_{ij} = b_{ji}$, если использовать преобразование плоского вращения

---

*) В $\widetilde{\mathbf{A}}$ элемент a_{ji} сразу заменяем равным ему элементом a_{ij}, с равенством же $b_{ji} = b_{ij}$ в $\widetilde{\mathbf{B}}$ пока не торопимся.

по формуле (4.29) на угол α такой, что

$$\operatorname{tg} 2\alpha = \frac{2a_{ij}}{a_{ii} - a_{jj}}$$

(для определенности считают $\alpha \in \left(-\frac{\pi}{4}, \frac{\pi}{4}\right]$).

Ясно, что нет необходимости находить непосредственно угол α, поскольку нужные для выполнения преобразований числа c и s можно получить через значение $\operatorname{tg} 2\alpha$ по формулам тригонометрии. При этом сразу отметим, что наибольшие требования к точности в описываемом методе предъявляются именно на стадии вычисления c и s, так как здесь возможны наибольшие потери точности, а искажение c и s нарушает ортогональность матриц $\mathbf{T}$, что ведет к неустранимым погрешностям (метод вращений, итерационный по форме, не является итерационным по существу: ему не присуща самоисправляемость методов последовательных приближений).

Проделав соответствующие элементарные выкладки и возвратившись к n-мерному случаю, запишем теперь совокупность формул, определяющую **один шаг метода вращений Якоби**. Для того чтобы не перегружать эти формулы лишними индексами, будем считать, что преобразуется матрица $\mathbf{A}$ в матрицу $\mathbf{B}$ согласно (4.29), хотя на самом деле на k-м шаге должно применяться преобразование (4.30) к матрице $\mathbf{B}_{k-1} = \left(b_{ml}^{(k-1)}\right)$ с результатом $\mathbf{B}_k = \left(b_{ml}^{(k)}\right)$.

Итак, пусть a_{ij} – ***ключевой элемент*** преобразуемой матрицы $\mathbf{A}$. Матрица $\mathbf{B}$, подобная $\mathbf{A}$, формируется следующим образом:

1. Вычисляют $p := 2a_{ij}$, $q := a_{ii} - a_{jj}$, $d := \sqrt{p^2 + q^2}$.

2. Если $q \neq 0$, то $r := |q|/(2d)$, $c := \sqrt{0.5 + r}$, $s := \sqrt{0.5 - r} \cdot \operatorname{sign}(pq)$ (если $|p| << |q|$, то лучше $s := |p| \cdot \operatorname{sign}(pq)/(2cd)$),
если же $q = 0$, то $c = s := \sqrt{2}/2$.

3. Вычисляют:
$b_{ii} := c^2 a_{ii} + s^2 a_{jj} + 2csa_{ij}$,
$b_{jj} := s^2 a_{ii} + c^2 a_{jj} - 2csa_{ij}$
(новые диагональные элементы).

4. Полагают $b_{ij} = b_{ji} := 0$
(или для контроля вычисляют $b_{ij} = b_{ji} := \left(c^2 - s^2\right)a_{ij} + cs\left(a_{jj} - a_{ii}\right)$);

5. При $m = 1,2,...,n$ таких, что $m \neq i$, $m \neq j$, вычисляют

$$b_{im} = b_{mi} := ca_{mi} + sa_{mj},$$
$$b_{jm} = b_{mj} := -sa_{mi} + ca_{mj}. \qquad (4.31)$$

6. Для всех остальных пар индексов m,l принимают $b_{ml} := a_{ml}$.

Конечно, в реальных вычислениях, если это считать основой алгоритма, не все записанное здесь следует выполнять, а именно, не нужно делать последних переприсвоений, а также должна учитываться симметрия получающейся матрицы $\mathbf{B}$.

Убедимся теперь, что *если в качестве ключевого или, в иной терминологии* [44], ***обреченного элемента*** *на каждом шаге преобразований подобия по указанным формулам брать максимальный по модулю элемент преобразуемой матрицы, то в пределе получится диагональная матрица.*

Для доказательства этого, т.е. сходимости последовательности $(\mathbf{B}_k)$ к Λ, используется (см. приложение 1) норма Фробениуса

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i,j=1}^{n} a_{ij}^2}.$$

Проследим за поведением норм матриц (точнее, квадратов этих норм), получающихся из матриц $\mathbf{B}_k$ заменой диагональных элементов нулями. Такие матрицы, определяемые внедиагональными элементами матриц $\mathbf{B}_k$, будем обозначать $\mathbf{V}_{\mathbf{B}_k}$. При этом опять для упрощения записей будем пока рассматривать переход от $\mathbf{A}$ к $\mathbf{B}$.

Найдем выражение суммы квадратов внедиагональных элементов матрицы

$$\mathbf{B} = \begin{pmatrix} a_{11} & \dots & b_{1i} & \dots & b_{1j} & \dots & a_{1n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ b_{i1} & \dots & b_{ii} & \dots & 0 & \dots & b_{in} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ b_{j1} & \dots & 0 & \dots & b_{jj} & \dots & b_{jn} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & \dots & b_{ni} & \dots & b_{nj} & \dots & a_{nn} \end{pmatrix},$$

принадлежащих изменяющимся по сравнению с $\mathbf{A}$ строке i и столбцу j, через элементы матрицы $\mathbf{A}$. Имеем (с учетом (4.31), (4.28) и условия обнуления элементов $b_{ij} = b_{ji}$):

$$\sum_{m \neq i} b_{im}^2 + \sum_{m \neq j} b_{mj}^2 = \sum_{m \neq i,j} \left(c^2 a_{mi}^2 + 2csa_{mi}a_{mj} + s^2 a_{mj}^2 \right) +$$
$$+ b_{ij}^2 + \sum_{m \neq j,i} \left(s^2 a_{mi}^2 - 2csa_{mi}a_{mj} + c^2 a_{mj}^2 \right) + b_{ji}^2 =$$
$$= \sum_{m \neq i,j} \left[\left(c^2 + s^2 \right) a_{mi}^2 + \left(c^2 + s^2 \right) a_{mj}^2 \right] = \sum_{m \neq i,j} \left(a_{mi}^2 + a_{mj}^2 \right).$$

Аналогичные суммы квадратов j-й строки и i-го столбца, в силу симметрии, дадут такое же выражение. Это означает, что если полученное равенство удвоить и дополнить левую и правую части суммой квадратов всех остальных внедиагональных элементов матрицы **A** (служащих соответствующими элементами и матрицы **B**), то в левой части будет стоять сумма квадратов всех внедиагональных элементов матрицы **B**, а в правой части – сумма квадратов всех внедиагональных элементов матрицы **A**, кроме a_{ji}^2 и a_{ij}^2. Следовательно, справедливо равенство

$$\|\mathbf{V_B}\|_F^2 = \|\mathbf{V_A}\|_F^2 - 2a_{ij}^2, \qquad (4.32)$$

говорящее об *убывании сумм квадратов внедиагональных элементов в рассматриваемом процессе преобразований подобия.*

Пусть $|a_{ij}| = \max\limits_{m\neq l}|a_{ml}|$. Тогда можно считать, что $a_{ij}^2 = \max\limits_{m\neq l}a_{ml}^2$ не меньше, чем среднее значение множества из $n^2 - n$ квадратов всех внедиагональных элементов n-мерной матрицы **A**, т.е. величины

$$\frac{1}{n(n-1)}\sum_{m\neq l}a_{ml}^2 = \frac{1}{n(n-1)}\|\mathbf{V_A}\|_F^2.$$

Подставляя полученную оценку $a_{ij}^2 \geq \frac{1}{n(n-1)}\|\mathbf{V_A}\|_F^2$ в (4.32), приходим к неравенству

$$\|\mathbf{V_B}\|_F^2 \leq \left(1 - \frac{2}{n(n-1)}\right)\|\mathbf{V_A}\|_F^2.$$

На основании этого для последовательности матриц $\mathbf{B}_k$, представляемых в виде

$$\mathbf{B}_k = \operatorname{diag}\left(b_{ii}^{(k)}\right) + \mathbf{V}_{\mathbf{B}_k},$$

можно записать:

$$\|\mathbf{V}_{\mathbf{B}_k}\|_F^2 \leq \left(1 - \frac{2}{n(n-1)}\right)\|\mathbf{V}_{\mathbf{B}_{k-1}}\|_F^2 \leq \cdots \leq \left(1 - \frac{2}{n(n-1)}\right)^k \|\mathbf{V_A}\|_F^2 \xrightarrow[k\to\infty]{} 0.$$

Значит, при указанном способе выбора ключевого элемента последовательность подобных матриц $\mathbf{B}_k$ сходится к диагональной матрице Λ из собственных значений, по крайней мере, со скоростью геометрической прогрессии.

Описанный выше классический вариант метода вращений Якоби, как показывают более тонкие оценки, на самом деле имеет асимптотически квадратичную скорость сходимости [15, 44, 53]. Однако при больших размерностях n его реализация наталкивается на существенные потери машинных ресурсов, связанные с поиском наибольшего по модулю ключевого элемента. Поэтому чаще применяется более медленно, но все-таки тоже асимптотически квадратично сходящийся ***циклический метод Якоби с***

барьерами. Стратегия выбора ключевого элемента здесь такова: устраивается циклический перебор всех над- или поддиагональных элементов матрицы $\mathbf{A}$ (точнее, $\mathbf{B}_{k-1}$) для их использования в роли обреченного, но при этом пропускаются элементы, абсолютные величины которых меньше некоторого заданного положительного числа – барьера. Этот барьер может быть переменным (уменьшающимся по какому-либо осмысленному принципу).

Следует отметить успешную работу таких итерационных процессов и в случаях, когда исходная матрица имеет кратные и, что хуже, близкие собственные числа.

В завершение остается вспомнить, что в соответствии со свойствами 7, 8 и проведенными рассуждениями за собственные векторы $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ матрицы $\mathbf{A}$ (имеющие единичную евклидову норму) могут быть приближенно приняты столбцы результирующей матрицы, получающейся справа от $\mathbf{A}$ в цепочке преобразований подобия

$$\mathbf{B}_1 = \mathbf{T}^T_{i_0 j_0} \mathbf{A} \mathbf{T}_{i_0 j_0},$$

$$\mathbf{B}_2 = \mathbf{T}^T_{i_1 j_1} \mathbf{B}_1 \mathbf{T}_{i_1 j_1} = \mathbf{T}^T_{i_1 j_1} \mathbf{T}^T_{i_0 j_0} \mathbf{A} \mathbf{T}_{i_0 j_0} \mathbf{T}_{i_1 j_1},$$

. . . . . .

$$\mathbf{B}_k = \mathbf{T}^T_{i_{k-1} j_{k-1}} \ldots \mathbf{T}^T_{i_0 j_0} \mathbf{A} \mathbf{T}_{i_0 j_0} \ldots \mathbf{T}_{i_{k-1} j_{k-1}},$$

т.е. матрицы

$$\mathbf{T}_k := \mathbf{T}_{i_0 j_0} \cdot \mathbf{T}_{i_1 j_1} \cdot \ldots \cdot \mathbf{T}_{i_{k-1} j_{k-1}}$$

при некотором $k = K$. Разумеется, в реальном вычислительном процессе нужно обойтись без матричных умножений.

Сигналом для окончания процесса вращений Якоби может служить, например, малость $\left\| \mathbf{V}_{\mathbf{B}_k} \right\|_F$ или $|a_{ij}|$.

4.5. ПОНЯТИЕ ОБ LU-АЛГОРИТМЕ ДЛЯ НЕСИММЕТРИЧНЫХ ЗАДАЧ

Чаще (по крайней мере, в несимметричном случае) алгоритмы приближенного решения полных проблем собственных значений основываются на приведении данных матриц к подобным им матрицам не диагонального, а треугольного вида. Наиболее простой из таких алгоритмов вычисления собственных чисел опирается на хорошо известное LU-разложение матриц (см. главу 2).

Пусть данная $n \times n$-матрица $\mathbf{A}$ представлена в виде $\mathbf{A} = \mathbf{LU}$, где $\mathbf{L}$ и $\mathbf{U}$ – соответственно нижняя и верхняя треугольные матрицы.

Обозначим*) $\mathbf{A}_1 := \mathbf{UL}$, тогда $\mathbf{U} = \mathbf{A}_1\mathbf{L}^{-1}$. Подставив это выражение матрицы $\mathbf{U}$ в равенство $\mathbf{A} = \mathbf{LU}$, получаем новое представление $\mathbf{A}$:

$$\mathbf{A} = \mathbf{L}\mathbf{A}_1\mathbf{L}^{-1}, \tag{4.33}$$

которое говорит о подобии матриц $\mathbf{A}$ и $\mathbf{A}_1$, т.е. о равенстве их собственных чисел $\lambda_{\mathbf{A}}$ и $\lambda_{\mathbf{A}_1}$.

Если матрица $\mathbf{A}_1$ может быть, как и $\mathbf{A}$, представлена в виде произведения нижней $\mathbf{L}_1$ и верхней $\mathbf{U}_1$ треугольных, т.е. $\mathbf{A}_1 = \mathbf{L}_1\mathbf{U}_1$, то, положив $\mathbf{A}_2 := \mathbf{U}_1\mathbf{L}_1$ и выразив отсюда $\mathbf{U}_1 = \mathbf{A}_2\mathbf{L}_1^{-1}$, аналогично предыдущему получаем

$$\mathbf{A}_1 = \mathbf{L}_1\mathbf{A}_2\mathbf{L}_1^{-1}. \tag{4.34}$$

Следовательно, $\mathbf{A}_1$ подобна $\mathbf{A}_2$ и ,значит, $\lambda_{\mathbf{A}_1} = \lambda_{\mathbf{A}_2}$.

Суперпозиция этих двух преобразований, т.е. подстановка (4.34) в (4.33), дает выражение $\mathbf{A}$ через $\mathbf{A}_2$:

$$\mathbf{A} = \mathbf{L}\mathbf{L}_1\mathbf{A}_2\mathbf{L}_1^{-1}\mathbf{L}^{-1} = \mathbf{L}\mathbf{L}_1\mathbf{A}_2(\mathbf{L}\mathbf{L}_1)^{-1},$$

непосредственно утверждающее равенство собственных чисел $\lambda_{\mathbf{A}}$ и $\lambda_{\mathbf{A}_2}$.

Такой процесс построения теоретически бесконечной последовательности подобных матриц и составляет основу ***LU***-(иначе, ***LR***-) ***алгоритма*****). Он определяется фактически двумя формулами:

$$\mathbf{A}_k = \mathbf{L}_k\mathbf{U}_k, \qquad \mathbf{A}_{k+1} = \mathbf{U}_k\mathbf{L}_k, \tag{4.35}$$

где $\mathbf{A}_0 := \mathbf{A}$, $k = 0,1,2,\ldots$, причем первая из этих формул означает процедуру треугольной факторизации матрицы $\mathbf{A}_k$ на k-м шаге, а вторая - простое умножение верхней треугольной матрицы на нижнюю.

Доказано [41, 53], что при ряде ограничений на данную матрицу $\mathbf{A}$ (простейшим из которых является, в частности, требование, чтобы все ее собственные числа были различны по модулю) итерационный процесс (4.35) осуществим, и формируемая им последовательность $(\mathbf{A}_k)$ сходится к треугольной матрице вида

$$\begin{pmatrix} \lambda_1 & * & * & \ldots & * \\ 0 & \lambda_2 & * & \ldots & * \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & \lambda_n \end{pmatrix} \quad \text{или} \quad \begin{pmatrix} \lambda_1 & 0 & 0 & \ldots & 0 \\ * & \lambda_2 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ * & * & * & \ldots & \lambda_n \end{pmatrix}$$

---

*) Напомним, что произведение матриц, вообще говоря, некоммутативно.

**) Алгоритм предложен Рутисхаузером (1958 г.).

в зависимости от того, фиксируется единичная диагональ при LU-факторизации у матрицы $\mathbf{L}$ или у $\mathbf{U}$ соответственно. К сожалению, эти ограничения трудно назвать конструктивными, и реализующие LU-алгоритм программы больше опираются на эмпирику. Осуществимости, устойчивости и ускорения сходимости процесса (4.35) обычно добиваются (если это в принципе возможно) путем подходящих сдвигов матриц и перестановок их элементов; соответствующие исследования и рекомендации по этому поводу можно найти в [53].

Пример 4.1. Рассмотрим, как ведет себя LU-алгоритм (4.35), примененный к нахождению собственных чисел матрицы $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 6 & 1 \end{pmatrix}$.

Выполнив LU-разложение*), получим

$$\mathbf{A}_0 := \mathbf{A} = \mathbf{L}_0\mathbf{U}_0 = \begin{pmatrix} 2 & 0 \\ 6 & -2 \end{pmatrix}\begin{pmatrix} 1 & 0.5 \\ 0 & 1 \end{pmatrix}.$$

Перемножая матрицы $\mathbf{L}_0$ и $\mathbf{U}_0$ в обратном порядке, строим матрицу

$$\mathbf{A}_1 := \mathbf{U}_0\mathbf{L}_0 = \begin{pmatrix} 5 & -1 \\ 6 & -2 \end{pmatrix}.$$

Факторизуя эту матрицу аналогично предыдущему, имеем

$$\mathbf{A}_1 = \mathbf{L}_1\mathbf{U}_1 = \begin{pmatrix} 5 & 0 \\ 6 & -0.8 \end{pmatrix}\begin{pmatrix} 1 & -0.2 \\ 0 & 1 \end{pmatrix},$$

откуда

$$\mathbf{A}_2 := \mathbf{U}_1\mathbf{L}_1 = \begin{pmatrix} 3.8 & 0.16 \\ 6 & -0.8 \end{pmatrix}.$$

Следующий шаг дает

$$\mathbf{A}_2 = \mathbf{L}_2\mathbf{U}_2 = \begin{pmatrix} 3.8 & 0 \\ 6 & -1.0526... \end{pmatrix}\begin{pmatrix} 1 & 0.0421... \\ 0 & 1 \end{pmatrix},$$

$$\mathbf{A}_3 := \mathbf{U}_2\mathbf{L}_2 = \begin{pmatrix} 4.0526... & 0.0443... \\ 6 & -1.0526... \end{pmatrix}.$$

Как видим, диагональные элементы матрицы $\mathbf{A}_2$ отличаются от точных значений собственных чисел $\lambda_1 = 4$, $\lambda_2 = -1$ на 0.2, а матрица $\mathbf{A}_3$ позволяет указать значения λ_1 и λ_2 с погрешностью ≈ 0.05.

---

*) По формулам из п 2.3 $u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj}$ при $i \le j$, $l_{ij} = \frac{1}{u_{jj}}\left(a_{ij} - \sum_{k=1}^{j-1} l_{ik}u_{kj}\right)$ при $i > j$, используемым поочередно

Если в этом же примере фиксировать единичную диагональ у матриц $\mathbf{L}_k$, то процесс (4.35) будет развиваться следующим образом*) :

$$\mathbf{A} = \tilde{\mathbf{L}}_0\tilde{\mathbf{U}}_0 = \begin{pmatrix} 1 & 0 \\ 3 & 1 \end{pmatrix}\begin{pmatrix} 2 & 1 \\ 0 & -2 \end{pmatrix}, \qquad \tilde{\mathbf{A}}_1 := \tilde{\mathbf{U}}_0\tilde{\mathbf{L}}_0 = \begin{pmatrix} 5 & 1 \\ -6 & -2 \end{pmatrix};$$

$$\tilde{\mathbf{A}}_1 = \tilde{\mathbf{L}}_1\tilde{\mathbf{U}}_1 = \begin{pmatrix} 1 & 0 \\ -1.2 & 1 \end{pmatrix}\begin{pmatrix} 5 & 1 \\ 0 & -0.8 \end{pmatrix}, \qquad \tilde{\mathbf{A}}_2 := \tilde{\mathbf{U}}_1\tilde{\mathbf{L}}_1 = \begin{pmatrix} 3.8 & 1 \\ 0.96 & -0.8 \end{pmatrix};$$

$$\tilde{\mathbf{A}}_2 = \tilde{\mathbf{L}}_2\tilde{\mathbf{U}}_2 = \begin{pmatrix} 1 & 0 \\ 0.2526... & 1 \end{pmatrix}\begin{pmatrix} 3.8 & 1 \\ 0 & -1.0526... \end{pmatrix},$$

$$\tilde{\mathbf{A}}_3 := \tilde{\mathbf{U}}_2\tilde{\mathbf{L}}_2 = \begin{pmatrix} 4.0526... & 1 \\ -0.2659... & -1.0526... \end{pmatrix}.$$

Диагонали матриц $\mathbf{A}_k$ и $\tilde{\mathbf{A}}_k$, несущие приближения к собственным числам $\mathbf{A}$, при одних и тех же значениях k полностью совпадают, но во втором случае не так заметно стремление к нулю поддиагональных элементов, хотя относительная скорость убывания модулей наддиагональных элементов $\mathbf{A}_k$ и поддиагональных элементов $\tilde{\mathbf{A}}_k$ примерно одинакова.

Одним из серьезных факторов, ограничивающих сферу применения LU-алгоритмов, является их недостаточно хорошая численная устойчивость (улучшение этого параметра путем перестановок строк и столбцов сильно отражается на экономичности метода). Этот фактор может играть особенно существенную роль на фоне возможной неустойчивости самой несимметричной проблемы собственных значений.

Ярким примером матрицы, для которой задача нахождения собственных чисел является неустойчивой, служит $n \times n$-матрица [41]

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

имеющая число 0 собственным значением n-й кратности. Введем возмущение ε в левый нижний элемент матрицы $\mathbf{A}$. Характеристическим урав-

*) Для LU-разложения здесь поочередно используются формулы (см. п.2.3)

$$l_{ij} = a_{ij} - \sum_{k=1}^{j-1} l_{ik}u_{kj} \quad (i \geq j), \qquad u_{kj} = \frac{1}{l_{ii}}\left(a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj}\right) \quad (i < j).$$

нением для возмущенной матрицы

$$\mathbf{A}_\varepsilon = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & 1 \\ \varepsilon & 0 & 0 & \dots & 0 & 0 \end{pmatrix}$$

будет уравнение

$$\begin{vmatrix} -\lambda & 1 & 0 & \dots & 0 & 0 \\ 0 & -\lambda & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & -\lambda & 1 \\ \varepsilon & 0 & 0 & \dots & 0 & -\lambda \end{vmatrix} = 0 .$$

Раскрывая определитель по элементам первого столбца, получаем:

$$-\lambda \cdot \begin{vmatrix} -\lambda & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & -\lambda & 1 \\ 0 & 0 & \dots & 0 & -\lambda \end{vmatrix} + (-1)^{n+1}\varepsilon \cdot \begin{vmatrix} 1 & 0 & \dots & 0 & 0 \\ -\lambda & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & -\lambda & 1 \end{vmatrix} = 0 \quad \Leftrightarrow$$

$$\Leftrightarrow \quad -\lambda(-\lambda)^{n-1} + (-1)^{n+1}\varepsilon = 0 \quad \Leftrightarrow \quad \lambda^n - \varepsilon = 0.$$

Следовательно, матрица $\mathbf{A}_\varepsilon$ имеет n различных, в общем, комплексных собственных значений $\lambda_i = \sqrt[n]{\varepsilon}$, $i = 1, 2, \dots, n$. Если взять, например, $n = 100$, а $\varepsilon = 10^{-100}$, то $|\lambda_i| = 0.1$, т.е. чрезвычайно малое, неощутимое для вычислительной машины искажение всего одного элемента данной специфической матрицы приводит к существенному изменению ее спектра.

Разумеется, большинство важных в приложениях задач на собственные значения не так плохи. Однако, обозначив и, возможно, намеренно утрировав проблему, призовем читателя к осторожности в применениях уже рассмотренных методов и интерпретации их результатов, а также к пониманию необходимости построения более устойчивых методов численного решения несимметричных спектральных алгебраических задач (см. следующий пункт). Численная устойчивость всех описываемых здесь методов подробно изучается в [53].

4.6. QR-АЛГОРИТМ

В идейном плане от схематично описанного выше LU-алгоритма мало чем отличается так называемый **QR-*алгоритм***[*] [1, 6, 15, 16, 26, 41, 44, 53]. При $k=0,1,2,\ldots$, начиная с $\mathbf{A}_0 := \mathbf{A}$, здесь строят последовательность матриц $(\mathbf{A}_k)$ по формулам

$$\mathbf{A}_k = \mathbf{Q}_k \mathbf{R}_k, \qquad \mathbf{A}_{k+1} = \mathbf{R}_k \mathbf{Q}_k, \tag{4.36}$$

первая из которых означает разложение матрицы $\mathbf{A}_k$ в произведение ортогональной $\mathbf{Q}_k$ и правой треугольной $\mathbf{R}_k$ (такое разложение существует для любой квадратной матрицы [6]), а вторая – перемножение полученных в результате факторизации $\mathbf{A}_k$ матриц $\mathbf{Q}_k$ и $\mathbf{R}_k$ в обратном порядке.

Аналогично предыдущему (см. п.4.5) на основе свойства ортогональных матриц $\mathbf{Q}_k^T = \mathbf{Q}_k^{-1}$ в соответствии с (4.36) можно записать представление данной матрицы $\mathbf{A}$ в виде

$$\mathbf{A} = \mathbf{Q}_0\mathbf{Q}_1\ldots\mathbf{Q}_{k-1}\mathbf{Q}_k\mathbf{A}_{k+1}\mathbf{Q}_k^T\mathbf{Q}_{k-1}^T\ldots\mathbf{Q}_1^T\mathbf{Q}_0^T$$

или, иначе,

$$\mathbf{A} = (\mathbf{Q}_0\mathbf{Q}_1\ldots\mathbf{Q}_{k-1}\mathbf{Q}_k)\mathbf{A}_{k+1}(\mathbf{Q}_0\mathbf{Q}_1\ldots\mathbf{Q}_k)^{-1}. \tag{4.37}$$

Следовательно, любая из матриц последовательности $(\mathbf{A}_k)$ ортогонально подобна матрице $\mathbf{A}$.

При определенных ограничениях, одним из которых опять выступает требование, чтобы матрица $\mathbf{A}$ не имела равных по модулю собственных значений, генерируемая процессом (4.36) последовательность матриц $(\mathbf{A}_k)$ сходится к матрице правой треугольной формы с диагональю из собственных чисел. Скорость обнуления поддиагональных частей матриц $\mathbf{A}_k$ линейна и зависит, как и во многих ранее рассмотренных методах, от отношений $\dfrac{|\lambda_i|}{|\lambda_j|}$ при $i > j$ (по-прежнему считаем $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|$). Наличие комплексно сопряженных пар собственных чисел у данной вещественной матрицы $\mathbf{A}$ не является, вообще говоря, препятствием для применения QR-алгоритма; просто в этом случае предельной матрицей для последовательности $(\mathbf{A}_k)$ будет матрица квазитреугольного (иначе, блочно-треугольного) вида. Каждой комплексной паре собственных чисел в такой матрице будет соответствовать диагональный 2×2-блок, причем сходимость здесь наблюдается по форме матрицы, а не поэлементно (т.е. элементы внутри этих блоков могут изменяться без видимой зависимости от k при сохранении неизменными их собственных чисел).

---

[*] Этот метод предложен почти одновременно российским математиком Кублановской (1961 г.) и англичанином Фрэнсисом (1962 г.).

Обычно QR-алгоритм (4.36) применяют не к исходной матрице $\mathbf{A}$, а к подобной ей ***правой почти треугольной матрице*** $\mathbf{B}$, называемой также ***матрицей Хессенберга***[*], вида

$$\mathbf{B}=\begin{pmatrix} b_{11} & b_{12} & \dots & b_{1,n-1} & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2,n-1} & b_{2n} \\ 0 & b_{32} & \dots & b_{3,n-1} & b_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & b_{n-1,n-1} & b_{n-1,n} \\ 0 & 0 & \dots & b_{n,n-1} & b_{nn} \end{pmatrix}.$$

В основе преобразования $\mathbf{A}$ к виду $\mathbf{B}$ (определяемому условием $b_{ij}=0$ при $j<i-1$) лежит ***преобразование Хаусхолдера*** или, иначе, ***преобразование отражения***, осуществляемое с помощью ***матрицы отражения (Хаусхолдера)***

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

где $\mathbf{W}$ – произвольный вектор-столбец, но такой, что его евклидова норма равна единице. В силу этого требования к вектору $\mathbf{W}$,

$$\|\mathbf{W}\|_2^2=(\mathbf{W},\mathbf{W})=\mathbf{W}^T\mathbf{W}=1,$$

и с учетом симметричности матрицы $\mathbf{H}$, вытекающей из симметричности матрицы

$$\mathbf{W}\mathbf{W}^T=\begin{pmatrix} W_1 \\ W_2 \\ \dots \\ W_n \end{pmatrix}\cdot(W_1,W_2,\dots,W_n)=\begin{pmatrix} W_1^2 & W_1W_2 & \dots & W_1W_n \\ W_2W_1 & W_2^2 & \dots & W_2W_n \\ \dots & \dots & \dots & \dots \\ W_nW_1 & W_nW_2 & \dots & W_n^2 \end{pmatrix},$$

имеем:

$$\mathbf{H}\mathbf{H}^T=\mathbf{H}^2=\mathbf{E}-4\mathbf{W}\mathbf{W}^T+4\mathbf{W}\mathbf{W}^T\mathbf{W}\mathbf{W}^T=\mathbf{E}.$$

Следовательно, матрица отражения ортогональна, и ,значит, матрицы $\mathbf{A}$ и $\mathbf{B}$, связанные соотношением

$$\mathbf{B}=\mathbf{H}\mathbf{A}\mathbf{H}\quad\left(=\mathbf{H}\mathbf{A}\mathbf{H}^T=\mathbf{H}\mathbf{A}\mathbf{H}^{-1}\right),$$

являются подобными[**].

Теперь нетрудно сообразить, как нужно распорядиться свободой задания элементов векторов $\mathbf{W}$ при построении матриц отражения, чтобы за

---

[*] Герхард Хессенберг (1874–1925) – немецкий математик. Иногда почти треугольной формой матрицы называют упомянутые матрицы блочно-треугольного вида [41].

[**] Матрица отражения обладает рядом других интересных свойств; в частности, ее название связано с тем, что линейное преобразование, осуществляемое такой матрицей, оставляет без изменений векторы, ортогональные вектору $\mathbf{W}$, а коллинеарные ему векторы переводит в противоположные ("отражает").

конечное число шагов преобразований Хаусхолдера произвольно заданную матрицу $\mathbf{A}$ привести к форме Хессенберга $\mathbf{B}$.

А именно, можно показать, что начатым с $\mathbf{B}_1 := \mathbf{A}$ процессом

$$\mathbf{B}_{m+1} = \mathbf{H}_m \mathbf{B}_m \mathbf{H}_m, \qquad m = 1, 2, \ldots, n-2, \tag{4.38}$$

где $\mathbf{H}_m = \mathbf{E} - 2\mathbf{W}_m \mathbf{W}_m^T$, данная $n \times n$-матрица $\mathbf{A}$ за $n-2$ шага будет приведена к виду $\mathbf{B}$, т.е. матрица $\mathbf{B} := \mathbf{B}_{n-1}$ подобна $\mathbf{A}$, если задающие матрицы Хаусхолдера $\mathbf{H}_m$ векторы $\mathbf{W}_m$ по данной матрице $\mathbf{A}$ строить следующим образом*).

При $m=1$ вектор $\mathbf{W}_1$ определяется равенством

$$\mathbf{W}_1^T = \mu_1 (0, a_{21} - s_1, a_{31}, \ldots, a_{n1}), \tag{4.39}$$

где $s_1 = \operatorname{sign}(-a_{21}) \cdot \sqrt{\sum_{i=2}^{n} a_{i1}^2}$, $\mu_1 = \dfrac{1}{\sqrt{2 s_1 (s_1 - a_{21})}}$. Такое задание $\mathbf{W}_1$ обеспечивает ортогональность симметричной матрицы

$$\mathbf{H}_1 = \mathbf{E} - 2\mathbf{W}_1 \mathbf{W}_1^T$$

и одновременное получение с ее помощью нужных $n-2$ нулей в первом столбце матрицы

$$\mathbf{B}_2 = \mathbf{H}_1 \mathbf{B}_1 \mathbf{H}_1 \quad (= \mathbf{H}_1 \mathbf{A} \mathbf{H}_1).$$

Вектор $\mathbf{W}_2$ по матрице $\mathbf{B}_2$ строится совершенно аналогично, только фиксируются нулевыми не одна, а две первые его координаты, и определяющую роль играют теперь не первый, а второй столбец матрицы $\mathbf{B}_2$ и его третий элемент. При этом у матрицы $\mathbf{B}_3 = \mathbf{H}_2 \mathbf{B}_2 \mathbf{H}_2$ окажется $n-3$ нулевых элемента во втором столбце и сохранятся полученные на предыдущем шаге нули в первом столбце.

Этот процесс очевидным образом может быть продолжен до исчерпания и без особого труда может быть описан общими формулами типа формул (4.39), для чего нужно лишь ввести обозначения для элементов последовательности матриц $\mathbf{B}_m$ (например, с помощью верхних индексов m).

Рассмотрим на простом числовом примере, как приводится матрица к форме Хессенберга, когда для этого требуется только один шаг преобразований Хаусхолдера.

Пример 4.2. Дана матрица $\mathbf{A} = \begin{pmatrix} 5 & 1 & -3 \\ 3 & 0 & -2 \\ -4 & -1 & 1 \end{pmatrix}$. Найти матрицу $\mathbf{B}$, подобную матрице $\mathbf{A}$ и имеющую форму Хессенберга.

---

*) Логику таких построений, основанную на сохранении евклидовой нормы столбца, см., например, в [53].

Решение проводим по формулам (4.38), (4.39) при $n = 3$, которые для данного случая можно записать так (в естественном для выполнения порядке):

$$s = \operatorname{sign}(-a_{21}) \cdot \sqrt{a_{21}^2 + a_{31}^2}\ ;\ \mu = \frac{1}{\sqrt{2s(s - a_{21})}};$$

$$\mathbf{W}^T = \mu(0, a_{21} - s, a_{31}); \qquad \mathbf{H} = \mathbf{E} - 2\mathbf{W}\mathbf{W}^T;\ \ \mathbf{B} = \mathbf{H}\mathbf{A}\mathbf{H}.$$

Имеем:

$$s = -\sqrt{3^2 + (-4)^2} = -5; \qquad \mu = \frac{1}{\sqrt{2(-5)(-5-3)}} = \frac{1}{4\sqrt{5}};$$

$$2\mathbf{W}\mathbf{W}^T = \frac{1}{40}\begin{pmatrix} 0 \\ 8 \\ -4 \end{pmatrix}(0{,}8, -4) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1.6 & -0.8 \\ 0 & -0.8 & 0.4 \end{pmatrix};$$

$$\mathbf{H} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -0.6 & 0.8 \\ 0 & 0.8 & 0.6 \end{pmatrix}; \qquad \mathbf{H}\mathbf{A} = \begin{pmatrix} 5 & 1 & -3 \\ -5 & -0.8 & 2 \\ 0 & -0.6 & -1 \end{pmatrix};$$

$$\mathbf{B} = \begin{pmatrix} 5 & -3 & -1 \\ -5 & 2.08 & 0.56 \\ 0 & -0.44 & -1.08 \end{pmatrix}.$$

Замечание 4.9. При решении симметричных проблем собственных значений методом вращений на первой стадии решения также часто применяют преобразования Хаусхолдера. Абсолютно те же формулы (4.38), (4.39) приведут симметричную матрицу $\mathbf{A}$ к подобной ей матрице $\mathbf{B}$ трехдиагонального вида (частному случаю формы Хессенберга), что значительно повышает эффективность последующих преобразований плоских вращений Якоби.

Замечание 4.10. Если начать процесс построения матриц отражения $\mathbf{H}_m$ не по формулам (4.39), а по аналогичным им, но с ключевым элементом a_{11}, т.е. полагая

$$\mathbf{W}_1^T = \mu_1(a_{11} - s, a_{21}, \ldots, a_{n1}),$$

$$s_1 = \operatorname{sign}(-a_{11}) \cdot \sqrt{\sum_{i=1}^{n} a_{i1}{}^2}\ ,\ \mu_1 = \frac{1}{\sqrt{2s_1(s_1 - a_{11})}},$$

то за $n - 1$ шаг ортогональных преобразований

$$\mathbf{R}_{m+1} = \mathbf{H}_m\mathbf{R}_m; \quad m = 1, 2, \ldots, n-1; \quad \mathbf{R}_1 := \mathbf{A}$$

можно получить разложение матрицы $\mathbf{A}$ в произведение ортогональной $\mathbf{H}:=\mathbf{H}_{n-1}\ldots\mathbf{H}_2\mathbf{H}_1$ и правой треугольной $\mathbf{R}:=\mathbf{R}_n$, так как результирующее равенство $\mathbf{R}=\mathbf{HA}$ равносильно равенству $\mathbf{A}=\mathbf{HR}$ в силу ортогональности и симметричности $\mathbf{H}$. Приведение матрицы $\mathbf{A}$ такими преобразованиями к треугольному виду составляет основу ***метода отражений решения линейных алгебраических систем***. Из равносильности равенств $\mathbf{Ax}=\mathbf{b}$, $\mathbf{HAx}=\mathbf{Hb}$ и $\mathbf{Rx}=\mathbf{Hb}$ легко понять, что для решения СЛАУ этим методом нужно над вектором свободных членов $\mathbf{b}$ выполнять те же преобразования, что и над матрицей коэффициентов $\mathbf{A}$, после чего нужно будет сделать только обратный ход, как в методе Гаусса. Метод отражений решения СЛАУ в полтора раза экономичнее метода вращений (см. п.2.7) и практически не уступает последнему по устойчивости к накоплению ошибок округлений.

Вообще говоря, весь QR-алгоритм (4.36) от начала и до конца может быть построен на базе описанной выше процедуры преобразований Хаусхолдера, направленной сразу на триангуляризацию в соответствии с замечанием 4.10. Однако такой подход значительно сужает границы применимости алгоритма и ухудшает его скоростные качества.

Обычно на втором этапе применяют другое ортогональное преобразование – ***преобразование плоских вращений Гивенса***. Определяющая эти преобразования матрица $\mathbf{G}_{ij}=\left(g_{ml}\right)_{m,l=1}^{n}$ при фиксированных i,j – индексах ключевого элемента преобразуемой матрицы – имеет точно такую же структуру, как и матрица плоских вращений Якоби $\mathbf{T}_{ij}$ (ср.(4.27)), только здесь, следуя [41], двумерную подматрицу из элементов, стоящих на пересечении i-х и j-х строк и столбцов, возьмем в виде

$$\hat{\mathbf{G}}_{ij}=\begin{pmatrix} s & c \\ -c & s \end{pmatrix}.$$

Как и прежде, числа s и c связываем соотношением $s^2+c^2=1$ (это позволяет интерпретировать их как синус и косинус некоторого угла θ), обеспечивающим ортонормированность матриц $\mathbf{G}_{ij}$.

Первый полный шаг преобразования Гивенса, применяемого к матрице Хессенберга $\mathbf{B}$ n-го порядка в рамках QR-алгоритма (4.36), состоит из $n-1$ элементарных подшагов, имеющих целью последовательное обнуление поддиагональных элементов в столбцах от первого до $(n-1)$-го. В результате этого получается разложение матрицы $\mathbf{B}$ в произведение ортогональной и треугольной, что требуется первой формулой (4.36) при $k=1$, $\mathbf{A}_1:=\mathbf{B}$.

Чтобы определить s и c на первом промежуточном шаге, рассмотрим произведение матриц*)

$$\mathbf{G}_1\mathbf{B}=\begin{pmatrix} s & c & 0 & \dots & 0 \\ -c & s & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}\begin{pmatrix} b_{11} & b_{12} & b_{13} & \dots & b_{1n} \\ b_{21} & b_{22} & b_{23} & \dots & b_{2n} \\ 0 & b_{32} & b_{33} & \dots & b_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & b_{nn} \end{pmatrix}=$$

$$=\begin{pmatrix} sb_{11}+cb_{21} & sb_{12}+cb_{22} & sb_{13}+cb_{23} & \dots & sb_{1n}+cb_{2n} \\ -cb_{11}+sb_{21} & -cb_{12}+sb_{22} & -cb_{13}+sb_{23} & \dots & -cb_{1n}+sb_{2n} \\ 0 & b_{32} & b_{33} & \dots & b_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & b_{nn} \end{pmatrix}.$$

Беря $c=\cos\theta$ и $s=\sin\theta$ такими, что $\operatorname{tg}\theta=\dfrac{b_{11}}{b_{21}}$, будем иметь $b_{11}\cos\theta=b_{21}\sin\theta$, т.е. $-c_1b_{11}+sb_{21}=0$. Значит, результат первого промежуточного шага – матрица $\mathbf{B}_1=\mathbf{G}_1\mathbf{B}$, получающаяся при таких c и s, не будет содержать ненулевых элементов под диагональю в первом столбце.

Второй промежуточный шаг совершается аналогично: матрица $\mathbf{B}_2=\mathbf{G}_2\mathbf{B}_1$ получается из предыдущей $\mathbf{B}_1$ с помощью матрицы Гивенса $\mathbf{G}_2$, отличающейся от $\mathbf{G}_1$ тем, что подматрица $\begin{pmatrix} s & c \\ -c & s \end{pmatrix}$ смещается на одну позицию вдоль диагонали и угол поворота подбирается так, чтобы в матрице $\mathbf{B}_2$ обнулить элемент $b_{32}^{(2)}$.

Продолжая этот процесс преобразований Гивенса далее, в итоге получим правую треугольную матрицу

$$\mathbf{B}_{n-1}=\mathbf{G}_{n-1}\mathbf{G}_{n-2}\dots\mathbf{G}_2\mathbf{G}_1\mathbf{B}.$$

Последнее равенство можно переписать в виде

$$\left(\mathbf{G}_{n-1}\mathbf{G}_{n-2}\dots\mathbf{G}_2\mathbf{G}_1\right)^{-1}\mathbf{B}_{n-1}=\mathbf{B},$$

который позволяет считать выполненным требуемое в (4.36) при $k=1$ разложение

$$\mathbf{B}=\mathbf{Q}_1\mathbf{R}_1,$$

где $\mathbf{Q}_1:=\left(\mathbf{G}_{n-1}\mathbf{G}_{n-2}\dots\mathbf{G}_2\mathbf{G}_1\right)^{-1}=\mathbf{G}_1^T\dots\mathbf{G}_{n-1}^T$ – ортогональная, а

*) Так как в каждом столбце матрицы Хессенберга нужно "убивать" только по одному элементу, то в данном случае матрицы элементарных вращений Гивенса можно помечать только одним индексом

$\mathbf{R}_1 := \mathbf{B}_{n-1}$ – правая треугольная матрицы. При этом матрица

$$\mathbf{A}_2 := \mathbf{R}_1\mathbf{Q}_1 = (\mathbf{G}_{n-1}\ldots\mathbf{G}_1)\mathbf{B}(\mathbf{G}_{n-1}\ldots\mathbf{G}_1)^{-1}, \qquad (4.40)$$

являющаяся результатом первого полного шага QR-алгоритма (примененного к $\mathbf{B}$), сохраняет не только спектр данной матрицы, но и форму Хессенберга [16, 41], благодаря чему приведение исходной матрицы $\mathbf{A}$ к почти треугольному виду $\mathbf{B}$ достаточно сделать только один раз.

Очевидно, скалярные параметры $c_j=\cos\theta_j$ и $s_j=\sin\theta_j$ матриц Гивенса $\mathbf{G}_j$, с помощью которых осуществляется переход от матрицы Хессенберга $\mathbf{B}=\left(b_{ij}\right)_{i,j=1}^{n}$ "транзитом" через матрицы Хессенберга $\mathbf{B}_j=\left(b_{im}^{(j)}\right)_{i,m=1}^{n}$ к матрице Хессенберга $\mathbf{A}_2$, можно вычислять на j -м промежуточном шаге ($j=1, 2, \ldots, n-1$) по формулам

$$c_j = \frac{1}{\sqrt{1+t_j^2}}, \qquad s_j=t_jc_j,$$

где

$$t_j=\frac{b_{jj}^{(j-1)}}{b_{j+1,j}^{(j-1)}}\left(=\operatorname{tg}\theta_j\right), \qquad b_{ij}^{(0)} := b_{ij}$$

(если знаменатель в выражении t_j равен нулю или по модулю меньше некоторого существенно малого порогового значения, то можно считать $c_j=0$, $s_j=1$, т.е. $\mathbf{G}_j := \mathbf{E}$).

Пример 4.3. Преобразованиями Гивенса выполним один шаг QR-алгоритма для матрицы $\mathbf{B}=\begin{pmatrix} 5 & -3 & -1 \\ -5 & 2.08 & 0.56 \\ 0 & -0.44 & -1.08 \end{pmatrix}$, полученной в результате преобразований Хаусхолдера в предыдущем примере.

При $j=1$ последовательно находим:

$$t_1 = \frac{5}{-5} = -1, \qquad c_1 = \frac{1}{\sqrt{1+(-1)^2}} = \frac{1}{\sqrt{2}}, \qquad s_1 = -\frac{1}{\sqrt{2}};$$

$$\mathbf{G}_1=\begin{pmatrix} -0.5\sqrt{2} & 0.5\sqrt{2} & 0 \\ -0.5\sqrt{2} & -0.5\sqrt{2} & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$\mathbf{B}_1 = \mathbf{G}_1\mathbf{B}=\begin{pmatrix} -5\sqrt{2} & 2.54\sqrt{2} & 0.78\sqrt{2} \\ 0 & 0.46\sqrt{2} & 0.22\sqrt{2} \\ 0 & -0.44 & -1.08 \end{pmatrix}.$$

При j=2 вычисляем (округляя до 10^{-6}):

$$t_2 = \frac{0.46\sqrt{2}}{-0.44} = -1.478496, \quad c_2 = 0.560248, \quad s_2 = -0.828325;$$

значит,

$$\mathbf{G}_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -0.828325 & 0.560248 \\ 0 & -0.560248 & -0.828325 \end{pmatrix},$$

$$\mathbf{R}_1 := \mathbf{B}_2 = \mathbf{G}_2\mathbf{B}_1 = \begin{pmatrix} -7.071068 & 3.592102 & 1.103087 \\ 0 & -0.785366 & -0.862782 \\ 0 & 0 & 0.720283 \end{pmatrix},$$

Следовательно,

$$\mathbf{Q}_1 = \mathbf{G}_1^T\mathbf{G}_2^T = \begin{pmatrix} -0.707107 & -0.585714 & -0.396155 \\ 0.585714 & -0.396155 & 0.396155 \\ 0 & 0.560248 & -0.828325 \end{pmatrix},$$

и, согласно (4.40), матрица

$$\mathbf{A}_2 = \mathbf{R}_1\mathbf{Q}_1 = \begin{pmatrix} 7.103946 & 3.336597 & 3.310554 \\ -0.460000 & -0.172245 & 0.403537 \\ 0 & 0.403537 & -0.596628 \end{pmatrix}$$

есть искомый результат первого полного шага QR-алгоритма. Она имеет те же собственные числа, что **B** и **A**, сохраняет форму Хессенберга и модули ее поддиагональных элементов меньше, чем у матрицы **B**, т.е. она более близка к подобной **B** матрице треугольного вида, на диагонали которой должны быть собственные числа данной матрицы ($\lambda_1 \approx 7.693$, $\lambda_2 \approx -1.205$, $\lambda_3 \approx -0.435$).

Замечание 4.11. Весь QR-алгоритм можно было бы построить на базе одних только преобразований Гивенса, т.е. не приводя исходную матрицу **A** к форме Хессенберга (или к трехдиагональному виду, если **A** симметрична) другими преобразованиями. В таком случае стала бы заметной разница между преобразованиями Якоби и Гивенса. Суть этой разницы в следующем: если для преобразований Якоби понятия "ключевой элемент" и "обреченный элемент" совпадают, то для преобразований Гивенса это, вообще говоря, не так. В общем случае при вращениях Гивенса угол поворота θ в фиксированной индексами i, j плоскости вращения подбирается так, чтобы аннулировать какой-нибудь элемент, стоящий либо в одном столбце, либо в одной строке с ключевым элементом a_{ij}. Такие преобразования теряют свойство минимальности суммы квадратов внедиагональных элементов, имевшее место в преобразованиях Якоби для симметричных матриц, но позволяют (Гивенс, 1954г. [53]) привести симметричную мат-

рицу к трехдиагональному виду существенно быстрее, чем это требуется для выполнения одного цикла преобразований в методе вращений Якоби*). Приведение несимметричных матриц к форме Хессенберга методом Гивенса требует большего числа арифметических операций, чем это нужно для такого приведения методом Хаусхолдера, поэтому обычно для этих целей отдают предпочтение последнему.

Приведенных выше сведений, в принципе, вполне достаточно, чтобы находить QR-алгоритмом все хорошо отделимые вещественные собственные числа вещественных матриц, реализуя равенство (4.37) при некотором k, хотя здесь и нет должного для этого обоснования**). Однако в такой непосредственной форме QR-алгоритм не применяется ввиду его медленной сходимости. Для ускорения сходимости в процесс (4.36) вводят ***сдвиги***, о роли которых немало говорилось при изучении метода обратных итераций (см. п.4.3). При выборе параметров сдвигов в QR-алгоритме учитывается подмеченная ранее целесообразность использования для этого приближений к собственным числам. В данном случае принимается во внимание, что формируемая QR-алгоритмом последовательность матриц $\mathbf{A}_k$ в пределе дает правую треугольную матрицу с диагональю из собственных чисел (когда все собственные числа матрицы $\mathbf{A}$ вещественны). Следовательно, есть основания утверждать, что последовательность элементов $a_{nn}^{(k)}$ может рассматриваться при $k=1, 2, \ldots$ как последовательность приближений к какому-то определенному собственному числу матрицы $\mathbf{A}$ и служить соответствующей последовательностью параметров переменных сдвигов, ускоряющих процесс обнуления поддиагональных элементов.

Таким образом, по крайней мере, в случае, когда данная матрица $\mathbf{A}$ имеет только вещественные собственные значения, QR-алгоритм будет сходиться более быстро (квадратично), если при каждом k преобразования Гивенса применять не к матрице $\mathbf{A}_k$, а к матрице $\widetilde{\mathbf{A}}_k := \mathbf{A}_k - a_{nn}^{(k)}\mathbf{E}$. При этом каждый раз спектр матрицы смещается на величину произведенного сдвига (см. свойство 2 в п.4.1), что может учитываться двояко: либо параметры сдвигов суммируются и затем сумма прибавляется к найденным в итоге значениям, либо каждый раз делается обратный сдвиг, т.е. вместо формул (4.36) используются формулы:

$$\mathbf{A}_k - a_{nn}^{(k)}\mathbf{E} = \mathbf{Q}_k\mathbf{R}_k$$

(QR-факторизация матрицы $\mathbf{A}_k - a_{nn}^{(k)}E$),

---

*) Для вычисления собственных значений трехдиагональных матриц разработан не рассматриваемый здесь довольно эффективный ***метод бисекций*** [11, 15, 53], который также можно считать одним из способов локализации собственных чисел.

**) Доказательство сходимости QR-алгоритма и его модификаций см. в [26, 53]. В [44] изучается сходимость принятого там за основу QL-алгоритма.

$$\mathbf{A}_{k+1} = \mathbf{R}_k\mathbf{Q}_k + a_{nn}^{(k)}\mathbf{E}$$

(перемножение в обратном порядке и обратный сдвиг).

Важно учесть, что большая экономия вычислительных затрат при реализации QR-алгоритма получается в результате включения сюда ***процедуры исчерпывания***. Нетрудно показать, что если на каком-то шаге $k=k_0$ последняя строка матрицы n-го порядка $\mathbf{A}_{k_0}$ стала нулевой (с некоторой точностью), то это позволяет считать $a_{nn}^{(k_0)}$ собственным числом, а другие ее собственные значения находить, работая далее только с подматрицей $(n-1)$-го порядка, получающейся из $\mathbf{A}_{k_0}$ отбрасыванием последних строки и столбца.

Если у данной матрицы возможны комплексные собственные значения, применяют более сложный процесс двойных сдвигов [26, 41, 53], позволяющий преодолеть ситуацию, когда в конце диагоналей матриц $\mathbf{A}_k$ "прорисовывается" 2×2-блок, соответствующий паре комплексно сопряженных собственных чисел.

Нахождение собственных векторов в рамках QR-алгоритма (и других методов, основанных на асимптотической триангуляризации) не является такой простой задачей, как это было в методе вращений Якоби для симметричных матриц, диагонализируемых в процессе ортогональных преобразований. Однако при известном собственном числе соответствующий ему собственный вектор эффективно может быть найден рассмотренным в п.4.3 методом обратных итераций (см. формулы (4.19),(4.20)). При этом обратные итерации обычно применяются не к исходной матрице $\mathbf{A}$, а к матрице $\mathbf{B}$, подобной $\mathbf{A}$ и имеющей форму Хессенберга. Если приведение $\mathbf{A}$ к виду $\mathbf{B}$ выполнялось преобразованиями Хаусхолдера (4.38), то $\mathbf{B} = \mathbf{HAH}$, где $\mathbf{H}$ – результирующая матрица $n-2$ элементарных вращений, и значит, согласно свойству 7 п.4.4, найдя собственный вектор $\mathbf{y}$ матрицы $\mathbf{B}$, искомый собственный вектор $\mathbf{x}$ данной матрицы $\mathbf{A}$ получаем равенством $\mathbf{x} = \mathbf{Hy}$.

УПРАЖНЕНИЯ

4.1. Дана матрица $\mathbf{A} = \begin{pmatrix} 30 & -12 & 53 \\ -42 & 19 & -78 \\ -28 & 12 & -51 \end{pmatrix}$.

а) Степенным методом найти несколько последовательных приближений к доминирующему собственному числу матрицы $\mathbf{A}$ и к соответствующему собственному вектору. Зная, что искомое собственное число есть

$\lambda_1 = -5$, проверить, насколько эффективно здесь применение δ^2- процесса Эйткена (см. замечание 4.6).

б) Методом обратных итераций найти младшую собственную пару $\{\lambda_3, \mathbf{x}_3\}$ данной матрицы $\mathbf{A}$. Можно ли утверждать, что $\lambda_3 = \Lambda + \lambda_1$, где Λ – наибольшее по модулю собственное число матрицы $\mathbf{A} - \lambda_1 \mathbf{E}$?

4.2. Найдя грубые приближения к собственным числам матрицы

$$\mathbf{A} = \begin{pmatrix} 5 & 2 & -3 \\ 4 & 5 & -4 \\ 6 & 4 & -4 \end{pmatrix}$$

степенным методом, уточнить эти значения обратными итерациями со сдвигами (см. (4.24)).

4.3.а) Проанализировать сходимость степенного метода в случае, когда λ_1 – кратное вещественное наибольшее по модулю собственное число n-мерной матрицы простой структуры (см. формулы (4.10), (4.11)). Как можно найти все соответствующие ему собственные векторы в зависимости от показателя кратности?

б) Что можно сказать о поведении последовательности отношений (4.10), если $\lambda_1 = -\lambda_2$, и $|\lambda_2| > |\lambda_3| \geq \ldots \geq |\lambda_n| \quad (\lambda_i \in R)$?

в) Рассмотреть и объяснить поведение степенного метода в случае, когда данная матрица $\mathbf{A}$ – диагональная.

4.4. Найти все собственные пары матрицы

$$\mathbf{A} = \begin{pmatrix} 4 & 2 & -1 \\ 2 & 4 & 1 \\ -1 & 1 & 3 \end{pmatrix}:$$

а) методом скалярных произведений (для нахождения второй собственной пары использовать формулы (4.14), (4.15));

б) RQI-алгоритмом, начиная его с различных векторов.

4.5. Сравнить два подхода к нахождению всех собственных чисел матрицы

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$$

методом вращений Якоби:

а) применяя его непосредственно к данной матрице;

б) предварительно приведя ее к трехдиагональному виду преобразованиями Хаусхолдера.

4.6. Для нахождения собственных пар симметричных положительно определенных матриц построить LU - алгоритм на базе $\mathbf{U}^T\mathbf{U}$ (или $\mathbf{L}\mathbf{L}^T$)-разложения Холецкого (см. главу 2). Опробовать его на матрицах

$$\mathbf{A}=\begin{pmatrix} 5 & -2 \\ -2 & 2 \end{pmatrix} \qquad \text{и} \qquad \mathbf{B}=\begin{pmatrix} 6 & 1 & 0 \\ 1 & 5 & -2 \\ 0 & -2 & 1 \end{pmatrix}.$$

Сохраняют ли получаемые на каждом шаге такого алгоритма подобные $\mathbf{B}$ матрицы трехдиагональную структуру?

4.7. Дана матрица $\mathbf{A}=\begin{pmatrix} 3 & -1 \\ -2 & 2 \end{pmatrix}$. Сделать по три шага:

а) LU-алгоритма;

б) QR-алгоритма на основе преобразований Гивенса;

в) QR-алгоритма на основе преобразований Хаусхолдера (с учетом замечания 4.10).

Сравнить полученные приближения к собственным числам матрицы $\mathbf{A}$ по точности (найдя сначала ее точные значения с помощью характеристического уравнения) и по вычислительным затратам.

4.8. Матрицу $\mathbf{A}$ из упражнения 4.5 привести к трехдиагональному виду преобразованием Гивенса (отличным от преобразования Якоби, см. замечание 4.11).

4.9. Для $n\times n$-матрицы сделать приблизительный подсчет количества арифметических операций, приходящихся на:

а) один шаг степенного метода;

б) один шаг метода обратных итераций (без сдвигов);

в) один полный цикл метода вращений Якоби;

г) полный цикл приведения матрицы к форме Хессенберга преобразованиями Хаусхолдера.

4.10. Методом отражений решить систему

$$\begin{cases} x+2y+z=0, \\ 2x-y-2z=6, \\ 2x+y-z=9 \end{cases}$$

(см. замечание 4.10).

4.11. Вывести формулы для приведения $n\times n$-матрицы Хессенберга к треугольной форме плоскими поворотами на углы φ_j $(j=1.2,\ldots,n-1)$ с помощью матриц $\mathbf{T}_j:=\mathbf{T}_{jj+1}$ вида (4.27). Как отличаются углы φ_j от углов θ_j, неявно присутствующих в представлении (4.40)?

ГЛАВА 5 МЕТОДЫ РЕШЕНИЯ НЕЛИНЕЙНЫХ СКАЛЯРНЫХ УРАВНЕНИЙ

Рассматривается задача вычисления действительных корней алгебраических и трансцендентных уравнений. Изучаются как универсальные методы (дихотомии, Ньютона, секущих) решения нелинейных уравнений, так и специальные подходы к нахождению корней многочленов. Большое внимание уделено методу простых итераций и базирующимся на нем Δ^2-процессу Эйткена и методу Вегстейна. На примере логистического уравнения (с параметром) исследуется возможное поведение итерационных последовательностей в случае нарушения одного из достаточных условий сходимости, дается представление о бифуркациях решений и циклов.

5.1. ЛОКАЛИЗАЦИЯ КОРНЕЙ

Будем рассматривать ***задачу приближенного нахождения нулей функции одной переменной***, иначе, задачу нахождения корней уравнения вида

$$f(x)=0, \qquad (5.1)$$

где $f: R_1 \to R_1$ – алгебраическая или трансцендентная функция. Такие уравнения называют ***скалярными, числовыми, конечными*** и т.п. Методам их решения посвящена обширная литература; кроме перечисленных здесь учебных и научных изданий, эту тему можно встретить почти в любом учебнике математического анализа или высшей математики, а во многих учебных пособиях по программированию даются "рецепты" решения таких задач*). Однако не все авторы учебников по вычислительной математике считают нужным включать рассматриваемую тему самостоятельным разделом. И в этом есть резон. Действительно, многие наиболее фундаментальные методы решения скалярных уравнений можно рассматривать как частные случаи соответствующих методов решения систем нелинейных уравнений со многими неизвестными и даже, более того, нелинейных операторных уравнений (как правило, в банаховых пространствах). Такой путь изучения методов более короток, но и более труден. Учитывая, что одномерный случай более прост и легко интерпретируется геометрически

*) Можно понять преподавателей, обучающих алгоритмическим языкам: методы решения скалярных уравнений предоставляют для этого благодатный материал. К сожалению, программирование формул без понимания их сути часто лишь компрометирует вычислительную математику.

(что немаловажно для нелинейных задач), а также то, что обычно теоретические результаты, полученные для скалярных уравнений, затем переносятся (не без потерь) на системы и операторные уравнения, изучаемые методы будут гораздо лучше поняты, если в них сначала как следует разобраться на объектах вида (5.1).

В общем случае, если речь идет не об отдельных достаточно узких классах уравнений, например, изучавшихся в школьном курсе математики, можно говорить лишь о приближенном вычислении корней уравнений (5.1), т.е. таких значений аргумента $x=\xi$, при которых равенство

$$f(\xi)=0$$

истинно. При этом под близостью приближенного значения $\bar{x}$ к корню ξ уравнения (5.1), как правило, понимают выполнение неравенства

$$|\xi-\bar{x}|<\varepsilon$$

при малых $\varepsilon>0$, хотя часто бывает важным контролировать не абсолютную погрешность приближенного равенства $\bar{x}\approx\xi$, а относительную, т.е. величину $|\xi-\bar{x}|/|\bar{x}|$, например, когда величина $|\bar{x}|$ близка к нулю.

Нелинейная функция $f(x)$ в своей области определения $D(f)\subseteq R_1$ может иметь конечное или бесконечное количество нулей или не иметь их вовсе. Большинство же методов нахождения нулей требует знания промежутков (возможно, малых), где имеется и притом единственный нуль функции*). Если такие конкретные промежутки не предоставляются постановщиком задачи, то на первый план выходят: выявление ситуации с наличием и количеством корней уравнения (5.1), нахождение области их расположения, получение отрезков, на которых имеется точно по одному корню. Иными словами, ставятся ***подзадачи существования и единственности, нахождения границ и локализации корней***. Эти подзадачи обычно решаются в комплексе средствами математического анализа. Но и численные методы здесь часто выступают в помощь математическому анализу: как будет видно из дальнейшего, многие теоремы сходимости итерационных методов можно считать локальными теоремами существования и единственности.

Для функций общего вида нет универсальных способов решения поставленных подзадач. Так что здесь, как говорится, все средства хороши.

Если функция $f(x)$ такова, что без особого труда можно построить ее график, этим следует воспользоваться, чтобы представить ситуацию с количеством и расположением нулей $f(x)$, выделяя те промежутки оси абсцисс, где график $y=f(x)$ пересекает Ox. (Знание графика много дает и для понимания поведения тех или иных процессов вычисления приближе-

*) Перефразируя Конфуция, можно сказать: трудно найти корень на бесконечном промежутке, особенно когда его там нет.

ний к корням.) Может оказаться, что построение графика $y=f(x)$ вызывает затруднения, но исходное уравнение (5.1) очевидным образом представляется в виде

$$f_1(x)=f_2(x)$$

и функции $f_1(x)$ и $f_2(x)$ таковы, что легко строятся графики $y=f_1(x)$ и $y=f_2(x)$. Тогда задача определения количества корней и областей их единственности решается отслеживанием точек пересечения этих графиков и выделением на оси абсцисс тех промежутков, которым принадлежат проекции таких точек. Описанный прием называют ***графическим способом локализации*** (иначе, ***отделения***, ***изоляции***) ***корней***.

Пример 5.1. Найдем промежутки изоляции корней уравнения

$$x^2 - \sin x - 1 = 0.$$

Представив это уравнение в виде $x^2 - 1 = \sin x$, строим схематично графики функций $y = x^2 - 1$ и $y = \sin x$ (рис. 5.1).

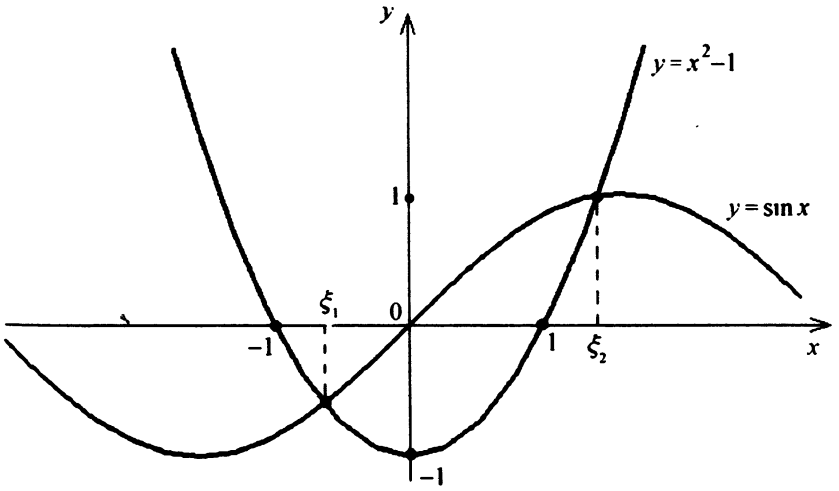


Рис.5.1. Графическая локализация корней уравнения $x^2 - \sin x - 1 = 0$

Совместное рассмотрение графиков позволяет сделать заключение, что данное уравнение имеет два корня: $\xi_1 \in [-1;0]$ и $\xi_2 \in [1;\pi]$.

Убедиться в том, что на данном отрезке $[a;b]$ (например, грубо определенном графическим способом) действительно имеется нуль непрерывной функции $f(x)$, можно ***аналитическим способом***, в основе которого лежит известное утверждение математического анализа.

Теорема 5.1 (***Больцано-Коши***). *Если непрерывная на отрезке* $[a;b]$ *функция* $f(x)$ *на концах его имеет противоположные знаки, т.е.*

$$f(a)f(b)<0, \qquad (5.2)$$

то на интервале $(a;b)$ *она хотя бы один раз обращается в нуль*[*].

Очевидна слабость теоремы 1 при ее применении к поставленной задаче: она не дает ответа на вопрос о количестве корней на отрезке $[a;b]$ в случае выполнения условия (5.2) (их может быть нечетное число) и не позволяет утверждать, что на отрезке $[a;b]$ нет корней, когда условие (5.2) не выполнено, так как в этом случае их может оказаться на $[a;b]$ четное количество (нетрудно представить себе соответствующие "картинки" возможных поведений графиков функций на отрезке).

Результат, сформулированный в виде теоремы 1, можно значительно усилить, если требование непрерывности функции $f(x)$ на $[a;b]$ дополнить требованием монотонности ее на этом отрезке.

Теорема 5.2. *Непрерывная строго монотонная функция* $f(x)$ *имеет и притом единственный нуль на отрезке* $[a;b]$ *тогда и только тогда, когда на его концах она принимает значения разных знаков.*

Последняя теорема позволяет не только принимать, но и отвергать те или иные промежутки из области определения $D(f)$ данной функции на предмет дальнейшего поиска ее нулей, если известно о ее монотонном поведении на этих промежутках и определены знаки значений функции на их концах.

Реально установить монотонность на данном отрезке можно для дифференцируемой функции, потребовав знакопостоянство ее производной на всем отрезке. Для таких функций основой решения задачи локализации корней уравнения (5.1) может служить следующая теорема.

---

[*] Допустимо считать, что a и/или b могут принимать значения $-\infty$ и/или $+\infty$ соответственно. В таком случае $f(a)$ и/или $f(b)$ следует понимать в предельном смысле, также допуская возможность бесконечных предельных значений f определенного знака.

Теорема 5.3. *Пусть* $f \in C^1_{[a;b]}$. *Тогда если* $f'(x)$ *не меняет знак на* $(a;b)$, *то условие* (5.2) *является необходимым и достаточным для того, чтобы уравнение* (5.1) *имело и притом единственный корень на* $[a;b]$.

Так как производная может менять знак только в точках, где она равна нулю или не существует, а также в граничных точках области определения функции, то в случаях (к сожалению, редких), когда уравнение $f'(x)=0$ легко решается, вопрос о числе и расположении корней уравнения (5.1) не вызывает трудностей.

Пример 5.2. Выясним, сколько корней у уравнения $x^2e^x=\pi$ и где они расположены.

Обозначим $f(x):=x^2e^x-\pi$. Тогда $f'(x)=x(x+2)e^x$. Очевидно, $f'(x)=0$ только при $x=0$ и $x=-2$. Поскольку $f(x)$ и $f'(x)$ определены и непрерывны на всей числовой оси, точки -2 и 0 – единственные на Ox такие, в которых может происходить смена убывания функции $y=f(x)$ на возрастание или наоборот. Поэтому, найдя знаки значений (в том числе, бесконечных) $\lim\limits_{x\to-\infty} f(x)$, $f(-2)$, $f(0)$ и $\lim\limits_{x\to+\infty} f(x)$, т.е. заполнив таблицу знаков

$$f:\ \begin{array}{c|c|c|c|c} -\infty & -2 & 0 & +\infty \\ \hline - & - & + & + \end{array},$$

можно на основании теоремы 5.3 утверждать, что данное уравнение имеет единственный корень, и этот корень положителен. Выяснив еще знаки $f(x)$ в точках $x=1$ (минус) и $x=2$ (плюс), область поиска корня данного уравнения с бесконечного промежутка $[0;+\infty)$ сужаем до промежутка единичной длины $[1;2]$.

В ситуациях, далеких от рассмотренной идеальной, часто поступают следующим образом. Всю область определения (если она конечна) или какую-нибудь ее часть, вызывающую по тем или иным соображениям интерес, разбивают на отрезки точками x_i, расположенными на условно небольшом расстоянии h одна от другой (сюда включаются также граничные точки области определения). Вычислив значения $f(x)$ во всех этих точках (или только определив знаки $f(x_i)$), сравнивают их в соседних точках, т.е. проверяют, не выполнится ли на отрезке $[x_{i-1};x_i]$ условие $f(x_{i-1})f(x_i)\le 0$. Если заранее известно количество корней в исследуемой области, то, измельчая шаг поиска h, таким процессом можно либо все их

локализовать, либо довести процесс до состояния, позволяющего утверждать, что возможно наличие пар корней, не различимых с точностью $h = \varepsilon$. Этот хорошо приспособленный для вычислительных машин ***способ перебора*** является дорогим в смысле затрат на получение многочисленных пробных значений функции и не дает гарантий выявления количества и локализации всех корней в общем случае, что ограничивает сферу его применения.

Одна из проблем, в которую упирается решение задачи локализации корней, – это практическая невозможность точного вычисления значений функций. Она уже обсуждалась в главе 1. Вернемся к ней еще раз.

Из-за ограниченности разрядной сетки ЭВМ даже алгебраические функции вычисляются приближенно, вычисление же трансцендентных функций составляет тему отдельного разговора; ясно, что их значения за редкими исключениями по определению могут записываться лишь приближенно. Погрешности, с которыми вычисляются значения функции, порождают блуждания (иначе, флуктуации, случайные отклонения) ординат графика около средних значений. Если рассматривать небольшой участок графика какой-либо функции, построенной путем изображения всех ее точек, вычисленных на компьютере при дискретном изменении аргумента с очень мелким шагом h, с последующим соединением этих точек отрезками прямых, то окажется, что этот реальный график имеет пилообразную форму (рис. 5.2).

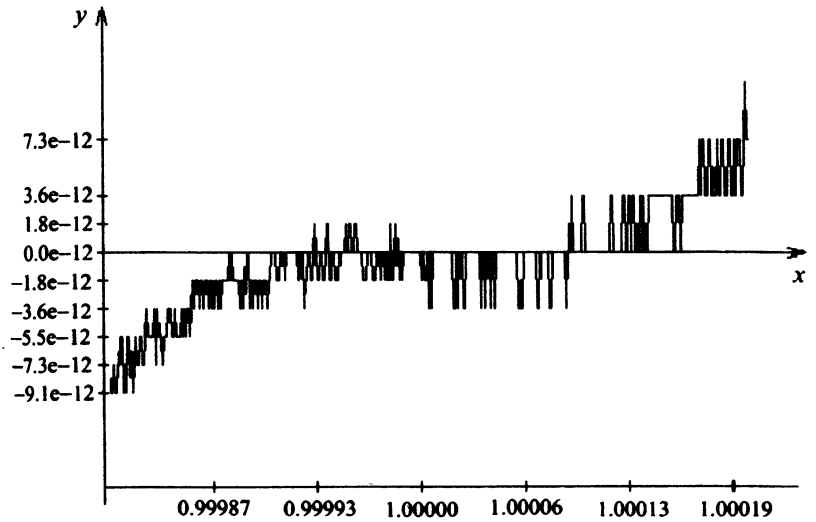


Рис. 5.2. Участок графика функции $y=x^3-3x^2+3x-1$, *полученный на компьютере DX5-133 вычислением ее значений с точностью* $O\left(10^{-12}\right)$ *с шагом* $O\left(10^{-6}\right)$

Эти вычислительные погрешности ограничивают точность, с которой можно находить корень. В результате их влияния та или иная процедура сужения промежутка локализации корня рано или поздно приведет к такому промежутку, называемому ***промежутком (отрезком, интервалом) неопределенности корня***, что любое число из этого промежутка с одинаковым успехом может быть принято за корень уравнения (ср. с промежутком $(\xi - \text{б.п.}\,\xi,\ \xi + \text{б.п.}\,\xi)$, который фигурировал в п.1.6 при рассмотрении затронутой проблемы в несколько иной постановке).

Наличие флуктуаций значений функций, порождающих невозможность точного вычисления корней нелинейных уравнений, следует всегда иметь в виду, применяя те или иные численные методы. Если не с самого начала итерационного процесса, то с приближением момента, когда погрешность метода становится сравнимой с погрешностью вычисления значений функций или выражений, входящих в расчетные формулы метода, целесообразно подключение приема Гарвика, упоминавшегося ранее (см. замечание 4.4).

5.2. МЕТОД ДИХОТОМИИ. МЕТОД ХОРД

Пусть функция $f(x)$ определена и непрерывна при всех $x \in [a;b]$ и на $[a;b]$ меняет знак, т.е. $f(a)f(b)<0$. Тогда согласно теореме 5.1 уравнение (5.1) имеет на $(a;b)$ хотя бы один корень. Возьмем произвольную точку $c \in (a;b)$. Будем называть в этом случае отрезок $[a;b]$ ***промежутком существования корня***, а точку c – ***пробной точкой***. Поскольку речь здесь идет лишь о вещественнозначных функциях вещественной переменной, то вычисление значения $f(c)$ приведет к какой-либо одной из следующих взаимоисключающих ситуаций:

а) $f(a)f(c)<0$; б) $f(c)f(b)<0$; в) $f(c)=0$.

Применительно к рассматриваемой задаче их можно интерпретировать так*):

а) корень находится на интервале $(a;c)$;

б) корень находится на интервале $(c;b)$;

в) точка c является искомым корнем.

Таким образом, одно вычисление значения функции позволяет уменьшить промежуток $[a;b]$ существования корня (ситуация а) или б)) или указать его значение (ситуация в), маловероятная в смысле "прямого попадания" пробной точкой c в корень, но вполне реальная в смысле выполнения при-

*) Из-за допустимости неединственности корня в этой интерпретации уже нет взаимоисключаемости ситуаций.

ближенного равенства $f(c) \approx 0$, когда длина промежутка существования корня близка к длине промежутка его неопределенности). Ясно, что в зависимости от того, имеет место ситуация а) или б), описанная процедура одного шага сужения промежутка существования нуля непрерывной функции $f(x)$ может быть применена к промежутку $[a;c] \subset [a;b]$ или к $[c;b] \subset [a;b]$ соответственно и далее повторяться циклически. Такой простой и легко программируемый процесс называется ***методом дихотомии*** (от греческого слова, означающего деление на две части), ***методом бисекции***, ***методом вилки***, ***методом проб***. Если способ задания пробных точек c определен так, что последовательность длин получающихся в этом процессе промежутков существования корня стремится к нулю, то методом дихотомии можно найти какой-либо корень уравнения (5.1) с наперед заданной точностью.

Наиболее употребительным частным случаем метода дихотомии является ***метод половинного деления***, реализующий самый простой способ выбора пробной точки – деление промежутка существования корня пополам. Выполнить приближенное вычисление с точностью ε корня уравнения (5.1) методом половинного деления при условии, что $f(x)$ непрерывна на $[a;b]$ и $f(a)f(b)<0$, можно, например, по следующей схеме:

Шаг 0. Задать концы отрезка a и b, функцию f, малое число $\varepsilon > 0$ (допустимую абсолютную погрешность корня или полудлину его промежутка неопределенности), малое число $\delta > 0$ (допуск, связанный с реальной точностью вычисления значений данной функции);
вычислить (или ввести) $f(a)$.

Шаг 1. Вычислить $c := 0.5(a+b)$.

Шаг 2. Если $b - a<2\varepsilon$, положить $\xi := \approx c$ (ξ – корень) и остановиться.

Шаг 3. Вычислить $f(c)$.

Шаг 4. Если $f(c) < \delta$, положить $\xi := \approx c$ и остановиться.

Шаг 5. Если $f(a)f(c) < 0$, положить $b := c$ и вернуться к шагу 1;
иначе положить $a := c$, $f(a) := f(c)$ и вернуться к шагу 1.

Замечание 5.1. В упрощенных вариантах схем реализации метода половинного деления обходятся без введения допуска δ. В таком случае в шаге 4 вместо неравенства $|f(c)| < \delta$ используют равенство $f(c) = 0$. Тогда разветвление алгоритма, диктуемое шагами 4 и 5, можно производить сравнением с нулем (с тремя исходами) произведения $f(a)f(c)$.

За один шаг метода половинного деления промежуток существования корня сокращается ровно вдвое. Поэтому, если за k-е приближение этим методом к корню ξ уравнения (5.1) примем точку x_k, являющуюся серединой полученного на k-м шаге отрезка $[a_k;b_k]$ в результате последовательного сужения данного отрезка $[a;b]$, полагая $a_1 := a$, $b_1 := b$, то придем к неравенству

$$|\xi - x_k| < \frac{b-a}{2^k} \qquad \forall k \in N \qquad (5.3)$$

(априори, ξ – любая точка интервала $(a_k;b_k)$, и расстояние от нее до середины этого интервала не превосходит половины его длины. Это как раз и видим в (5.3) при $k = 1$).

Неравенство (5.3), с одной стороны, позволяет утверждать, что последовательность (x_k) имеет предел – искомый корень ξ уравнения (5.1); с другой стороны, являясь априорной оценкой абсолютной погрешности приближенного равенства $x_k \approx \xi$, дает возможность подсчитать число шагов (итераций) метода половинного деления, достаточное для получения корня ξ с заданной точностью ε, для чего нужно лишь найти наименьшее натуральное k, удовлетворяющее неравенству

$$\frac{b-a}{2^k} < \varepsilon.$$

Используемый в методе половинного деления способ фиксирования пробной точки можно охарактеризовать как пассивный, ибо он осуществляется по заранее жестко заданному плану и никак не учитывает вычисляемые на каждом шаге значения функции. Логично предположить, что в семействе методов дихотомии можно достичь несколько лучших результатов, если отрезок $[a;b]$ делить точкой c на части не пополам, а пропорционально величинам ординат $f(a)$ и $f(b)$ графика данной функции $f(x)$. Это означает, что точку c есть смысл находить как абсциссу точки пересечения оси Ox с прямой, проходящей через точки $A(a;f(a))$ и $B(b;f(b))$, иначе, с хордой AB дуги $A\xi B$ (рис. 5.3).

Запишем уравнение прямой, проходящей через две данные точки A и B:

$$\frac{y - f(a)}{f(b) - f(a)} = \frac{x-a}{b-a}.$$

Отсюда, полагая $y = 0$ (уравнение оси Ox), $x = c$ (обозначение искомой точки пересечения прямой AB с осью Ox), находим

$$c = a - \frac{f(a)(b-a)}{f(b) - f(a)}. \qquad (5.4)$$

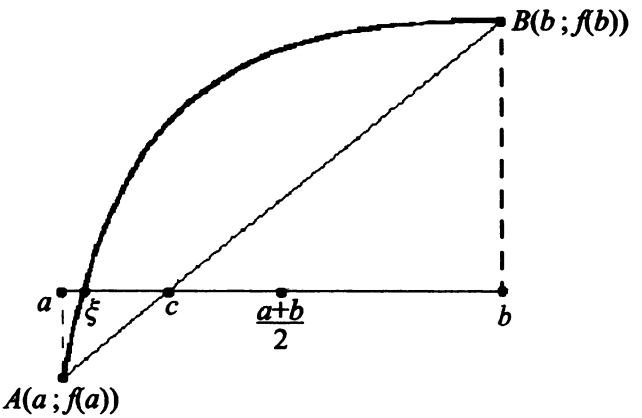


Рис. 5.3. Дуга графика функции $f(x)$ и стягивающая ее хорда

Метод, получающийся из метода дихотомии таким фиксированием пробной точки, называют ***методом хорд***, ***методом пропорциональных частей***, ***методом линейной интерполяции***. Все названия метода вполне естественны и отражают различные подходы к его выводу или интерпретации. Иногда (в последнее время реже) используют еще и название ***правило ложного положения*** или ***regula falsi*** [1, 23, 43].

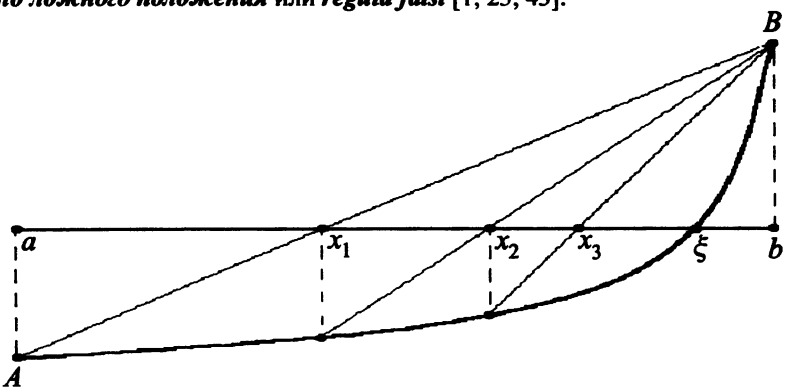


Рис. 5.4. Приближения к корню нелинейного уравнения по методу хорд

Существует несколько версий реализации метода хорд. Одна из них – подсчет значений c по формуле (5.4) в рамках алгоритма типа рассмотренного выше алгоритма половинного деления, где следует положить $f(b) := f(c)$ при $f(a)f(c) < 0$. Длина промежутка локализации корня при этом может не стремиться к нулю, поэтому обычно счет ведется до совпа-

дения значений c на двух соседних итерациях с точностью ε (лучше с точностью $\frac{m\varepsilon}{M-m}$, если $0 < m \le |f'(x)| \le M \quad \forall x \in [a;b]$).

Так как для линейной функции $f(x)$ метод хорд дает корень ξ точно за один шаг при любой длине отрезка $[a;b]$, то можно рассчитывать на его довольно быструю сходимость, если $f(x)$ близка к линейной. При определенных достаточно жестких условиях можно доказать соответствующие утверждения о монотонной сходимости, получить более точные оценки и упростить алгоритм (см., например, [21]). Однако в общем случае, если на функцию $f(x)$ не накладывать дополнительных ограничений, может оказаться, что метод хорд будет проигрывать в скорости методу половинного деления; чтобы убедиться в этом, достаточно взглянуть на рис. 5.4, демонстрирующий возможное поведение нескольких приближений по методу хорд.

5.3. ТИПЫ СХОДИМОСТЕЙ ИТЕРАЦИОННЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Чтобы более объективно судить о скорости сходимости тех или иных итерационных методов, введем следующие понятия [22 и др.].

Пусть некоторый итерационный процесс генерирует последовательность $(x_k)_{k=0}^{\infty}$, имеющую пределом x^*.

Определение 5.1. *Сходимость последовательности (x_k) к x^* называется **линейной** (соответственно, итерационный процесс – **линейно сходящимся**), если существует такая постоянная $C \in (0;1)$ и такой номер k_0, что*

$$|x^* - x_{k+1}| \le C|x^* - x_k| \qquad \forall k \ge k_0, \tag{5.5}$$

*и **сверхлинейной**, если существует такая положительная последовательность $(C_k)_{k=0}^{\infty}$, что $C_k \underset{k\to\infty}{\to} 0$ и*

$$|x^* - x_{k+1}| \le C_k|x^* - x_k| \qquad \forall k \in N_0. \tag{5.6}$$

Определение 5.2. *Говорят, что последовательность (x_k) **сходится к x^* по меньшей мере с p-м порядком** (соответственно, **итерационный процесс имеет по меньшей мере p-й порядок**), если найдутся такие константы $C > 0$ и $p \ge 1$, что*

$$\left|x^* - x_{k+1}\right| \le C\left|x^* - x_k\right|^p \qquad (5.7)$$

при всех $k \in N_0$, *начиная с некоторого* $k = k_0$.

Фиксируя в определении 5.2 значение $p = 1$, видим, что линейно сходящийся процесс можно называть ***процессом первого порядка***; значению $p = 2$ в (5.7) соответствует ***квадратично сходящийся процесс***, $p = 3$ означает ***кубическую сходимость***.*)

К линейной сходимости применяют также термин ***сходимость со скоростью геометрической прогрессии***. Объяснение ему можно найти в том, что определяющее линейную сходимость неравенство (5.5) между абсолютными погрешностями $(k+1)$-го и k-го приближений к предельной точке x^* означает существование последовательности положительных чисел ε_k, мажорирующих эти погрешности и связанных соотношением $\varepsilon_{k+1} = C\varepsilon_k$, т.е. являющихся членами геометрической прогрессии со знаменателем $C = \varepsilon_{k+1}/\varepsilon_k$ (=const). Отсюда следует также естественность в определении 5.1 условия $C < 1$, чтобы последовательность погрешностей была убывающей, иначе и речи не может быть о сходимости (в определении 5.2 для предельного случая $p = 1$ также следует ограничить C единицей; при $p > 1$ в этом, вообще говоря, нет необходимости; проанализируйте, почему?).

Если требуемой в неравенстве (5.5) константы C не удается найти, но установлено неравенство (5.6) с $C_k \underset{k\to\infty}{\to} C \in (0;1)$, то в этом случае говорят об ***асимптотически линейной сходимости*** (пример такой сходимости дает степенной метод в п.4.4 или рассматриваемый далее в п.5.10 метод Бернулли). Аналогично можно определить ***асимптотически*** p***-й порядок***.

Имеются и другие понятия и термины, позволяющие более тонко классифицировать сходимость итерационных последовательностей и рассматривать ее как бы под разными углами зрения (см., например, [42, 43, 50, 52]). Как правило, они вводятся в более общем случае конечномерных или бесконечномерных нормированных пространств. Ничто не мешает и данные здесь определения распространить на многомерный случай, достаточно лишь элементы последовательности (x_k) и предельный элемент x^* считать n-мерными векторами $(n \ge 1)$ или матрицами, а вместо модуля использовать норму.

---

*) В качестве известных примеров методов высоких порядков в пространстве R_n следует обратить внимание на изученный в п.3.6 класс итерационных процессов обращения матриц.

Среди нескольких способов охарактеризовать скорость сходимости итерационных последовательностей наиболее четко оформились два способа: опирающиеся на ***q-сходимость*** (от англ. quotient – частное) и на ***r-сходимость*** (от англ. root – корень). Происхождение этих терминов можно связать соответственно с признаками Даламбера (через отношение) и Коши (через арифметический корень), применяемыми для установления абсолютной сходимости ряда

$$x_0 + (x_1 - x_0) + (x_2 - x_1) + \ldots + (x_k - x_{k-1}) + \ldots . \qquad (5.8)$$

что равносильно установлению сходимости данной последовательности (x_k), так как сходимость ряда (5.8) означает существование предела его частичных сумм

$$S_1 = x_0, \quad S_2 = x_1, \quad S_3 = x_2, \ \ldots \ , \quad S_{k+1} = x_k, \ldots .$$

Четкие определения и различия между этими двумя типами сходимостей можно найти в [42]; более существенно эти различия проявляются в многомерном случае. Здесь же главное понимать, что представление о порядке сходимости того или иного метода важно для возможности сравнить его с другими; более точно для этого подходит знание порядка q-сходимости (что и определено выше), порядок же r-сходимости говорит лишь о наличии такой последовательности положительных чисел, которая, сходясь к нулю с этим порядком, мажорирует последовательность величин $|x^* - x_k|$. Не вникая в тонкости, можно сказать, что обычно, изучая итерационный метод, устанавливают факт сходимости итерационной последовательности (x_k) к искомому элементу x^* и получают апостериорные и априорные оценки погрешности, а о порядке метода (в том или ином смысле, не всегда уточняя, в каком) судят или на основе неравенства типа (5.7), или по априорной оценке погрешности вида

$$|x^* - x_k| \le C\nu^{p^k}, \qquad (5.9)$$

где $C > 0$, $\nu \in (0;1)$ – некоторые константы, а $p \ge 1$ – порядок метода, или по неравенству вида

$$|x_{k+1} - x_k| \le C|x_k - x_{k-1}|^p, \qquad (5.10)$$

показывающему скорость сближения членов итерационной последовательности и являющемуся ключевым для установления сходимости и получения оценок погрешностей. Чаще всего разные способы приводят к одному и тому же значению p, хотя это и не гарантировано.

Коснемся еще одного аспекта понятия сходимости итерационного метода. В приведенных выше определениях отождествлялись сходимость итерационного процесса и сходимость итерационной последовательности, порождаемой этим процессом; при этом негласно считалось, что последовательность (x_k) уже как бы фиксирована заданием начальной точки x_0.

Больший же интерес представляет сходимость множества всевозможных итерационных последовательностей, генерируемых итерационным методом при варьировании x_0 в границах некоторой области. Итерационные методы, дающие в пределе решение данной задачи при любом начальном приближении x_0, называются ***глобально сходящимися***. Если же сходимость итерационной последовательности (x_k) к искомому элементу x^* имеет место лишь при задании x_0 из некоторой, вообще говоря, достаточно малой окрестности x^*, то соответствующий итерационный метод называют ***локально сходящимся***.

Так, рассмотренные выше методы дихотомии можно отнести к глобально сходящимся методам, так как с их помощью всегда можно получить какой-нибудь из корней уравнения $f(x) = 0$, если начать итерационный процесс с отрезка $[a;b]$ любой длины, входящего в область непрерывности $f(x)$, лишь бы было выполнено условие $f(a)f(b) < 0$. При этом, как видно из оценки (5.3), имеющей форму (5.9), метод половинного деления нужно считать линейно сходящимся методом, т.е. он сходится со скоростью геометрической прогрессии со знаменателем $q = 0.5$ и имеет [56] ***среднюю скорость сходимости*** $-\ln q = \ln 2$. Метод хорд также является методом первого порядка и в зависимости от свойств $f(x)$ может иметь как бóльшую, так и меньшую, чем $\ln 2$, среднюю скорость (часто бóльшую).

5.4. МЕТОД НЬЮТОНА

Одним из популярнейших итерационных методов решения нелинейных уравнений, что связано с его идейной простотой и быстрой сходимостью, является ***метод Ньютона***[*]. Правило построения итерационной последовательности (x_k) здесь получают или из геометрических соображений, откуда другое название этого метода – ***метод касательных***, говорящее само за себя, или из аналитических путем подмены данной нелинейной функции ее линейной моделью на основе формулы конечных приращений Лагранжа или формулы Тейлора, в связи с чем метод Ньютона также называют ***методом линеаризации***. В любом случае говорить о нахождении нуля функции $f(x)$ методом Ньютона можно лишь в предположении, что данная функция обладает достаточной гладкостью.

Для простоты будем считать, что функция $f(x)$ дважды дифференцируема на отрезке $[a;b]$, содержащем корень ξ уравнения (5.1).

---

[*] В зарубежной литературе его часто называют ***методом Ньютона–Рафсона*** [22, 38, 43, 50]

Пусть $x_k \in [a;b]$ – уже известный член последовательности приближений к ξ, полученный конструируемым методом (или заданное начальное приближение x_0 при $k=0$). Для любого x из $[a;b]$ можно записать формальное представление $f(x)$ по формуле Тейлора

$$f(x) = f(x_k) + f'(x_k)(x - x_k) + \frac{1}{2} f''(\Theta_k)(x - x_k)^2, \qquad (5.11)$$

где $\Theta_k \in [a;b]$ – некоторая точка между x и x_k.

Так как корень ξ – потенциально произвольная точка отрезка $[a;b]$, то разложение (5.11) справедливо и для $x=\xi$, т.е. существует точка $\Theta_k = \overline{\Theta_k}$ такая, что

$$f(\xi) = f(x_k) + f'(x_k)(\xi - x_k) + \frac{1}{2} f''(\overline{\Theta_k})(\xi - x_k)^2.$$

Но $f(\xi) = 0$, и если точка $\overline{\Theta_k}$ известна, то корень ξ можно точно найти из квадратного уравнения

$$f(x_k) + f'(x_k)(\xi - x_k) + \frac{1}{2} f''(\overline{\Theta_k})(\xi - x_k)^2 = 0. \qquad (5.12)$$

Считая, что значение x_k близко к ξ, т.е. разность $(\xi - x_k)$ по модулю достаточно мала, можно рассчитывать, что величина $(\xi - x_k)^2$ будет тем более малой. На этом основании отбросим в (5.12) последнее слагаемое и подменим квадратное уравнение (5.12) линейным уравнением. Естественно, что при этом будет найден не корень ξ, а некоторая другая точка, которую обозначим x_{k+1}.

Таким образом, итерационный процесс Ньютона определяется линейным уравнением

$$f(x_k) + f'(x_k)(x_{k+1} - x_k) = 0 \qquad (5.13)$$

или в явном виде формулой

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \qquad (5.14)$$

где $k=0,1,2,\ldots$, и предполагается, что по крайней мере на элементах последовательности (x_k) первая производная данной функции в нуль не обращается*).

Если в равенстве (5.13) фиксированную точку x_{k+1} заменить переменной x, а 0 в правой части – переменной y, то в полученном легко уз-

*) Как видим, процесс вычислений по методу Ньютона не требует знания второй производной. Можно обойтись без нее и при выводе (другим способом), но при этом значительно усложнилось бы изучение метода.

нать уравнение касательной к кривой $y = f(x)$, проведенной к ней в точке $(x_k f(x_k))$. Отсюда **геометрический смысл метода Ньютона**: *приближения к корню ξ совершаются по абсциссам точек пересечения касательных к графику данной функции, проводимых в точках, соответствующих предыдущим приближениям* (рис. 5.5).

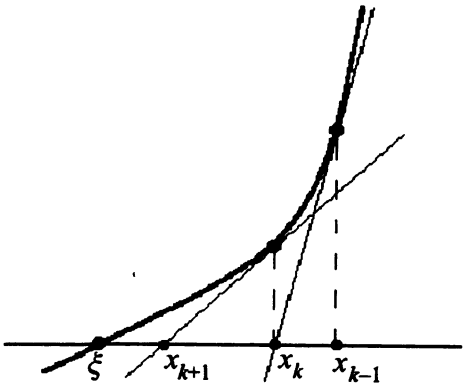


Рис. 5.5. Приближения к корню нелинейного уравнения методом касательных

Изучение сходимости метода Ньютона (осуществимость процесса вычисления элементов последовательности (x_k) по формуле (5.14) в пределах заданного отрезка, сходимость к корню ξ данного уравнения, порядок метода, оценки погрешности и критерии окончания процесса построения приближений, условия на выбор начального элемента последовательности) проводится при более ограничительных требованиях к данной функции $f(x)$.

Интуитивно ясно (из вида формулы (5.14), рассуждений при ее выводе и из ее геометрического смысла), что сходимость (x_k) к ξ будет тем быстрее и говорить о ней можно тем увереннее, чем ближе функция $f(x)$ к линейной и чем круче ее график пересекает ось абсцисс; так что есть смысл потребовать от $f(x)$, чтобы по модулю вторая ее производная была ограниченна сверху, а первая снизу. При этих условиях обратимся сначала к исследованию б ы с т р о т ы сходимости итерационного метода Ньютона (5.14) в предположении, что факт его осуществимости и сходимости к корню ξ сомнений не вызывает.

Теорема 5.4. *Пусть функция* $f(x)$ *удовлетворяет условиям*

(A) $$\begin{bmatrix} |f'(x)| \ge \alpha > 0 \\ |f''(x)| \le \beta < \infty \end{bmatrix} \qquad \forall x \in [a,b].$$

Тогда если члены последовательности (x_k), *определяемые методом Ньютона* (5.14), *при любом фиксированном* $k \in N_0$ *принадлежат отрезку* $[a;b]$ *и эта последовательность сходится на* $[a;b]$ *к корню* ξ *уравнения* (5.1), *то справедливы неравенства* ($\forall k \in N_0$):

$$|\xi - x_{k+1}| \le \frac{\beta}{2\alpha}|\xi - x_k|^2; \qquad (5.15)$$

$$|\xi - x_{k+1}| \le \frac{\beta}{2\alpha}|x_{k+1} - x_k|^2. \qquad (5.16)$$

(Первое из этих неравенств в соответствии с определением 5.2 позволяет считать (5.14) ***методом второго порядка***, а второе, являясь апостериорной оценкой погрешности, может служить в качестве критерия останова процесса вычислений).

Доказательство. Отметим, что требование принадлежности точек x_k отрезку $[a;b]$ дает право пользоваться оговоренным условиями (A) поведением данной функции в этих точках и их малых окрестностях.

Подставляя в правую часть формулы (5.12) вместо нуля левую часть равенства (5.13) (оба равенства истинны для рассматриваемых точек x_k, x_{k+1}), имеем:

$$f'(x_k)\xi - f'(x_k)x_k + \frac{1}{2}f''(\overline{\Theta_k})(\xi - x_k)^2 = f'(x_k)x_{k+1} - f'(x_k)x_k.$$

Это равенство можно записать в виде точной связи между ошибками k-го и $(k+1)$-го приближений*) :

$$\xi - x_{k+1} = -\frac{f''(\overline{\Theta_k})}{2f'(x_k)}(\xi - x_k)^2,$$

из которой, переходя к модулям и привлекая условия (A), получаем первое из доказываемых неравенств.

Для доказательства второго неравенства сначала установим связь между невязкой $(k+1)$-го приближения и разностью соседних (k-го и $(k+1)$-го) приближений. С этой целью в формулу Тейлора (5.11) подставим $x = x_{k+1}$. Имеем

---

*) Для элемента x_{k+1} последовательности (x_k) приближений к корню ξ уравнения $f(x) = 0$ величину $\xi - x_{k+1}$ называют ***ошибкой***, $x_{k+1} - x_k$ – ***поправкой***, а $f(x_{k+1})$ – ***невязкой***.

$$f(x_{k+1}) = f(x_k) + f'(x_k)(x_{k+1} - x_k) + \frac{1}{2} f''(\tilde{\Theta}_k)(x_{k+1} - x_k)^2,$$

и воспользуемся тем, что согласно (5.13), первые два слагаемых правой части равенства в совокупности дают нуль. Таким образом,

$$f(x_{k+1}) = \frac{1}{2} f''(\tilde{\Theta}_k)(x_{k+1} - x_k)^2,$$

и значит,

$$|f(x_{k+1})| \le \frac{\beta}{2} |x_{k+1} - x_k|^2. \qquad (5.17)$$

Применим теперь формулу Лагранжа к разности $f(\xi) - f(x_{k+1})$. Согласно этой формуле, между точками ξ и x_{k+1} найдется точка τ_{k+1} такая, что

$$f(\xi) - f(x_{k+1}) = f'(\tau_{k+1})(\xi - x_{k+1}).$$

Принимая во внимание, что $f(\xi) = 0$, а также условия (А), получаем неравенство между абсолютными величинами невязок и ошибок приближений:

$$|\xi - x_{k+1}| \le \frac{1}{\alpha} |f(x_{k+1})|. \qquad (5.18)$$

Усилив последнее с помощью неравенства (5.17), приходим к доказываемой оценке (5.16).

Теорема полностью доказана.

Замечание 5.2. Неравенство (5.18), справедливое для приближений x_{k+1} к нулю ξ дифференцируемой функции $f(x)$ независимо от способа их получения, обосновывает контроль точности по невязкам, а именно, оправдывает критерий окончания итерационного процесса

$$|f(x_k)| < \varepsilon \Rightarrow \xi := x_k \text{ с точностью } \varepsilon,$$

если в окрестности ξ выполняется требование $|f'(x)| > 1$, и говорит о необходимости учитывать множитель $\frac{1}{\alpha}$, т.е. добиваться выполнения неравенства

$$|f(x_k)| < \alpha\varepsilon, \qquad \text{если } \alpha < 1.$$

Чтобы выяснить, при каких условиях на выбор начального приближения x_0 начинающаяся с него последовательность (x_k), генерируемая методом Ньютона (5.14), будет сходиться к корню ξ, проитерируем формально неравенство (5.15). Имеем:

при $k = 0$ $\qquad |\xi - x_1| \le \frac{\beta}{2\alpha} |\xi - x_0|^2;$

при $k=1$ $\quad |\xi - x_2| \le \frac{\beta}{2\alpha}|\xi - x_1|^2 \le \frac{\beta}{2\alpha}\left(\frac{\beta}{2\alpha}\right)^2 |\xi - x_0|^{2^2}$;

при $k=2$ $\quad |\xi - x_3| \le \frac{\beta}{2\alpha}|\xi - x_2|^2 \le \frac{\beta}{2\alpha}\left(\frac{\beta}{2\alpha}\right)^2\left(\frac{\beta}{2\alpha}\right)^{2^2} |\xi - x_0|^{2^3}$;

далее по индукции получаем

$$|\xi - x_k| \le \left(\frac{\beta}{2\alpha}\right)^{1+2+2^2+\cdots+2^{k-1}} |\xi - x_0|^{2^k} =$$

$$= \left(\frac{\beta}{2\alpha}\right)^{2^k-1} |\xi - x_0|^{2^k} = \frac{2\alpha}{\beta}\left(\frac{\beta}{2\alpha}\right)^{2^k} |\xi - x_0|^{2^k},$$

т.е. при любых $k \in N$

$$|\xi - x_k| \le \frac{2\alpha}{\beta}\left(\frac{\beta}{2\alpha}|\xi - x_0|\right)^{2^k}. \qquad (5.19)$$

Отсюда следует, что

$$x_k \to \xi, \quad \text{если} \quad \frac{\beta}{2\alpha}|\xi - x_0| < 1.$$

Таким образом, появилась возможность судить о том, насколько далеко от корня ξ можно брать начальное приближение x_0 в зависимости от свойств данной функции $f(x)$, и по априорной оценке (5.19) заранее подсчитывать число итераций, достаточное для вычисления корня с заданной точностью, если есть оценка близости x_0 к ξ.

Теорема 5.5. *Пусть для функции* $f(x)$ *на отрезке* $[a;b]$ *выполнены условия* (А).

Тогда если интервал $J := \left(\xi - \frac{2\alpha}{\beta}; \xi + \frac{2\alpha}{\beta}\right)$ *содержится в* $[a;b]$, *то при произвольном выборе* x_0 *из* J *для определяемой методом Ньютона* (5.14) *последовательности* (x_k):

1) $x_k \in J \quad \forall k \in N$;

2) $\exists \lim\limits_{k\to\infty} x_k = \xi \quad$ *и* $\quad f(\xi) = 0$;

3) *справедливо утверждение теоремы* 5.4 *и оценка* (5.19).

Доказательство. Прежде всего заметим, что условие $x_0 \in J$ равносильно неравенству

$$v := \frac{\beta}{2\alpha}|\xi - x_0| < 1, \qquad (5.20)$$

к которому мы пришли выше, анализируя неравенство (5.19) (рис. 5.6).

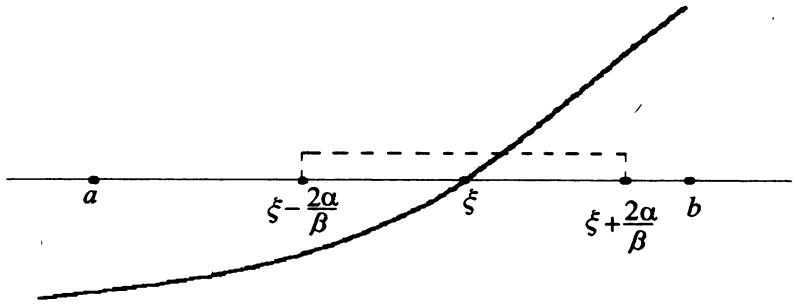


Рис. 5.6. Промежуток выбора начальной точки в методе Ньютона

Покажем осуществимость процесса (5.14) в J.

Так как $x_0 \in J \subset [a;b]$, то, согласно условиям (А), $f'(x_0) \neq 0$, и значение x_1 по формуле (5.14) может быть получено. Более того, справедливо неравенство (5.15) при $k = 0$. Следовательно,

$$|\xi - x_1| \le \frac{\beta}{2\alpha}|\xi - x_0|^2 = \frac{2\alpha}{\beta}\nu^2 < \frac{2\alpha}{\beta},$$

т.е. $x_1 \in J$. Аналогично из индукционного предположения, что $x_k \in J$ при некотором $k \in N$, т.е. что $|\xi - x_k| < \frac{2\alpha}{\beta}$, используя (5.15), получаем:

$$|\xi - x_{k+1}| \le \frac{\beta}{2\alpha}|\xi - x_k|^2 < \frac{\beta}{2\alpha}\cdot\left(\frac{2\alpha}{\beta}\right)^2 = \frac{2\alpha}{\beta} \quad \Rightarrow \quad x_{k+1} \in J.$$

Итак, согласно принципу математической индукции, все элементы последовательности (x_k) лежат в $J \subset [a;b]$, и значит, можно воспользоваться неравенством (5.19), из которого тут же следует, что $\xi = \lim x_k$, и заключение предыдущей теоремы.

То, что точка ξ – середина заявленного в теореме интервала J – есть корень уравнения (5.1), если заранее это неизвестно, устанавливается переходом к пределу в формуле (5.14) (с учетом условия $f'(x) \neq 0$ $\forall x \in [a;b]$).

Слабым местом только что доказанной теоремы 5.5 является то, что точкой отсчета при построении промежутка применимости метода Ньютона служит корень ξ, который как раз и неизвестен. Более естественно за центр такого промежутка принимать некоторую конкретную точку x_0 – начальное приближение, полагая по тем или иным соображениям (например, геометрическим), что в ее окрестности должен быть корень. Имеется

ряд теорем подобного типа (см. [34, 35, 40, 43, 47] и др.). Как правило, им присуща некоторая громоздкость и трудная проверяемость условий.

Заменив условия (А) на требование знакопостоянства первой и второй производных данной функции, означающих монотонность и определенную выпуклость ее графика, докажем простую теорему несколько иного плана.

Теорема 5.6 [21]. *Пусть на отрезке* $[a;b]$ *функция* $f(x)$ *имеет первую и вторую производные постоянного знака и пусть*

$$f(a)f(b)<0.$$

Тогда если точка x_0 *выбрана на* $[a;b]$ *так, что*

$$f(x_0)f''(x_0)>0, \tag{5.21}$$

то начатая с нее последовательность (x_k), *определяемая методом Ньютона* (5.14), *монотонно сходится к корню* $\xi\in(a;b)$ *уравнения* (5.1).

Доказательство опирается на теорему Вейерштрасса о сходимости монотонной ограниченной последовательности.

Положим, для определенности, что

$$f'(x)>0 \qquad \text{и} \qquad f''(x)>0 \quad \forall x\in[a;b].$$

Тогда $f(a)<0$, $f(b)>0$, и в качестве начальной точки x_0, удовлетворяющей условию (5.21),*) можно взять любую точку из промежутка $(\xi;b]$ (наличие корня ξ, единственного в $(a;b)$, условиями теоремы обеспечено), и при этом x_0 из $(\xi;b]$ будет $f(x_0)>0$.

Покажем ограниченность последовательности (x_k) снизу и ее монотонное убывание. Из равенства (5.12) при $k=0$, т.е. из

$$f(x_0)+f'(x_0)(\xi-x_0)+\frac{1}{2}f''(\Theta_0)(\xi-x_0)^2=0,$$

в силу положительности последнего слагаемого, следует, что

$$f(x_0)+f'(x_0)(\xi-x_0)<0.$$

Отсюда, учитывая положительность $f'(x)$, имеем

$$\xi<x_0-\frac{f(x_0)}{f'(x_0)}.$$

Это неравенство показывает, что

$$\xi<x_1=x_0-\frac{f(x_0)}{f'(x_0)}<x_0$$

(значит, $f(x_1)>0$). По индукции устанавливаем, что

$$\xi<x_{k+1}<x_k \qquad \forall k\in N_0,$$

---

*) Называемому иногда *условием Фурье.*

т.е. последовательность (x_k) монотонно убывает и ограничена снизу самим корнем ξ, следовательно, имеет предел. Сходимость (x_k) именно к корню, как и в предыдущей теореме, получаем переходом к пределу в равенстве (5.14).

Остальные комбинации знаков производных рассматриваются аналогично. Теорема доказана.

Замечание 5.3. Нарушение условия Фурье (5.21) на выбор начального приближения x_0 при выполненных требованиях к знакопостоянству производных может отразиться лишь на первом приближении: его перебросит через корень ξ на другую часть отрезка $[a;b]$ (рис. 5.7). Если при этом x_1 не окажется за пределами отрезка $[a;b]$, то далее итерационный процесс Ньютона пойдет монотонно (сформулируйте самостоятельно аналог теоремы 5.6 без условия Фурье).

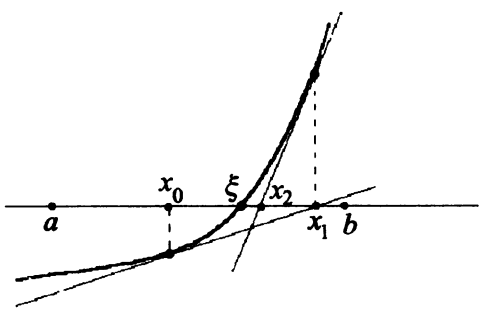


Рис.5.7. Поведение первых приближений по методу Ньютона при нарушении условия Фурье на выбор x_0

Обратимся вновь к зафиксированной теоремами 5.4 и 5.5 квадратичной сходимости метода Ньютона, которая имеет место и в условиях теоремы 5.6. О поведении такого процесса можно получить наглядное представление из оценки (5.19), схематично записанной в виде

$$|\xi - x_k| \le C \cdot \nu^{2^k},$$

где $\nu \in (0;1)$ (см. (5.20)), а $C\left(=\dfrac{2\alpha}{\beta}\right)$ – некоторая положительная постоянная. Допустим, что функция $f(x)$ и приближение x_0 к корню ξ таковы, что $C=1$, а $\nu = 0.1$, т.е. абсолютные погрешности убывают по закону $|\xi - x_k| \le 0.1^{2^k}$. Подставляя сюда $k = 1,2,3,\ldots$, имеем: $|\xi - x_1| \le 10^{-2}$, $|\xi - x_2| \le 10^{-4}$, $|\xi - x_3| \le 10^{-8}$, $|\xi - x_4| \le 10^{-16}$ и т.д.

Как видим, *квадратично сходящийся процесс в идеале*, т.е. если он реализуется точно, *должен давать удвоение числа верных знаков на каждой итерации, начиная с некоторой*. Такой высокий темп установления верных цифр искомого корня не только позволяет получить корень с большой точностью сравнительно небольшим количеством вычислений, но и обеспечивает хорошую численную устойчивость метода, а также меньшую критичность к правилу окончания итерационного процесса (часто здесь применяют упрощенное правило останова

$$|x_k - x_{k-1}| \le \varepsilon \quad \Rightarrow \quad \xi \approx x_k).$$

5.5. ПРИМЕНЕНИЕ МЕТОДА НЬЮТОНА К ВЫЧИСЛЕНИЮ ЗНАЧЕНИЙ ФУНКЦИЙ

Элементарные функции чаще всего вычисляются с помощью приближения их подходящими многочленами. В некоторых же случаях для этих целей применяют итерационные методы, в частности, базирующиеся на методе Ньютона.

Пусть требуется найти значение заданной функции φ в заданной точке a. Считая a произвольной точкой из области $D(\varphi)$ или какой-либо ее подобласти, функциональное соответствие

$$x = \varphi(a)$$

зададим неявно уравнением

$$F(a,x) = 0 \qquad (5.22)$$

таким, чтобы: 1) оно было локально эквивалентным (в окрестности точки a) данному; 2) функция F была дифференцируема по второму аргументу; 3) функции F и F' были легко вычислимы.

При каждом фиксированном a уравнение (5.22) можно считать уравнением типа (5.1) и получать приближенно его корень – требуемое значение $x = \varphi(a)$ – методом Ньютона (5.14). Для уравнения (5.22) формула (5.14) принимает вид

$$x_{k+1} = x_k - \frac{F(a,x_k)}{F'_x(a,x_k)}, \qquad (5.23)$$

где $k = 0, 1, 2, \ldots$, а x_0 – задаваемое начальное приближение к $\varphi(a)$.

В качестве более конкретного примера применения такого подхода выведем из формулы (5.23) ***правило Ньютона вычисления арифметических корней***.

Пусть a – данное положительное число, а $n \ge 2$ – данный натуральный показатель корня.

Очевидна связь между задачей вычисления вещественного значения

$$x = \sqrt[n]{a}$$

и задачей нахождения положительного корня уравнения

$$x^n - a = 0.$$

Приняв $F(a,x) := x^n - a$, находим $F'_x(a,x) = nx^{n-1}$, и согласно (5.23), процесс приближений к $\sqrt[n]{a}$ определяем формулой

$$x_{k+1} = x_k - \frac{x_k^n - a}{nx_k^{n-1}}$$

или в другом виде

$$x_{k+1} = \frac{1}{n}\left[(n-1)x_k + \frac{a}{x_k^{n-1}}\right],$$

где $k = 0, 1, 2, \ldots$, а $x_0 > 0$ задается.

Еще из глубокой древности известен частный случай полученного правила Ньютона – ***процесс Герона***

$$x_{k+1} = \frac{1}{2}\left(x_k + \frac{a}{x_k}\right),$$

применяемый для извлечения квадратных корней. Здесь $F(a,x) = x^2 - a$, $F'(a,x) = 2x$, $F''(a,x) = 2$, т.е. при $x > 0$ выполняются предварительные условия теоремы 6, и для монотонной сходимости (x_k) к $\sqrt{a}$ достаточно лишь удовлетворить условию Фурье (5.21), т.е. взять x_0 таким, чтобы было $x_0^2 > a$. Легко убедиться, что это условие будет выполнено, если взять $x_0 = 2^{0.5m}$, если m – четное, и $x_0 = 2^{0.5(m+1)}$, если m – нечетное, где $m \in Z$ такое, что $a = 2^m q$ и $q \in [0.5;1]$ (a "зажимается" между двумя соседними целыми степенями двойки: $2^{m-1} \le a < 2^m$)*).

Другим примером использования метода Ньютона в форме (5.23) может служить вычисление обратной величины данного числа a, иначе говоря, выполнение операции деления с помощью других арифметических операций.

---

*) Не заботясь о монотонности последовательности приближений, обычно ограничиваются заданием $x_0 = 2^{[0.5m]}$.

Аналогично предыдущему зададим искомое $x = \frac{1}{a}$ как корень уравнения

$$a - \frac{1}{x} = 0.$$

Подставляя $F(a,x) = a - \frac{1}{x}$ и $F'_x(a,x) = \frac{1}{x^2}$ в (5.23), после упрощения получаем итерационный процесс без делений*)

$$x_{k+1} = x_k(2 - ax_k), \quad k = 0, 1, 2, \ldots. \tag{5.24}$$

Здесь также считаем a и x положительными, и очевидно, что для любых $x > 0$

$$F' > 0, \text{ а } F'' = -\frac{2}{x^3} < 0,$$

т.е. условия теоремы 5.6 будут выполнены, если взять $x_0 \in \left(0, \frac{1}{a}\right)$.

Непосредственное изучение процесса (5.24) показывает, что область выбора начального приближения x_0 может быть в два раза расширена (правда монотонность приближений при этом уже не гарантируется).

Действительно, рассмотрим связь между поправками $(k+1)$-го и k-го приближений:

$$\frac{1}{a} - x_{k+1} = \frac{1}{a} - 2x_k + ax_k^2 = a\left(\frac{1}{a} - x_k\right)^2$$

(равенство типа (5.7), подтверждающее, что (5.24) – процесс второго порядка). Отсюда последовательным итерированием приходим к равенству

$$\frac{1}{a} - x_k = \frac{1}{a}(1 - ax_0)^{2^k},$$

из которого следует, что сходимость (x_k) к $\frac{1}{a}$ имеет место в случае, когда $|1 - ax_0| < 1$, т.е. при $x_0 \in \left(0; \frac{2}{a}\right)$.

Обычно за начальное приближение, удовлетворяющее условию $x_0 \in \left(0; \frac{1}{a}\right)$, берут число 2^{-m}, где $m \in Z$ то же, что и в предыдущем примере.

---

*) Легко увидеть аналогию между этим процессом и рассмотренным ранее процессом Шульца второго порядка для обращения матриц (ср (5 24) с (3 34) при $m = 1$).

5.6. МОДИФИКАЦИИ МЕТОДА НЬЮТОНА. МЕТОД СЕКУЩИХ

Вновь обратимся к теоремам 5.4–5.6 о сходимости метода Ньютона. Их условия предполагают неравенство нулю производной данной функции $f(x)$ на промежутке $[a;b]$, где применяется метод. А это означает, что они регламентируют применение метода Ньютона только для нахождения простых нулей функции $f(x)$, поскольку для кратного корня ξ уравнения (5.1) имеет место равенство $f'(\xi)=0$. Действительно, пусть ξ – m-кратный корень $(m\ge 2)$; тогда функция $f(x)$ представима*) в виде $f(x)=(x-\xi)^m f_1(x)$ и ее производная $f'(x)=(x-\xi)^{m-1}\times$ $\times[mf_1(x)+(x-\xi)f_1'(x)]$ обращается в нуль при $x=\xi$.

Согласно (5.14), формально нужно, чтобы производная не равнялась нулю в точках x_k последовательности приближений, в предельной же точке ξ допустимо обращение производной в нуль. Как показывает пример 5.3, итерационный процесс Ньютона может сходиться и в этом случае, т.е. когда ξ является кратным корнем уравнения (5.1), но сходимость при этом – только линейная.

Если заведомо известно число m – показатель кратности корня ξ, то для ускорения сходимости метода Ньютона в формулу (5.14) рекомендуется ввести корректирующий множитель m:

$$x_{k+1}=x_k-m\cdot\frac{f(x_k)}{f'(x_k)}\,. \qquad (5.25)$$

Такую модификацию будем называть ***методом Ньютона–Шрёдера*****). Доказательство сверхлинейной сходимости этого метода можно найти в [47], где он называется ***методом Ньютона с параметром***, а также в [50] и в [43], где имеется ссылка на содержащую этот метод работу Э.Шрёдера.

Пример 5.3. Для функции $f(x)=(x-1)^2$ корень $\xi=1$ – двукратный. Подстановка этой $f(x)$ и $f'(x)=2(x-1)$ в формулу (5.14) определяет процесс

$$x_{k+1}=x_k-\frac{x_k-1}{2}, \qquad (5.26)$$

---

*) Это представление здесь принимается за определение m-й кратности корня ξ. Часто ξ считают корнем кратности m, если

$f(\xi)=f'(\xi)=\ldots=f^{(m-1)}(\xi)=0$, а $f^{(m)}(\xi)\ne 0$ (см., например, [47]).

) В [23] первое введение параметра m в формулу (5.14) приписывается Е.Бодевигу (1949 г.), в связи с чем (5.25) там называют ***методом Ньютона–Бодевига*.

или проще,

$$x_{k+1} = \frac{1}{2}(x_k + 1).$$

Если начать его с $x_0 = 2$, то на каждой последующей итерации будем получать все более близкие к $\xi = 1$ значения:

$$x_1 = 1.5, \quad x_2 = 1.25, \quad x_3 = 1.125, \quad x_4 = 1.0625, \quad \ldots.$$

Вычитая 1 из обеих частей (5.26), приходим к равенству

$$x_{k+1} - 1 = \frac{1}{2}(x_k - 1),$$

означающему, что сходимость (x_k) к 1 – точно линейная. В то же время введение множителя m=2 в (5.26) в соответствии с (5.25) приводит к стационарной последовательности $x_{k+1} = 1 \quad \forall k \in N_0$.

Не следует думать, что из (5.25) всегда будет получаться $x_{k+1} = \xi$. Рассмотрим менее утрированный **пример**.

Функция $f(x) = x(x-1)^2$ с тем же двухкратным корнем $\xi = 1$ с помощью формул (5.14) и (5.25) порождает следующие процессы:

метод Ньютона

$$x_{k+1} = \frac{2x_k^2}{3x_k - 1};$$

$x_0 = 2,$
$x_1 = 1.6,$
$x_2 = 1.347368,$
$x_3 = 1.193517,$
...

метод Ньютона–Шрёдера

$$x_{k+1} = \frac{x_k(x_k + 1)}{3x_k - 1};$$

$x_0 = 2,$
$x_1 = 1.2,$
$x_2 = 1.000116,$
$x_3 = 1.000000.$

Налицо эффективность коррекции метода Ньютона введением в него показателя кратности корня.

Цель всех последующих видоизменений основной формулы (5.14) метода Ньютона – уменьшение вычислительных затрат, связанных с необходимостью вычисления производной на каждом итерационном шаге.

Самый простой выход на этом пути – использование на каждом шаге одного и того же шагового множителя $\frac{1}{f'(x_0)}$, т.е. счет по формуле

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_0)}, \quad k = 0, 1, 2, \ldots \qquad (5.27)$$

Такой метод называют ***модифицированным*** или ***упрощенным методом Ньютона***[*]. Он имеет очевидную геометрическую интерпретацию: в начальной точке x_0 проводится касательная к графику $y = f(x)$ (первый шаг основного и модифицированного методов Ньютона совпадают), а во всех последующих точках $x_1, x_2, \ldots$ проводятся прямые, параллельные этой касательной (рис. 5.8).

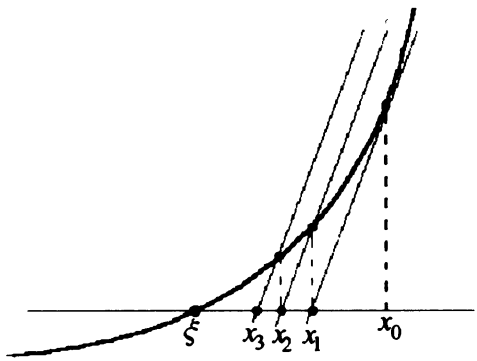


Рис.5.8. Приближения к корню нелинейного уравнения модифицированным (упрощенным) методом Ньютона

При такой модификации метод Ньютона утрачивает высокую скорость сходимости (процесс не реагирует на изменение наклона кривой к оси абсцисс при приближении к корню) и вместо квадратичной имеет лишь скорость сходимости геометрической прогрессии, что станет очевидным несколько позже (см. п.5.7).

На получение сверхлинейной скорости сходимости при видоизменении метода Ньютона (5.14) можно надеяться в случае, когда $f'(x_k)$ при каждом $k \in N$ подменяется не одним и тем же числом $f'(x_0)$, а некоторым близким к $f'(x_k)$ значением, которое может быть найдено (при каждом k свое) через значения данной функции. Для таких аппроксимаций[**] $f'(x_k)$ можно использовать, например, определение производной. Имеем:

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

и при малых h (произвольного знака) получаем приближенное равенство

$$f'(x_k) \approx \frac{f(x_k + h) - f(x_k)}{h}. \qquad (5.28)$$

---

[*] Иногда его называют еще ***огрубленный метод Ньютона*** [47].

[**] Approximare (лат.) – приближаться

позволяющее производную приближенно подменять так называемым ***разностным отношением***. Подстановка (5.28) в (5.14) приводит к итерационной формуле

$$x_{k+1} = x_k - \frac{f(x_k)\cdot h}{f(x_k + h) - f(x_k)}, \qquad (5.29)$$

где k=0, 1, 2, ..., а h – малый параметр, которым должен распорядиться вычислитель.

Ясно, что при каждом k в формуле (5.28) может быть свое значение h, т.е. в формуле (5.29) вместо постоянного параметра h имеет смысл использовать связанный с номером итерации параметр h_k, т.е. вести вычисления по формуле

$$x_{k+1} = x_k - \frac{f(x_k)h_k}{f(x_k + h_k) - f(x_k)}. \qquad (5.30)$$

Итерационный метод, определяемый формулами (5.29) или (5.30), назовем ***разностным методом Ньютона***[*].

Так как равенство (5.28) можно сделать сколь угодно точным за счет малости шага h разностного отношения (теоретически; практически это далеко не так из-за потерь точности при вычитании близких чисел), то по непрерывности можно утверждать асимптотически квадратичную скорость сходимости разностного метода Ньютона при определенных условиях.

Рассмотрим соображения, которыми следует руководствоваться при задании последовательности параметров h_k в разностном методе Ньютона (5.30). В любом случае будем исходить из постулата, что с ростом k значения $|h_k|$ должны убывать, чтобы при приближении x_k к корню ξ производная $f'(x_k)$ все более точно аппроксимировалась разностным отношением $\left[f(x_k + h_k) - f(x_k)\right]/h_k$.

Первое, что можно здесь предложить, так это задать какое-либо значение h_0, а каждое последующее значение параметра получать рекуррентным равенством $h_{k+1} = \delta h_k$, где $\delta \in (0;1)$ – некоторое фиксированное число. Например, можно положить $h_0 = 0.1$, $h_1 = 0.01$, $h_2 = 0.001$ и т.д. Очевиден недостаток такого подхода – отсутствие связи между скоростью сходимости (x_k) к ξ и скоростью убывания $|h_k|$ (может оказаться, что x_k еще не имеет достаточной близости к ξ, а значение $|h_k|$ настолько мало, что значения $f(x_k + h_k)$ и $f(x_k)$ реально не различимы; противоположная ситуация чревата большой потерей скорости сходимости или, еще хуже, нарушением канонического развития итерационного процесса).

---

[*] Другие названия – ***конечно-разностный*** [22] и ***дискретный*** [42] ***метод Ньютона*** Сходимость этого метода изучается в [22], там же можно найти рекомендации по сопряжению шага дискретизации h с точностью машинных вычислений

Если учесть, что при зафиксированных в теореме 5.4 условиях (A) $f(x_k) \to 0$ с той же скоростью, что и $x_k \to \xi$ (см. неравенство (5.18)), есть смысл полагать в (5.30) $h_k := f(x_k)$. Разумеется, это можно делать на той стадии итерационного процесса, когда значения $|f(x_k)|$ уже достаточно малы (иначе теряет силу (5.28)). При таких h_k формула (5.30) принимает вид

$$x_{k+1} = x_k - \frac{(f(x_k))^2}{f(x_k + f(x_k)) - f(x_k)} \qquad (5.31)$$

и называется ***методом Стеффенсена***. Подчеркнем еще раз его сугубо локальный характер сходимости, но зато *сходимость* эта *квадратичная* [43, 47].

При приблизительно равных затратах на вычисление значений данной функции и ее производной ни один из рассмотренных выше вариантов разностного метода Ньютона не дает выигрыша по сравнению с основным методом (5.14), поскольку каждый из них требует два вычисления функции на каждом итерационном шаге, не увеличивая при этом скорость сходимости. Построим такую модификацию метода Ньютона, развивая далее его разностный аналог (5.30), в которой на один шаг итерации приходилось бы только одно вычисление функции.

Опираясь на то, что необходимым условием сходимости некоторой последовательности x_k к пределу ξ, как это следует из (5.8), является сходимость к нулю последовательности разностей $(x_k - x_{k-1})$ (причем с той же скоростью, см. (5.10)), положим в (5.30)

$$h_k := x_{k-1} - x_k \text{ , откуда } x_{k-1} = x_k + h_k .$$

В результате этого из (5.30) получаем итерационный процесс

$$x_{k+1} = x_k - \frac{f(x_k)(x_{k-1} - x_k)}{f(x_{k-1}) - f(x_k)}, \qquad (5.32)$$

где $k = 1, 2, 3, \ldots$, а x_0 и x_1 должны задаваться.

Формула (5.32) определяет новый метод как ***двухшаговый*** (результат $(k+1)$-го шага зависит от результатов k-го и $(k-1)$-го шагов) и на каждой итерации требует вычисления только одного значения функции, другое же значение, фигурирующее в этой формуле, передается с предыдущего шага. Сравнив (5.32) с формулой (5.4), полученной из геометрических соображений, легко понять, что x_{k+1} есть абсцисса точки пересечения с осью Ox прямой, проведенной через точки $(x_{k-1}, f(x_{k-1}))$ и $(x_k, f(x_k))$, т.е. секущей (рис. 5.9). Отсюда название этого метода – ***метод секущих****).

*) В [5] методом секущих называют метод хорд (с фиксированным концом).

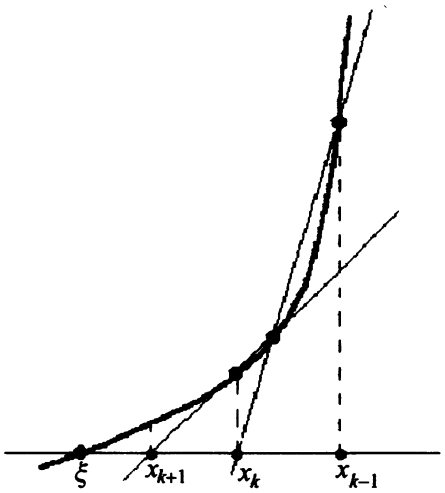


Рис.5.9. Приближения к корню методом секущих

Можно сказать, что метод секущих и метод хорд определяются совершенно однотипными формулами, но порождающие их идеологии различны, что сказывается на свойствах и скорости сходимости генерируемых ими последовательностей приближений.

Выясним, по какому закону убывают погрешности приближений, получаемых методом секущих (5.32) в условиях (А) теоремы 5.4.

Вычитая равенство (5.32) из равенства $\xi = \xi$ и применяя в знаменателе формулу Лагранжа, имеем:

$$\xi - x_{k+1} = \xi - x_k + \frac{f(x_k)(x_{k-1} - x_k)}{f(x_{k-1}) - f(x_k)} = \xi - x_k + \frac{f(x_k)}{f'(\mu_k)}, \qquad (5.33)$$

где μ_k – некоторая точка между приближениями x_k и x_{k-1}. Далее воспользуемся формулой Тейлора, согласно которой можно записать

$$f(x) = f(\xi) + f'(\xi)(x - \xi) + \frac{1}{2} f''(\theta)(x - \xi)^2,$$

откуда при $x = x_k$, $\theta = \theta_k$ с учетом того, что ξ – корень уравнения (5.1), следует представление

$$f(x_k) = f'(\xi)(x_k - \xi) + \frac{1}{2} f''(\theta_k)(x_k - \xi)^2.$$

Подставим это в (5.33) и вынесем в правой части общий множитель $\xi - x_k$:

$$\xi - x_{k+1} = \frac{\xi - x_k}{f'(\mu_k)} \cdot \left[f'(\mu_k) - f'(\xi) + \frac{1}{2} f''(\theta_k)(\xi - x_k) \right].$$

Вновь применяя формулу Лагранжа – теперь к разности производных – и переходя к модулям, получаем неравенство

$$|\xi - x_{k+1}| \le \frac{|\xi - x_k|}{|f'(\mu_k)|} \cdot \left[|f''(\nu_k)| \cdot |\xi - \mu_k| + \frac{1}{2} |f''(\theta_k)| |\xi - x_k| \right].$$

Так как для сходящейся к ξ нестационарной последовательности (x_k) справедливы неравенства

$$|\xi - \mu_k| < |\xi - x_{k-1}|, \quad |\xi - x_k| < |\xi - x_{k-1}|,$$

то в итоге можно записать искомую связь погрешностей в виде

$$|\xi - x_{k+1}| < \frac{3\beta}{2\alpha} |\xi - x_k| \cdot |\xi - x_{k-1}|, \qquad (5.34)$$

где α и β – константы из условий (А).

Ясно, что неравенство (5.34) характеризует *метод секущих* как *сверхлинейно сходящийся процесс*. Конкретный порядок метода секущих устанавливается следующим образом*).

Обозначим для краткости

$$\varepsilon_k := |\xi - x_k|, \quad C := \frac{3\beta}{2\alpha}$$

и, используя записанное в этих обозначениях неравенство (5.34)

$$\varepsilon_{k+1} < C\varepsilon_k \varepsilon_{k-1},$$

получаем последовательно несколько первых оценок ε_k через степени ε_0 (полагая по определению $\varepsilon_1 < \varepsilon_0$). Имеем:

при k=1 $\quad \varepsilon_2 < C\varepsilon_1\varepsilon_0 < \frac{1}{C}(C\varepsilon_0)^2$;

при k=2 $\quad \varepsilon_3 < C\varepsilon_2\varepsilon_1 < \frac{1}{C}(C\varepsilon_0)^3$;

при k=3 $\quad \varepsilon_4 < C\varepsilon_3\varepsilon_2 < \frac{1}{C}(C\varepsilon_0)^5$;

при k=4 $\quad \varepsilon_5 < C\varepsilon_4\varepsilon_3 < \frac{1}{C}(C\varepsilon_0)^8$;

при k=5 $\quad \varepsilon_6 < C\varepsilon_5\varepsilon_4 < \frac{1}{C}(C\varepsilon_0)^{13}$ и т.д.

Обратив внимание на то, что показатели в правых частях этих неравенств подчиняются закону "каждый последующий есть сумма двух предыдущих", нетрудно доказать в общем виде, что

$$\varepsilon_k < \frac{1}{C}(C\varepsilon_0)^{\Phi_k}, \qquad (5.35)$$

*) Близкое к данному изложение можно найти в [43].

где (Φ_k) – последовательность ***чисел Фибоначчи***[*], определяемая рекуррентным соотношением

$$\Phi_{k+1} = \Phi_{k-1} + \Phi_k, \quad k = 1, 2, 3, \ldots, \quad \Phi_0 = \Phi_1 = 1 .$$

Для общего члена Φ_k этой последовательности известна ***формула Бинэ***:

$$\Phi_k = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^{k+1} - \left(-\frac{1+\sqrt{5}}{2.}\right)^{-(k+1)}\right].$$

Поскольку с ростом k роль второго члена в выражении Φ_k становится ничтожной, для достаточно больших значений k на основе (5.35) можно считать справедливым асимптотическое неравенство

$$|\xi - x_k| < C_1 \cdot \nu^{\left(\frac{1+\sqrt{5}}{2}\right)^k},$$

где C_1 и ν – некоторые новые постоянные (которые легко выписать). Полученная оценка вида (5.9) позволяет утверждать, что *метод секущих имеет порядок по крайней мере* $\frac{1+\sqrt{5}}{2} \approx 1.618$.

Замечание 5.4. Для многошаговых итерационных методов иногда вводят специфическое понятие порядка сходимости. Согласно [22], *j-шаговый итерационный метод, генерирующий сходящуюся к ξ последовательность (x_k), называется* ***j-шагово сходящимся с порядком p****, если* $|\xi - x_{k+j}| \le C|\xi - x_k|^p$. Так как из неравенства (5.34) после его усиления получается

$$|\xi - x_{k+1}| < \frac{3\beta}{2\alpha}|\xi - x_{k-1}|^2,$$

то можно сказать, что *метод секущих сходится двухшагово квадратично.*

Высокий порядок скорости сходимости метода секущих в сочетании с минимальными вычислительными затратами – одно вычисление значения функции на один итерационный шаг – выводит этот метод на первое место по эффективности решения скалярных уравнений вида (5.1) среди прочих итерационных методов. Это подтверждается как теоретическими, так и практическими наблюдениями[**].

При применении метода секущих возникают вопросы, связанные с началом итерационного процесса и с его окончанием. Поскольку касатель-

---

[*] Фибоначчи – псевдоним знаменитого итальянского математика Леонардо Пизанского (1180–1240 гг).

[**] При сравнении с методом Ньютона и другими методами, использующими производные, в [43] предлагается приравнивать работу по вычислению значений функций и ее производных. Единица такой работы в [43] называется ***горнером***, а в [50] – ***единицей объема информационного запроса***.

ная к кривой есть предельное положение секущей, выбор начальной точки x_0 в методе секущих нужно осуществлять по тому же принципу, что и в методе касательных, например, привлекая условие Фурье (5.21); вторая же из начальных точек x_1, требуемая в двухшаговом методе (5.32), может быть взята в непосредственной близости от x_0 (понятие близости здесь, разумеется, условно), желательно между точкой x_0 и искомым корнем ξ.

Окончание счета по методу секущих, учитывая его быструю сходимость, можно контролировать с помощью проверок на малость модулей невязок $|f(x_k)|$ или поправок $|x_k - x_{k-1}|$. Однако главное здесь – это суметь вовремя остановить процесс вычислений, не дожидаясь момента, когда погрешности вычислений начнут превосходить погрешность метода вследствие вычитания приближенно вычисляемых близких значений $f(x_{k-1})$ и $f(x_k)$ в знаменателе расчетной формулы (5.32). В этом плане, т.е. в численной устойчивости, метод секущих проигрывает методу Ньютона.

Если почти все рассмотренные выше методы можно отнести к классу методов линеаризации, имея ввиду, что в их основе лежит подмена исходной нелинейной модели линейной, построенной тем или иным способом, то следующим шагом должно быть построение классов методов "параболизации". Квадратичная модель (парабола) может быть получена, например, по формуле Тейлора или квадратичной интерполяцией, но в любом случае соответствующие итерационные формулы, хотя и дают более быстро сходящиеся последовательности приближений к корню, либо содержат старшие производные данной функции, либо являются слишком громоздкими и сложными как для исследования, так и для их применения.

Более важно обратить внимание на локальную сходимость таких простых и достаточно быстро сходящихся методов, как метод Ньютона и метод секущих, условия которой не так часто удается обеспечить. В связи с этим встает задача построения ***гибридных алгоритмов*** на базе двух или нескольких методов, соединяя быструю сходимость одних с глобальной сходимостью других. Принципы комбинирования методов в таких алгоритмах могут быть различными: можно "стартовать" с глобально сходящегося "медленного" метода и подключить быстросходящийся метод на финише для уточнения значения корня, а можно сразу начать процесс вычислений "быстрым" методом, но проводить корректировку получаемых им значений, пользуясь глобально сходящимся методом. Последний подход порождает, например, следующий простейший гибридный алгоритм.

Метод Ньютона – метод половинного деления

Шаг 0. Задать начальное приближение x_0, положить $k := 0$.

Шаг 1. Вычислить $\tilde{x}_k = x_k - \dfrac{f(x_k)}{f'(x_k)}$.

Шаг 2. Если $|f(\tilde{x}_k)| < |f(x_k)|$, то $x_{k+1} := \tilde{x}_k$,

иначе $\tilde{x}_k := \frac{1}{2}(x_k + \tilde{x}_k)$ и возвратиться к началу шага 2.

Шаг 3. Проверить на останов (работа алгоритма либо прекращается с $\xi :\approx x_{k+1}$, либо продолжается переходом к шагу 1 с $k := k+1$).

Приведенный алгоритм учитывает, что метод Ньютона вырабатывает локально правильное направление (убывания функции), но продвижение в этом направлении может оказаться чрезмерным, что и корректируется с помощью деления отрезка пополам, если не выполняется ***условие релаксации*** $|f(\tilde{x}_k)| < |f(x_k)|$ в шаге 2 (рис. 5.10).

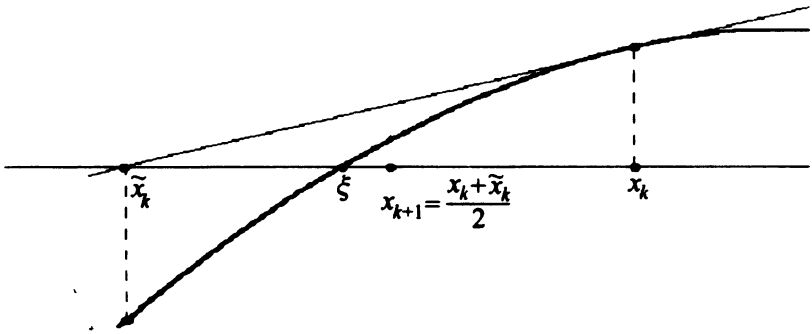


Рис.5.10. Иллюстрация одного шага гибридного метода Ньютона – половинного деления

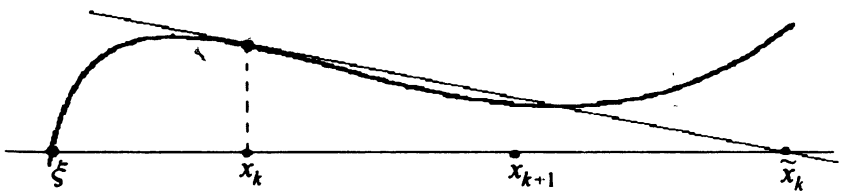


Рис.5.11. Пример поведения гибридного метода Ньютона – половинного деления, когда x_{k+1} дальше от корня ξ, чем x_k

Такой гибрид трудно считать глобально сходящимся (рассмотрите, например, его дальнейшее поведение в ситуации, изображенной на рис. 5.11), но он позволяет расширить границы применимости метода Ньютона и хоть как-то вести процесс поиска корня в условиях неопределенности знаков производных. Разумеется, этот алгоритм весьма схемати-

чен и требует некоторых усилий на его деталировку. Особенно важно решить, как в тех или иных случаях выполнить шаг 3. Большую роль здесь играет правильное сопряжение задаваемой точности решения задачи с погрешностью метода (иначе, с остаточной погрешностью) и с точностью выполнения арифметических операций на используемой ЭВМ (вычислительной погрешностью), а также с точностью вычисления значений функций, что, вообще говоря, не одно и то же. Важные сведения о таких критериях окончания процессов поиска корней с привязкой их к реальным компьютерам можно почерпнуть в книге [22].

5.7. ЗАДАЧА О НЕПОДВИЖНОЙ ТОЧКЕ. МЕТОД ПРОСТЫХ ИТЕРАЦИЙ

Нельзя не заметить, что все расчетные формулы, определяющие уже изученные методы решения скалярных уравнений (5.1), такие как метод Ньютона (5.14) и его модификации (5.25), (5.27), (5.29), (5.31), имеют вид

$$x_{k+1} = \varphi(x_k), \quad k = 0, 1, 2, \ldots , \tag{5.36}$$

где $\varphi(x)$ – некоторая функция, для каждого метода своя, так или иначе связанная с исходной функцией $f(x)$. Попытаемся понять, каким требованиям должна удовлетворять функция $\varphi(x)$, чтобы последовательность (x_k), определяемая этим самым общим одношаговым итерационным способом (5.36), называемым ***методом простых итераций***[*], была сходящейся, и как построить функцию $\varphi(x)$ по функции $f(x)$, чтобы эта последовательность сходилась к корню данного уравнения (5.1).

Сразу отметим, что функцию $\varphi(x)$ будем считать непрерывной в исследуемой области оси Ox. Поэтому, если определяемая формулой (5.36) последовательность (x_k) окажется сходящейся к некоторому числу ξ, то, переходя к пределу в равенстве (5.36), получаем

$$\xi = \varphi(\xi), \tag{5.37}$$

т.е. $\xi = \lim\limits_{k\to\infty} x_k$ – корень уравнения

$$x = \varphi(x). \tag{5.38}$$

Решение уравнений такого вида, наряду с (5.1), представляет самостоятельный интерес; нахождение их корней называется ***задачей о неподвижной точке***. Это название связано с тем, что точка ξ при отображении с помощью φ из R_1 в R_1 остается на месте (если разумеется, таковая существует).

---

[*] Другие названия. ***метод итераций***, ***метод последовательных приближений***. Далее, как и в гл. 3, будем также использовать аббревиатуру МПИ

Существование и единственность корня уравнения (5.38) основывается на ***принципе сжимающих отображений*** или, иначе, ***принципе неподвижной точки***.

Определение 5.3. *Непрерывная функция* $\varphi(x)$ *называется* ***сжимающей*** (***функцией сжатия***) *на отрезке* $[a;b]$, *если*:

1) $\varphi(x)\in[a;b] \quad \forall x\in[a;b]$;

2) $\exists q\in(0;1)$: $|\varphi(x_2)-\varphi(x_1)|\le q|x_2-x_1| \quad \forall x_1,x_2\in[a;b]$.

Графическое толкование применения сжимающего отображения φ (как функции множества) к промежутку сжатия $[a;b]$ предоставляет рис. 5.12.

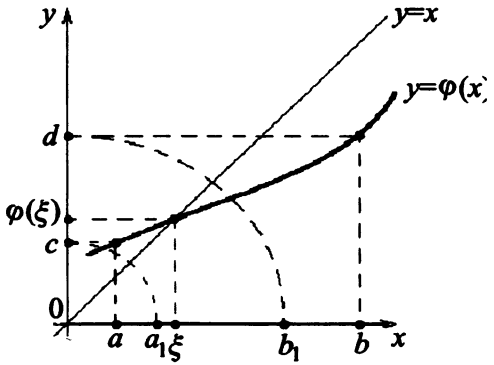


Рис.5.12. Сжатие отрезка $[a;b]$ *возрастающей функцией* $\varphi(x)$

Как видно из этого рисунка, если $[a;b]$ рассматривать как область определения функции сжатия $\varphi(x)$, то соответствующая ей область значений $[\varphi(a);\varphi(b)]$ на оси ординат (отрезок $[c;d]$), будучи перенесенным на ось абсцисс (отрезок $[a_1;b_1]$), целиком содержится в $[a;b]$. Применяя к $[a_1;b_1]$ те же рассуждения, что и к $[a;b]$, получим $[a_2;b_2]\subset[a_1;b_1]$ и т.д. В итоге образуется бесконечная последовательность вложенных отрезков

$$[a;b]\supset[a_1;b_1]\supset[a_2;b_2]\supset\cdots\supset[a_k;b_k]\supset\cdots,$$

причем их длины убывают по закону

$$b_k-a_k\le q^k(b-a)\underset{k\to\infty}{\longrightarrow}0.$$

Следовательно, в условиях сжатия эта последовательность имеет единственную общую точку (ξ), которая переходит сама в себя, т.е. является неподвижной точкой отображения φ. При этом, очевидно, последова-

тельность (a_k) левых концов этих промежутков монотонно сходится к ξ слева, а последовательность (b_k) правых концов – справа. Так как при условии возрастания $\varphi(x)$, как это показано на рис. 5.12, в этом процессе

$$a_1 = \varphi(a),\quad a_2 = \varphi(a_1),\quad a_3 = \varphi(a_2),\quad \dots,$$
$$b_1 = \varphi(b),\quad b_2 = \varphi(b_1),\quad b_3 = \varphi(b_2),\quad \dots,$$

то можно утверждать, что МПИ (5.36) будет давать монотонно сходящуюся к ξ последовательность, если ее начинать с $x_0 = a$ или с $x_0 = b$. Так же монотонно возрастающая и монотонно убывающая последовательности приближений будут получаться по формуле (5.36) и в случаях, когда за x_0 будет браться любая точка из промежутков $[a,\xi)$ и $(\xi,b]$ соответственно.

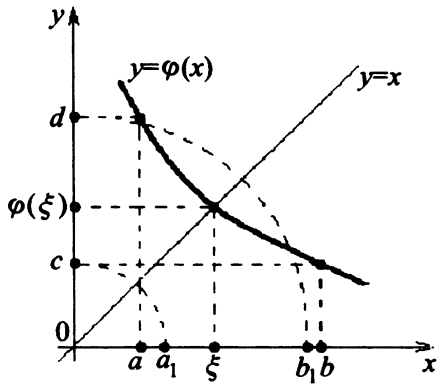


Рис.5.13. Сжатие отрезка $[a;b]$ убывающей функцией $\varphi(x)$

При условии убывания сжимающей функции $\varphi(x)$, т.е. в случае, изображенном на рис. 5.13, начинающиеся с концов a и b промежутка сжатия последовательности выстраиваются следующим образом:

$$a,\quad b_1 = \varphi(a),\quad a_2 = \varphi(b_1),\quad b_3 = \varphi(a_2),\quad \dots,$$
$$b,\quad a_1 = \varphi(b),\quad b_2 = \varphi(a_1),\quad a_3 = \varphi(b_2),\quad \dots.$$

Каждая из них сходится к неподвижной точке ξ, и элементы каждой из этих последовательностей с удалением от начала дают все более хорошие приближения то с недостатком, то с избытком. Такую сходимость к ξ имеет и любая другая последовательность (x_k), получаемая по формуле (5.36) при любом $x_0 \in [a;b]$. Отсюда другой термин, применяемый к неподвижной точке ξ, – ***центр итерации*** [43].

Более удобно иллюстрировать геометрически поведение итерационной последовательности (x_k), определяемой МПИ (5.36), не отмечая значения $\varphi(x_k)$ на оси ординат, а отражая их на ось абсцисс с помощью биссектрисы координатного угла $y = x$. Такие иллюстрации для случаев монотонного возрастания (ломаная типа "ступеньки") и монотонного убывания (ломаная типа "спираль") сжимающей функции $\varphi(x)$ показаны на рис. 5.14 и 5.15 соответственно (обоснуйте!).

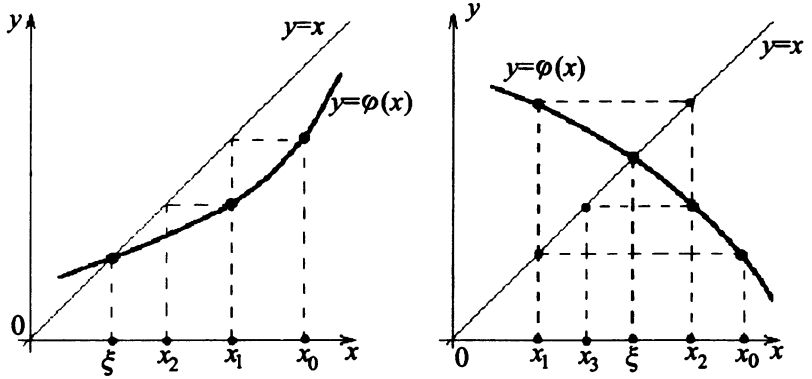


Рис.5.14, 5.15. Монотонные и двусторонние приближения к корню методом простых итераций

Итогом проведенных выше рассуждений является следующий **вывод**:

если на некотором промежутке $[a;b]$ *функция* $\varphi(x)$ *удовлетворяет условиям сжатия, зафиксированным определением 5.3, то* : 1) *уравнение* (5.38) *имеет и притом единственный корень* $\xi \in [a;b]$; 2) *к этому корню со скоростью геометрической прогрессии сходится определяемая МПИ* (5.36) *последовательность* (x_k), *начинающаяся с любого* $x_0 \in [a;b]$, *причем скорость сходимости тем выше, чем меньше* ***коэффициент сжатия****) $q \in (0;1)$; 3) *если функция* $\varphi(x)$ *монотонно возрастает на* $[a;b]$, *то приближения* x_k *к* ξ *также будут монотонными, если же* $\varphi(x)$ *убывает, то процесс* (5.36) *порождает двусторонние приближения к корню* ξ.

*) Иначе, ***константа Липшица***, в соответствии с применяемым к неравенству $|\varphi(x_2) - \varphi(x_1)| \le q|x_2 - x_1|$ термином ***условие Липшица*** (или ***Коши-Липшица***, если заведомо $q \in (0;1)$).

Этот вывод имеет, в основном, качественный характер. Дальнейшие исследования будут направлены на выявление конструктивных требований к $\varphi(x)$, обеспечивающих для нее выполнение условий сжатия, и получение оценок близости генерируемых МПИ (5.36) приближений x_k к неподвижной точке ξ.

Теорема 5.7. *Пусть функция $\varphi(x)$ определена и дифференцируема на отрезке $[a;b]$. Тогда если выполняются условия:*

1) $\varphi(x) \in [a;b] \quad \forall x \in [a;b]$,

2) $\exists q : |\varphi'(x)| \le q < 1 \quad \forall x \in (a;b)$,

то уравнение (5.38) *имеет и притом единственный на $[a;b]$ корень ξ; к этому корню сходится определяемая методом простых итераций* (5.36) *последовательность (x_k), начинающаяся с любого $x_0 \in [a;b]$; при этом справедливы оценки погрешности $(\forall k \in N)$:*

$$|\xi - x_k| \le \frac{q}{1-q}|x_k - x_{k-1}| , \qquad (5.39)$$

$$|\xi - x_k| \le \frac{q^k}{1-q}|x_1 - x_0| . \qquad (5.40)$$

Доказательство. Взяв произвольное x_0 из $[a;b]$, согласно условию 1) заключаем, что $x_1 = \varphi(x_0)$ также принадлежит $[a;b]$. По индукции можно утверждать, что все члены последовательности (x_k), генерируемой МПИ (5.36), благодаря условию 1), не могут покинуть отрезок $[a;b]$.

Вычтем из равенства (5.36) такое же равенство

$$x_k = \varphi(x_{k-1})$$

и к правой части полученного равенства

$$x_{k+1} - x_k = \varphi(x_k) - \varphi(x_{k-1})$$

применим формулу Лагранжа; согласно которой на интервале, определяемом точками x_{k-1} и x_k (а значит, на интервале $(a;b)$), найдется точка θ_k такая, что

$$\varphi(x_k) - \varphi(x_{k-1}) = \varphi'(\theta_k)(x_k - x_{k-1}) .$$

Следовательно,

$$x_{k+1} - x_k = \varphi'(\theta_k)(x_k - x_{k-1})$$

и, в силу условия 2), справедливо неравенство

$$|x_{k+1} - x_k| \le q|x_k - x_{k-1}|, \qquad (5.41)$$

которое можно расценивать как выполнение главного условия сжатия на элементах итерационной последовательности (x_k) с коэффициентом сжатия q. С другой стороны, неравенство (5.41) указывает на факт и скорость сближения членов этой последовательности.

Для разностей соседних более удаленных от начала членов последовательности (x_k) на основе (5.41) получаем:

$$\begin{aligned} &|x_{k+2} - x_{k+1}| \le q|x_{k+1} - x_k| \le q^2|x_k - x_{k-1}|; \\ &|x_{k+3} - x_{k+1}| \le q|x_{k+2} - x_{k+1}| \le q^3|x_k - x_{k-1}|; \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &|x_{k+i} - x_{k+i-1}| \le q^i|x_k - x_{k-1}| \quad \forall i \in N_0, \quad k \in N. \end{aligned} \tag{5.42}$$

Используя эти неравенства, оценим близость между x_{k+m} (где $m \in N$) и x_k, вычитая и прибавляя все промежуточные члены x_{k+m-1}, $x_{k+m-1}, x_{k+m-2}, \ldots, x_{k+1}$ и применяя свойство "модуль суммы не превосходит суммы модулей". Имеем:

$$|x_{k+m} - x_k| \le |x_{k+m} - x_{k+m-1}| + |x_{k+m-1} - x_{k+m-2}| + \ldots +$$
$$+ |x_{k+2} - x_{k+1}| + |x_{k+1} - x_k| \le \left(q^m + q^{m+1} + \ldots + q^2 + q\right)|x_k - x_{k-1}| =$$
$$= \frac{q - q^{m+1}}{1-q}|x_k - x_{k-1}|.$$

Но, в свою очередь,

$$|x_k - x_{k-1}| \le q^{k-1}|x_1 - x_0|, \tag{5.43}$$

что можно установить либо итерируя неравенство (5.41) при k=1, 2,..., либо непосредственно из (5.42), полагая k=1, а затем $i = k - 1$. Подставляя (5.43) в полученную выше оценку

$$|x_{k+m} - x_k| \le \frac{q}{1-q}\left(1 - q^m\right)|x_k - x_{k-1}|, \tag{5.44}$$

имеем

$$|x_{k+m} - x_k| \le \frac{q^k}{1-q}\left(1 - q^m\right)|x_1 - x_0|. \tag{5.45}$$

Поскольку правая часть неравенства (5.45) при фиксированном $m \in N$ и $k \to \infty$ стремится к нулю, (x_k) – последовательность Коши и, в силу замкнутости отрезка, имеет предел $\xi \in [a;b]$. Так как дифференцируемая функция непрерывна, то этот предел – корень уравнения (5.38). Его единственность на $[a;b]$ доказывается от противного: предположив, что наряду с ξ есть другой корень $\tau \in [a;b]$, т.е. имеет место равенство $\tau = \varphi(\tau)$, а значит, $\xi - \tau = \varphi(\xi) - \varphi(\tau)$, по формуле Лагранжа получаем

$$\exists \theta \in (a;b) : \xi - \tau = \varphi'(\theta)(\xi - \tau);$$

последнее же равенство возможно лишь при $\xi = \tau$, поскольку по условию производная не может быть равна единице.

Перейдя к пределу в неравенствах (5.44) и (5.45) при $m \to \infty$, получаем оценки (5.39) и (5.40) соответственно, что и завершает доказательство теоремы.

Анализируя условия и доказательство теоремы 5.7, нельзя не заметить, что основным требованием к функции $\varphi(x)$, обеспечивающим сходимость МПИ (5.36) с оценками (5.39), (5.40), является условие малости модуля производной. Требование же $\varphi(x) \in [a;b]$ нужно лишь постольку, поскольку оно должно обеспечить попадание значений $\varphi(x)$ на промежуток, где выполняются другие требования. Очевидно, что достаточно его выполнения только на элементах итерационной последовательности (x_k) и практически это может проверяться непосредственно в процессе счета по формуле (5.36). Теоретически же, когда $[a;b]$ – не вся числовая ось (тогда надобности в этом условии нет), имеется несколько возможностей заменить требование $\varphi(x) \in [a;b]$ более конструктивным, т.е. априори проверяемым условием.

Будем исходить из того, что имеется некоторая точка x_0, которую можно взять в качестве начального приближения и в окрестности которой функция $\varphi(x)$ дифференцируема и имеет малую по модулю производную. Тогда существование и единственность корня ξ уравнения (5.38) и сходимость к нему начатого с x_0 процесса (5.36) можно связать с величиной r радиуса этой окрестности точки x_0.

Теорема 5.8. *Пусть на отрезке* $[x_0 - r; x_0 + r]$ *функция* $\varphi(x)$ *определена, дифференцируема и ее производная удовлетворяет неравенству*

$$|\varphi'(x)| \le q < 1 . \qquad (5.46)$$

Тогда если величина r *такова, что*

$$|x_0 - \varphi(x_0)| \le r(1-q), \qquad (5.47)$$

то на $[x_0 - r; x_0 + r]$ *имеется единственный корень* ξ *уравнения* (5.38), *и к нему сходится начатый с* x_0 *метод простых итераций* (5.36) *с оценками погрешности* (5.39), (5.40).

Доказательство. Так как $1-q<1$, то, в силу (5.47), $x_1 = \varphi(x_0) \in [x_0 - r; x_0 + r]$. Предположив, что $x_k = \varphi(x_{k-1})$ принадлежит r-окрестности точки x_0 при некотором $k \in N$, покажем, что там же будет и $x_{k+1} = \varphi(x_k)$. Действительно, поскольку, согласно предположению, на

отрезке $[x_0 - x_k; x_0 + x_k] \subset [x_0 - r; x_0 + r]$ функция $\varphi(x)$ определена и дифференцируема, то на этом отрезке найдется точка μ_k такая, что

$$x_{k+1} - x_0 = \varphi(x_k) - \varphi(x_0) + \varphi(x_0) - x_0 =$$
$$= \varphi'(\mu_k)(x_k - x_0) + \varphi(x_0) - x_0.$$

Отсюда, переходя к модулям, получаем неравенство

$$|x_{k+1} - x_0| \le qr + r(1-q) = r,$$

означающее, что $x_{k+1} \in [x_0 - r; x_0 + r]$.

Таким образом, на элементах итерационной последовательности (x_k) выполняется первое требование теоремы 5.7 для отрезка $[a;b]$ с $a := x_0 - r$, $b := x_0 + r$, и вместе с требованием (5.46) оно обеспечивает справедливость заключения теоремы.

Полученные для МПИ простые оценки погрешности (5.39) и (5.40) можно использовать в практических вычислениях как для завершения итерационного процесса (5.36) по правилу:

$$|x_k - x_{k-1}| \le \frac{1-q}{q}\varepsilon \quad \Rightarrow \quad \xi := x_k(\pm\varepsilon), \tag{5.48}$$

так и для предварительной прикидки числа итераций, достаточного для получения корня с заданной точностью ε:

$$\frac{q^k}{1-q}|x_1 - x_0| \le \varepsilon \quad \Leftrightarrow \quad k \ge \frac{1}{\ln q}\ln\frac{\varepsilon(1-q)}{|x_0 - \varphi(x_0)|}$$

(или $k \ge \dfrac{\ln \varepsilon - \ln r}{\ln q}$ в условиях теоремы 5.8).

Замечание 5.5. Легко видеть, что часто применяемый на практике простой критерий окончания процесса итераций (5.36) по выполнении неравенства $|x_k - x_{k-1}| \le \varepsilon$ обоснован в двух случаях: когда $-1 < \varphi'(x_k) \le 0$, т.е. МПИ дает двусторонние приближения (корень ξ всегда "зажат" между любыми двумя соседними приближениями, расстояние между которыми служит эффективной оценкой погрешности, см. рис. 5.15), и когда $0 \le \varphi'(x_k) \le q < 1$, но при этом $q \le \dfrac{1}{2}$ (тогда $\dfrac{1-q}{q} \ge 1$ и (5.48) будет заведомо выполнено). Использование этого упрощенного критерия окончания при значениях q, близких к концам интервала $(0;1)$, чревато либо лишними итерациями, либо недоитерированием.

Обратимся к связи между уравнением (5.1) и задачей о неподвижной точке – уравнением (5.38). Переписав (5.38) в виде

$$x - \varphi(x) = 0, \tag{5.49}$$

можно сказать, что это есть уравнение (5.1) с $f(x) := x - \varphi(x)$, и применять к (5.49) все рассмотренные в предыдущих пунктах рассуждения и методы.

Приведение уравнения (5.1) к виду (5.38) можно осуществлять множеством способов, но при этом всегда следует помнить, что это приведение нужно выполнять так, чтобы полученное уравнение соответствующего вида было не только эквивалентным (5.1), но и пригодным для проведения итераций, т.е. удовлетворяло оговоренным в теоремах 5.7, 5.8 условиям. Кроме того, попутно могут учитываться такие требования к получающемуся при этом методу итераций, как простота расчетной формулы, быстрота сходимости (малость q), характер сходимости (монотонность или двусторонность приближений). Если уравнение (5.1) имеет несколько корней, то для нахождения каждого из них формируется своя задача о неподвижной точке.

Иногда преобразование уравнения (5.1) к виду (5.38) не вызывает больших затруднений и выполняется непосредственно. Например, уравнение

$$4 - 2x - \sin x = 0$$

достаточно записать в виде

$$x = 2 - 0.5 \sin x,$$

чтобы сказать, что оно имеет и притом единственное в R решение, к которому сходится при любом $x_0 \in R$ последовательность

$$x_{k+1} = 2 - 0.5 \sin x_k$$

со скоростью геометрической прогрессии со знаменателем $q \le 0.5$, поскольку функция $\varphi(x) = 2 - 0.5 \sin x$ имеет производную $\varphi'(x) = -0.5 \cos x$, абсолютная величина которой не превосходит 0.5 при любых $x \in R$.

В общем случае переход от (5.1) к (5.38) осуществляют так: умножают левую и правую части уравнения (5.1) на отличный от нуля параметр $-\lambda$ и к обеим частям прибавляют по x; в результате получается равносильное (5.1) уравнение

$$x = x - \lambda f(x), \qquad (5.50)$$

которое имеет вид (5.38) с $\varphi(x) := x - \lambda f(x)$. Далее параметр λ подбирается таким, чтобы производная $\varphi'(x) = 1 - \lambda f'(x)$ в нужной области была малой по модулю (а если надо, то чтобы еще имела определенный знак).

Конкретные рекомендации по фиксированию λ в (5.50) могут быть даны в случае, когда, например, известны оценки сверху и снизу для производной исходной функции $f'(x)$. А именно, пусть

$$0 < \alpha \le f'(x) \le \gamma < \infty$$

(если производная $f'(x)$ отрицательна, можно заменить уравнение $f(x)=0$ на уравнение $-f(x)=0$, т.е. работать с функцией $-f(x)$). Тогда, соответственно,

$$1-\lambda\gamma \le \varphi'(x) \le 1-\lambda\alpha, \qquad (5.51)$$

и значит,

$$|\varphi'(x)| \le q(\lambda) := \max\{|1-\lambda\alpha|, |1-\lambda\gamma|\}.$$

Анализируя двойное неравенство (5.51), можно увидеть, что при любых $\lambda \in \left(0; \frac{2}{\gamma}\right)$ будет $q(\lambda)<1$. В частности, при $\lambda = \frac{1}{\gamma}$ имеет место неравенство

$$0 \le \varphi'(x) \le 1-\frac{\alpha}{\gamma} < 1,$$

обеспечивающее монотонную сходимость соответствующего МПИ со скоростью, определяемой оценками (5.39), (5.40) при $q = 1-\frac{\alpha}{\gamma}$. Оптимальным же значением λ является $\lambda = \lambda_0 := \frac{2}{\alpha+\gamma}$. При этом значении λ границы неравенства (5.51) таковы:

$$1-\lambda\gamma = \frac{\alpha-\gamma}{\alpha+\gamma}, \quad 1-\lambda\alpha = \frac{\gamma-\alpha}{\alpha+\gamma},$$

т.е. максимум $|\varphi'(x)|$, равный $q(\lambda_0) = \frac{\gamma-\alpha}{\alpha+\gamma}$, достигается на каждом из элементов двухэлементного множества $\{|1-\lambda\alpha|, |1-\lambda\gamma|\}$.

В отсутствие нужных оценок для $f'(x)$ можно предложить следующие рассуждения. Если известно, что искомый корень находится в окрестности заданной точки x_0, где производная меняется не очень быстро, возьмем λ таким, чтобы $\varphi'(x_0)=0$. Тогда по непрерывности производная должна остаться малой в окрестности x_0, т.е. можно рассчитывать на сходимость получающегося при этом итерационного процесса. Имеем:

$$1-\lambda f'(x_0)=0 \;\Rightarrow\; \lambda = \frac{1}{f'(x_0)}.$$

Подставляя это λ в (5.50), получаем уравнение

$$x = x - \frac{f(x)}{f'(x_0)},$$

а соответствующий ему МПИ (5.36) имеет вид

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_0)},$$

в котором узнаем ***модифицированный метод Ньютона***[*] (5.27).

5.8. УСКОРЕНИЕ СХОДИМОСТИ ПОСЛЕДОВАТЕЛЬНЫХ ПРИБЛИЖЕНИЙ

Как явствует из предыдущего пункта, МПИ (5.36) имеет лишь линейную сходимость, причем в случаях, когда производная функции $\varphi(x)$ близка к единице, эта сходимость может быть весьма медленной.

Один путь ускорения сходимости МПИ – построение ***нестационарных процессов*** на основе МПИ. На стадии приведения уравнения (5.1) к задаче о неподвижной точке (5.38) можно подменить (5.1) не однопараметрическим семейством уравнений (5.50), а последовательностью уравнений

$$x = x - \lambda_k f(x), \quad k = 0, 1, 2, \ldots,$$

и в соответствующем таким уравнениям МПИ

$$x_{k+1} = x_k - \lambda_k f(x_k) \qquad (5.52)$$

параметры λ_k подбирать так, чтобы при этом учитывалась информация, получаемая на предыдущем шаге. Если, например, выбор λ_k подчинить соображениям, что $\varphi'(x_k) = 1 - \lambda_k f'(x_k) = 0$, т.е. взять $\lambda_k = \frac{1}{f'(x_k)}$, то (5.52) будет определять основной метод Ньютона, сходящийся, как уже известно (см. п.5.4), квадратично. Можно выбирать и другие стратегии фиксирования λ_k, но вряд ли здесь следует ожидать нечто принципиально новое, что нельзя более естественно получить с помощью формулы Тейлора и интерполяционных формул.[**]

Другой путь – это алгоритмическое построение последовательностей, так или иначе "паразитирующих" на последовательности приближений МПИ (5.36), т.е. получаемых с помощью несложных арифметических манипуляций над несколькими членами последовательности (x_k) и в ре-

---

[*] В [41, 42] метод простых итераций, примененный к уравнению (5.50), т.е. процесс вида

$$x_{k+1} = x_k - \alpha f(x_k),$$

при условии, что $\operatorname{sign}\alpha = \operatorname{sign} f'(x_k)$, называют ***методом хорд*** (или параллельных хорд), и модифицированный метод Ньютона считают частным случаем этого метода.

[**] Довольно обширная и глубокая теория итерационных методов, в основу которой положено понятие ***итерационной функции***, содержится в монографии известного американского математика Дж. Трауба [50].

зультате имеющих более быструю сходимость. Для всех таких методов характерны *многошаговость, экономичность* (поскольку более быстрая сходимость по сравнению с базовой последовательностью достигается без дополнительного вычисления значений функций), а также *сложность исследования* условий и скорости сходимости, отсюда – *отсутствие эффективных априорных оценок погрешностей*. Возможны ситуации, когда новый метод окажется сходящимся, в то время как базовый для него МПИ расходится.

Рассмотрим два таких метода ускорения сходимости последовательности (5.36), не делая попыток их строгого обоснования, а ограничиваясь рациональными рассуждениями при их выводе, а также наглядными примерами, демонстрирующими их эффективность. *Наличие неподвижной точки* ξ *и дифференцируемость функции* $\varphi(x)$ *далее всюду предполагается.*

а) Δ^2- процесс Эйткена*)

Пусть (x_k) – последовательность, получаемая по формуле (5.36). Вычитая (5.36) из (5.37), имеем

$$\xi - x_{k+1} = \varphi(\xi) - \varphi(x_k),$$

а уменьшив здесь индекс на единицу, получаем

$$\xi - x_k = \varphi(\xi) - \varphi(x_{k-1}).$$

К правым частям этих равенств применим формулу Лагранжа, согласно которой найдутся точки c_k и c_{k-1} такие, что

$$\varphi(\xi) - \varphi(x_k) = \varphi'(c_k)(\xi - x_k)$$

и

$$\varphi(\xi) - \varphi(x_{k-1}) = \varphi'(c_{k-1})(\xi - x_{k-1}).$$

Таким образом, имеют место следующие связи между ошибками соседних приближений:

$$\xi - x_{k+1} = \varphi'(c_k)(\xi - x_k), \qquad \xi - x_k = \varphi'(c_{k-1})(\xi - x_{k-1}).$$

Предположим, что в той окрестности корня ξ, в которой находятся точки x_{k-1} и x_k, производная $\varphi'(x)$ меняется не очень быстро. Это допущение позволяет считать, что

$$\varphi'(c_k) \approx \varphi'(c_{k-1}) \approx \eta$$

(где η некоторое число), и значит,

$$\xi - x_{k+1} \approx \eta(\xi - x_k), \qquad \xi - x_k \approx \eta(\xi - x_{k-1}).$$

---

*) Метод (читается: "дельта-два-процесс") заложен в публикациях A.Aitken, относящихся к 1926–1955 гг. Первоначально был предназначен для улучшения метода Бернулли решения алгебраических уравнений (см. далее п.5.10)

Беря отношение этих приближенных равенств, избавляемся от η:

$$\frac{\xi - x_{k+1}}{\xi - x_k} \approx \frac{\xi - x_k}{\xi - x_{k-1}}, \qquad (5.53)$$

и разрешаем полученное приближенное уравнение относительно ξ:

$$\xi \approx \frac{x_{k+1}x_{k-1} - x_k^2}{x_{k+1} - 2x_k + x_{k-1}}.$$

Последнее выражение можно использовать на завершающем этапе применения метода простых итераций (5.36), чтобы получить более точное приближение к корню ξ с помощью трех последних членов последовательности (x_k). В развитие же метода обозначим правую часть этого приближенного равенства через $\widetilde{x}_{k+1}$ и придадим его выражению другой вид:

$$\widetilde{x}_{k+1} := \frac{x_{k+1}x_{k-1} - x_k^2}{x_{k+1} - 2x_k + x_{k-1}} = x_{k+1} - \frac{(x_{k+1} - x_k)^2}{x_{k+1} - 2x_k + x_{k-1}}.$$

Более коротко это записывается так:

$$\widetilde{x}_{k+1} = x_{k+1} - \frac{(\Delta x_k)^2}{\Delta^2 x_{k-1}}, \qquad (5.54)$$

где $\Delta x_k := x_{k+1} - x_k$, $\Delta^2 x_{k-1} := \Delta x_k - \Delta x_{k-1} = x_{k+1} - 2x_k + x_{k-1}$ – так называемые ***конечные разности первого и второго порядков*** соответственно. Отсюда название (5.54) Δ^2- ***преобразование*** или Δ^2- ***процесс Эйткена***.*)

Организация вычислений на основе этого преобразования может быть различной. Наиболее целесообразным считается применение Δ^2–***ускорения*** (5.54) через два шага МПИ на третий**). Поскольку в этом комбинированном методе нахождения корня ξ уравнения (5.38) участвует хорошо изученный МПИ (5.36), для останова процесса вычислений можно использовать в подходящей его фазе вполне надежный критерий (5.48).

Примером реализации такого метода может служить следующий алгоритм.

---

*) Вместо (5.54) для Δ^2- преобразования Эйткена используют и другое, эквивалентное (5.54), представление $\widetilde{x}_{k+1} = x_{k-1} - \frac{(\Delta x_{k-1})^2}{\Delta^2 x_{k-1}}$.

**) Применение Δ^2- преобразования менее чем через два шага МПИ не обосновано, так как оно должно применяться к трем последовательным членам л и н е й н о сходящейся последовательности.

Δ^2–алгоритм Эйткена

Шаг 0. Ввод x_0 (начального приближения), $\varphi(x)$ (исходной функции), q (оценки модуля производной), ε (допустимой абсолютной погрешности).

Шаг 1. Вычисление значений: $x_1 := \varphi(x_0)$, $x_2 := \varphi(x_1)$.

Шаг 2. Δ^2- ускорение: $\tilde{x}_2 := \dfrac{x_0 x_2 - x_1^2}{x_2 - 2x_1 + x_0}$.

Шаг 3. Вычисление контрольного значения: $x_3 := \varphi(\tilde{x}_2)$.

Шаг 4. Проверка на точность: если $|x_3 - \tilde{x}_2| > \dfrac{1-q}{q}\varepsilon$, то положить $x_0 := \tilde{x}_2$, $x_1 := x_3$, вычислить $x_2 := \varphi(x_1)$ и вернуться к шагу 2.

Шаг 5. Положить $\xi \approx x_3$ (с точностью ε).

Шаг ускорения по методу Эйткена (5.54) на базе последовательности (x_k), получаемой МПИ (5.36), имеет простую геометрическую интерпретацию (рис. 5.16 и рис. 5.17 для случаев $-1 < \varphi'(x) < 0$ и $0 < \varphi'(x) < 1$ соответственно).

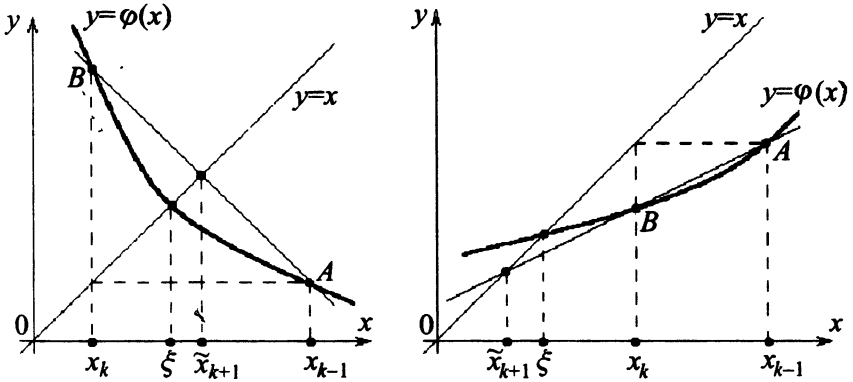


Рис. 5.16, 5.17. Геометрическая иллюстрация одного шага ускорения по методу Эйткена (случаи убывающей и возрастающей функций $\varphi(x)$)

Если провести прямую (хорду, секущую) через точки $A(x_{k-1}, \varphi(x_{k-1}))$ и $B(x_k, \varphi(x_k))$ кривой $y = \varphi(x)$, то абсцисса точки пересечения этой прямой с прямой $y=x$ как раз и будет определяемая Δ^2-преобразованием Эйткена (5.54) точка $\tilde{x}_{k+1}$ оси Ox.

Действительно, в силу (5.36), координаты точек A и B можно записать иначе: $A(x_{k-1},x_k)$, $B(x_k,x_{k+1})$. Значит, уравнение прямой (AB) имеет вид

$$\frac{y-x_k}{x_{k+1}-x_k}=\frac{x-x_{k-1}}{x_k-x_{k-1}}.$$

Рассматривая его совместно с уравнением $y=x$, т.е. заменяя в нем y и x на $\tilde{x}_{k+1}$, получаем то же выражение

$$\tilde{x}_{k+1}=\frac{x_{k+1}x_{k-1}-x_k^2}{x_{k+1}-2x_k+x_{k-1}},$$

от которого пришли к (5.54).

Такая интерпретация Δ^2- процесса Эйткена наталкивает на мысль о его возможной связи с изученным ранее методом секущих (5.32).

Полагая $f(x):=x-\varphi(x)$, применим к этой функции формулу (5.32), считая при этом, что требуемые в (5.32) значения элементов последовательности (x_k) получаются не по той же формуле (5.32), а с помощью МПИ (5.36). Имеем:

$$x_k-\frac{f(x_k)(x_{k-1}-x_k)}{f(x_{k-1})-f(x_k)}=x_k-\frac{(x_k-\varphi(x_k))(x_{k-1}-x_k)}{x_{k-1}-\varphi(x_{k-1})-x_k+\varphi(x_k)}=$$

$$=x_k-\frac{(x_k-x_{k+1})(x_{k-1}-x_k)}{x_{k+1}-2x_k+x_{k-1}}=\frac{x_{k+1}x_{k-1}-x_k^2}{x_{k+1}-2x_k+x_{k-1}}=\tilde{x}_{k+1}.$$

Таким образом, один шаг Δ^2- ускорения МПИ по методу Эйткена совпадает с одним шагом метода секущих, примененного к той же последовательности МПИ.

Сделанное наблюдение в свете известных о методе секущих сведений позволяет с большой осторожностью судить об эффекте ускорения сходимости, который может принести использование метода Эйткена. Надежность его, очевидно, выше в случае, когда $\varphi'(x)\in(-1;0)$, применение более актуально, если $q>\frac{1}{2}$, и ускорение тем эффективней, чем меньше $|\varphi''(x)|$ в окрестности корня ξ.

Применяя метод Эйткена, не следует забывать о проблеме своевременного прерывания счета из-за потерь точности при вычитании близких чисел. Подключение Δ^2- ускорения на ранней стадии МПИ, когда x_0 далеко от ξ, может привести к расходимости процесса, по крайней мере, в случае, когда $\varphi'(x)>0$ (представьте эту ситуацию, глядя на рис. 5.17). В то же время иногда с помощью метода Эйткена можно получить сходимость в условиях, когда МПИ (5.36) расходится (см., например, рис. 5.18, где $|\xi-x_k|>|\xi-x_{k-1}|$, но $|\xi-\tilde{x}_{k+1}|<|\xi-x_k|$).

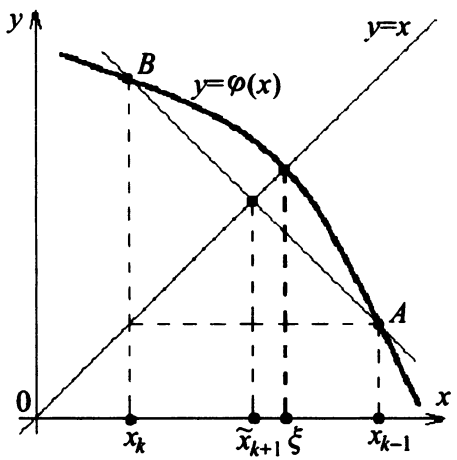


Рис. 5.18. Демонстрация возможной сходимости последовательности Эйткена $(\tilde{x}_k)$ при расходимости последовательности МПИ (x_k)

Замечание 5.6. Δ^2- преобразование Эйткена применимо не только к последовательности приближений (5.36), но и к любым другим последовательностям, сходящимся со скоростью геометрической прогрессии. Действительно, пусть (x_k) – некоторая последовательность, линейно сходящаяся к предельной точке ξ. Тогда можно считать, что разность $\xi - x_k$ изменяется по закону геометрической прогрессии, т.е. существуют такие постоянная $\nu \in (0;1)$ и слабо изменяющаяся варианта $C_k \approx C$, что

$$\xi - x_{k-1} \approx C\nu^{k-1}, \qquad \xi - x_k \approx C\nu^k, \qquad \xi - x_{k+1} \approx C\nu^{k+1}.$$

Отсюда получаем приближенные равенства

$$\frac{\xi - x_k}{\xi - x_{k-1}} \approx \nu, \qquad \frac{\xi - x_{k+1}}{\xi - x_k} \approx \nu,$$

следствием которых является равенство (5.53), в итоге приводящее к формуле (5.54). Такой подход к выводу Δ^2- метода Эйткена позволяет использовать его при решении других задач (см., например, замечание 4.6).

Применение Δ^2- преобразования Эйткена к последовательностям, сходящимся квадратично, эффекта ускорения не дает [50].

б) метод Вегстейна*)

При выводе ***метода Вегстейна*** решения задачи о неподвижной точке (5.38) будем использовать как аналитические, так и геометрические соображения.

Пусть уже найдены: $\bar{x}_k$ – элемент строящейся здесь последовательности, и $x_{k+1} = \varphi(\bar{x}_k)$ – точка, соответствующая одному шагу МПИ, примененного к точке $\bar{x}_k$. Независимо от того, сходится начатый с $\bar{x}_k$ МПИ (рис. 5.19, где $|\xi - x_{k+1}| < |\xi - \bar{x}_k|$) или расходится (рис. 5.20 с $|\xi - x_{k+1}| > |\xi - \bar{x}_k|$), отрезок AB, параллельный оси Ox и имеющий концами точки $A(\bar{x}_k, \varphi(\bar{x}_k))$ и $B(x_{k+1}, x_{k+1})$, можно разделить точкой C так, чтобы она принадлежала вертикальной прямой $x = \xi$ (при этом во втором случае речь идет о делении отрезка внешним образом).

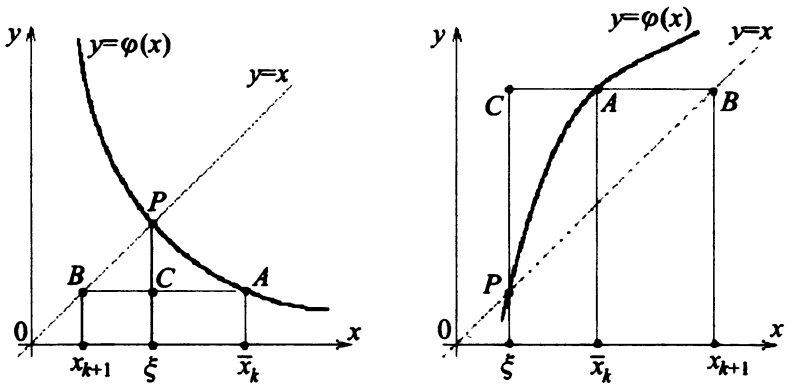


Рис. 5.19, 5.20. К построению метода Вегстейна

При любых комбинациях направлений возрастания и выпуклости графика функции $y = \varphi(x)$ в окрестности неподвижной точки ξ имеет место равенство длин отрезков $BC = PC$. Различаются два случая: когда $BC = \xi - x_{k+1}$ и когда $BC = x_{k+1} - \xi$. По формуле Лагранжа соответственно имеем

$$PC = \varphi(\xi) - \varphi(\bar{x}_k) = \varphi'(\theta_k)(\xi - \bar{x}_k)$$

или

*) Ссылки на первую публикацию этого метода (1958 г.) имеются в [36, 38] (в [38] его называют ***усовершенствованным методом последовательных приближений***). К сожалению, описание метода Вегстейна пока не заняло достойного места в отечественной учебной литературе по вычислительной математике. Нет упоминания о нем и в монографии [50].

$$PC = \varphi(\bar{x}_k) - \varphi(\xi) = \varphi'(\theta_k)(\bar{x}_k - \xi).$$

В любом случае можно утверждать, что существует точка $\theta_k \in (\bar{x}_k; \xi)$ или $\theta_k \in (\xi; \bar{x}_k)$ такая, что

$$\xi - x_{k+1} = \varphi'(\theta_k)(\xi - \bar{x}_k).$$

Разрешая это линейное уравнение относительно ξ, находим

$$\xi = x_{k+1} - \frac{x_{k+1} - \bar{x}_k}{1 - \dfrac{1}{\varphi'(\theta_k)}}. \qquad (5.55)$$

Если бы значение $\varphi'(\theta_k)$ было известно, то тем самым задача о неподвижной точке (5.38) была бы решена точно. Заменим это неизвестное значение $\varphi'(\theta_k)$ аппроксимирующим его разностным отношением:

$$\varphi'(\theta_k) \approx \frac{\varphi(\bar{x}_k) - \varphi(\bar{x}_{k-1})}{\bar{x}_k - \bar{x}_{k-1}} = \frac{x_{k+1} - x_k}{\bar{x}_k - \bar{x}_{k-1}}.$$

Подставляя приближенное значение $\dfrac{1}{\varphi'(\theta_k)} \approx \dfrac{\bar{x}_k - \bar{x}_{k-1}}{x_{k+1} - x_k}$ в (5.55), вместо корня ξ получаем приближение к нему

$$\bar{x}_{k+1} = x_{k+1} - \frac{(x_{k+1} - x_k)(x_{k+1} - \bar{x}_k)}{(x_{k+1} - x_k) - (\bar{x}_k - \bar{x}_{k-1})}. \qquad (5.56)$$

Эта итерационная формула, где $k = 1, 2, 3, \ldots$, совместно с формулой

$$x_{k+1} = \varphi(\bar{x}_k) \qquad (k = 0, 1, 2, \ldots) \qquad (5.57)$$

и начальными значениями $\bar{x}_0 := x_0$, $\bar{x}_1 := x_1$ полностью определяет метод Вегстейна для задачи (5.38).

Значение $\bar{x}_2$, получаемое по формуле Вегстейна (5.56) при заданных начальных значениях $\bar{x}_0$ и $\bar{x}_1$, совпадает со значением $\tilde{x}_2$, вычисляемым Δ^2- процессом Эйткена. Далее, т.е. при $k \geq 2$, процессы (5.54) и (5.56) различаются. Учитывая, что МПИ является составной частью метода Вегстейна, в случаях, когда $|\varphi'(x)| \leq q < 1$, можно заканчивать процесс вычислений, как и в методе Эйткена, по выполнении критерия (5.48).

Таким образом, для реализации метода (5.56)-(5.57) может быть предложен, например, следующий алгоритм.

алгоритм Вегстейна

Шаг 0. Ввести x_0 (начальное приближение), $\varphi(x)$ (исходную функцию), q (оценку модуля производной), ε (допустимую абсолютную погрешность).

Шаг 1. Вычислить $x_1 := \varphi(x_0)$; положить $\bar{x}_0 := x_0$, $\bar{x}_1 := x_1$.

Шаг 2. Вычислить $x_2 := \varphi(\bar{x}_1)$.

Шаг 3. Проверить на точность: если $|x_2 - \bar{x}_1| > \varepsilon(1-q)/q$, то вычислить $\bar{x}_2 := \dfrac{x_2\bar{x}_0 - x_1\bar{x}_1}{x_2 + \bar{x}_0 - x_1 - \bar{x}_1}$; переприсвоить значения $x_0 := x_1$, $\bar{x}_0 := \bar{x}_1$, $x_1 := x_2$, $\bar{x}_1 := \bar{x}_2$ и вернуться к шагу 2.

Шаг 4. Положить $\xi :\approx x_2$ (с точностью ε).

Разумеется, проверку на точность в подобном алгоритме можно устраивать иную (что просто необходимо, если метод Вегстейна применяется в случаях, когда $|\varphi'(x)| > 1$). Если нет угрозы большой потери точности из-за вычитания близких чисел, то заканчивать работу алгоритма Вегстейна лучше выводом значения $\xi :\approx \bar{x}_2$. Для вычисления значения $\bar{x}_2$ в этом алгоритме применена равносильная (5.56) формула

$$\bar{x}_{k+1} = \frac{x_{k+1}\bar{x}_{k-1} - x_k\bar{x}_k}{x_{k+1} + \bar{x}_{k-1} - x_k - \bar{x}_k},$$

имеющая несколько отличную от (5.56) структуру.

Как показывают многочисленные эксперименты с уравнениями вида $x = \varphi(x)$, особый интерес среди которых вызывают случаи, когда простые итерации дают расходящиеся последовательности (x_k), метод Вегстейна имеет определенные преимущества перед методом Эйткена по количеству обращений к вычислению значений $\varphi(x)$ для получения корня с заданной точностью. Чаще всего метод Вегстейна еще и позволяет в более широких пределах варьировать выбор начальной точки x_0. Результаты сравнения этих двух методов на нескольких таких уравнениях представлены в табл. 5.1. В двух ее последних столбцах указано количество вычислений значений функции $\varphi(x)$ (горнеров), потребовавшееся для достижения точки $x^* \approx \xi$ (приближенного значения корня) такой, что $|x^* - \varphi(x^*)| < 10^{-7}$. (Приведенные данные получены Ковалевым П.В. на IBM PC/AT - 286).

Таблица 5.1

Функция $\varphi(x)$	Корень ξ (точность 10^{-6})	Значение производной $\varphi'(\xi)$ (точность 10^{-2})	Начальное приближение x_0	Число горнеров	
				метод Эйткена	метод Вегстейна
x^3+2	−1.521379	6.94	−2	18	8
			−0.5	16	9
			1	190	10
$(1-x^2)^2$	0.524889	−1.52	0	6	5
			1.3	8	5
	1.490216	7.28	1.7	20	8
			4	расходится	14
$\frac{1-x}{x}$	0.618034	−2.62	0.8	10	7
	−1.618034	−0.38	−3	6	5
			−0.1	10	8
$\frac{1}{x}$	1	−1	1.2	6	5
			10	14	10
$e^{x \ln x}$	1	1	0.2	20	16
			1.2	20	15

Обратим внимание, что лишь для одного корня из семи в данной таблице можно говорить о сходимости МПИ.

5.9. НЕЛИНЕЙНЫЕ УРАВНЕНИЯ С ПАРАМЕТРОМ. БИФУРКАЦИИ*)

Материал п.5.7 можно интерпретировать так: нелинейная непрерывная математическая модель (5.38) некоего явления изучается путем построения и исследования соответствующей дискретной модели (5.36). Связь между этими моделями на отрезке $[a;b]$ устанавливается при выполнении двух следующих условий:

$$\varphi(x)\in[a;b] \qquad \forall x\in[a;b] \qquad \text{(отображение в себя)}$$

и (5.58)

$$|\varphi'(x)|<1 \qquad \forall x\in(a;b) \qquad \text{(сжатие).}$$

А именно, согласно теореме 5.7, эти условия являются достаточными для существования и единственности на $[a;b]$ решения ξ непрерывной задачи (5.38), причем оно может быть получено как предел последовательности (x_k) (т.е. как решение дискретной задачи (5.36)), начинающейся с лю-

*) Сведения, излагаемые в этом пункте, имеют ознакомительный характер и опираются на статью [3]. Из других литературных источников, посвященных рассматриваемым здесь вопросам, отметим еще книгу [58].

бой точки $x_0 \in [a;b]$. Последнее можно расценить как устойчивость в данном смысле решения ξ дискретной модели (5.36).

Продолжим изучение взаимосвязи непрерывной и дискретной одномерных нелинейных моделей в следующем русле.

Во-первых, возьмем за основу и будем рассматривать некоторую конкретную дискретную модель, а для ее исследования привлечем соответствующую непрерывную модель.

Во-вторых, попытаемся выяснить, к чему может привести нарушение условий сходимости МПИ, т.е. условий (5.58), применительно к данной модели, и какие "тайны" могут скрываться за термином *нелинейность*.

Преследуя эти цели, введем в дискретное (5.36) и непрерывное (5.38) уравнения вещественный параметр λ, т.е. будем изучать связь между моделями вида

$$x_{k+1} = \varphi(x_k, \lambda), \qquad k = 0, 1, 2, \ldots; \qquad x_0 \in [a;b] \tag{5.59}$$

и

$$x = \varphi(x, \lambda), \qquad x \in [a;b]. \tag{5.60}$$

Заметим, что не так редки ситуации, когда в приложениях математики первичными являются именно дискретные модели, а их непрерывные аналоги нужны для того, чтобы воспользоваться хорошо развитой теорией математического анализа и плодами вычислительной математики, которая, в основном, построена по принципу "непрерывная задача → дискретная аппроксимация".

Представим себе следующую весьма идеализированную картину. Пусть на некоторой ограниченной территории, например, на острове, может прокормиться не более N животных определенного вида, и пусть в начальный момент наблюдений за ними их количество было $g \in (0; N)$. Будем считать, что животные ежегодно приносят потомство, и скорость размножения характеризуется некоторым параметром $\alpha > 0$. Тогда, если через g_k обозначить численность животных в k-й год после начала наблюдения, то можно предположить, что закон ежегодного изменения численности популяции грубо описывается моделью

$$g_{k+1} = \alpha g_k (N - g_k), \qquad k = 0, 1, 2, \ldots . \tag{5.61}$$

В пользу принятия такой модели говорят следующие рассуждения. Если значение g_0 начальной численности мало, то второй сомножитель в начале процесса почти постоянен, и все зависит от коэффициента роста α: при малых α, т.е. при низкой скорости размножения численность животных будет снижаться и в конце концов популяция гибнет. Если же $g_k \in [0; N]$ возрастает и приближается к максимально возможному значению N, то за счет близости к нулю второго сомножителя численность популяции естественно начнет снижаться.

Чтобы облегчить исследование модели (5.61), упростим ее заменой переменных. Переписав (5.61) в виде

$$\frac{g_{k+1}}{N} = N\alpha \cdot \frac{g_k}{N}\left(1 - \frac{g_k}{N}\right)$$

и положив $x_k := \frac{g_k}{N}$, $\lambda := N\alpha$, приходим к уравнению*)

$$x_{k+1} = \lambda x_k(1 - x_k), \qquad (5.62)$$

где $k = 0, 1, 2, \ldots$, а значения x_k в соответствии со смыслом задачи должны принадлежать отрезку $[0;1]$.

На равенство (5.62) можно смотреть как на МПИ (5.59), применяемый к задаче о неподвижной точке вида (5.60), т.е. к задаче о корнях уравнения

$$x = \lambda x(1 - x) \qquad (5.63)$$

в области $x \in [0;1]$.

Условившись не использовать далее в обозначениях функции ее явную зависимость от параметра λ, положим

$$\varphi(x) := \lambda x(1 - x)$$

и преобразуем эту квадратичную функцию к виду

$$\varphi(x) = \frac{\lambda}{4} - \lambda\left(x - \frac{1}{2}\right)^2.$$

Из последнего следует, что $\max \varphi(x) = \frac{\lambda}{4}$ при $x = \frac{1}{2}$ и что $\varphi(x)$ отображает отрезок $[0;1]$ в $\left[0;\frac{\lambda}{4}\right]$. Значит, при $\lambda \in [0;4]$ функция $\varphi(x)$ осуществляет на отрезке $[0;1]$ отображение в себя, т.е. при этих λ элементы последовательности (x_k), получаемой с помощью равенства (5.62), при любом $x_0 \in [0;1]$ не выйдут за пределы $[0;1]$, другими словами, определены при любом $k = 0, 1, 2, \ldots$**).

Для производной данной функции $\varphi(x)$ имеем:

$$\varphi'(x) = \lambda(1 - 2x) \qquad \text{и} \qquad \max_{x \in [0;1]} |\varphi'(x)| = \lambda.$$

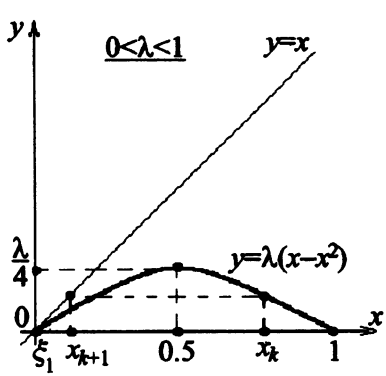
Следовательно, при $\lambda < 1$ отображение $\varphi(x)$ является сжимающим на $[0;1]$ и имеет единственную неподвижную точку $\xi_1 \in [0;1]$, а именно

---

*) Это конкретное уравнение называют ***логистическим*** [58].

**) При $\lambda > 4$ нарушается соответствие между дискретной (5.62) и непрерывной (5.63) моделями. Например, при $x_0 = \frac{1}{2}$ уже $x_1 = \frac{\lambda}{4} > 1$, т.е. не принадлежит множеству, на котором определена функция $\varphi(x)$

$\xi_1 = 0$, которая является пределом последовательности (x_k) при любом начальном значении $x_0 \in [0;1]$ (популяция гибнет по причине недостаточной скорости воспроизводства). Геометрическая иллюстрация этого случая, соответствующего выполнению условий теоремы 5.7, показана на рис.5.21.



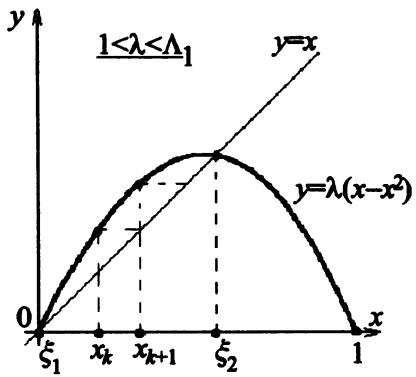


Рис. 5.21, 5.22. Сходимость МПИ (5.62) *к корню* $\xi_1 = 0$ *логистического уравнения* (5.63) *при* $\lambda \in (0;1)$ *и к корню* $\xi_2 = \dfrac{\lambda - 1}{\lambda}$ *при* $\lambda \in (1;\Lambda_1)$

При $1 \le \lambda \le 4$ нарушается одно из условий сходимости МПИ: $\varphi(x)$ не является функцией сжатия. Что же это влечет?

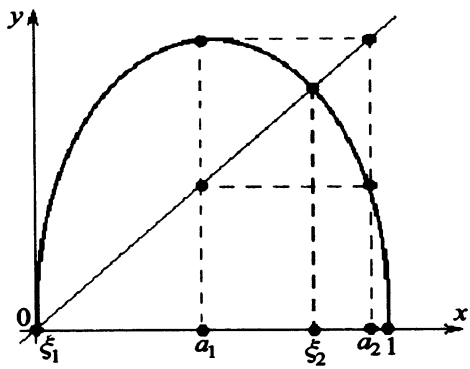


Рис. 5.23. Предельное поведение последовательности $x_{k+1} = \varphi(x_k)$, *где* $\varphi(x) := \lambda(x - x^2)$ *при* $\lambda \in (\Lambda_1;\Lambda_2)$

Очевидно, уравнение (5.63) по-прежнему сохраняет решение $\xi_1 = 0 \in [0;1]$. Но при переходе λ через 1 это решение теряет устойчивость и появляется второе решение $\xi_2 = \dfrac{\lambda - 1}{\lambda} \in [0;1]$, которое следует считать устойчивым, поскольку теперь именно к нему будет сходиться любая последовательность, определяемая начатым с $x_0 \in [0;1]$ МПИ (5.62) (рис. 5.22). Произошло явление, которое носит название ***бифуркация решений***: вместо одного решения на рассматриваемом промежутке стало два решения*).

Сходимость (x_k) к ξ_2 будет наблюдаться не при всех $\lambda \in [1;4]$. Оказывается, существует число $\Lambda_1 > 1$ такое, при переходе λ через которое начнет происходить ***зацикливание*** последовательности (x_k). А именно, какое бы ни взяли $x_0 \in (0;1)$, начатая с него и продолжаемая по формуле (5.62) последовательность будет обладать тем свойством, что все ее четные члены будут иметь пределом одно число, а нечетные – другое. Это означает, что найдутся числа $a_1, a_2 \in (0;1)$ (при каждом λ свои) такие, что $a_2 = \varphi(a_1)$ и $a_1 = \varphi(a_2)$, причем $a_1 \neq a_2 \neq \xi_i \ (i = 1,2)$ (см. рис. 5.23). В этом случае говорят, что дискретное отображение (5.62) имеет ***устойчивый цикл периода 2*** и обозначают его S^2. Относя это к исходной модельной задаче с животными, можно сказать, что при значениях $\lambda > \Lambda_1$ численность популяции будет меняться периодически с периодом в два "года".

Зная ситуацию качественно, нетрудно найти точно пороговое значение Λ_1, при котором появляется устойчивый цикл S^2.

Действительно, если известно, что на $(0;1)$ имеются точки a_1, a_2 такие, что $a_2 = \varphi(a_1)$, $a_1 = \varphi(a_2)$, то значит, $a_2 = \varphi(\varphi(a_2))$, $a_1 = \varphi(\varphi(a_1))$, т.е. a_1 и a_2 – неподвижные точки отображения $\varphi^{(2)}(x) := \varphi(\varphi(x))$, иначе, – корни уравнения $x = \varphi(\varphi(x))$.

Так как эта суперпозиция сохраняет старые неподвижные точки $\xi_1 = 0$ и $\xi_2 = (\lambda - 1)/\lambda$, го уравнение

$$x = \lambda(\lambda x(1-x))(1 - \lambda x(1-x))$$

должно иметь четыре корня, из которых два известны (рис. 5.24). Исключив из этого уравнения известные корни, приходим к квадратному уравнению

$$x^2 - \frac{\lambda + 1}{\lambda} x + \frac{\lambda + 1}{\lambda^2} = 0.$$

---

*) Bifurcus (лат.) – раздвоенный. Термин введен К. Якоби в 1834 году Теория бифуркаций заложена Анри Пуанкаре в конце XIX века.

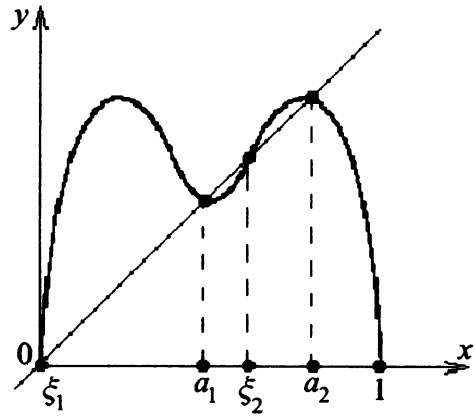


Рис. 5.24. График функции $\varphi^{②}(x) := \varphi(\varphi(x))$ *при* $\varphi(x) := \lambda\left(x - x^2\right)$, $\lambda \in (\Lambda_1, \Lambda_2)$ *и ее неподвижные точки* (ξ_1, ξ_2, a_1, a_2)

Положительное значение λ, которое служит границей области положительности дискриминанта $D = \left(\lambda^2 - 2\lambda - 3\right)/\lambda^2$ этого уравнения, как раз и есть искомое значение Λ_1, начиная с которого появляются новые устойчивые неподвижные точки a_1, a_2, т.е. цикл S^2. Очевидно, это $\Lambda_1 = 3$. Устойчивость неподвижных для $\varphi^{②}(x)$ точек $a_{1,2} = \left(\lambda + 1 \pm \sqrt{\lambda^2 - 2\lambda - 3}\right)/2\lambda$ в том смысле, что они становятся точками четно-нечетного притяжения для последовательности (x_k), устанавливается непосредственной проверкой условия $\left|\left.\dfrac{d\varphi^{②}(x)}{dx}\right|_{x=a_{1,2}}\right| < 1$.

Дальнейшее увеличение λ в дискретном уравнении (5.62) ведет к тому, что начиная с некоторого Λ_2 зацикливание будет иметь более сложный характер: при каждом $\lambda \in (\Lambda_2, \Lambda_3)$, где $\Lambda_2 > 3$, $\Lambda_3 < 4$, найдутся числа a_1, a_2, a_3, a_4 (зависящие от λ) такие, что $a_2 = \varphi(a_1)$, $a_3 = \varphi(a_2)$, $a_4 = \varphi(a_3)$, $a_1 = \varphi(a_4)$, и члены последовательности (x_k) будут поочередно все сильнее притягиваться к этим числам, с какого $x_0 \in (0;1)$ ни начинался бы процесс (5.62). Говорят, что в этом случае имеет место устойчивый цикл периода 4, т.е. S^4.

Такой процесс образования новых циклов происходит с увеличением параметра λ и далее. Точки Λ_1, Λ_2, Λ_3, ... , в которых имеет место зарождение циклов S^2, S^4, S^8, ..., называются ***точками бифуркации удвоения периода***.

Этому процессу бифуркаций удвоения периода можно придать наглядный вид, если отобразить на графике зависимость ***элементов цикла*** – значений устойчивых неподвижных точек отображений $\varphi^{(n)}(x)$ (иначе, точек притяжения подпоследовательностей получаемой посредством МПИ (5.62) последовательности (x_k)) – от значений параметра λ при $\lambda > 1$ (рис. 5.25).

Последовательность (Λ_n) точек бифуркации удвоения периода обладает определенной закономерностью:

$$\frac{\Lambda_n - \Lambda_{n-1}}{\Lambda_{n+1} - \Lambda_n} \xrightarrow[n\to\infty]{} \delta = 4.66920\ldots .$$

Имеет постоянный предел, равный величине $\alpha = 2.50290\ldots$, также отношение d_{n-1}/d_n, где через d_n обозначено расстояние от точки $x = 0.5$ до ближайшего элемента цикла $S^{2^{n-1}}$, соответствующего такому значению λ, при котором $x = 0.5$ является элементом того же цикла (рис. 5.25). Числа δ и α называют ***постоянными Фейгенбаума*** в честь открывшего эти закономерности американского математика (1978 г.).

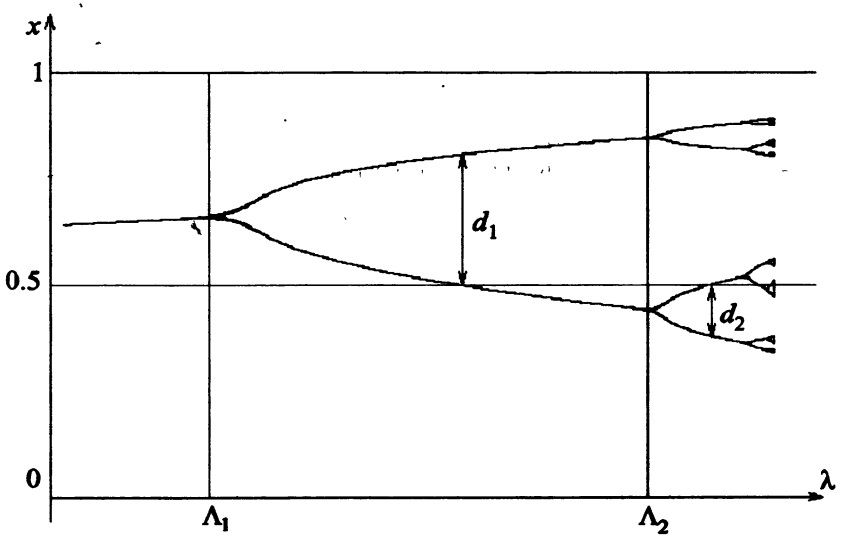


Рис. 5.25. Бифуркационная диаграмма циклов периода 2^n

Верхней границей значений λ, при которых получаемая с помощью (5.62) последовательность (x_k) ведет себя указанным образом, т.е. имеются циклы S^{2^n}, является значение $\lambda^* \approx 3.57$. Дальнейшее увеличение λ приводит к срыву цикличности. В каком-то диапазоне значений $\lambda > \lambda^*$ будет наблюдаться бесконечное хаотическое блуждание точек последовательности (x_k) в пределах промежутка $(0;1)$, с какого бы $x_0 \in (0;1)$ она не начиналась. Затем снова из хаоса возникают устойчивые циклы, происходят бифуркации удвоения периода и опять срыв в хаос. Такие чередования циклического (с разными периодами, например, $12 \cdot 2^n$, $10 \cdot 2^n$, $6 \cdot 2^n$, $8 \cdot 2^n$, $7 \cdot 2^n$, $5 \cdot 2^n$ и др.) и хаотического поведения последовательности (x_k) имеют место в процессе увеличения λ почти вплоть до предельного значения $\lambda = 4$. При этом самый большой (по λ) промежуток цикличности после циклов вида S^{2^n} будет иметь цикл периода $3 \cdot 2^n$ (при $\lambda \gtrsim 3.829$), играющий особую роль в теории бифуркаций.

Как зарождается порядок в хаосе, каковы связи между циклами разных периодов, какую роль играет последовательность обхода элементов цикла, что можно сказать об устойчивости тех или иных циклов – эти и другие вопросы возникают перед математиками, изучающими нелинейные отображения. Непростые ответы на них, достаточно наглядные в одномерном случае, позволяют понять природу многих сложных явлений (например, оценить принципиальные возможности долгосрочного прогнозирования погоды).

5.10. О МЕТОДАХ РЕШЕНИЯ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ. МЕТОД БЕРНУЛЛИ

Пусть требуется найти один, несколько или все корни многочлена с действительными коэффициентами

$$P_n(x) := a_0 x^n + a_1 x^{n-1} + \ldots + a_{n-1} x + a_n \,, \qquad (5.64)$$

т.е. решить в каком-то указанном смысле уравнение

$$P_n(x) = 0. \qquad (5.65)$$

Если речь идет о нахождении только действительных корней и, особенно, если нужно найти не все, а только некоторые из них, то есть резон в применении к алгебраическому уравнению (5.65) какого-нибудь из рассмотренных выше методов решения нелинейных скалярных уравнений (5.1) с $P_n(x)$ в роли $f(x)$. При этом следует обратить внимание на то, что $P_n(x)$ является определенной на всей действительной оси бесконечно дифференцируемой функцией, и здесь можно применять методы, использующие производные $P_n(x)$. Поскольку дифференцирование понижает

степень многочлена, вычисление значений производных потребует даже меньше арифметических действий, чем вычисление значений исходного многочлена. Отсюда – целесообразность нахождения отдельных корней многочлена (в том числе и комплексных, см. [21, 23, 38, 40, 47]) методом Ньютона или гибридным алгоритмом, его использующим.

Применение итерационных методов решения нелинейных скалярных уравнений (5.65), а также многих других методов требует многократного вычисления значений многочленов (5.64). Эта промежуточная задача для многочленов намного проще, чем задача вычисления значений трансцендентных функций: нужно всего лишь при заданном $x = x_0$ простым перемножением находить степени x_0^i при $i = 1, 2, \ldots, n$ (это $n-1$ умножение), затем умножить их на коэффициенты (n умножений) и сложить результаты (n сложений). Однако в практике вычислений используют более эффективный способ вычисления значений многочленов – ***схему Горнера***, позволяющую почти вдвое уменьшить количество умножений*). Выведем этот способ.

Согласно теореме Безу, при любом x_0 найдутся числа b_i $(i = 0, 1, 2, \ldots, n)$ такие, что

$$a_0 x^n + a_1 x^{n-1} + \ldots + a_{n-1} x + a_n \equiv$$

$$\equiv (x - x_0)\left(b_0 x^{n-1} + b_1 x^{n-2} + \ldots + b_{n-2} x + b_{n-1}\right) + b_n,$$

причем $b_n = P_n(x_0)$. Выполняя умножение в правой части тождества и приравнивая коэффициенты при одинаковых степенях x, получим совокупность $n+1$ равенств:

$$b_0 = a_0, \quad b_1 = a_1 + x_0 b_0, \quad b_2 = a_2 + x_0 b_1, \quad \ldots \quad , \quad b_n = a_n + x_0 b_{n-1}.$$

Таким образом, после n умножений (на одно и то же число x_0) и n сложений приходим к искомому результату $b_n = P_n(x_0)$.

Однотипность вычислений, производимых в схеме Горнера, позволяет организовать их в цикл, определяемый формулой

$$b_i = a_i + x_0 b_{i-1},$$

где $i = 1, 2, \ldots, n$, $b_0 := a_0$, со значением $P_n(x_0) = b_n$ на его выходе.

Числа $b_0, b_1, \ldots, b_n$ далее будем называть ***коэффициентами схемы Горнера***. При ручном счете эти коэффициенты удобно записывать в процессе вычислений под соответствующими размещенными в один ряд коэффициентами данного многочлена (5.64) в виде следующей таблицы:

	a_0	a_1	a_2	$\cdots$	a_{n-1}	a_n
x_0	b_0	b_1	b_2	$\cdots$	b_{n-1}	b_n

*) Такой способ был известен в Китае еще в средние века и назывался ***Тяньюань***, а затем в начале XIX века был "переоткрыт" в Европе англичанином Горнером и итальянцем Руффини.

Пример 5.4. Доказать, что число 2 является единственным рациональным корнем многочлена $P_5(x) := x^5 - x^4 - 3x^3 + 2x + 4$, причем простым.

Все доказательство можно отразить следующей таблицей:

	1	−1	−3	0	2	4
2	1	1	−1	−2	−2	0
1	1	2	1	−1	−3	
−1	1	0	−1	−1	−1	
2	1	3	5	8	14	
−2	1	−1	1	−4	6	

В ее первой строке находятся коэффициенты данного многочлена $P_5(x)$, во второй строке – коэффициенты схемы Горнера, примененной к $P_5(x)$ при $x_0 = 2$, означающие, что $P_5(2) = 0$ и $P_5(x) = (x-2)P_4(x)$, где $P_4(x) = x^4 + x^3 - x^2 - 2x - 2$ (см. тождество, лежащее в основе вывода схемы Горнера). Остальные строки – результаты применения схемы Горнера к многочлену $P_4(x)$ (что отражено в таблице подчеркиванием) в точках $\pm 1, \pm 2$, являющихся делителями свободного члена многочлена с целыми коэффициентами $P_4(x)$ со старшим коэффициентом 1. Так как $P_4(1) = -3$, $P_4(-1) = -1$, $P_4(2) = 14$, и $P_4(-2) = 6$, то $P_4(x)$ не имеет рациональных корней, а значит, и $P_5(x)$ не имеет других рациональных корней, кроме 2, причем 2 не может быть двухкратным корнем.

Как видно из приведенного примера, схему Горнера удобно использовать для понижения степени алгебраического уравнения выделением линейного множителя, соответствующего известному вещественному корню. Поскольку в подавляющем большинстве случаев эти корни бывают известны лишь приближенно, при понижении степени алгебраического уравнения неизбежна потеря точности, с которой могут быть найдены последующие корни. Это обстоятельство заставляет с осторожностью относиться к выигрышу в вычислительных затратах, достигаемому таким понижением степени, когда находятся несколько корней многочлена последовательно корень за корнем.

Существует способ вычисления корня многочлена (5.64) последовательно цифра за цифрой непосредственно по схеме Горнера, применяемой к специальным образом преобразованному многочлену; в [20] такой способ называется ***методом Горнера***.

Более распространен метод выделения множителей*) [8, 23, 34, 35, 40]. Последовательное выделение линейного множителя этим методом

*) На самом деле, имеется несколько таких методов, из которых наиболее известен ***метод Лина*** (***предпоследнего остатка***), который и имеется здесь в виду.

базируется на применении схемы Горнера в точках последовательности приближений $x_0, x_1, x_2, \ldots$ к искомому корню ξ многочлена $P_n(x)$. Это оказывается равнозначным простым итерациям вида

$$x_{k+1} = \frac{P_n(0)x_k}{P_n(0) - P_n(x_k)}, \quad k = 0, 1, 2, \ldots,$$

что позволяет использовать известные критерии сходимости МПИ. При нахождении пары комплексно сопряженных корней выделяют к в а д р а т и ч н ы й множитель. Для этих целей нужна эффективная схема деления с остатком многочлена на квадратный трехчлен, которую нетрудно получить по аналогии с выводом схемы Горнера.

В случаях, когда вычисление корней многочлена ориентировано на применение итерационных методов решения нелинейных уравнений общего вида, встает вопрос о сужении области поиска корней. Здесь опять полезную службу может сослужить схема Горнера. Справедливо утверждение: *если все коэффициенты схемы Горнера, примененной к многочлену $P_n(x)$ в точке $x = \beta > 0$, положительны, то правее β на оси Ox действительных корней $P_n(x)$ нет.* Это следует из тождества

$$P_n(x) \equiv (x - \beta)\left(b_0 x^{n-1} + b_1 x^{n-2} + \ldots + b_{n-2}x + b_{n-1}\right) + b_n ,$$

показывающего, что $P_n(x) > 0$ для любых $x \geq \beta > 0$, в силу положительности всех b_i $(i = 0, 1, 2, \ldots, n)$ и разности $x - \beta$.

Так, в примере 5.4, глядя на приведенную там таблицу, можно сказать, что многочлен $P_4(x)$ (а значит, и $P_5(x)$) не имеет действительных корней, больших 2.

Есть более конструктивные способы указания границ всех корней многочлена.

Например, можно утверждать, что *за верхнюю границу положительных корней многочлена $P_n(x)$ с $a_0 > 0$ можно принять число*

$$R = 1 + \sqrt[m]{\frac{B}{a_0}},$$

где m – индекс первого отрицательного коэффициента в ряду $a_1, a_2, \ldots, a_n$ (иначе, разность между показателем степени многочлена и показателем степени первого отрицательного члена), а *B – максимум модулей всех отрицательных коэффициентов*[*].

---

[*] В одних литературных источниках такое нахождение правой границы действительных корней называют ***методом Лагранжа*** [20, 21], в других – ***методом Маклорена*** [23].

Действительно, заменив в $P_n(x)$ неотрицательные коэффициенты $a_1, a_2, \ldots, a_{m-1}$ нулями, а все последующие – числом $-B$, при $x > 1$ имеем:

$$P_n(x) \geq a_0 x^n - B\left(x^{n-m} + \ldots + x + 1\right) =$$

$$= a_0 x^n - B\frac{x^{n-m+1} - 1}{x-1} > a_0 x^n - \frac{Bx^{n-m+1}}{x-1} = \qquad (5.66)$$

$$= \frac{x^{n-m+1}}{x-1}\left[a_0 x^{m-1}(x-1) - B\right] > \frac{x^{n-m+1}}{x-1}\left[a_0(x-1)^m - B\right].$$

Так как последнее выражение в этой цепочке равенств и неравенств равно нулю при $x = R$ и больше нуля при $x > R$, то значит, $P_n(x) > 0$ при всех $x \geq R$, т.е. правее R действительных корней нет.

Любой метод нахождения верхней границы положительных корней можно приспособить для нахождения нижней (левой) границы отрицательных корней. Для этого достаточно преобразовать многочлен $P_n(x)$ заменой $t = -x$ и для положительных корней многочлена $(-1)^n P_n(t)$ найти верхнюю границу R^*; тогда число $-R^*$ будет искомой нижней границей*). Так же, используя известный метод нахождения верхней границы в многочленах $P_n\left(\frac{1}{x}\right)$ и $(-1)^n P_n\left(-\frac{1}{x}\right)$, можно найти нижнюю границу положительных и верхнюю границу отрицательных корней многочлена $P_n(x)$, т.е. отделить его корни от нуля.

Зачастую более хорошие результаты показывает ***метод Вестерфильда*** получения симметричных границ расположения всех корней многочлена [23]. Согласно ему, *модули всех корней* (в том числе и комплексных) *приведенного многочлена* $P_n(x)$ (т.е. при $a_0 = 1$) *лежат в круге, радиус которого не превосходит суммы двух наибольших из чисел* $\sqrt[m]{|a_m|}$, *где* $m = 1, 2, \ldots, n$.

Решение проблемы изоляции корней алгебраического уравнения не имеет особой специфики, выделяющей ее из более общего случая нелинейных скалярных уравнений. Более продвинутым здесь является решение вопроса о количестве действительных корней. Можно указать простые способы выяснения возможного количества положительных и отрицательных корней по числу перемен знаков в последовательностях коэффициентов $P_n(x)$ и $P_n(-x)$.

---

*) Множитель $(-1)^n$ поставлен ради положительности старшего коэффициента преобразованного многочлена.

Так, ***теорема Декарта*** *говорит о том, что число положительных корней уравнения* (5.65) *с учетом их кратностей равно числу перемен знаков в последовательности коэффициентов* $a_0, a_1, \ldots, a_n$ (без учета нулевых коэффициентов) *или на четное число меньше* [21].

Более точный ответ на вопрос о числе действительных корней алгебраического уравнения можно получить с помощью широко известной в алгебре ***теоремы Штурма*** (см. [8, 20, 21] и др.). Если при этом уже затрачены усилия на составление системы Штурма, то ее можно использовать и для нахождения промежутков изоляции действительных корней.

Имеются и другие способы нахождения границ действительных и комплексных корней алгебраических уравнений, выяснения количества положительных и отрицательных корней, а также их изоляции.

Одним из наиболее эффективных методов нахождения в с е х или почти всех корней алгебраического уравнения, как вещественных, так и комплексных, является ***метод Лобачевского***, предложенный выдающимся русским математиком в 1834 году*). Основная идея метода заключается в последовательном применении операции квадрирования корней. Суть ее такова.

С помощью обобщенной теоремы Виета легко показать, что если корни $\xi_1, \xi_2, \ldots, \xi_n$ уравнения (5.65) сильно отличатся по модулю (что, кстати, гарантирует их вещественность), а именно, $|\xi_1| >> |\xi_2| >> \cdots >> |\xi_n|$, то

$$\xi_1 \approx -\frac{a_1}{a_0}, \quad \xi_2 \approx -\frac{a_2}{a_1}, \quad \ldots \quad , \quad \xi_n \approx -\frac{a_n}{a_{n-1}}.$$

При возведении их в натуральную степень будут получаться все более удаленные друг от друга числа. Одна из макроопераций метода Лобачевского (***квадрирование корней***) состоит в том, что от данного уравнения с корнями $\xi_1, \xi_2, \ldots, \xi_n$ переходят к новому уравнению той же степени с коэффициентами

$$A_i = a_i^2 + 2\sum_{j=1}^{i}(-1)^j a_{i-j}a_{i+j} \quad (i = 0, 1, 2, \ldots, n) \qquad (5.67)$$

и корнями $\mu_i = -\xi_i^2 \ (i = 1, 2, \ldots, n)$. После ее многократного применения, когда в (5.67) будет практически сведена на нет роль второго слагаемого (удвоенной суммы парных произведений), извлечением корней соответствующих степеней можно приближенно найти модули корней исходного уравнения.

---

*) Этот метод называют также ***методом Лобачевского-Греффе*** или ***методом Данделена*** в честь швейцарского математика Греффе и французского математика Данделена, причастных к одним из первых версий метода. Впоследствии метод Лобачевского неоднократно совершенствовался.

Более подробно о методе Лобачевского, его реализациях, разных ситуациях, с которыми можно встретиться при его применении, см. в книгах [8, 21, 23, 40]. Одна из последних версий этого не самого простого метода содержится в брошюре А.А. Беланова [7] (ориентированной, правда, на программируемые микрокалькуляторы, а не на ЭВМ).

Рассмотрим простой способ приближенного вычисления наибольшего по модулю действительного корня алгебраического уравнения (5.65), который был опубликован И. Бернулли в 1732 г. и называется ***методом Бернулли***.

Запишем уравнение (5.65) в виде

$$x^n = c_1 x_1^{n-1} + c_2 x_2^{n-2} + \ldots + c_{n-1} x + c_n .$$

С помощью коэффициентов $c_i = -\dfrac{a_i}{a_0}$ $(i = 1, 2, \ldots, n)$ этого уравнения будем строить последовательность $\left(u_{n+k}\right)_{k=1}^{\infty}$ по рекуррентной формуле

$$u_{n+k} = c_1 u_{n+k-1} + c_2 u_{n+k-2} + \ldots + c_{n-1} u_{k+1} + c_n u_k , \qquad (5.68)$$

начиная этот процесс при $k = 1, 2, \ldots, n$ со значений

$$u_1 = 0, \quad u_2 = 0, \quad \ldots \quad , \quad u_{n-1} = 0, \quad u_n = k$$

(как это рекомендовано Хильдебрандом).

На основе обобщенной теоремы Виета можно выяснить, что эта последовательность обладает замечательным свойством: если $\xi_1, \xi_2, \ldots, \xi_n$ – корни многочлена $P_n(x)$, то

$$\begin{aligned}
&\xi_1 + \xi_2 + \ldots + \xi_n = u_{n+1}, \\
&\xi_1^2 + \xi_2^2 + \ldots + \xi_n^2 = u_{n+2}, \\
&\ldots\ldots\ldots\ldots\ldots\ldots \\
&\xi_1^k + \xi_2^k + \ldots + \xi_n^k = u_{n+k} \\
&\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned}$$

Взяв отношение последующего члена u_{n+k+1} к предыдущему u_{n+k}, выраженных через степени корней, имеем:

$$\frac{u_{n+k+1}}{u_{n+k}} = \frac{\xi_1^{k+1} + \xi_2^{k+1} + \ldots + \xi_n^{k+1}}{\xi_1^k + \xi_2^k + \ldots + \xi_n^k} = \frac{\xi_1^{k+1}\left[1 + \left(\dfrac{\xi_2}{\xi_1}\right)^{k+1} + \ldots + \left(\dfrac{\xi_n}{\xi_1}\right)^{k+1}\right]}{\xi_1^k\left[1 + \left(\dfrac{\xi_2}{\xi_1}\right)^{k} + \ldots + \left(\dfrac{\xi_n}{\xi_1}\right)^{k}\right]}. \qquad (5.69)$$

Если $|\xi_1| > |\xi_i| \quad \forall i = 2,\ldots,n$, то, очевидно, $\frac{u_{n+k+1}}{u_{n+k}} \underset{k\to\infty}{\to} \xi_1$. К такому же результату придем и в случае, когда ξ_1 – не простой, а m-кратный корень (выражения в квадратных скобках в числителе и в знаменателе (5.69) имеют пределом m).

Если сходимость последовательности отношений $\frac{u_{n+k+1}}{u_{n+k}}$ не обнаруживается, но существует предел последовательности величин

$$\frac{u_{n+k+2}}{u_{n+k}} = \xi_1^2 \frac{1+\left(\frac{\xi_2}{\xi_1}\right)^{k+2}+\ldots+\left(\frac{\xi_n}{\xi_1}\right)^{k+2}}{1+\left(\frac{\xi_2}{\xi_1}\right)^{k}+\ldots+\left(\frac{\xi_n}{\xi_1}\right)^{k}},$$

то в этом случае данное уравнение имеет два действительных корня $\xi_1 = -\xi_2$, наибольших по модулю. При достаточно больших k (желательно, четных) они могут быть найдены приближенным равенством $\xi_{1,2} \approx \pm\sqrt{\frac{u_{n+k+2}}{u_{n+k}}}$. Хаотическое поведение последовательности $\left(\frac{u_{n+k+2}}{u_{n+k}}\right)$ говорит о том, что превалирующими являются комплексные корни. Незначительной доработкой метод Бернулли можно приспособить и для этого случая (см. [23]).

Как видно, метод Бернулли весьма близок к степенному методу решения частичной проблемы собственных значений, более тщательно рассмотренному в п.4.2. Зная в тонкостях один из этих методов, можно многое сказать о поведении другого.

УПРАЖНЕНИЯ

5.1. Доказать существование и единственность корня уравнения

$$(x-1)^2 e^x - 7 = 0.$$

5.2. Найти промежуток локализации (единичной длины) отрицательного корня уравнения $x^3 - x^2 + 4 = 0$. За сколько шагов метода половинного деления можно уточнить корень до 0.1? до 0.01? до 10^{-6}?

5.3. С точностью до 0.01 решить уравнение $\sqrt{|x-4|} - x + 1 = 0$:

а) методом половинного деления; б) методом хорд.

5.4. С точностью до 0.001 найти положительный корень уравнения

$$x^4 - 2x - 4 = 0:$$

а) методом Ньютона; б) методом секущих.

Уточнить корень, применив к результату б) два шага метода Стеффенсена.

5.5. Подготовить алгоритм вычисления значения функции $y = \sqrt[7]{x}$ в точке $x = 100$ с точностью $\varepsilon = 10^{-6}$, пользуясь правилом Ньютона. Сделать два приближения.

5.6. На основе метода Ньютона записать итерационный процесс, который позволял бы вычислять значения $\frac{1}{\sqrt{a}}$ при заданных вещественных значениях a, не производя делений.

5.7. Составить гибридный алгоритм "секущих-половинного деления". Изучить его поведение на уравнении $x^4 + (x-2)^2 - 8 = 0$, строя приближения к положительному корню из разных пар начальных точек x_0, x_1.

5.8. Дать обоснованное заключение о факте и скорости сходимости последовательности

$$x_{k+1} = \frac{1}{4}x_k^4 - \frac{1}{2}x_k + 1, \quad k = 0, 1, 2, \ldots, \quad x_0 = 0$$

на отрезке $[0;1]$. С какой точностью можно приблизиться к $\lim\limits_{k\to\infty} x_k$ за 10 шагов? Записать итерационный процесс Ньютона, имеющий тот же предел.

5.9. Записать сходящийся процесс простых итераций для нахождения корня уравнения $x^3 - 2x^2 - 4x - 7 = 0$, изолированного на промежутке $[3;4]$.

5.10. Подготовить расчетные формулы для нахождения корня уравнения $2 - \lg x - x = 0$ методом простых итераций. Выбрав начальное приближение, подсчитать, за сколько итерационных шагов можно гарантировать получение корня с точностью $\varepsilon = 10^{-6}$. Сделать четыре шага МПИ и по два полных шага Δ^2 -процесса Эйткена и метода Вегстейна.

5.11. Сравнить эффективность следующих двух подходов к вычислению значений $P_n(x_k)$ и $P_n'(x_k)$ при нахождении алгебраического корня уравнения $P_n(x)=0$ методом Ньютона $x_{k+1}=x_k-\dfrac{P_n(x_k)}{P_n'(x_k)}$:

а) вычислив степени x_k^i, использовать их затем в $P_n(x_k)$ и $P_n'(x_k)$;

б) применить схему Горнера сначала к многочлену $P_n(x)$ в точке x_k, затем к его неполному частному – результату первого ее применения (правомерность этого подхода обосновать!).

5.12. Придумать алгоритм, позволяющий найти наибольший корень многочлена $P_n(x)$ с заданной точностью только одними проверками на положительность коэффициентов схемы Горнера.

5.13. а) Доказать, что если R_1 и R_2 – правые границы действительных корней многочленов $Q(x)$ и $T(x)$ соответственно, то за верхнюю границу действительных корней многочлена $P(x):=Q(x)+T(x)$ можно принять число $R=\max\{R_1,R_2\}$.

б) Для многочлена $P_5(x)=x^5-2x^4+15x^2-32x+1$ найти границы положительных корней сначала непосредственно по формуле (5.66), затем на основе а), выполняя подходящее представление $P_5(x)$ суммой $Q(x)+T(x)$ так, чтобы получить как можно более точную границу.

в) Найти границу модулей всех корней данного многочлена $P_5(x)$ методом Вестерфильда.

5.14. Пусть ξ_1, ξ_2 – вещественные корни уравнения $x^2-px+q=0$. Убедиться в справедливости равенств

$$\xi_1+\xi_2=u_3, \xi_1^2+\xi_2^2=u_4, \quad \xi_1^3+\xi_2^3=u_5,$$

где u_i – элементы последовательности Бернулли-Хильдебранда, задаваемой формулой (5.68).

5.15. Дано уравнение $x^4-3x^3-7x^2+15x+18=0$. Сделать 5–6 приближений к его старшему корню методом Бернулли.

ГЛАВА 6 МЕТОДЫ РЕШЕНИЯ СИСТЕМ НЕЛИНЕЙНЫХ УРАВНЕНИЙ

Рассматривается ряд методов решения систем алгебраических и трансцендентных уравнений. Среди них метод простых итераций, метод Ньютона в разных модификациях, метод Брауна. Показывается связь между данной задачей и задачей безусловной минимизации функции нескольких переменных. Проводится сравнение методов на примере решения конкретной системы. С единых позиций изучается сходимость метода Ньютона и метода, получаемого из него применением итерационного процесса Шульца для приближенного обращения матриц Якоби.

6.1. ВЕКТОРНАЯ ЗАПИСЬ НЕЛИНЕЙНЫХ СИСТЕМ. МЕТОД ПРОСТЫХ ИТЕРАЦИЙ

Пусть требуется решить систему уравнений

$$\begin{cases} f_1(x_1, x_2, \ldots, x_n) = 0, \\ f_2(x_1, x_2, \ldots, x_n) = 0, \\ \cdots\cdots\cdots\cdots \\ f_n(x_1, x_2, \ldots, x_n) = 0, \end{cases} \tag{6.1}$$

где $f_1, f_2, \ldots, f_n$ – заданные, вообще говоря, нелинейные (среди них могут быть и линейные) вещественнозначные функции n вещественных переменных $x_1, x_2, \ldots, x_n$.

Обозначив

$$\mathbf{x} := \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad F(\mathbf{x}) := \begin{pmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_n(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} f_1(x_1, x_2, \ldots x_n) \\ f_2(x_1, x_2, \ldots x_n) \\ \vdots \\ f_n(x_1, x_2, \ldots x_n) \end{pmatrix}, \quad \mathbf{0} := \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

данную систему (6.1) можно записать одним уравнением

$$F(\mathbf{x}) = \mathbf{0} \tag{6.1'}$$

относительно векторной функции F векторного аргумента $\mathbf{x}$. Таким образом, исходную задачу можно рассматривать как задачу о нулях нелинейного отображения $F: R_n \to R_n$. В этой постановке она является прямым обобщением основной задачи предыдущей главы – задачи построения методов нахождения нулей одномерных нелинейных отображений. Фактиче-

ски это та же задача, только в пространствах большей размерности. Поэтому можно как заново строить методы ее решения на основе разработанных выше подходов, так и осуществлять формальный перенос выведенных для скалярного случая расчетных формул. В любом случае следует позаботиться о правомочности тех или иных операций над векторными переменными и векторными функциями, а также о сходимости получаемых таким способом итерационных процессов. Часто теоремы сходимости для этих процессов являются тривиальным обобщением соответствующих результатов, полученных для методов решения скалярных уравнений. Однако не все результаты и не все методы можно перенести со случая $n = 1$ на случай $n \geq 2$. Например, здесь уже не будут работать методы дихотомии, поскольку множество векторов не упорядочено. В то же время переход от $n = 1$ к $n \geq 2$ вносит в задачу нахождения нулей нелинейного отображения свою специфику, учет которой приводит к новым методам и к различным модификациям уже имеющихся. В частности, большая вариативность методов решения нелинейных систем связана с разнообразием способов, которыми можно решать линейные алгебраические задачи, возникающие при пошаговой линеаризации данной нелинейной вектор-функции $F(\mathbf{x})$.

Начнем изучение методов решения нелинейных систем с наиболее простого метода.

Пусть система (6.1) имеет вид (преобразована к виду):

$$\begin{cases} x_1 = \varphi_1(x_1, x_2, \ldots, x_n), \\ x_2 = \varphi_2(x_1, x_2, \ldots, x_n), \\ \cdots\cdots\cdots\cdots \\ x_n = \varphi_n(x_1, x_2, \ldots, x_n), \end{cases} \tag{6.2}$$

или иначе, в компактной записи,

$$\mathbf{x} = \Phi(\mathbf{x}), \tag{6.2'}$$

где

$$\Phi(\mathbf{x}) := \begin{pmatrix} \varphi_1(\mathbf{x}) \\ \varphi_2(\mathbf{x}) \\ \vdots \\ \varphi_n(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} \varphi_1(x_1, x_2, \ldots, x_n) \\ \varphi_2(x_1, x_2, \ldots, x_n) \\ \vdots \\ \varphi_n(x_1, x_2, \ldots, x_n) \end{pmatrix}.$$

Для этой ***задачи о неподвижной точке*** нелинейного отображения $\Phi : R_n \to R_n$ запишем формально рекуррентное равенство

$$\mathbf{x}^{(k+1)} = \Phi\left(\mathbf{x}^{(k)}\right), \quad k = 0, 1, 2, \ldots, \tag{6.3}$$

которое определяет ***метод простых итераций*** (***МПИ***) (или ***метод последовательных приближений***) для задачи (6.2).

Если начать процесс построения последовательности $\left(\mathbf{x}^{(k)}\right)$ с некоторого вектора $\mathbf{x}^{(0)} = \left(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}\right)^T$ и продолжить по формуле (6.3), то при определенных условиях эта последовательность со скоростью геометрической прогрессии будет приближаться к вектору $\mathbf{x}^*$ – неподвижной точке отображения $\Phi(\mathbf{x})$. А именно, справедлива следующая теорема.

Теорема 6.1. *Пусть функция* $\Phi(\mathbf{x})$ *и замкнутое множество* $M \subseteq D(\Phi) \subseteq R_n$ *таковы, что:*

1) $\Phi(\mathbf{x}) \in M \quad \forall \mathbf{x} \in M$;

2) $\exists q < 1$: $\left\|\Phi(\mathbf{x}) - \Phi(\tilde{\mathbf{x}})\right\| \le q \cdot \left\|\mathbf{x} - \tilde{\mathbf{x}}\right\| \quad \forall \mathbf{x}, \tilde{\mathbf{x}} \in M$.

Тогда $\Phi(\mathbf{x})$ *имеет в* M *единственную неподвижную точку* $\mathbf{x}^*$; *последовательность* $\left(\mathbf{x}^{(k)}\right)$, *определяемая МПИ* (6.3), *при любом* $\mathbf{x}^{(0)} \in M$ *сходится к* $\mathbf{x}^*$ *и справедливы оценки*

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \le \frac{q}{1-q}\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| \le \frac{q^k}{1-q}\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\| \quad \forall k \in \mathrm{N}.$$

Доказательство этой теоремы почти полностью повторяет доказательство теоремы 5.7 и даже несколько проще его. Однако и практическая ценность такой теоремы не так велика из-за неконструктивности ее условий. В случаях, когда имеется хорошее начальное приближение $\mathbf{x}^{(0)}$ к решению $\mathbf{x}^*$, больший интерес для приложений может представить следующий аналог теоремы 5.8.

Теорема 6.2. *Пусть* $\Phi(\mathbf{x})$ *дифференцируема**) *в замкнутом шаре**\) $S\left(\mathbf{x}^{(0)}, r\right) \subseteq D(\Phi)$, *причем* $\exists q \in (0;1)$: $\sup\limits_{\mathbf{x} \in S}\left\|\Phi'(\mathbf{x})\right\| \le q$. *Тогда если центр* $\mathbf{x}^{(0)}$ *и радиус* r *шара* S *таковы, что* $\left\|\mathbf{x}^{(0)} - \Phi\left(\mathbf{x}^{(0)}\right)\right\| \le r(1-q)$, *то справедливо заключение теоремы* 6.1 *с* $M = S$.

Как видим, теорема 6.2 – это просто соответствующим образом отредактированная для многомерного случая теорема 5.8. Переложение доказательства теоремы 5.8 на этот случай менее тривиально из-за того, что

---

*) Здесь и далее всюду под дифференцируемостью понимается дифференцируемость по Фреше. См приложение 2.

**) Т е. на множестве S точек $\mathbf{x} \in R_n$ таких, что $\left\|\mathbf{x} - \mathbf{x}^{(0)}\right\| \le r$.

в R_n при $n>1$ не работает дифференциальная теорема о среднем в привычном виде формулы конечных приращений Лагранжа. Но если воспользоваться теоремой о среднем, например, в виде неравенства (см. [29], [42] и др.)

$$\|F(\mathbf{x})-F(\tilde{\mathbf{x}})\| \le \sup_{0<t<1} \|F'(\tilde{\mathbf{x}}+t(\mathbf{x}-\tilde{\mathbf{x}}))\| \cdot \|\mathbf{x}-\tilde{\mathbf{x}}\|,$$

то доказательство теоремы 6.2 становится более прозрачным. Другой путь доказательства теоремы 6.2 – покоординатное применение одномерной формулы Лагранжа и использование конкретных матричных и векторных пар норм [17, 21].

Если потребовать непрерывную дифференцируемость $\Phi(\mathbf{x})$, то более просто перейти от теоремы (6.1) к теореме 6.2, применив следующее утверждение [51].

Лемма 6.1. *Пусть функция* $\Phi: R_n \to R_n$ *непрерывно дифференцируема на множестве* $M \subseteq D(\Phi)$ *и пусть* $\|\Phi'(\mathbf{x})\| \le L \quad \forall \mathbf{x} \in M$. *Тогда* $\Phi(\mathbf{x})$ *удовлетворяет на множестве* M *условию Липшица*

$$\|\Phi(\mathbf{x})-\Phi(\tilde{\mathbf{x}})\| \le L\|\mathbf{x}-\tilde{\mathbf{x}}\|.$$

Запись МПИ (6.3) в развернутом виде, т.е. совокупность рекуррентных равенств

$$\begin{cases} x_1^{(k+1)} = \varphi_1\left(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}\right), \\ x_2^{(k+1)} = \varphi_2\left(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}\right), \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ x_n^{(k+1)} = \varphi_n\left(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}\right), \end{cases}$$

напоминает изученный ранее (см. гл.3) МПИ для СЛАУ, который укладывается в эту схему, если все функции φ_j – линейные. Учитывая, что в линейном случае, как правило, по сравнению с МПИ более эффективен метод Зейделя (см. (3.12) и теорему 3.7), здесь тоже может оказаться полезной подобная модификация. А именно, вместо (6.3) можно реализовать следующий ***метод покоординатных итераций***:

$$\begin{cases} x_1^{(k+1)} = \varphi_1\left(x_1^{(k)}, x_2^{(k)}, \ldots, x_{n-1}^{(k)}, x_n^{(k)}\right), \\ x_2^{(k+1)} = \varphi_2\left(x_1^{(k+1)}, x_2^{(k)}, \ldots, x_{n-1}^{(k)}, x_n^{(k)}\right), \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x_n^{(k+1)} = \varphi_n\left(x_1^{(k+1)}, x_2^{(k+1)}, \ldots, x_{n-1}^{(k+1)}, x_n^{(k)}\right). \end{cases} \qquad (6.4)$$

Заметим, что как и для линейных систем, отдельные уравнения в методе (6.4) неравноправны, т.е. перемена местами уравнений системы (6.2) может изменить в каких-то пределах число итераций и вообще ситуацию со сходимостью последовательности итераций. Чтобы применить метод простых итераций (6.3) или его зейделеву модификацию (6.4) к исходной системе (6.1), нужно, как и в скалярном случае, сначала тем или иным способом привести ее к виду (6.2). Это можно сделать, например, умножив (6.1′) на некоторую неособенную $n \times n$-матрицу $-\mathbf{A}$ и прибавив к обеим частям уравнения $-\mathbf{A}F(\mathbf{x}) = \mathbf{0}$ вектор неизвестных $\mathbf{x}$. Полученная система

$$\mathbf{x} = \mathbf{x} - \mathbf{A}F(\mathbf{x})$$

эквивалентна данной и имеет вид задачи о неподвижной точке (6.2′). Проблема теперь состоит лишь в подборе матричного параметра $\mathbf{A}$ такого, при котором вектор-функция $\Phi(\mathbf{x}) := \mathbf{x} - \mathbf{A}F(\mathbf{x})$ обладала бы нужными свойствами.

6.2. МЕТОД НЬЮТОНА, ЕГО РЕАЛИЗАЦИИ И МОДИФИКАЦИИ

Пусть $(\mathbf{A}_k)$ – некоторая последовательность невырожденных вещественных $n \times n$-матриц. Тогда, очевидно, последовательность задач

$$\mathbf{x} = \mathbf{x} - \mathbf{A}_k F(\mathbf{x}), \quad k = 0, 1, 2, \ldots,$$

имеет те же решения, что и исходное уравнение (6.1′), и для приближенного нахождения этих решений можно формально записать итерационный процесс

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \mathbf{A}_k F\left(\mathbf{x}^{(k)}\right), \quad k = 0, 1, 2, \ldots, \tag{6.5}$$

имеющий вид метода простых итераций (6.3) при $\Phi(\mathbf{x}) := \Phi_k(\mathbf{x}) := \mathbf{x} - \mathbf{A}_k F(\mathbf{x})$. В случае $\mathbf{A}_k \equiv \mathbf{A}$ это, как показано в конце предыдущего пункта, – действительно МПИ с линейной сходимостью последовательности $\left(\mathbf{x}^{(k)}\right)$. Если же $\mathbf{A}_k$ различны при разных k, то формула (6.5) определяет большое семейство итерационных методов с матричными параметрами $\mathbf{A}_k$. Рассмотрим некоторые из методов этого семейства.

Положим $\mathbf{A}_k := \left[F'\left(\mathbf{x}^{(k)}\right)\right]^{-1}$, где

$$F'(\mathbf{x}) = J(\mathbf{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \dots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \dots & \frac{\partial f_2}{\partial x_n} \\ \dots & \dots & \dots & \dots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \dots & \frac{\partial f_n}{\partial x_n} \end{pmatrix}$$

– матрица Якоби вектор-функции $F(\mathbf{x})$. Подставив это $\mathbf{A}_k$ в (6.5), получаем явную формулу ***метода Ньютона***

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \left[F'\left(\mathbf{x}^{(k)}\right)\right]^{-1} F\left(\mathbf{x}^{(k)}\right), \qquad (6.6)$$

обобщающего на многомерный случай скалярный метод Ньютона (5.14). Эту формулу, требующую обращения матриц на каждой итерации, можно переписать в неявном виде:

$$F'\left(\mathbf{x}^{(k)}\right)\left(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right) = -F\left(\mathbf{x}^{(k)}\right). \qquad (6.7)$$

Применение (6.7) предполагает при каждом $k = 0,1,2,\dots$ решение линейной алгебраической системы

$$F'\left(\mathbf{x}^{(k)}\right)\mathbf{p}^{(k)} = -F\left(\mathbf{x}^{(k)}\right)$$

относительно векторной ***поправки*** $\mathbf{p}^{(k)} = \left(p_1^{(k)}, p_2^{(k)}, \dots, p_n^{(k)}\right)^T$, а затем прибавление этой поправки к текущему приближению для получения следующего:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{p}^{(k)}.$$

К решению таких линейных систем можно привлекать самые разные методы как прямые, так и итерационные (см. гл. 2,3) в зависимости от размерности n решаемой задачи и специфики матриц Якоби $J\left(\mathbf{x}^{(k)}\right)$ (например, можно учитывать их симметрию, разреженность и т.п.).

Сравнивая (6.7) с формальным разложением $F(\mathbf{x})$ в ряд Тейлора

$$F(\mathbf{x}) = F\left(\mathbf{x}^{(k)}\right) + F'\left(\mathbf{x}^{(k)}\right)\left(\mathbf{x} - \mathbf{x}^{k}\right) + \frac{1}{2!}F''\left(\mathbf{x}^{(k)}\right)\left(\mathbf{x} - \mathbf{x}^{(k)}\right)^2 + \dots,$$

видим, что последовательность $\left(\mathbf{x}^{(k)}\right)$ в методе Ньютона получается в результате подмены при каждом $k = 0, 1, 2, \dots$ нелинейного уравнения $F(\mathbf{x}) = \mathbf{0}$ или, что то же (при достаточной гладкости $F(\mathbf{x})$), уравнения

$$F\left(\mathbf{x}^{(k)}\right) + F'\left(\mathbf{x}^{(k)}\right)\left(\mathbf{x} - \mathbf{x}^{(k)}\right) + \frac{1}{2!}F''\left(\mathbf{x}^{(k)}\right)\left(\mathbf{x} - \mathbf{x}^{(k)}\right)^2 + \dots = \mathbf{0}$$

линейным уравнением

$$F\left(\mathbf{x}^{(k)}\right) + F'\left(\mathbf{x}^{(k)}\right)\left(\mathbf{x} - \mathbf{x}^{(k)}\right) = \mathbf{0},$$

т.е. пошаговой линеаризацией*). Как следствие этого факта, более тщательно изученного в R_1 (см. гл.5 п.4), можно рассчитывать, что при достаточной гладкости $F(\mathbf{x})$ и достаточно хорошем начальном приближении $\mathbf{x}^{(0)}$ сходимость порождаемой методом Ньютона последовательности $\left(\mathbf{x}^{(k)}\right)$ к решению $\mathbf{x}^*$ будет квадратичной и в многомерном случае. Имеется ряд теорем, устанавливающих это при тех или иных предположениях (см. [6, 21, 29, 40 и др.]). В частности, одна из таких теорем приводится ниже (теорема 6.5 в п.6).

Новым, по сравнению со скалярным случаем, фактором, осложняющим применение метода Ньютона к решению n-мерных систем, является необходимость решения n-мерных линейных задач на каждой итерации (обращения матриц в (6.6) или решения СЛАУ в (6.7)), вычислительные затраты на которые растут с ростом n, вообще говоря, непропорционально быстро. Уменьшение таких затрат – одно из направлений модификации метода Ньютона.

Если матрицу Якоби $F'(\mathbf{x})$ вычислить и обратить лишь один раз – в начальной точке $\mathbf{x}^{(0)}$, то от метода Ньютона (6.6) придем к ***модифицированному методу Ньютона***

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \left[F'\left(\mathbf{x}^{(0)}\right)\right]^{-1} F\left(\mathbf{x}^{(k)}\right). \qquad (6.8)$$

Этот метод требует значительно меньших вычислительных затрат на один итерационный шаг, но итераций при этом может потребоваться значительно больше для достижения заданной точности по сравнению с основным методом Ньютона (6.6), поскольку, являясь частным случаем МПИ (с $A:=\left[F'\left(\mathbf{x}^{(0)}\right)\right]^{-1}$), он имеет лишь скорость сходимости геометрической прогрессии**).

Компромиссный вариант – это вычисление и обращение матриц Якоби не на каждом итерационном шаге, а через несколько шагов (иногда такие методы называют ***рекурсивными*** [50]).

Например, простое чередование основного (6.6) и модифицированного (6.8) методов Ньютона приводит к итерационной формуле

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \mathbf{A}_k F\left(\mathbf{x}^{(k)}\right) - \mathbf{A}_k F\left(\mathbf{x}^{(k)} - \mathbf{A}_k F\left(\mathbf{x}^{(k)}\right)\right), \qquad (6.9)$$

---

*) Обратим внимание на некоторую сознательную некорректность использования здесь термина *«линейный»*, идущую от «школьной» привычки называть функцию $y = ax + b$ линейной, хотя она не является линейным оператором, ибо не удовлетворяет условию однородности . В [22] в таких случаях используется термин *«аффинная аппроксимация»*.

**) Независимо от МПИ линейная сходимость модифицированного метода Ньютона (6.8) обосновывается в п.6 теоремой 6.6.

где $\mathbf{A}_k := \left[F'\left(\mathbf{x}^{(k)}\right)\right]^{-1}$, $k = 0, 1, 2, \ldots$. За $\mathbf{x}^{(k)}$ здесь принимается результат последовательного применения одного шага основного, а затем одного шага модифицированного метода, т.е. ***двухступенчатого процесса***

$$\begin{cases} \mathbf{z}^{(k)} = \mathbf{x}^{(k)} - \mathbf{A}_k F\left(\mathbf{x}^{(k)}\right), \\ \mathbf{x}^{(k+1)} = \mathbf{z}^{(k)} - \mathbf{A}_k F\left(\mathbf{z}^{(k)}\right). \end{cases} \tag{6.10}$$

Доказано, что такой процесс при определенных условиях порождает кубически сходящуюся последовательность $\left(\mathbf{x}^{(k)}\right)$.

Задачу обращения матриц Якоби на каждом k-м шаге метода Ньютона (6.6) можно попытаться решать не точно, а приближенно. Для этого можно применить, например, итерационный процесс Шульца (см. п.3.6), ограничиваясь минимумом – всего одним шагом процесса второго порядка, в котором за начальную матрицу принимается матрица, полученная в результате предыдущего $(k-1)$-го шага. Таким образом, приходим к ***методу Ньютона с последовательной аппроксимацией обратных матриц***:

$$\begin{cases} \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \mathbf{A}_k F\left(\mathbf{x}^{(k)}\right), \\ \Psi_{\mathbf{k}} = \mathbf{E} - F'\left(\mathbf{x}^{(k+1)}\right)\mathbf{A}_k, \quad \mathbf{A}_{k+1} = \mathbf{A}_k + \mathbf{A}_k\Psi_{\mathbf{k}}, \end{cases} \tag{6.11}$$

где $k = 0, 1, 2, \ldots$, а $\mathbf{x}^{(0)}$ и $\mathbf{A}_0$ – начальные вектор и матрица $\left(\approx \left[F'\left(\mathbf{x}^{(0)}\right)\right]^{-1}\right)$ соответственно*) . Этот метод (будем называть его более коротко ААМН – ***аппроксимационный аналог метода Ньютона***) имеет простую схему вычислений – поочередное выполнение векторных в первой строке и матричных во второй строке его записи (6.11) операций. Скорость его сходимости почти так же высока, как и у метода Ньютона. Как будет показано в п.6.6, последовательность $\left(\mathbf{x}^{(k)}\right)$ может квадратично сходиться к решению $\mathbf{x}^*$ уравнения $F(\mathbf{x}) = \mathbf{0}$ (при этом матричная последовательность $\left(\mathbf{A}_k\right)$ также квадратично сходится к $\mathbf{A}^* := \left[F'\left(\mathbf{x}^*\right)\right]^{-1}$, т.е. в нормально развивающемся итерационном процессе (6.11) должна наблюдаться достаточно быстрая сходимость $\left(\|\Psi_k\|\right)$ к нулю).

Применение той же последовательной аппроксимации обратных матриц к простейшему рекурсивному методу Ньютона (6.9) или, что то же,

---

*) Требования к степени близости $\mathbf{A}_0$ к $\left[F'\left(\mathbf{x}^{(0)}\right)\right]^{-1}$, наряду с другими требованиями, гарантирующими сходимость метода, см. далее в теореме 6.7 и следствии 6.1.

к двухступенчатому процессу (6.10) определяет его аппроксимационный аналог

$$\begin{cases} \mathbf{z}^{(k)} = \mathbf{x}^{(k)} - \mathbf{A}_k F\left(\mathbf{x}^{(k)}\right), \quad \mathbf{x}^{(k+1)} = \mathbf{z}^{(k)} - \mathbf{A}_k F\left(\mathbf{z}^{(k)}\right), \\ \Psi_k = \mathbf{E} - F'\left(\mathbf{x}^{(k+1)}\right)\mathbf{A}_k, \quad \mathbf{A}_{k+1} = \mathbf{A}_k + \mathbf{A}_k \Psi_k, \end{cases} \tag{6.12}$$

который, как и (6.9), также можно отнести к методам третьего порядка. Доказательство кубической сходимости этого метода требует уже более жестких ограничений на свойства $F(\mathbf{x})$ и близость $\mathbf{x}^{(0)}$ к $\mathbf{x}^*$, $\mathbf{A}_0$ к $\left[F'\left(\mathbf{x}^{(0)}\right)\right]^{-1}$, чем в предыдущем методе. Заметим, что к улучшению сходимости здесь может привести повышение порядка аппроксимации обратных матриц, например, за счет добавления еще одного слагаемого в формуле для подсчета $\mathbf{A}_{k+1}$ (см. (3.34)):

$$\mathbf{A}_{k+1} = \mathbf{A}_k + \mathbf{A}_k \Psi_k + \mathbf{A}_k \Psi_k^2 .$$

На базе метода Ньютона (6.6) можно построить близкий к нему по поведению итерационный процесс, не требующий вычисления производных. Сделаем это, заменив частные производные в матрице Якоби $J(\mathbf{x})$ разностными отношениями, т.е. подставив в формулу (6.5) вместо $\mathbf{A}_k$ матрицу $\left[J\left(\mathbf{x}^{(k)}, \mathbf{h}^{(k)}\right)\right]^{-1}$, где

$$J(\mathbf{x}, \mathbf{h}) := \left(\frac{f_i\left(x_1, \ldots, x_j + h_j, \ldots, x_n\right) - f_i\left(x_1, \ldots, x_j, \ldots, x_n\right)}{h_j} \right)_{i,j=1}^{n} .$$

При удачном задании последовательности малых векторов $\mathbf{h}^{(k)} = \left(h_1^{(k)}, \ldots, h_n^{(k)}\right)^T$ (постоянной или сходящейся к нулю) полученный таким путем ***разностный*** (или иначе, ***дискретный***) метод Ньютона

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \left[J\left(\mathbf{x}^{(k)}, \mathbf{h}^{(k)}\right)\right]^{-1} F\left(\mathbf{x}^{(k)}\right) \tag{6.13}$$

имеет сверхлинейную, вплоть до квадратичной, скорость сходимости и обобщает на многомерный случай метод (5.29). При задании векторного параметра $\mathbf{h}$ – шага дискретизации – следует учитывать точность машинных вычислений (*macheps*), точность вычисления значения функций f_i, средние значения получаемых приближений (см. по этому поводу [22]).

Можно связать задание последовательности $\left(\mathbf{h}^{(k)}\right)$ с какой-либо сходящейся к нулю векторной последовательностью, например, с последовательностью *невязок* $\left(F\left(\mathbf{x}^{(k)}\right)\right)$ или *поправок* $\left(\mathbf{p}^{(k)}\right)$. Так, полагая

$h_j^{(k)} := x_j^{(k-1)} - x_j^{(k)}$, где $j = 1,\ldots,n$, а $k = 1, 2, \ldots$, приходим к ***простейшему методу секущих*** – обобщению скалярного метода секущих (5.32):

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \left[B\left(\mathbf{x}^{(k)}, \mathbf{x}^{(k-1)}\right)\right]^{-1} F\left(\mathbf{x}^{(k)}\right), \qquad (6.14)$$

где

$$B\left(\mathbf{x}^{(k)}, \mathbf{x}^{(k-1)}\right) := \left(\frac{f_i\left(\mathbf{x}_1^{(k)}, \ldots, \mathbf{x}_j^{(k-1)}, \ldots, \mathbf{x}_n^{(k)}\right) - f_i\left(\mathbf{x}_1^{(k)}, \ldots, \mathbf{x}_j^{(k)}, \ldots, \mathbf{x}_n^{(k)}\right)}{\mathbf{x}_j^{(k-1)} - \mathbf{x}_j^{(k)}}\right)_{i,j=1}^{n},$$

$k = 1, 2, 3, \ldots$.

Этот метод является двухшаговым и требует задания двух начальных точек $\mathbf{x}^{(0)}$ и $\mathbf{x}^{(1)}$. Как было показано в гл.5, при $n = 1$ сходимость метода (6.14) имеет порядок $\frac{1+\sqrt{5}}{2}$. Можно рассчитывать на такую же скорость и в многомерном случае.

К методу секущих так же, как и к методу Ньютона, можно применить пошаговую аппроксимацию обратных матриц на основе метода Шульца. Расчетные формулы этой модификации легко выписать, заменив в совокупности формул ААМН (6.11) матрицу $F'\left(\mathbf{x}^{(k+1)}\right)$ на матрицу $B\left(\mathbf{x}^{(k+1)}, \mathbf{x}^{(k)}\right)$ из (6.14).

Замечание 6.1. Если в одномерном случае разные подходы к линеаризации $f(\mathbf{x})$ приводили к одной и той же формуле секущих (5.32), то для функции n переменных $F(\mathbf{x})$ известно несколько разных обобщений этой формулы в зависимости от того, на какую основу положена линеаризация $F(\mathbf{x})$ в текущей точке. Предложенный здесь простейший метод секущих (6.14) является одним из семейства методов секущих, базирующихся на аппроксимации матриц Якоби. Линейная интерполяция $F(\mathbf{x})$ в R_n может привести к ряду других методов секущих (см., например, [42]). Среди множества таких методов особый интерес представляет ***метод секущих Бройдена*** [22]. В простейшем варианте его можно реализовать по формулам

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \mathbf{B}_k^{-1} F\left(\mathbf{x}^{(k)}\right),$$

$$\mathbf{B}_{k+1} = \mathbf{B}_k + \frac{F\left(\mathbf{x}^{(k+1)}\right)\left(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right)^T}{\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\|_2^2}. \qquad (6.15)$$

В этом процессе, если его начать с $\mathbf{B}_0 :\approx F'\left(\mathbf{x}^{(0)}\right)$, матрицы $\mathbf{B}_k$, не являющиеся прямыми аппроксимациями матриц Якоби $F'\left(\mathbf{x}^{(k)}\right)$, а получаемые в результате уточнения («пересчета») соответствующих матриц предыдущего шага, в какой-то мере близки к $F'\left(\mathbf{x}^{(k)}\right)$, однако в общем случае нельзя утверждать сходимость $\left(\mathbf{B}_k\right)$ к $F'\left(\mathbf{x}^*\right)$ (что может сыграть даже полезную роль при вырождении матриц $F'\left(\mathbf{x}^*\right)$).

Замечание 6.2. Для останова процесса вычислений в быстросходящихся методах таких, как метод Ньютона, методы секущих и т.п., часто вполне успешно применяют простой критерий:

$$\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| < \varepsilon \quad \Rightarrow \quad stop, \quad \mathbf{x}^* :\approx \mathbf{x}^{(k)}. \tag{6.16}$$

Это можно объяснить двумя причинами. Во-первых, оценки погрешности здесь довольно «дороги». Имеется в виду как их получение (особенно для различных модификаций базовых методов), так и их реальное применение. Во-вторых, в силу своей быстрой сходимости, к моменту достижения требуемой малости нормы поправки эти методы набирают такую инерцию, что зачастую «проскакивают» установленный порог точности, т.е. выход по критерию (6.16) дает значение $\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\|$ значительно (иногда на несколько порядков) меньшее, чем ε, см. численный пример в п.6.5.*) Отслеживать факт сходимости в процессе итераций для того, чтобы реагировать на возможную расходимость в случаях, когда заранее не обеспечены условия сходимости применяемого метода, можно с помощью текущих проверок на уменьшение от шага к шагу поправок и невязок, т.е. выполнение неравенств

$$\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| < \left\|\mathbf{x}^{(k-1)} - \mathbf{x}^{(k-2)}\right\| \qquad \text{и} \qquad \left\|F\left(\mathbf{x}^{(k)}\right)\right\| < \left\|F\left(\mathbf{x}^{(k-1)}\right)\right\|.$$

---

*) Как отмечается в [22], для метода Ньютона установлены неравенства

$$0.5\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| \le \left\|\mathbf{x}^* - \mathbf{x}^{(k-1)}\right\| \le 2\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\|,$$

подтверждающие высказанные соображения. В соответствии с этими неравенствами срабатывание критерия (6.16) для метода Ньютона означает, что уже вектор $\mathbf{x}^{(k-1)}$ может быть принят за решение $\mathbf{x}^*$, но поскольку подсчитан вектор $\mathbf{x}^{(k)}$, полагаем

$$\mathbf{x}^* :\approx \mathbf{x}^{(k)}$$

6.3. МЕТОД БРАУНА

В отличие от пошаговой линеаризации векторной функции $F(\mathbf{x})$, приведшей к методу Ньютона (6.6), Брауном (1966 г.)*) предложено проводить на каждом итерационном шаге поочередную линеаризацию компонент вектор – функции $F(\mathbf{x})$, т.е. линеаризовать в системе (6.1) сначала функцию f_1, затем f_2 и т.д., и последовательно решать получаемые таким образом уравнения. Чтобы не затенять эту идею громоздкими выкладками и лишними индексами, рассмотрим вывод расчетных формул метода Брауна в двумерном случае.

Пусть требуется найти решение системы

$$\begin{cases} f(x,y)=0, \\ g(x,y)=0, \end{cases} \tag{6.17}$$

и пусть уже получены приближения x_k, y_k.

Подменим первое уравнение системы (6.17) линейным, полученным по формуле Тейлора для функции двух переменных:

$$f(x,y) \approx f(x_k,y_k) + f'_x(x_k,y_k)(x-x_k) + f'_y(x_k,y_k)(y-y_k) = 0.$$

Отсюда выражаем x (обозначим этот результат через $\tilde{x}$):

$$\tilde{x} = x_k - \frac{1}{f'_x(x_k,y_k)}\left[f(x_k,y_k) + f'_y(x_k,y_k)(y-y_k)\right]. \tag{6.18}$$

При $y = y_k$ находим значение $\tilde{x}_k$ переменной $\tilde{x}$:

$$\tilde{x}_k = x_k - \frac{f(x_k,y_k)}{f'_x(x_k,y_k)},$$

которое будем считать лишь промежуточным приближением (т.е. не x_{k+1}), поскольку оно не учитывает второго уравнения системы (6.17).

Подставив в $g(x,y)$ вместо x переменную $\tilde{x} = \tilde{x}(y)$, придем к некоторой функции $G(y) := g(\tilde{x},y)$ только одной переменной y. Это позволяет линеаризовать второе уравнение системы (6.17) с помощью формулы Тейлора для функции одной переменной:

$$g(\tilde{x},y) \approx G(y_k) + G'(y_k)(y-y_k) = 0. \tag{6.19}$$

При нахождении производной $G'(y)$ нужно учесть, что $G(y) = g(\tilde{x}(y),y)$ есть сложная функция одной переменной y, т.е. применить формулу полной производной

$$G'(y) = g'_x(\tilde{x},y)\cdot\tilde{x}'_y + g'_y(\tilde{x},y). \tag{6.20}$$

Дифференцируя по y равенство (6.18), получаем

$$\tilde{x}'_y = -\frac{f'_y(x_k,y_k)}{f'_x(x_k,y_k)}.$$

*) Ссылки на первоисточники можно найти в [42].

Подстановка последнего в (6.20) при $y = y_k$, $\tilde{x} = \tilde{x}_k$ дает

$$G'(y_k) = -g'_k(\tilde{x}_k, y_k) \cdot \frac{f'_y(x_k, y_k)}{f'_x(x_k, y_k)} + g'_y(\tilde{x}_k, y_k).$$

При известных значениях $G(y_k) = g(\tilde{x}_k, y_k)$ и $G'(y_k)$ теперь можно разрешить линейное уравнение (6.19) относительно y (назовем полученное значение y_{k+1}):

$$y_{k+1} = y_k - \frac{G(y_k)}{G'(y_k)} = y_k - \frac{g(\tilde{x}_k, y_k) f'_x(x_k, y_k)}{f'_x(x_k, y_k) g'_y(\tilde{x}_k, y_k) - f'_y(x_k, y_k) g'_x(\tilde{x}_k, y_k)}.$$

Заменяя в (6.18) переменную y найденным значением y_{k+1}, приходим к значению x_{k+1}:

$$x_{k+1} = \tilde{x}(y_{k+1}) = x_k - \frac{1}{f'_x(x_k, y_k)}\left[f(x_k, y_k) + f'_y(x_k, y_k)(y_{k+1} - y_k)\right].$$

Таким образом, реализация метода Брауна решения двумерных нелинейных систем вида (6.17) сводится к следующему.

При выбранных начальных значениях x_0, y_0 каждое последующее приближение по ***методу Брауна*** находится при $k = 0,1,2,...$ с помощью совокупности формул

$$\tilde{x}_k = x_k - \frac{f(x_k, y_k)}{f'_x(x_k, y_k)},$$

$$q_k = \frac{g(\tilde{x}_k, y_k) \cdot f'_x(x_k, y_k)}{f'_x(x_k, y_k) g'_y(x_k, y_k) - f'_y(x_k, y_k) g'_x(\tilde{x}_k, y_k)},$$

$$p_k = \frac{f(x_k, y_k) - q_k f'_y(x_k, y_k)}{f'_x(x_k, y_k)},$$

$$x_{k+1} = x_k - p_k, y_{k+1} = y_k - q_k,$$

счет по которым должен выполняться в той очередности, в которой они записаны.

Вычисления в методе Брауна естественно заканчивать, когда выполнится неравенство $\max\{|p_{k-1}|, |q_{k-1}|\} < \varepsilon$ (с результатом $(x^*, y^*) := (x_k, y_k)$).

В ходе вычислений следует контролировать немалость знаменателей расчетных формул. Заметим, что функции f и g в этом методе неравноправны, и перемена их ролями может изменить ситуацию со сходимостью.

Указывая на наличие *квадратичной сходимости* метода Брауна, в [42] отмечают, что рассчитывать на его бóльшую по сравнению с методом Ньютона эффективность в смысле вычислительных затрат можно лишь в случае, когда фигурирующие в нем частные производные заменяются разностными отношениями.

6.4. О РЕШЕНИИ НЕЛИНЕЙНЫХ СИСТЕМ МЕТОДАМИ СПУСКА

Общий недостаток всех рассмотренных выше методов решения систем нелинейных уравнений – это сугубо локальный характер сходимости, затрудняющий их применение в случаях (довольно типичных), когда имеются проблемы с выбором хороших начальных приближений. Помощь здесь может прийти со стороны численных методов оптимизации – ветви вычислительной математики, обычно выделяемой в самостоятельную дисциплину. Для этого нужно поставить задачу нахождения решений данной нелинейной системы как оптимизационную или, иначе, экстремальную задачу. Ради геометрической интерпретации проводимых ниже рассуждений и их результатов ограничимся, как и в предыдущем пункте, рассмотрением системы, состоящей из двух уравнений с двумя неизвестными, т.е. системы (6.17).

Из функций f и g системы (6.17) образуем новую функцию

$$\Phi(x,y) := f^2(x,y) + g^2(x,y). \qquad (6.21)$$

Так как эта функция неотрицательна, то найдется точка*) (x^*, y^*) такая, что

$$\Phi(x,y) \geq \Phi(x^*, y^*) \geq 0 \quad \forall (x,y) \in R_2,$$

т.е. $(x^*, y^*) = \arg\min\limits_{x \in R_2} \Phi(x,y)$. Следовательно, если тем или иным способом удается получить точку (x^*, y^*), минимизирующую функцию $\Phi(x,y)$, и если при этом окажется, что $\min\limits_{(x,y)\in R_2} \Phi(x,y) = \Phi(x^*, y^*) = 0$, то (x^*, y^*) – искомое решение системы (6.17), поскольку

$$\Phi(x^*, y^*) = 0 \quad \Leftrightarrow \quad \begin{cases} f(x^*, y^*) = 0, \\ g(x^*, y^*) = 0. \end{cases}$$

Последовательность точек (x_k, y_k) – приближений к точке (x^*, y^*) минимума $\Phi(x,y)$ – обычно получают по рекуррентной формуле

$$\begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} = \begin{pmatrix} x_k \\ y_k \end{pmatrix} + \alpha_k \begin{pmatrix} p_k \\ q_k \end{pmatrix}, \quad k = 0, 1, 2, \ldots, \qquad (6.22)$$

где $(p_k, q_k)^T$ – вектор, определяющий ***направление минимизации***, а α_k – скалярная величина, характеризующая величину шага минимизации (***шаговый множитель***). Учитывая геометрический смысл задачи минимизации функции двух переменных $\Phi(x,y)$ – «спуск на дно» поверхности $z = \Phi(x,y)$

*) Вообще говоря, не единственная.

(см. рис. 6.1), итерационный метод (6.22) можно назвать ***методом спуска***, если вектор $(p_k, q_k)^T$ при каждом k является ***направлением спуска*** (т.е. существует такое $\alpha > 0$, что $\Phi(x_k + \alpha p_k, y_k + \alpha q_k) < \Phi(x_k, y_k)$) и если множитель α_k подбирается так, чтобы выполнялось ***условие релаксации*** $\Phi(x_{k+1}, y_{k+1}) < \Phi(x_k, y_k)$, означающее переход на каждой итерации в точку с меньшим значением минимизируемой функции.

Итак, при построении численного метода вида (6.22) минимизации функции $\Phi(x,y)$ следует ответить на два главных вопроса: как выбирать направление спуска $(p_k, q_k)^T$ и как регулировать длину шага в выбранном направлении с помощью скалярного параметра – шагового множителя α_k. Приведем наиболее простые соображения по этому поводу.

При выборе направления спуска естественным является выбор такого направления, в котором минимизируемая функция убывает наиболее быстро. Как известно из математического анализа функций нескольких переменных, направление наибольшего возрастания функции в данной точке показывает ее градиент в этой точке. Поэтому примем за направление спуска вектор

$$\begin{pmatrix} p_k \\ q_k \end{pmatrix} := -\operatorname{grad}\Phi(x_k, y_k) = -\begin{pmatrix} \Phi'_x(x_k, y_k) \\ \Phi'_y(x_k, y_k) \end{pmatrix}$$

– антиградиент функции $\Phi(x,y)$. Таким образом, из семейства методов (6.22) выделяем ***градиентный метод***

$$\begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} := \begin{pmatrix} x_k \\ y_k \end{pmatrix} - \alpha_k \begin{pmatrix} \Phi'_x(x_k, y_k) \\ \Phi'_y(x_k, y_k) \end{pmatrix}. \qquad (6.23)$$

Оптимальный шаг в направлении антиградиента – это такой шаг, при котором значение $\Phi(x_{k+1}, y_{k+1})$ – наименьшее среди всех других значений $\Phi(x,y)$ в этом фиксированном направлении, т.е. когда точка (x_{k+1}, y_{k+1}) является точкой условного минимума. Следовательно, можно рассчитывать на наиболее быструю сходимость метода (6.23), если полагать в нем

$$\alpha_k = \arg\min_{\alpha>0} \Phi\left(x_k - \alpha\Phi'_x(x_k, y_k), y_k - \alpha\Phi'_y(x_k, y_k)\right). \qquad (6.24)$$

Такой выбор шагового множителя, называемый ***исчерпывающим спуском***, вместе с формулой (6.23) определяет ***метод наискорейшего спуска***.

Геометрическая интерпретация этого метода хорошо видна из рис. 6.1, 6.2. Характерны девяностоградусные изломы траектории наискорейшего спуска, что объясняется исчерпываемостью спуска и свойством

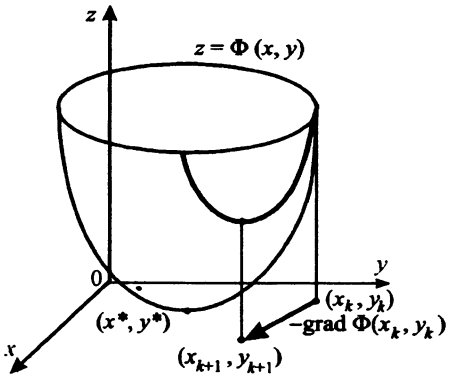


Рис. 6.1. Пространственная интерпретация метода наискорейшего спуска для функции (6.21)

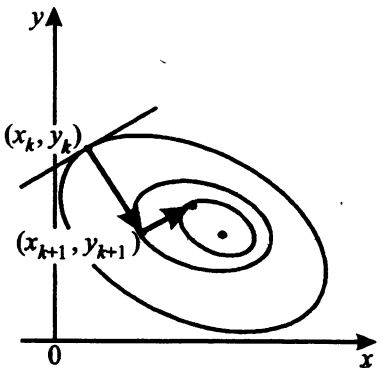


Рис. 6.2. Траектория наискорейшего спуска для функции (6.21)

градиента (а значит, и антиградиента) быть перпендикулярным к касательной к линии уровня в соответствующей точке*).

Наиболее типичной является ситуация, когда найти точно (аналитическими методами) оптимальное значение α_k не удается. Следовательно, приходится делать ставку на применение каких-либо численных методов одномерной минимизации и находить α_k в (6.24) лишь приближенно.

Несмотря на то, что задача нахождения минимума функции одной переменной $\varphi_k(\alpha) = \Phi\left(x_k - \alpha\Phi'_x(x_k, y_k), y_k - \alpha\Phi'_y(x_k, y_k)\right)$ намного проще, чем решаемая задача, применение тех или иных численных методов нахождения значений $\alpha_k = \arg\min\varphi_k(\alpha)$ с той или иной точностью требует вычисления нескольких значений минимизируемой функции. Так как это нужно делать на каждом итерационном шаге, то при большом числе шагов реализация метода наискорейшего спуска в чистом виде является

*) Представив образно процесс движения текущей точки по траектории наискорейшего спуска на участке от положения (x_k, y_k) до (x_{k+1}, y_{k+1}), можно отметить, что спуск характеризуется пересечением линий уровня $\Phi(x, y) = c$ все с меньшими значениями c до тех пор, пока не произойдет касание некоторой линии уровня (сплошь заполняющих $D(\Phi)$); дальнейшее движение в этом направлении приведет к пересечению линий уровня $\Phi(x, y) = c$ с увеличивающимися значениями параметра c.

достаточно высокозатратной. Существуют эффективные схемы приближенного вычисления квазиоптимальных α_k, в которых учитывается специфика минимизируемых функций (типа сумм квадратов функций) [21].

Зачастую успешной является такая стратегия градиентного метода, при которой шаговый множитель α_k в (6.23) берется либо сразу достаточно малым постоянным, либо предусматривается его уменьшение, например, делением пополам для удовлетворения условию релаксации на очередном шаге. Хотя каждый отдельный шаг градиентного метода при этом, вообще говоря, далек от оптимального, такой процесс по числу вычислений функции может оказаться более эффективным, чем метод наискорейшего спуска.

Главное достоинство градиентных методов решения нелинейных систем – глобальная сходимость. Нетрудно доказать, что процесс градиентного спуска приведет к какой-либо точке минимума функции из любой начальной точки. При определенных условиях найденная точка минимума будет искомым решением исходной нелинейной системы.

Главный недостаток – медленная сходимость. Доказано, что сходимость таких методов – лишь линейная, причем, если для многих методов, таких как метод Ньютона, характерно ускорение сходимости при приближении к решению, то здесь имеет место скорее обратное. Поэтому есть резон в построении гибридных алгоритмов, которые начинали бы поиск искомой точки – решения данной нелинейной системы – глобально сходящимся градиентным методом, а затем производили уточнение каким-то быстросходящимся методом, например, методом Ньютона (разумеется, если данные функции обладают нужными свойствами).

Разработан ряд методов решения экстремальных задач, которые соединяют в себе низкую требовательность к выбору начальной точки и высокую скорость сходимости. К таким методам, называемым ***квазиньютоновскими***, можно отнести, например, ***метод переменной метрики (Дэвидона-Флетчера-Пауэлла)***, ***симметричный*** и ***положительно определенный методы секущих*** (на основе формулы пересчета Бройдена, см. замечание 6.1), а также уже упоминавшийся ранее применительно к СЛАУ (см. гл. 3) ***метод сопряженных градиентов***.

При наличии негладких функций в решаемой задаче следует отказаться от использования производных или их аппроксимаций и прибегнуть к так называемым ***методам прямого поиска*** (***циклического покоординатного спуска***, ***Хука и Дживса***, ***Розенброка*** и т.п.). Описание упомянутых и многих других методов такого типа можно найти в учебной и в специальной литературе, посвященной решению экстремальных задач (см., например, [12, 22, 37]).

Замечание 6.3. Для разных семейств численных методов минимизации могут быть рекомендованы свои критерии останова итерационного процесса. Например, учитывая, что в точке минимума дифференцируемой

функции должно выполняться необходимое условие экстремума, на конец счета градиентным методом можно выходить, когда станет достаточно малой норма градиента. Если же принять во внимание, что минимизация применяется к решению нелинейной системы, то целесообразно отслеживать близость к нулю значений минимизируемой неотрицательной функции, т.е. судить о точности получаемого приближения по квадрату его евклидовой нормы невязки.

Замечание 6.4. Как отмечалось в начале этого пункта, ограничение размерности решаемой системы здесь делалось сугубо из иллюстративных соображений. Ничто не помешает развить рассматриваемый подход на случай n-мерной системы (6.1), сводя ее решение к экстремальной задаче

$$\Phi(x) := \sum_{i=1}^{n} f_i^2(x) \to \min.$$

6.5. ЧИСЛЕННЫЙ ПРИМЕР

Типичное поведение рассмотренных методов решения систем нелинейных уравнений отражает следующий численный пример.

Пусть ищется решение системы

$$\begin{cases} 20\ln(x-y) - x - y - 6 = 0, \\ 20\sin(0.7x - 0.7y) + 7x + 7y = 0 \end{cases} \tag{6.25}$$

в окрестности точки $x_0 = 0$, $y_0 = -1$.

Результаты применения к этой системе разных описанных выше итерационных процессов, начинающихся с данной точки (x_0, y_0) и заканчивающихся, как только выполнится неравенство

$$\max\{|x_k - x_{k-1}|, |y_k - y_{k-1}|\} < \varepsilon, \tag{6.26}$$

где k – номер итерации, представлены таблицами 6.1 и 6.2*). В них приведены: приближенные решения, полученные на k-й итерации с помощью семи перечисленных там методов, значения k, при которых сработал критерий (6.26) с $\varepsilon = 0.0001$ для табл. 6.1 и с $\varepsilon = 0.000001$ для табл. 6.2, а также векторы невязок, характеризующие некоторую меру близости указанного приближения к точному решению системы (6.25). При этом всюду шаг аппроксимации производных (начальный шаг в методах секущих и Брауна) брался равным ε в каждой компоненте.

*) Расчеты проведены Кислухиным Д.А.

Таблица 6.1

№ п/п	Метод	Приближенное решение $(x_k, y_k)^T$	Число итераций k	Невязка $(f(x_k, y_k), g(x_k, y_k))^T$
1	Метод Ньютона	–0.46584782 –1.67846885	3	–0.000000164961 –0.000000089180
2	Разностный метод Ньютона	–0.46584782 –1.67846886	4	–0.000000000504 –0.000000000268
3	Метод секущих	–0.46584782 –1.67846885	4	–0.000000293536 –0.000000158314
4	Метод Ньютона с аппроксимацией обратных матриц	–0.46584829 –1.67846812	3	–0 000020367752 –0.000009436432
5	Метод Брауна (с аппроксимацией производных*))	–0.46584558 –1.67846688	4	0.000000052534 0.000031958146
6	Модифицированный (упрощенный) метод Ньютона	–0.46585337 –1 67845758	5	–0.000276258352 –0 000114675484
7	Метод наискорейшего спуска	–0.46570007 –1.67823855	30	–0.001739733248 0.001882572819

Таблица 6.2

№ п/п	Метод	Приближенное решение $(x_k, y_k)^T$	Число итераций k	Невязка $(f(x_k, y_k), g(x_k, y_k))^T$
1	Метод Ньютона	–0.46584782 –1.67846886	4	0.000000000001 –0.000000000002
2	Разностный метод Ньютона	–0.46584782 –1.67846886	4	–0.000000000497 –0.000000000261
3	Метод секущих	–0.46584782 –1.67846886	5	0.000000000001 –0.000000000002
4	Метод Ньютона с аппроксимацией обратных матриц	–0.46584782 –1 67846886	4	0.000000000098 –0.000000000048
5	Метод Брауна (с аппроксимацией производных)	–0 46584782 –1 67846886	5	–0 000000000000 –0 000000037367
6	Модифицированный (упрощенный) метод Ньютона	–0 46584784 –1 67846880	8	–0.000001418985 –0 000000589060
7	Метод наискорейшего спуска	–0.46574667 –1 67846711	36	–0.000012735642 0 000014735135

*) Аппроксимация производных на k-й итерации осуществлялась с шагом $h_k = \min\{|p_k|, |q_k|\}$.

Число итераций и, соответственно, число нулей перед первыми значащими цифрами в невязках подкрепляет сделанные выше заявления о быстрой сходимости методов 1-5 и справедливость высказанных в замечании 6.2 соображений о надежности и даже некоторой грубости простого критерия останова (6.16) (для системы (6.25) реализованного в виде (6.26)). Сравнивая четвертые знаки после запятой у приближенных решений, полученных методом наискорейшего спуска и, например, методом Ньютона, в таблице 6.1, а также, аналогично, шестые знаки в таблице 6.2, видим, что для медленно сходящихся методов примененный критерий уже не является столь надежным (из процесса итерирования в градиентных методах, как уже упоминалось в замечании 6.3, обычно выходят по своим критериям, например, по малости нормы градиента).

6.6. СХОДИМОСТЬ МЕТОДА НЬЮТОНА И НЕКОТОРЫХ ЕГО МОДИФИКАЦИЙ

Многие утверждения о сходимости метода Ньютона восходят к известным результатам Л.В. Канторовича*), перенесшего метод Ньютона на нелинейные операторные уравнения в банаховых пространствах, в связи с чем и метод в таком общем случае часто называют ***методом Ньютона-Канторовича***. В основе этих результатов лежит принцип мажорирования операторного уравнения некоторым скалярным уравнением, через корни которого оценивается абсолютная погрешность приближений, получаемых в итерационном процессе (см. [29], а также [40] и др.). Таким путем получают условия квадратичной сходимости основного (6.6) и линейной сходимости модифицированного (упрощенного) (6.8) методов Ньютона. Непосредственное применение подобной методики к анализу сходимости методов более общего вида, в частности, ААМН (6.11), затруднительно. Поэтому изберем другой путь исследования сходимости итерационных последовательностей. Вообще говоря, здесь будет эксплуатироваться та же идея скалярного мажорирования, только в несколько ином виде.

Прежде всего, попытаемся выяснить, какие условия нужно наложить на скалярную последовательность (p_k), мажорирующую последовательности норм поправок $\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}$ и невязок $F(\mathbf{x}^{(k)})$, чтобы последовательность n-мерных векторов $\mathbf{x}^{(k)}$ сходилась к нулю вектор-функции $F(\mathbf{x})$ с заданным порядком μ (≥ 1) скорости сходимости независимо от способа получения элементов последовательности $(\mathbf{x}^{(k)})$.

---

*) Канторович Леонид Витальевич (1912-1986) – российский академик, приобретший мировую известность своими основополагающими работами в области линейного программирования и применения функционального анализа в вычислительной математике.

Теорема 6.3. **I.** *Пусть непрерывная векторная функция* $F: M \subseteq R_n \to R_n$*) *и последовательность векторов* $\mathbf{x}^{(k)} \in M$ *таковы, что при всех* $k \in N_0$ *выполняются условия*:

1) $\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| \le \lambda p_k$;

2) $\left\|F\left(\mathbf{x}^{(k)}\right)\right\| \le p_k$,

где числа p_k *определяются рекуррентным равенством*

$$p_{k+1} = G_0 p_k^{\mu}, \quad k = 0, 1, 2, \ldots, \tag{6.27}$$

а $\lambda > 0$, $G_0 > 0$, $p_0 > 0$ *и* $\mu > 1$ *– некоторые числовые параметры.*

II. *Тогда если* $\nu := G_0 p_0^{\mu-1} < 1$ *и замкнутый шар* $S\left(\mathbf{x}^{(0)},\ r := \lambda p_0 \sum_{i=0}^{\infty} \nu^{\frac{\mu^i - 1}{\mu - 1}}\right)$ *содержится в* M, *то все члены последовательности* $\left(\mathbf{x}^{(k)}\right)$ *принадлежат* S, *последовательность* $\left(\mathbf{x}^{(k)}\right)$ *имеет предел* $\mathbf{x}^* \in S$ *такой, что* $F\left(\mathbf{x}^*\right) = \mathbf{0}$; *при этом быстрота сходимости* $\left(\mathbf{x}^{(k)}\right)$ *к* $\mathbf{x}^*$ *характеризуется неравенством* $(\forall k \in N)$

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \le \frac{\lambda p_0}{1 - \nu^{\mu^k}} \nu^{\frac{\mu^k - 1}{\mu - 1}}. \tag{6.28}$$

Доказательство. Пользуясь равенством (6.27), выразим элементы последовательности (p_i) через ее начальный член p_0 и определенную в теореме величину ν:

$$p_i = G_0 p_{i-1}^{\mu} = G_0 \left(G_0 p_{i-2}^{\mu}\right)^{\mu} = G_0^{1+\mu} p_{i-2}^{\mu^2} = G_0^{1+\mu+\mu^2} \cdot p_{i-3}^{\mu^3} = \ldots =$$

$$= G_0^{1+\mu+\mu^2+\ +\mu^{i-1}} p_0^{\mu^i} = G_0^{\mu^i - 1} \cdot p_0^{\mu^i - 1} \cdot p_0^{1-\mu^i} \cdot p_0^{\mu^i} = p_0 \cdot \nu^{\frac{\mu^i - 1}{\mu - 1}}.$$

Следовательно, условие 1) доказываемой теоремы можно переписать в виде

$$\left\|\mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}\right\| \le \lambda p_0 \cdot \nu^{\frac{\mu^i - 1}{\mu - 1}}. \tag{6.29}$$

---

*) Без всяких прочих изменений можно заменить здесь $R_n \to R_n$ на более общий случай $R_n \to R_m$.

Посредством (6.29) теперь устанавливаем, что

$$\left\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(0)}\right\| \le \sum_{i=0}^{k}\left\|\mathbf{x}^{(i+1)}-\mathbf{x}^{(i)}\right\| \le \sum_{i=0}^{k}\lambda p_0 \cdot \nu^{\frac{\mu^i-1}{\mu-1}} \le r,$$

т.е. все члены заданной последовательности $\left(\mathbf{x}^{(k)}\right)$ принадлежат $S \subseteq M$. Покажем, что она удовлетворяет критерию Коши. С помощью того же неравенства (6.29) имеем

$$\left\|\mathbf{x}^{(k+m)}-\mathbf{x}^{(k)}\right\| \le \sum_{i=k}^{k+m-1}\left\|\mathbf{x}^{(i+1)}-\mathbf{x}^{(i)}\right\| \le \lambda p_0 \sum_{i=k}^{k+m-1}\nu^{\frac{\mu^i-1}{\mu-1}} =$$

$$= \lambda p_0 \cdot \nu^{\frac{\mu^k-1}{\mu-1}}\left(1+\nu^{\mu^k}+\nu^{\mu^k+\mu^{k+1}}+\ldots+\nu^{\mu^k+\mu^{k+1}+\ldots+\mu^{k+m-1}}\right) <$$

$$< \lambda p_0 \cdot \nu^{\frac{\mu^k-1}{\mu-1}}\left(1+\nu^{\mu^k}+\nu^{2\mu^k}+\ldots+\nu^{(m-1)\mu^k}\right) = \lambda p_0 \cdot \nu^{\frac{\mu^k-1}{\mu-1}} \cdot \frac{1-\nu^{m\mu^k}}{1-\nu^{\mu^k}}.$$

Полученное неравенство, рассматриваемое при фиксированном $m \in N$ и $k \to \infty$, говорит о фундаментальности $\left(\mathbf{x}^{(k)}\right)$ и существовании предельного вектора $\mathbf{x}^*$ в шаре S (в силу предполагаемой замкнутости S). С другой стороны, если в нем зафиксировать k и перейти к пределу при $m \to \infty$, сразу получается утверждаемая оценка (6.28). Подстановка выражения $p_k = p_0 \cdot \nu^{\frac{\mu^k-1}{\mu-1}}$ в условие 2) показывает, что $\left\|F\left(\mathbf{x}^{(k+1)}\right)\right\| \to 0$ при $k \to \infty$ (т.е. при $\mathbf{x}^{(k)} \to \mathbf{x}^*$), а это, в силу предполагаемой непрерывности $F(\mathbf{x})$, означает, что $\mathbf{x}^* := \lim\limits_{k\to\infty}\mathbf{x}^{(k)}$ есть решение уравнения $F(\mathbf{x}) = \mathbf{0}$. Теорема доказана.

Замечание 6.5. Нетрудно убедиться, что теорема 6.3 (а также следующая теорема 6.4) справедлива и при $\mu = 1$. При этом

$$\nu := G_0, \quad r := \frac{\lambda p_0}{1-\nu}, \quad \left\|\mathbf{x}^*-\mathbf{x}^{(k)}\right\| \le \frac{\lambda p_0}{1-\nu^k}\cdot\nu^k$$

(т.е. сходимость $\left(\mathbf{x}^{(k)}\right)$ к $\mathbf{x}^*$ в $S\left(\mathbf{x}^{(0)}, r\right)$ в этом случае – линейная).

Изменим требования к $\left(\mathbf{x}^{(k)}\right)$ и $F(\mathbf{x})$, фигурирующие в части **I** теоремы 6.3, так, чтобы осталось неизменной ее констатирующая часть **II**.

Теорема 6.4. *Пусть существуют такие последовательности положительных чисел* H_k *и* G_k, *удовлетворяющих условиям* $H_k \le H_0$, $G_k \le G_0$, *и число* $\mu > 1$, *что в предположении, что* $\mathbf{x}^{(k)} \in M$, *при всех* $k \in N_0$ *выполняются неравенства:*

$$1)\ \left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| \le H_k \cdot \left\|F\left(\mathbf{x}^{(k)}\right)\right\|;$$

$$2)\ \left\|F\left(\mathbf{x}^{(k+1)}\right)\right\| \le G_k \cdot \left\|F\left(\mathbf{x}^{(k)}\right)\right\|^{\mu}.$$

Тогда справедлива часть II теоремы 6.3 с $p_0 \ge \left\|F\left(\mathbf{x}^{(0)}\right)\right\|$, $\lambda := H_0$.

Доказательство. Определим последовательность (p_k) равенством $p_{k+1} = G_0 p_k^{\mu}$ и по индукции покажем, что эта последовательность мажорирует $\left(\left\|F\left(\mathbf{x}^{(k)}\right)\right\|\right)$ одновременно с доказательством принадлежности векторов $\mathbf{x}^{(k)}$ шару $S\left(\mathbf{x}^{(0)}, r\right)$.

По условию $\mathbf{x}^{(0)} \in S$ и $\left\|F\left(\mathbf{x}^{(0)}\right)\right\| \le p_0$. Сделаем индукционное предположение, что

$$\mathbf{x}^{(i)} \in S \quad \text{и} \quad \left\|F\left(\mathbf{x}^{(i)}\right)\right\| \le p_i \quad \forall i \in \{0, 1, \ldots, k\}.$$

Тогда

$$\left\|\mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}\right\| \le H_i \cdot \left\|F\left(\mathbf{x}^{(i)}\right)\right\| \le H_0 p_i = H_0 p_0 \cdot \nu^{\frac{\mu^i - 1}{\mu - 1}}.$$

Из этого следует (см. доказательство теоремы 6.3) неравенство $\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(0)}\right\| \le r$, означающее, что $\mathbf{x}^{(k+1)} \in S \subseteq M$. Таким образом, значение $F\left(\mathbf{x}^{(k+1)}\right)$ существует и

$$\left\|F\left(\mathbf{x}^{(k+1)}\right)\right\| \le G_k \left\|F\left(\mathbf{x}^{(k)}\right)\right\|^{\mu} \le G_0 p_k^{\mu} = p_{k+1}.$$

Теперь можно сказать, что условия теоремы 6.3 полностью выполнены, значит справедливо и ее заключение.

Прежде чем применить доказанные выше теоремы к конкретным методам типа (6.5), выведем из формулы Тейлора*)

$$F(\mathbf{x}) \equiv F\left(\mathbf{x}^{(0)}+\mathbf{h}\right) = F\left(\mathbf{x}^{(0)}\right) + F'\left(\mathbf{x}^{(0)}\right)\mathbf{h} + \frac{1}{2!}F''\left(\mathbf{x}^{(0)}\right)\mathbf{h}^2 + \\ +\ldots+\frac{1}{(l-1)!}F^{(l-1)}\left(\mathbf{x}^{(0)}\right)\mathbf{h}^{l-1} + \omega\left(\mathbf{x}^{(0)},\mathbf{h}\right) \tag{6.30}$$

с остаточным членом

$$\omega\left(\mathbf{x}^{(0)},\mathbf{h}\right) = \frac{1}{(l-1)!}\int_{\mathbf{x}^{(0)}}^{\mathbf{x}} F^{(l)}(\mathbf{z})(\mathbf{x}-\mathbf{z})^{l-1}\,d\mathbf{z} \tag{6.31}$$

простое неравенство для оценивания $\|F(\mathbf{x})\|$ в произвольной точке $\mathbf{x} \in M$ через значения F и F' в близкой к $\mathbf{x}$ точке $\mathbf{x}^{(0)} \in M$.

Лемма 6.2. *Пусть векторная функция $F(\mathbf{x})$ в области $M \in R_n$ дифференцируема по Фреше и ее производная удовлетворяет условию Липшица*:

$$\left\|F'(\mathbf{x}) - F'\left(\mathbf{x}^{(0)}\right)\right\| \le L\cdot\left\|\mathbf{x}-\mathbf{x}^{(0)}\right\| \quad \forall \mathbf{x},\mathbf{x}^{(0)} \in M.$$

Тогда при любых $\mathbf{x}$, $\mathbf{x}^{(0)}$ *из* M

$$\|F(\mathbf{x})\| \le \left\|F\left(\mathbf{x}^{(0)}\right) + F'\left(\mathbf{x}^{(0)}\right)\left(\mathbf{x}-\mathbf{x}^{(0)}\right)\right\| + \frac{L}{2}\left\|\mathbf{x}-\mathbf{x}^{(0)}\right\|^2. \tag{6.32}$$

Доказательство. Запишем формулу Тейлора (6.30) с остаточным членом (6.31) для случая $l = 1$:

$$F(\mathbf{x}) = F\left(\mathbf{x}^{(0)}\right) + \int_{\mathbf{x}^{(0)}}^{\mathbf{x}} F'(\mathbf{z})d\mathbf{z}.$$

Учитывая, что $\int_{\mathbf{x}^{(0)}}^{\mathbf{x}} F'\left(\mathbf{x}^{(0)}\right)d\mathbf{z} \equiv F'\left(\mathbf{x}^{(0)}\right)\left(\mathbf{x}-\mathbf{x}^{(0)}\right)$, ее можно преобразовать к виду

$$F(\mathbf{x}) = F\left(\mathbf{x}^{(0)}\right) + F'\left(\mathbf{x}^{(0)}\right)\left(\mathbf{x}-\mathbf{x}^{(0)}\right) + \int_{\mathbf{x}^{(0)}}^{\mathbf{x}}\left[F'(\mathbf{z}) - F'\left(\mathbf{x}^{(0)}\right)\right]d\mathbf{z}.$$

В последнем представлении $F(\mathbf{x})$ интеграл по отрезку $\left[\mathbf{x}^{(0)};\mathbf{x}\right]$ заменой

*) Выражения типа $F^{(l)}\mathbf{h}^l$ следует понимать как результат применения l –линейного оператора l -кратного дифференцирования $F^{(l)}$ к вектору $\mathbf{h}$ (см. приложение 2).

$\mathbf{z} = \mathbf{x}^{(0)} + \tau\left(\mathbf{x} - \mathbf{x}^{(0)}\right)$ сведем к интегралу по абстрактной переменной $\tau \in [0;1]$. Имеем:

$$F(\mathbf{x}) = F\left(\mathbf{x}^{(0)}\right) + F'\left(\mathbf{x}^{(0)}\right)\left(\mathbf{x} - \mathbf{x}^{(0)}\right) +$$
$$+ \int_0^1 \left[F'\left(\mathbf{x}^{(0)} + \tau\left(\mathbf{x} - \mathbf{x}^{(0)}\right)\right) - F'\left(\mathbf{x}^{(0)}\right)\right]\left(\mathbf{x} - \mathbf{x}^{(0)}\right)d\tau.$$

Отсюда, переходя к нормам, получаем доказываемое неравенство (6.32), предварительно оценив норму интеграла следующим образом:

$$\left\| \int_0^1 \left[F'\left(\mathbf{x}^{(0)} + \tau\left(\mathbf{x} - \mathbf{x}^{(0)}\right)\right) - F'\left(\mathbf{x}^{(0)}\right)\right]\left(\mathbf{x} - \mathbf{x}^{(0)}\right)d\tau \right\| \le$$
$$\le \int_0^1 \left\| F'\left(\mathbf{x}^{(0)} + \tau\left(\mathbf{x} - \mathbf{x}^{(0)}\right)\right) - F'\left(\mathbf{x}^{(0)}\right)\right\| \cdot \left\|\mathbf{x} - \mathbf{x}^{(0)}\right\| d\tau \le$$
$$\le \int_0^1 L\left\|\mathbf{x} - \mathbf{x}^{(0)}\right\|^2 \tau d\tau = \frac{L}{2}\left\|\mathbf{x} - \mathbf{x}^{(0)}\right\|^2.$$

Теорема 6.5. *Пусть функция $F(\mathbf{x})$ определена и дифференцируема по Фреше в некоторой открытой области $M \subseteq R_n$, причем:*

1) $\exists L > 0$: $\left\|F'(\mathbf{x}) - F'(\tilde{\mathbf{x}})\right\| \le L\left\|\mathbf{x} - \tilde{\mathbf{x}}\right\| \quad \forall \mathbf{x}, \tilde{\mathbf{x}} \in M$;*)

2) $\exists \left[F'(\mathbf{x})\right]^{-1}$ *и* $\exists C > 0$: $\left\|\left[F'(\mathbf{x})\right]^{-1}\right\| \le C \quad \forall \mathbf{x} \in M$.

Тогда если

$$\nu := 0.5LC^2 p_0 < 1, \qquad \text{где} \quad p_0 \ge \left\|F\left(\mathbf{x}^{(0)}\right)\right\|,$$

и замкнутый шар $S\left(\mathbf{x}^{(0)}, r := Cp_0 \sum_{i=0}^{\infty} \nu^{2^i - 1}\right)$ *целиком содержится в M, то все члены последовательности* $\left(\mathbf{x}^{(k)}\right)$, *определяемые методом Ньютона* (6.6), *начинающимся с заданного* $\mathbf{x}^{(0)}$, *лежат в $S \subseteq M$; последовательность* $\left(\mathbf{x}^{(k)}\right)$ *имеет предел* $\mathbf{x}^* \in S$, *служащий решением уравнения* $F(\mathbf{x}) = \mathbf{0}$; *справедлива оценка погрешности*

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \le \frac{Cp_0}{1 - \nu^{2^k}} \cdot \nu^{2^k - 1}.$$

*) Согласно лемме 6.1, для дважды непрерывно дифференцируемой функции $F(\mathbf{x})$ в качестве L можно брать оценку сверху величины $\left\|F''(\mathbf{x})\right\|$.

Доказательство. Непосредственно из равенства (6.6) получаем

$$\left\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\right\| \le \left\|\left[F'\left(\mathbf{x}^{(k)}\right)\right]^{-1}\right\| \cdot \left\|F\left(\mathbf{x}^{(k)}\right)\right\| \le C \cdot \left\|F\left(\mathbf{x}^{(k)}\right)\right\|,$$

т.е. требование 1) теоремы 6.4 с $H_k \equiv C$.

Далее обратимся к неравенству (6.32), установленному леммой 6.2. Положив в нем $\mathbf{x} = \mathbf{x}^{(k+1)}$, $\mathbf{x}^{(0)} = \mathbf{x}^{(k)}$, приведем (6.32) к виду

$$\left\|F\left(\mathbf{x}^{(k+1)}\right)\right\| \le \left\|F\left(\mathbf{x}^{(k)}\right)+F'\left(\mathbf{x}^{(k)}\right)\left(\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\right)\right\| + \frac{L}{2}\left\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\right\|^2. \quad (6.33)$$

Но в данном случае, т.е. для метода Ньютона,

$$F\left(\mathbf{x}^{(k+1)}\right)+F'\left(\mathbf{x}^{(k)}\right)\left(\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\right)=\mathbf{0},$$

поэтому

$$\left\|F\left(\mathbf{x}^{(k+1)}\right)\right\| \le \frac{L}{2}\left\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\right\|^2 \le \frac{1}{2}LC^2\left\|F\left(\mathbf{x}^{(k)}\right)\right\|^2.$$

Таким образом, выполнено и требование 2) теоремы 6.4 с $G_k = \frac{1}{2}LC^2$, $\mu = 2$.

Завершает доказательство подстановка постоянных $\lambda = H_0 = C$, $G_0 = \frac{1}{2}LC^2$, $\mu = 2$ в часть **II** теоремы 6.3.

Для модифицированного метода Ньютона требование обратимости матрицы Якоби в любой точке M заменим менее ограничительным требованием ее обратимости лишь в начальной точке $\mathbf{x}^{(0)}$.

Теорема 6.6. *Пусть для* $F:\left(M \subseteq R_n\right) \to R_n$:

1) $\exists F'(\mathbf{x}):\left(\exists L>0:\left\|F'(\mathbf{x})-F'(\tilde{\mathbf{x}})\right\| \le L\|\mathbf{x}-\tilde{\mathbf{x}}\| \quad \forall \tilde{\mathbf{x}} \in M\right) \quad \forall \mathbf{x} \in M$;

2) $\exists\left[F'\left(\mathbf{x}^{(0)}\right)\right]^{-1}, \quad \exists C_0>0:\left\|\left[F'\left(\mathbf{x}^{(0)}\right)\right]^{-1}\right\| \le C_0$.

Тогда если при $p_0 \ge \left\|F\left(\mathbf{x}^{(0)}\right)\right\|$ *величина*

$$t := LC_0^2 p_0 \le 0.125 \quad (6.34)$$

и замкнутый шар $S\left(\mathbf{x}^{(0)}, r := \frac{2C_0 p_0}{1+\sqrt{1-8t}}\right)$ *содержится в* M, *то начатый с* $\mathbf{x}^{(0)}$ *модифицированный метод Ньютона* (6.8) *сходится*

в S к решению $\mathbf{x}^*$ уравнения (6.1') с оценкой погрешности

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \le \frac{C_0 p_0}{1-\nu^k}\cdot \nu^k,$$

где[*)] $\nu := \frac{1}{2} \mp \sqrt{\frac{1}{4} - 2t}$.

Доказательство. Как и в предыдущей теореме, требование 1) теоремы 6.4 получается сразу же из формулы, определяющей метод:

$$\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| \le \left\|\left[F'\left(\mathbf{x}^{(0)}\right)\right]^{-1} F\left(\mathbf{x}^{(k)}\right)\right\| \le C_0 \left\|F\left(\mathbf{x}^{(k)}\right)\right\|.$$

Для оценки $\left\|F\left(\mathbf{x}^{(k+1)}\right)\right\|$ преобразуем неравенство (6.33) так, чтобы воспользоваться равенством нулю выражения $F\left(\mathbf{x}^{(k)}\right) + F'\left(\mathbf{x}^{(0)}\right)\left(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right)$ в соответствии с (6.8). Имеем

$$\left\|F\left(\mathbf{x}^{(k+1)}\right)\right\| \le \left\|F\left(\mathbf{x}^{(k)}\right) + F'\left(\mathbf{x}^{(k)}\right)\left(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right)\right. +$$

$$+ \left[F'\left(\mathbf{x}^{(k)}\right) - F'\left(\mathbf{x}^{(0)}\right)\right]\left(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right)\Big\| + \frac{L}{2}\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\|^2 \le$$

$$\le \left(\left\|F'\left(\mathbf{x}^{(k)}\right) - F'\left(\mathbf{x}^{(0)}\right)\right\| + \frac{L}{2}\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\|\right)\cdot\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| \le$$

$$\le L\left(\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(0)}\right\| + \frac{1}{2}\left(\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(0)}\right\| + \left\|\mathbf{x}^{(0)} - \mathbf{x}^{(k)}\right\|\right)\right)\cdot C_0 \cdot \left\|F\left(\mathbf{x}^{(k)}\right)\right\| \le$$

$$\le LC_0\left(r + \frac{1}{2}(r+r)\right)\cdot\left\|F\left(\mathbf{x}^{(k)}\right)\right\| = 2LC_0 r\cdot\left\|F\left(\mathbf{x}^{(k)}\right)\right\|.$$

Отсюда видно, что можно считать выполненным требование 2) теоремы 6.4 с $G_k = G_0 = 2LC_0 r$ и $\mu = 1$.

Подстановка постоянных $\mu = 1$, $\lambda = C_0$ в заключительную часть **II** теоремы 6.3 показывает (с учетом замечания 6.5), что в данном случае должно быть $r = \frac{C_0 p_0}{1-\nu}$, а $\nu = 2LC_0 r < 1$. Исключая из последних двух равенств параметр r, получаем квадратное относительно ν уравнение

$$\nu^2 - \nu + 2t = 0,$$

оба корня которого $\nu_{1,2} = 0{,}5 \mp \sqrt{0{,}25 - 2t}$ положительны и меньше 1, если только неотрицателен дискриминант $0{,}25 - 2t$, что обеспечивается условием (6.34). Подставляя эти значения ν, находим связанные с ним значения радиуса r шара S. Теперь справедливость заключения данной теоремы очевидна.

[*)] В выражениях r и ν одновременно берутся либо только верхние, либо только нижние знаки.

Обратимся, наконец, к обоснованию квадратичной сходимости метода Ньютона с последовательной аппроксимацией обратных матриц, т.е. ААМН (6.11).

Теорема 6.7. *Пусть функция* $F(\mathbf{x})$ *определена и дифференцируема в* $M \subseteq R_n$, *причем*

$$\exists L > 0: \quad \|F'(\mathbf{x}) - F'(\widetilde{\mathbf{x}})\| \le L\|\mathbf{x} - \widetilde{\mathbf{x}}\| \quad \forall \mathbf{x}, \widetilde{\mathbf{x}} \in M.$$

Тогда если вектор $\mathbf{x}^{(0)}$ *и матрица* $\mathbf{A}_0$ *таковы, что при некотором* $\lambda > 0$ *выполняются неравенства*

$$\left\|\mathbf{E} - F'\left(\mathbf{x}^{(0)}\right)\mathbf{A}_0\right\| \le L\lambda^2\left\|F\left(\mathbf{x}^{(0)}\right)\right\|, \tag{6.35}$$

$$\nu := 4L\lambda^2\left\|F\left(\mathbf{x}^{(0)}\right)\right\| \le 1 - \frac{\|\mathbf{A}_0\|}{2\lambda - \|\mathbf{A}_0\|} \tag{6.36}$$

и

$$S := \left\{\mathbf{x} \in R_n \,\middle|\, \left\|\mathbf{x} - \mathbf{x}^{(0)}\right\| \le \lambda\left\|F\left(\mathbf{x}^{(0)}\right)\right\| \cdot \sum_{i=0}^{\infty}\nu^i\right\} \subseteq M,$$

то начатый с данных $\mathbf{x}^{(0)}$, $\mathbf{A}_0$ *ААМН* (6.11) *сходится в* S *к решению* $\mathbf{x}^*$ *уравнения* $F(\mathbf{x}) = \mathbf{0}$ *и имеет место оценка погрешности*

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \le \frac{\lambda\left\|F\left(\mathbf{x}^{(0)}\right)\right\|}{1 - \nu^{2^k}} \cdot \nu^{2^k}. \tag{6.37}$$

Доказательство. Введем в рассмотрение последовательности положительных величин p_k, β_k, b_k, определяемых при $k = 0,1,2,\ldots$ равенствами

$$p_{k+1} = 4L\lambda^2 p_k, \quad p_0 := \left\|F\left(\mathbf{x}^{(0)}\right)\right\|; \tag{6.38}$$

$$\beta_k = 2L\lambda^2 p_k; \tag{6.39}$$

$$b_{k+1} = \beta_k^2, \quad b_0 := L\lambda^2 p_0. \tag{6.40}$$

Очевидно невозрастание этих последовательностей. Легко также видеть, что

$$b_k = L\lambda^2 p_k \quad \forall k \in N_0. \tag{6.41}$$

Действительно, предположив равенство (6.41) верным при некотором k, имеем

$$b_{k+1} = \left(2L\lambda^2 p_k\right)^2 = L\lambda^2 \cdot 4L\lambda^2 p_k^2 = L\lambda^2 p_{k+1},$$

т.е. то, что получили бы формальной заменой в (6.41) k на $k+1$.

Обозначим $\mathbf{B}_k := \mathbf{E} - F'\left(\mathbf{x}^{(k)}\right)\mathbf{A}_k$ (– невязка для $\mathbf{A}_k$ относительно $\left[F'\left(\mathbf{x}^{(k)}\right)\right]^{-1}$).

Докажем, что при любом $k \in N_0$ скалярные последовательности (p_k), (β_k), (b_k) мажорируют последовательности норм векторов $F\left(\mathbf{x}^{(k)}\right)$ и матриц Ψ_k, $\mathbf{B}_k$ соответственно и, вместе с тем, λ ограничивает сверху $\|\mathbf{A}_k\|$. С этой целью сделаем индукционное предположение, что одновременно выполняются неравенства:

$$\left\|F\left(\mathbf{x}^{(k)}\right)\right\| \le p_k, \quad \|\Psi_{k-1}\| \le \beta_{k-1}, \quad \|\mathbf{B}_k\| \le b_k \quad \text{и} \quad \|\mathbf{A}_k\| \le \lambda. \quad (6.42)$$

Тогда можно проделать следующие выкладки:

$$\|\Psi_k\| = \left\|\mathbf{E} - F'\left(\mathbf{x}^{(k+1)}\right)\mathbf{A}_k\right\| = \left\|\mathbf{B}_k + F'\left(\mathbf{x}^{(k)}\right)\mathbf{A}_k - F'\left(\mathbf{x}^{(k+1)}\right)\mathbf{A}_k\right\| \le$$
$$\le \|\mathbf{B}_k\| + L\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| \cdot \|\mathbf{A}_k\| \le b_k + L\lambda\left\|\mathbf{A}_k F\left(\mathbf{x}^{(k)}\right)\right\| \le$$
$$\le L\lambda^2 p_k + L\lambda^2 p_k = \beta_k$$

(см. (6.11), (6.42), (6.41), (6.39));

$$\|\mathbf{B}_{k+1}\| = \left\|\mathbf{E} - F'\left(\mathbf{x}^{(k+1)}\right)\mathbf{A}_{k+1}\right\| = \left\|\mathbf{E} - F'\left(\mathbf{x}^{(k+1)}\right)\mathbf{A}_k - F'\left(\mathbf{x}^{(k+1)}\right)\mathbf{A}_k\Psi_k\right\| =$$
$$= \left\|\Psi_k^2\right\| \le \|\Psi_k\|^2 \le \beta_k^2 = b_{k+1}$$

(см. (6.11), (6.40) и предыдущее неравенство);

$$\left\|F\left(\mathbf{x}^{(k+1)}\right)\right\| \le \left\|F\left(\mathbf{x}^{(k)}\right) - F'\left(\mathbf{x}^{(k)}\right)\mathbf{A}_k F\left(\mathbf{x}^{(k)}\right)\right\| + \frac{L}{2}\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\|^2 \le$$
$$\le \|\mathbf{B}_k\| \cdot \left\|F\left(\mathbf{x}^{(k)}\right)\right\| + \frac{L}{2}\|\mathbf{A}_k\|^2 \cdot \left\|F\left(\mathbf{x}^{(k)}\right)\right\|^2 \le b_k p_k + \frac{L}{2}\lambda^2 p_k^2 =$$
$$= L\lambda^2 p_k^2 + \frac{L}{2}\lambda^2 p_k^2 < 4L\lambda^2 p_k^2 = p_{k+1}$$

(см. (6.33), (6.11), (6.42), (6.41), (6.38)); также из (6.11) следует, что

$$\|\mathbf{A}_{k+1}\| \le \|\mathbf{A}_k\|\left(1 + \|\Psi_k\|\right) \le \|\mathbf{A}_{k-1}\|\left(1 + \|\Psi_{k-1}\|\right)\left(1 + \|\Psi_k\|\right) \le \ldots \le$$
$$\le \|\mathbf{A}_0\| \cdot \prod_{j=0}^{k}\left(1 + \|\Psi_j\|\right) \le \|\mathbf{A}_0\| \cdot \prod_{j=0}^{k}\left(1 + \beta_j\right);$$

но $\beta_k = \frac{1}{2}\nu^{2^k}$ (действительно, из предположения, что $\beta_{k-1} = \frac{1}{2}\nu^{2^{k-1}}$, получаем $\beta_k = 2L\lambda^2 p_k = \frac{1}{2}\left(4L\lambda^2 p_{k-1}\right)^2 = \frac{1}{2}(2\beta_{k-1})^2 = \frac{1}{2}\left(\nu^{2^{k-1}}\right)^2 = \frac{1}{2}\nu^{2^k}$),

поэтому далее можно продолжить оценивание

$$\|\mathbf{A}_{k+1}\| \le \|\mathbf{A}_0\| \cdot \prod_{j=0}^{k}\left(1+\frac{1}{2}\nu^{2^j}\right) \le \|\mathbf{A}_0\|\left(1+\frac{1}{2}\sum_{i=1}^{2^{k+1}-1}\nu^i\right) =$$

$$= \|\mathbf{A}_0\|\left(1+\frac{1}{2}\cdot\frac{\nu-\nu^{2^{k+1}}}{1-\nu}\right) \le \|\mathbf{A}_0\|\cdot\frac{2-\nu}{2-2\nu} \le \lambda$$

(последнее, в силу наложенного на ν условия (6.36)).

Так как $\left\|F\left(\mathbf{x}^{(0)}\right)\right\| = p_0$, согласно заданию (6.38), а $\|\mathbf{B}_0\| \le b_0$ по условию (6.35) и, кроме того, $\|\mathbf{A}_0\| \le \lambda$ (ибо в противном случае должно быть $\nu < 0$, что противоречило бы определению ν в (6.36)), то в соответствии с принципом математической индукции доказываемое мажорирование действительно имеет место при любом $k \in N_0$.

Итак, отберем только нужную для применения теоремы 6.3 информацию. Имеем:

$$\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| \le \|\mathbf{A}_k\| \cdot \left\|F\left(\mathbf{x}^{(k)}\right)\right\| \le \lambda p_k,$$

$$\left\|F\left(\mathbf{x}^{(k)}\right)\right\| \le p_k,$$

$$p_{k+1} = G_0 p_k^2, \qquad \text{где} \quad G_0 = 4L\lambda^2, \quad p_0 = \left\|F\left(\mathbf{x}^{(0)}\right)\right\|$$

(т.е. $\mu = 2$, $\nu = G_0 p_0 = 4L\lambda^2 p_0$).

Поскольку здесь на величину ν наложено более сильное, чем в теореме 6.3, ограничение (6.36), с этим ν заключительная часть теоремы 6.3 будет тем более верна. Теорема доказана.

Условие (6.36) теоремы 6.7 можно трактовать так: параметр $\lambda > 0$ должен удовлетворять кубическому неравенству

$$4Lp_0\lambda^3 - 2L\|\mathbf{A}_0\|p_0\lambda^2 - \lambda + \|\mathbf{A}_0\| \le 0.$$

Исследование этого неравенства приводит к ряду следствий [13]; наиболее простым из них (в некотором смысле) является

Следствие 6.1. *Квадратичная сходимость начатого с* $\mathbf{x}^{(0)}$, $\mathbf{A}_0$ *метода* (6.11) *обеспечивается выполнением условий*

$$\left\|\mathbf{E} - F'\left(\mathbf{x}^{(0)}\right)\mathbf{A}_0\right\| \le 0.109, \quad L\|\mathbf{A}_0\|^2\left\|F\left(\mathbf{x}^{(0)}\right)\right\| \le 0.0567,$$

$$S\left(\mathbf{x}^{(0)}, r := 2.46\cdot\|\mathbf{A}_0\|\cdot\left\|F\left(\mathbf{x}^{(0)}\right)\right\|\right) \subseteq M.^{*)}$$

*) Как показывают более тонкие исследования, фигурирующие здесь постоянные 0.109 и 0.0567 сильно занижены (см., например, [60], где в этой роли выступает одна постоянная 0.25).

Замечание 6.6. Приведенный в этом пункте цикл теорем можно значительно расширить и обобщить.

Во-первых, без каких-либо существенных изменений все эти утверждения могут быть доказаны для нелинейных операторных уравнений в банаховых пространствах.

Во-вторых, вместо условия Липшица на $F'(\mathbf{x})$ (которое, кстати, легко можно заменить требованием ограниченности $\|F''(\mathbf{x})\|$ в M постоянной L) можно вставить более слабое ***условие Гёльдера***

$$\|F'(\mathbf{x}) - F'(\tilde{\mathbf{x}})\| \le L\|\mathbf{x} - \tilde{\mathbf{x}}\|^{\alpha}, \quad \alpha \in (0;1],$$

частным случаем которого является условие Липшица. При этом порядок μ метода в теоремах типа теорем 6.5, 6.7 может быть установлен «плавающим» от 1 до $1+\alpha$ в зависимости от жесткости требований, накладываемых на исходные данные.

В-третьих, общие теоремы 6.3 и 6.4 позволяют исследовать сходимость методов более высоких порядков как с аппроксимацией обратных к матрицам Якоби матриц, так и без нее, например, методов третьего порядка (6.9) и (6.12).

Использование техники оценочных функций, ключ к которой можно найти в книге [31] Л. Коллатца, а также обобщение леммы 6.2 на случай, когда условие Липшица или Гельдера накладывается на производные более высоких порядков, позволяют указывать условия сходимости семейства методов вида

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - Q\left(\mathbf{x}^{(k)}, \mathbf{A}_k\right),$$

содержащего изученные здесь методы и ряд других методов.

В заключение отметим, что применение теорем сходимости типа теорем (6.5) - (6.7) и других подобных утверждений для конкретных нелинейных систем упирается, как правило, в проблему нахождения постоянных Липшица L для производных векторных функций. Даже в простейшем случае, когда L находится из условия $L \ge \|F''(\mathbf{x})\|$, требуется сделать оценку величины нормы функциональной матрицы Гессе $\left(\frac{\partial^2 f_m}{\partial x_i \partial y_j}\right)_{i,j,m=1}^{n}$ в некоторой n-мерной области M, что весьма непросто.

УПРАЖНЕНИЯ

6.1. Провести доказательство теоремы 6.1 по аналогии с доказательством теоремы 5.7.

6.2. Провести доказательство теоремы 6.2 по аналогии с доказательством теоремы 5.8, используя лемму 6.1.

6.3. Можно ли утверждать, что система

$$\begin{cases} x = 0.1\sin x + 0.3\cos y - 0.4, \\ y = 0.2\cos x - 0.1\sin y - 0.3 \end{cases}$$

имеет и притом единственное вещественное решение? Почему? Сделать 5 итераций МПИ (6.3), начиная процесс с нулевого начального вектора, и оценить погрешность с помощью априорной оценки теоремы 6.1. Найти решение с точностью $\varepsilon = 10^{-6}$, останавливая процесс вычислений на основе апостериорной оценки погрешности. Сделать 5 шагов метода покоординатных итераций (6.4). Сравнить результат с предыдущим.

6.4. Дана система

$$\begin{cases} x^3 - y^3 + 0.1 = 0, \\ xy - 0.95 = 0 \end{cases}$$

и точка $(x_0, y_0) = (1,1)$. Записать для этой системы основной (6.6) и модифицированный (6.8) методы Ньютона и сделать по 2-3 итерационных шага. Попытаться применить здесь какие-нибудь подходящие теоремы сходимости (см. теоремы 6.1, 6.2, 6.5, 6.6).

6.5. Сравнить по числу арифметических операций, приходящихся на реализацию одного итерационного шага при решении n-мерной нелинейной системы, следующие методы:

а) метод Ньютона в явной форме (6.6);

б) метод Ньютона в неявной форме (6.7);

в) аппроксимационный аналог метода Ньютона (6.11);

г) простейший метод секущих (6.14);

д) разностный метод Ньютона (6.13).

Подсчет вычислительных затрат вести, предполагая, что решение линейных систем в (6.7) и обращение матриц в (6.6), (6.14) и (6.13) производится методом Гаусса и что вычисление значений функций и производных при этом во внимание не принимается, т.е. эти значения считаются уже найденными.

6.6. Записать алгоритм решения нелинейных систем такой модификацией метода Ньютона, при которой частные производные в матрицах Якоби аппроксимируются симметричными разностными отношениями:

$$\frac{\partial f_i}{\partial x_j} \approx \frac{f_i\left(x_1,\ldots,x_j+h_j,\ldots,x_n\right)-f_i\left(x_1,\ldots,x_j-h_j,\ldots,x_n\right)}{2h_j}.$$

Увеличатся ли вычислительные затраты при переходе от (6.13) к такому симметричному разностному методу?

6.7. Дополнить рассмотрение численного примера п.6.5 (табл.6.1 и 6.2) применением методов третьего порядка (6.9) (или, что то же, (6.10)) и (6.12). Провести сравнительный анализ полученных результатов.

6.8. Записать аппроксимационные аналоги (типа ААМН (6.11)) для методов секущих (6.14) и (6.15). Протестировать новые методы на системе (6.25).

6.9. Вывести расчетные формулы метода Брауна (см. п. 6.3) для трехмерного случая. Опробовать полученные формулы на системе

$$\begin{cases} x^2+y^2-z=0, \\ x^2+y^2-z^2=0, \\ \ln x-\sqrt{y}+0.8=0, \end{cases}$$

взяв начальное приближение $x_0=0.5$, $y_0=0.5$, $z_0=0.5$. Проверить результаты тем же методом Брауна, но примененным к двумерной системе, к которой легко перейти от данной исключением z.

6.10. Составить гибридный алгоритм, осуществляющий поиск решения нелинейной системы сначала методом градиентного спуска, а затем методом Ньютона или какой-либо его быстросходящейся модификацией. Применить построенный алгоритм к системе из предыдущего упражнения при различных начальных точках (в частности, при $x_0=y_0=z_0=0$).

КРАТКИЕ СВЕДЕНИЯ О НОРМАХ ВЕКТОРОВ И МАТРИЦ. СХОДИМОСТЬ В КОНЕЧНОМЕРНЫХ ПРОСТРАНСТВАХ *)

Норма – это одна из важнейших скалярных характеристик векторов и матриц, удобная для применения, в частности, в вычислительной математике.

Будем сначала рассматривать множество R_n всех n-мерных вещественных векторов, образующее, как известно, линейное пространство. Это множество неупорядоченное, т.е. нельзя сказать какой вектор больше, а какой меньше. Чтобы получить возможность каким-то образом сравнивать векторы по величине, и вводится понятие нормы вектора. Норма обобщает понятие длины вектора, являющееся естественным и хорошо изученным в реальных пространствах R_2, R_3. Как и многие другие математические понятия, норма определяется аксиоматически.

Определение 1. ***Нормой вектора*** $\mathbf{x} \in R_n$ *называется такое действительное число, обозначаемое* $\|\mathbf{x}\|$, *что*:

1) $\|\mathbf{x}\| \geq 0$, *причем* $\|\mathbf{x}\| = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}$;

2) $\|\lambda\mathbf{x}\| = |\lambda|\|\mathbf{x}\| \quad \forall\lambda \in R_1$;

3) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\| \quad \forall\mathbf{y} \in R_n$.

Линейное пространство R_n с введенной в нем нормой называют ***нормированным пространством*** (точнее, вещественным нормированным пространством).

Легко видеть, что обычное понятие длины (модуля) вектора удовлетворяет всем аксиомам, определяющим норму, т.е. длина есть норма. Обратное неверно: определение 1 норму вектора однозначно не задает. Можно указать бесчисленное множество конструкций $\|\mathbf{x}\|$, удовлетворяющих всем трем аксиомам нормы вектора. Например, нетрудно проверить, что нормой вектора имеет право называться выражение

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}} \qquad (1)$$

*) Более полное изложение см., например, в книгах [16, 21, 29, 31, 42, 45, 51, 54].

при любом $p \in [1,\infty)$, где x_i при $i = 1,2,\ldots,n$ – компоненты вектора $\mathbf{x}$. Это так называемая l_p-***норма*** или ***норма Гёльдера*** [16]. Фиксированием значений параметра p отсюда можно получать конкретные привлекательные по тем или иным соображениям нормы.

Так, при $p = 2$ из (1) получаем ***евклидову норму вектора***

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2},$$

являющуюся наиболее естественным расширением понятия длины вектора на n-мерный случай.*) Множество R_n всех n-мерных векторов с введенной в нем евклидовой нормой называют ***евклидовым пространством*** и часто обозначают через E_n. Это пространство характерно тем, что в нем определено скалярное произведение векторов $\mathbf{x} = (x_1,\ldots,x_n)$ и $\mathbf{y} = (y_1,\ldots,y_n)$ равенством

$$(\mathbf{x},\mathbf{y}) := \sum_{i=1}^{n} x_i y_i$$

и при этом, очевидно, имеет место связь

$$\|\mathbf{x}\|_2 = \sqrt{(\mathbf{x},\mathbf{x})}.$$

При $p = 1$ из l_p-нормы получается ***норма-сумма***

$$\|\mathbf{x}\|_1 := \sum_{i=1}^{n} |x_i|,$$

а переход в (1) к пределу при $p \to \infty$ дает так называемую ***норму-максимум***

$$\|\mathbf{x}\|_\infty := \max_{i=1,n} |x_i|.$$

Указанные наиболее распространенные на практике три нормы $\|\mathbf{x}\|_2$, $\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_\infty$ называют еще соответственно *сферической*, *октаэдрической* и *кубической* по названию поверхности, определяемой уравнением $\|\mathbf{x}\| = \text{const}$ (в частности, $\|\mathbf{x}\| = 1$), когда $\mathbf{x}$ считается переменным трехмерным радиус-вектором.

Пусть теперь рассматривается множество всех $n \times n$-матриц, также образующих линейное пространство.

---

*) Разумеется, знак $|\cdot|$ здесь может быть опущен (чего никак нельзя делать в следующих двух частных случаях). Его сохранение требуется лишь тогда, когда вместо R_n рассматривается пространство C_n векторов с комплексными компонентами.

Определение 2. ***Нормой матрицы*** $\mathbf{A}$ *называется действительное число* $\|\mathbf{A}\|$, *удовлетворяющее условиям*:

1) $\|\mathbf{A}\| \geq 0$, *причем* $\|\mathbf{A}\| = 0 \Leftrightarrow \mathbf{A} = \mathbf{0}$;

2) $\|\lambda\mathbf{A}\| = |\lambda| \cdot \|\mathbf{A}\| \quad \forall \lambda \in R_1$;

3) $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$;

4) $\|\mathbf{AB}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{B}\|$

(B – *произвольная* $n \times n$-*матрица*).

Иногда норму матрицы определяют только с помощью первых трех условий (аксиом); в таком случае говорят, что определение 2 задает в пространстве матриц ***мультипликативную норму***.

Как и норму вектора, норму матрицы, отвечающую определению 2, можно вводить далеко не единственным способом. Из множества всевозможных норм матрицы наибольший интерес представляют такие, которые определенным образом соотносятся с векторными нормами, поскольку, чаще всего, матрицы и векторы рассматриваются в комплексе. Так, при умножении матрицы $\mathbf{A}$ на вектор $\mathbf{x}$ получается вектор $\mathbf{Ax}$, и естественно потребовать, чтобы матричная норма удовлетворяла ***условию согласованности***

$$\|\mathbf{Ax}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{x}\|.$$

Например, если для матрицы $\mathbf{A} = \left(a_{ij}\right)_{i,j=1}^{n}$ ввести норму равенством*)

$$\|\mathbf{A}\|_F := \sqrt{\sum_{i,j=1}^{n} \left|a_{ij}\right|^2},$$

то она будет удовлетворять всем четырем аксиомам определения 2 и согласована с евклидовой нормой вектора. Такую норму называют ***нормой Фробенуса*** [44], ***евклидовой*** или ***шуровской нормой*** [53], а также ***нормой Э. Шмидта*** [31].

Более сильным требованием к норме матрицы, чем условие согласованности, является условие подчиненности. А именно, норма $n \times n$-матрицы $\mathbf{A}$ называется ***подчиненной норме*** n-мерного вектора $\mathbf{x}$, если она задается равенством

$$\|\mathbf{A}\| := \max_{\mathbf{x}=1} \|\mathbf{Ax}\|.$$

Таким образом, при заданной векторной норме за подчиненную ей норму матрицы $\mathbf{A}$ принимается максимум норм векторов $\mathbf{Ax}$, когда $\mathbf{x}$ пробегает множество всех векторов, норма которых равна единице.

---

*) См предыдущую сноску

Очевидно, при всякой подчиненной норме для единичной матрицы **E** должно быть

$$\|\mathbf{E}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{E}\mathbf{x}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{x}\| = 1.$$

Так как

$$\|\mathbf{E}\|_F = \sqrt{\sum_{i=1}^{n} 1^2} = \sqrt{n},$$

то введенная выше норма Фробениуса, будучи согласованной с евклидовой нормой вектора, не является подчиненной ей. Нормой вещественной матрицы **A**, подчиненной евклидовой норме вектора, служит ***спектральная норма***:

$$\|\mathbf{A}\|_2 := \sqrt{\Lambda},$$

где Λ – наибольшее собственное число матрицы $\mathbf{A}^T\mathbf{A}$ (положительное, в силу симметричности $\mathbf{A}^T\mathbf{A}$).

Нормами матрицы $\mathbf{A} = \left(a_{ij}\right)_{i,j=1}^{n}$, подчиненными другим введенным выше нормам векторов, являются:

$$\|\mathbf{A}\|_1 := \max_j \sum_{i=1}^{n} |a_{ij}| \qquad \text{для } \|\mathbf{x}\|_1,$$

$$\|\mathbf{A}\|_\infty := \max_i \sum_{j=1}^{n} |a_{ij}| \qquad \text{для } \|\mathbf{x}\|_\infty.$$

В тех случаях, когда безразлично, какую из норм следует употребить, пишут просто $\|\mathbf{x}\|$ или $\|\mathbf{A}\|$, понимая под этим любые нормы (или любые подчиненные, или любые согласованные, что бывает ясно из рассмотрения объектов, с которыми оперируют). Так, например, для любой мультипликативной нормы матрицы справедливо неравенство

$$\left\|\mathbf{A}^k\right\| \le \|\mathbf{A}\|^k \quad \forall k \in N,$$

но под символом $\|\cdot\|$ в левой и правой частях должна пониматься одна и та же норма.

Пусть имеется последовательность $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(k)}, \ldots$, членами которой являются n-мерные векторы

$$\mathbf{x}^{(k)} = \left(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}\right).$$

Если все последовательности $\left(x_i^{(k)}\right)$ соответствующих компонент этих векторов имеют пределы, т.е. при всех $i = 1, 2, \ldots, n$ существуют $\lim\limits_{k\to\infty} x_i^{(k)} = x_i^*$, то вектор $\mathbf{x}^* = \left(x_1^*, x_2^*, \ldots, x_n^*\right)$ считают ***пределом векторной***

последовательности $\left(\mathbf{x}^{(k)}\right)_{k=1}^{\infty}$. Этот факт сходимости оформляют обычным образом:

$$\lim_{k\to\infty}\mathbf{x}^{(k)}=\mathbf{x}^{*} \quad \text{или} \quad \mathbf{x}^{(k)}\underset{k\to\infty}{\to}\mathbf{x}^{*}.$$

Так же покоординатно (поэлементно) понимается и сходимость последовательностей матриц. А именно, ***пределом последовательности матриц*** $\mathbf{A}^{(1)},\mathbf{A}^{(2)},\ldots,\mathbf{A}^{(k)},\ldots$ с элементами $a_{ij}^{(1)},a_{ij}^{(2)},\ldots,a_{ij}^{(k)},\ldots$ называется такая матрица $\mathbf{A}^{*}$, элементами которой являются числа $a_{ij}^{*}:=\lim\limits_{k\to\infty}a_{ij}^{(k)}$.

Необходимым и достаточным условием поэлементной сходимости последовательности векторов $\mathbf{x}^{(k)}$ к вектору $\mathbf{x}^{*}$, последовательности матриц $\mathbf{A}^{(k)}$ к матрице A^{*}, является условие ***сходимости по норме***:

$$\left\|\mathbf{x}^{(k)}-\mathbf{x}^{*}\right\|\to 0,\quad \left\|\mathbf{A}^{(k)}-\mathbf{A}^{*}\right\|\to 0$$

соответственно. При этом важно отметить, что все нормы в конечномерных пространствах (каковыми являются векторные и матричные пространства) эквивалентны. Это надо понимать так: если доказана сходимость векторной или матричной последовательности в одной нормировке, то сходимость к тому же пределу обеспечена и при любой другой нормировке.

ПРОИЗВОДНЫЕ ВЕКТОРНЫХ ФУНКЦИЙ *)

Пусть R_n и R_m – два вещественных нормированных пространства с одинаковой нормой, и пусть F – некоторое заданное, в общем случае нелинейное отображение из R_n в R_m, т.е. F – векторная функция, аргументами которой являются n-мерные векторы $\mathbf{x}$, а значениями – векторы $F(\mathbf{x})$ размерности m. Пусть, далее, $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ – фиксированная точка некоторой открытой области D, содержащейся в области определения отображения F, $\mathbf{h}$ – произвольный (текущий) n-мерный вектор такой, что $\mathbf{x} + \mathbf{h} \in D$, и наконец, пусть**) $F(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x}))$.

Определение 1. *Если существует такое линейное преобразование (отображение, оператор)* $\mathbf{A}$, *что приращение* F *представимо в виде*

$$F(\mathbf{x}+\mathbf{h}) - F(\mathbf{x}) = \mathbf{A}\mathbf{h} + \omega(\mathbf{x},\mathbf{h}), \qquad (1)$$

где $\dfrac{\|\omega(\mathbf{x},\mathbf{h})\|}{\|\mathbf{h}\|} \to 0$ *при* $\|\mathbf{h}\| \to 0$, *то отображение* F *называется* ***дифференцируемым по Фреше****, линейное преобразование* $\mathbf{A} = F'(\mathbf{x})$ *–* ***производной Фреше*** *или* ***сильной производной*****)*, а его значение* $\mathbf{A}\mathbf{h} = F'(\mathbf{x})\mathbf{h}$ *–* ***дифференциалом Фреше****.*

Очевидно, при $n = m = 1$ это определение совпадает с обычным определением дифференцируемости функций одной вещественной переменной, и производная Фреше в этом случае есть обычная производная.

Выясним структуру производной Фреше при $n, m \geq 1$. Для этого нужно найти выражения элементов $a_{ij}\left(i = \overline{1,m};\ j = \overline{1,n}\right)$ матрицы линейного преобразования $\mathbf{A}$, зависящие, вообще говоря, от $\mathbf{x}$. Преследуя эту цель, будем рассматривать определяющее производную равенство (1) на координатных векторах, что правомерно, в силу произвольности $\mathbf{h}$.

Положим сначала $\mathbf{h} := \mathbf{h}_1 := (\delta_1, 0, \ldots, 0)$, где δ_1 – произвольное действительное число. Тогда, учитывая, что

---

*) См [29, 42, 45, 51 и др.]

**) В пределах этого приложения вектор-столбцы фигурируют, в основном, в виде векторов-строк без символа транспонирования

) Наряду с производной Фреше определяют еще ***производную Гато или ***слабую производную***. Всякая функция, дифференцируемая по Фреше, дифференцируема и по Гато, но не наоборот.

$$\mathbf{x}+\mathbf{h}=\begin{pmatrix} x_1+\delta_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \qquad \mathbf{Ah}=\mathbf{Ah}_1=\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}\begin{pmatrix} \delta_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}=\begin{pmatrix} a_{11}\delta_1 \\ a_{21}\delta_1 \\ \vdots \\ a_{m1}\delta_1 \end{pmatrix},$$

и предполагая отображение F дифференцируемым, т.е. равенство (1) выполненным, получим векторное равенство

$$\begin{pmatrix} f_1(x_1+\delta_1,x_2,\dots,x_n)-f_1(x_1,x_2,\dots,x_n) \\ f_2(x_1+\delta_1,x_2,\dots,x_n)-f_2(x_1,x_2,\dots,x_n) \\ \dots\dots\dots\dots\dots\dots\dots \\ f_m(x_1+\delta_1,x_2,\dots,x_n)-f_m(x_1,x_2,\dots,x_n) \end{pmatrix}=\begin{pmatrix} a_{11}\delta_1 \\ a_{21}\delta_1 \\ \dots \\ a_{m1}\delta_1 \end{pmatrix}+\begin{pmatrix} \omega_1(\mathbf{x},\delta_1) \\ \omega_2(\mathbf{x},\delta_1) \\ \dots\dots \\ \omega_m(\mathbf{x},\delta_1) \end{pmatrix}, \quad (2)$$

где $\omega_i(\mathbf{x},\delta_1)$ – i-я компонента остаточного члена $\omega(\mathbf{x},\mathbf{h})$ дифференциала Фреше на векторе $\mathbf{h}_1$. Так как $\|\omega(\mathbf{x},\mathbf{h}_1)\| \geq |\omega_i(\mathbf{x},\mathbf{h}_1)|$, а $\|\mathbf{h}\|=\|\mathbf{h}_1\|=|\delta_1|$, то, очевидно,

$$\lim_{\|\mathbf{h}_1\|\to 0}\frac{\|\omega(\mathbf{x},\mathbf{h}_1)\|}{\|\mathbf{h}_1\|}=0 \Rightarrow \lim_{\delta_1\to 0}\frac{|\omega_i(\mathbf{x},\delta_1)|}{|\delta_1|}=0 \Rightarrow \lim_{\delta_1\to 0}\frac{\omega_i(\mathbf{x},\delta_1)}{\delta_1}=0.$$

Поэтому, деля i-е равенство, получающееся из (2) переходом к координатам, на число δ_1 и устремляя δ_1 к нулю, имеем

$$\lim_{\delta_1\to 0}\frac{f_i(x_1+\delta_1,x_2,\dots,x_n)-f_i(x_1,x_2,\dots,x_n)}{\delta_1}=\lim_{\delta_1\to 0}\frac{\omega_i(\mathbf{x},\delta_1)}{\delta_1}=a_{i1},$$

т.е. при $i=1,2,\dots,m$ справедливо представление

$$a_{i1}=\frac{\partial f_i(\mathbf{x})}{\partial x_1}.$$

Рассматривая равенство (1) на других координатных векторах $\mathbf{h}=\mathbf{h}_j=(0,\dots,\delta_j,\dots,0)$, аналогично предыдущему находим выражения всех элементов матрицы линейного преобразования $\mathbf{A}$, фигурирующей в определении 1:

$$a_{ij}=\lim_{\delta_j\to 0}\frac{f_i(x_1,\dots,x_j+\delta_j,\dots,x_n)-f_i(x_1,\dots,x_j,\dots,x_n)}{\delta_j}=\frac{\partial f_i(\mathbf{x})}{\partial x_j}$$

$$(i=1,2,\dots,m;\quad j=1,2,\dots,n).$$

Таким образом, производная Фреше отображения $F: R_n \to R_m$ представляет собой матрицу частных производных от каждой компоненты f_i векторной функции $F=(f_1,f_2,\dots,f_m)$ по каждой компоненте x_j векторного аргумента $\mathbf{x}$. Эта $m\times n$-матрица называется ***матрицей Якоби*** и имеет вид

$$F'(\mathbf{x}) := J(\mathbf{x}) := \begin{pmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \frac{\partial f_1(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \frac{\partial f_2(\mathbf{x})}{\partial x_1} & \frac{\partial f_2(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f_2(\mathbf{x})}{\partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f_m(\mathbf{x})}{\partial x_1} & \frac{\partial f_m(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{pmatrix}.$$

В частности, если F осуществляет отображение из R_n в R_1, т.е. если $F(\mathbf{x}) = f(x_1, x_2, \ldots, x_n)$ – функция n переменных, то

$$F'(\mathbf{x}) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n} \right) = \operatorname{grad} F.$$

Следовательно, ***градиент*** – это матрица-строка Якоби.

Приведем некоторые **свойства производной Фреше.**

1. *Если производная Фреше в точке* $\mathbf{x}$ *существует, то она единственна.*

2. *Наличие всех частных производных первого порядка у функции* $F(\mathbf{x})$ *еще не означает существования у* $F(\mathbf{x})$ *производной Фреше.*

3. *Отображение* F, *дифференцируемое по Фреше в точке* $\mathbf{x}$, *непрерывно в этой точке* (чего нельзя утверждать в случае слабой дифференцируемости).

4. *Справедливо цепное правило дифференцирования композиции отображений: если отображение* $F: R_n \to R_m$ *имеет производную в точке* $\mathbf{x}$, *а* $G: R_m \to R_l$ *– в точке* $F(\mathbf{x})$, *то их композиция* $H := G \cdot F$, *переводящая элементы пространства* R_n *в элементы пространства* R_l, *имеет производную в точке* $\mathbf{x}$, *причем* $H'(\mathbf{x}) = G'(F(\mathbf{x})) \cdot F'(\mathbf{x})$.

Множество $m \times n$-матриц образует линейное пространство $L(R_n, R_m)$ размерности $s = n \cdot m$, которое можно естественным образом нормировать. Поэтому к F', как к отображению из R_n в $L(R_n, R_m)$, можно применить данное выше определение 1 производной Фреше, в результате чего приходим к понятию второй производной Фреше нелинейного многомерного отображения F.

Определение 2. *Если отображение* $F: R_n \to R_m$ *имеет в точке* $\mathbf{x} \in D \subseteq R_n$ *дифференцируемую по Фреше производную* $F': R_n \to L(R_n, R_m)$, *то производная производной, т.е.* $(F'(\mathbf{x}))'$, *на-*

зывается ***второй производной Фреше*** *отображения* F *и обозначается* $F''(\mathbf{x})$.

Ясно, что $F''(\mathbf{x}) \in L(R_n, L(R_n, R_m))$, т.е. $F''(\mathbf{x})\mathbf{g}$ является линейным отображением ($m \times n$-матрицей) при каждом $\mathbf{g}$ из R_n, в то время как $F'(\mathbf{x})\mathbf{h}$ есть вектор из R_m при каждом $\mathbf{h}$ из R_n. Следовательно,

$$F''(\mathbf{x})\mathbf{gh} \in R_m \quad \forall \mathbf{g} \in R_n, \mathbf{h} \in R_n,$$

и значит, вторую производную можно интерпретировать как отображение из $R_n \times R_n$ в R_m. При этом заметим, что $F''(\mathbf{x})$ является билинейным отображением*) и обладает свойством симметрии

$$F''(\mathbf{x})\mathbf{hg} = F''(\mathbf{x})\mathbf{gh} \quad \forall \mathbf{g}, \mathbf{h} \in R_n.$$

Для того чтобы связать $F''(\mathbf{x})$ с частными производными функций $f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})$ – компонент вектора F, нужно так же, как и при выявлении структуры первой производной, рассмотреть равенство (1), примененное к $F'(\mathbf{x})$, на координатных векторах $\mathbf{h} = \mathbf{h}_i$, $\mathbf{g} = \mathbf{g}_i$. В результате получается, что вектор $F''(\mathbf{x})\mathbf{hg}$ при произвольных $\mathbf{h}$ и $\mathbf{g}$ из R_n имеет представление

$$F''(\mathbf{x})\mathbf{hg} = \left(\mathbf{g}^T H_1(\mathbf{x})\mathbf{h}, \mathbf{g}^T H_2(\mathbf{x})\mathbf{h}, \ldots, \mathbf{g}^T H_m(\mathbf{x})\mathbf{h}\right),$$

где $H_j(\mathbf{x})$ – так называемая ***матрица Гессе***, составленная из вторых частных производных, а именно:

$$H_j(\mathbf{x}) := \begin{pmatrix} \dfrac{\partial^2 f_j(\mathbf{x})}{\partial x_1^2} & \dfrac{\partial^2 f_j(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \dfrac{\partial^2 f_j(\mathbf{x})}{\partial x_1 \partial x_n} \\ \dfrac{\partial^2 f_j(\mathbf{x})}{\partial x_2 \partial x_1} & \dfrac{\partial^2 f_j(\mathbf{x})}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f_j(\mathbf{x})}{\partial x_2 \partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \dfrac{\partial^2 f_j(\mathbf{x})}{\partial x_n \partial x_1} & \dfrac{\partial^2 f_j(\mathbf{x})}{\partial x_n \partial x_2} & \cdots & \dfrac{\partial^2 f_j(\mathbf{x})}{\partial x_n^2} \end{pmatrix}$$

Аналогично могут быть введены и изучены производные высших порядков.

---

*) Отображение $B: R_n \times R_n \to R_m$ называется ***билинейным***, если оно линейно по каждому из своих аргументов, т.е. если

$$B(\alpha_1 g_1 + \alpha_2 g_2)h = \alpha_1 Bg_1 h + \alpha_2 Bg_2 h \quad \text{и} \quad Bg(\beta_1 h_1 + \beta_2 h_2) = \beta_1 Bgh_1 + \beta_2 Bgh_2.$$

ОБРАЗЦЫ ПОСТАНОВОК ЛАБОРАТОРНЫХ ЗАДАНИЙ

Лабораторная работа 1. **«Метод Гаусса и** LU**-разложение матриц»**

Дана система $\mathbf{Ax} = \mathbf{b}$, где:

вариант 1

$$\mathbf{A} = \begin{pmatrix} 14 & -8 & -21 & 12 \\ 10 & -6 & -15 & 9 \\ 35 & -20 & -56 & 32 \\ 25 & -15 & -40 & 24 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 19 \\ 14 \\ 53 \\ 39 \end{pmatrix};$$

вариант 2

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & -3 & -9 \\ 0 & 1 & -7 & -21 \\ 3 & 12 & -92 & -279 \\ 1 & 4 & -31 & -94 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 11 \\ 22 \\ 297 \\ 100 \end{pmatrix};$$

и т.д.

1. Решить систему методом Гаусса. Предусмотреть постолбцовый выбор главного элемента и итерационное уточнение решения до достижения точности $\varepsilon = 10^{-12}$ по евклидовой норме невязки в рамках применяемой схемы реализации метода.

2. Выполнить LU-разложение матрицы $\mathbf{A}$ и с его помощью получить $\det \mathbf{A}$ и решение $\mathbf{x}$ данной системы.

3. Найти матрицу $\mathbf{X} = \mathbf{A}^{-1}$ двумя способами:

а) решая подсистемы $\mathbf{Ax}^j = \mathbf{e}^j$ системы $\mathbf{AX} = \mathbf{E}$ (используя при этом фрагменты выполнения п.1);

б) применяя готовые формулы, полученные на основе LU-разложения.

4. Вычислить $\operatorname{cond} \mathbf{A}$ в различных простых нормах и охарактеризовать чувствительность данной системы к погрешностям исходных данных.

Лабораторная работа 2. **«Метод прогонки»**

Методом прогонки найти вектор $(u_1, u_2, \ldots, u_{10})$, являющийся решением уравнения (системы)

$$a_k u_{k-1} + b_k u_k + c_k u_{k+1} = f_k,$$

где $k = 1,2,\ldots,10$; $a_1 = c_{10} = 0$; коэффициенты a_k (при $k = 2,3,\ldots,10$), b_k (при $k = 1,2,\ldots,10$), c_k (при $k = 1,2,\ldots,9$) и f_k (при $k = 1,2,\ldots,10$) задаются следующей таблицей:

	a_k	b_k	c_k	f_k
Вариант 1	k	$3.1k$	$-2k$	$\dfrac{2.1k^2 + 7.2k + 2}{k^2 + 3k + 2}$
Вариант 2	$\dfrac{3}{k}$	$\dfrac{11}{10k}$	$\dfrac{2}{k}$	$30.5 - \dfrac{41.6}{k}$
...	...	...	...	...

Что можно сказать об устойчивости прогонки в данном конкретном случае?

Лабораторная работа 3. **«Прямое и итерационное решение симметричной линейной алгебраической системы»**

С точностью $\varepsilon = 10^{-6}$ найти решение системы

$$\sum_{j=1}^{6} a_{ij} x_j = b_j \quad (i = 1,2,\ldots,6)$$

с матрицей коэффициентов вида

$$\mathbf{A} = (a_{ij}) = \begin{pmatrix} p_1 & 0.1p_1 & 0 & 0 & q & 0 \\ 0.1p_1 & p_2 & 0.1p_2 & 0 & 0 & q \\ 0 & 0.1p_2 & p_3 & 0.1p_3 & 0 & 0 \\ 0 & 0 & 0.1p_3 & p_4 & 0.1p_4 & 0 \\ q & 0 & 0 & 0.1p_4 & p_5 & 0.1p_5 \\ 0 & q & 0 & 0 & 0.1p_5 & p_6 \end{pmatrix},$$

если:

	p_i $(i = 1,2,\ldots,6)$	b_i $(i = 1,2,\ldots,6)$	q
Вариант 1	i	1	–0.5
Вариант 2	$10 - i$	$25 - 9i$	2
...	...	...	...

Применить:

а) метод квадратных корней;

б) метод Якоби (сделав предварительно подсчет числа итераций, гарантирующего получение решения с заданной точностью);

в) метод Зейделя.

Провести сравнительный анализ примененных методов.

Попытаться уменьшить число итераций метода Зейделя, вводя релаксационный параметр и оптимизируя его значение экспериментальным путем.

Лабораторная работа 4. **«Численное решение алгебраических проблем собственных значений»**

Дана матрица:

вариант 1

$$\begin{pmatrix} 7 & -1 & -2 & 3 \\ -1 & 6 & 0 & 2 \\ -2 & 0 & 5 & 1 \\ 3 & 2 & 1 & 7 \end{pmatrix};$$

вариант 2

$$\begin{pmatrix} 5 & 2 & 0 & -1 \\ 2 & 7 & -3 & 1 \\ 0 & -3 & 9 & 4 \\ -1 & 1 & 4 & 8 \end{pmatrix};$$

и т.д.

1. Найти наибольшее по модулю собственное число и соответствующий ему собственный вектор

а) степенным методом;

б) методом скалярных произведений.

(В качестве начального взять вектор $(1;1;1;1)$).

2. Найти наименьшее по модулю собственное число и соответствующий ему собственный вектор

а) методом обратных итераций;

б) методом обратных итераций с отношениями Рэлея.

3. Решить полную проблему собственных значений методом вращений Якоби.

Точность $\varepsilon = 10^{-6}$ (в евклидовой норме).

Лабораторная работа 5. **«Методы решения скалярных уравнений»**

С точностью $\varepsilon = 10^{-6}$ найти каждый из корней уравнения:

вариант 1

$$4x\ln^2 x - 4\sqrt{1+x} + 5 = 0;$$

вариант 2

$$x^4 e^x + \sqrt[3]{x-1} - 2 = 0;$$

и т.д.

каждым из следующих четырех способов:

1) методом половинного деления;

2) методом хорд;

3) методом Ньютона;

4) методом секущих.

Сравнить методы по числу итераций и по вычислительным затратам. Что можно сказать об эффективности примененных методов?

Лабораторная работа 6. **«Скалярная задача о неподвижной точке»**

С точностью $\varepsilon = 10^{-6}$ решить уравнение:

вариант 1

$$e^{-0\,45x} - \sqrt{x-3} = 0;$$

вариант 2

$$(x-4)^3 + \ln x = 0;$$

и т.д.

а) методом простых итераций;

б) Δ^2-процессом Эйткена;

в) методом Вегстейна.

Предварительно привести уравнение к виду, пригодному для проведения итераций; доказать существование и единственность корня; выбрав начальное приближение, сделать априорную оценку количества шагов метода простых итераций.

Результаты представить по следующей форме:

Метод	Начальное приближение	Априорное число итераций	Фактическое число итераций	Полученный корень	Невязка
МПИ					
Эйткена					
Вегстейна					

Сравнить методы по требуемому количеству вычислений функций для получения решения с заданной точностью.

Лабораторная работа 7. **«Метод Бернулли вычисления корней многочлена»**

Дано уравнение:

вариант 1

$$x^4 + 1.1x^3 - 11.51x^2 - 2.331x - 0.117 = 0;$$

вариант 2

$$x^4 - 10.3x^3 - 15.75x^2 - 286.875x - 84.375 = 0;$$

и т.д.

1. Непосредственным применением метода Бернулли найти наибольший и наименьший по модулю корни.

2. Используя найденные корни, понизить по схеме Горнера степень многочлена и найти остальные корни.

Лабораторная работа 8. **«Решение систем нелинейных уравнений»**

Дана система и начальная точка:

вариант 1

$$\begin{cases} (x-y)^3 - 8(x+y) = 0, & x_0 = 2, \\ 2(x-y) + 15\ln(x+y) - 5 = 0, & y_0 = -0.5; \end{cases}$$

вариант 2

$$\begin{cases} 0.8x^2 + 2xy + 1.3y^2 + 20x - 15y = 0, & x_0 = 0.5, \\ e^{0.6y - 0.8x} - 1.14x - 1.52y = 0, & y_0 = 1; \end{cases}$$

и т.д.

Найти решение данной системы, исходя из данной начальной точки, следующими методами:

1) основным методом Ньютона (явным или неявным);

2) разностным методом Ньютона (с разными шагами дискретизации производной);

3) модифицированным (упрощенным) методом Ньютона;

4) методом Ньютона с аппроксимацией обратных матриц;

5) методом Брауна;

6) методом градиентного спуска.

Провести сравнение всех указанных методов решения нелинейных систем на основе конкретного вычислительного материала, полученного при задании точности $\varepsilon = 10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}$.

Литература

1. *Амосов А.А., Дубинский Ю.А., Копчёнова Н.В.* Вычислительные методы для инженеров. – М.: Высшая школа, 1994.

2. *Арушанян О.Б., Залёткин С.Ф.* Численное решение обыкновенных дифференциальных уравнений на Фортране. – М.: Изд.-во МГУ, 1990.

3. *Ахромеева Т.С., Курдюмов С.П., Малинецкий Г.Г.* Парадоксы мира нестационарных структур. В кн. "Компьютеры и нелинейные явления". – М.: Наука, 1988.

4. *Бабушка И., Витасек Э., Прагер М.* Численные процессы решения дифференциальных уравнений. – М.: Мир, 1969.

5. *Бахвалов Н.С.* Численные методы. – М: Наука, 1973.

6. *Бахвалов Н.С., Жидков Н.П., Кобельков Г.М.* Численные методы. – М.: Наука, 1987.

7. *Беланов А.А.* Решение алгебраических уравнений методом Лобачевского. – М.: Наука, 1989.

8. *Березин И.С., Жидков Н.П.* Методы вычислений. Т.2. – М.: Физматгиз, 1962.

9. *Боглаев Ю.П.* Вычислительная математика и программирование. – М.: Высшая школа, 1990.

10. *Брадис В.М.* Вычислительная работа в курсе математики средней школы. – М.: Изд.АПН РСФСР, 1962.

11. *Валях Е.* Последовательно-параллельные вычисления. – М.: Мир, 1985.

12. *Васильев Ф.П.* Численные методы решения экстремальных задач. – М.: Наука, 1988.

13. *Вержбицкий В.М.* Выбор параметров в теоремах сходимости одного аппроксимационного аналога метода Ньютона. «Ж. Вычисл. мат. и мат. физ.», 1975, 15, № 6, с.1594-1597.

14. *Вержбицкий В.М.* Обращение матриц и решение нелинейных систем. – Ижевск: Изд. ИМИ, 1980.

15. *Воеводин В.В.* Вычислительные основы линейной алгебры. – М. : Наука, 1977.

16. *Воеводин В.В., Кузнецов Ю.А.* Матрицы и вычисления. – М.: Наука, 1984.

17. *Волков Б.А.* Численные методы. – М: Наука, 1979.

18. *Гавурин М.К.* Нелинейные функциональные уравнения и непрерывные аналоги итерационных методов. "Изв. Вузов, Матем.", 1958, № 5(6), с.18-31.

19. *Гутер Р.С., Овчинский Б.В.* Элементы численного анализа и математической обработки результатов опыта. – М.: Наука, 1970.

20. *Данилина Н.И., Дубровская Н.С. и др.* Вычислительная математика. – М.: Наука, 1985.

21. *Демидович Б.П., Марон И.А.* Основы вычислительной математики. – М.: Наука, 1970.

22. *Дэннис Дж., Шнабель Р.* Численные методы безусловной оптимизации и решения нелинейных уравнений. – М.: Мир, 1988.

23. *Загускин В.Л.* Справочник по численным методам решения уравнений. – М.: Физматгиз, 1960.

24. *Иванов В.В.* Методы вычислений на ЭВМ: Справочное пособие. – Киев: Наукова думка, 1986.

25. *Икрамов Х.Д.* Численные методы линейной алгебры (решение линейных уравнений). "Математика, кибернетика", №4. – М.: Знание, 1987.

26. *Икрамов Х.Д.* Несимметричная проблема собственных значений. – М.: Наука, 1991.

27. *Ильин В.П., Кузнецов Ю.А.* Трехдиагональные матрицы и их приложения. – М.: Наука , 1985.

28. *Калиткин Н.Н.* Численные методы. – М: Наука, 1978.

29. *Канторович Л.В., Акилов Г.П.* Функциональный анализ. – М.: Наука, 1977.

30. *Коллатц Л.* Задачи на собственные значения с. техническими приложениями. – М. : Наука, 1968.

31. *Коллатц Л.* Функциональный анализ и вычислительная математика. – М.: Мир, 1969.

32. *Коллатц Л., Альбрехт Ю.* Задачи по прикладной математике. – М.: Мир, 1978.

33. *Косарев В.И.* 12 лекций по вычислительной математике. – М.: Изд-во МФТИ, 1995.

34. *Крылов В.И., Бобков В.В., Монастырный П.И.* Вычислительные методы. Т.1. – М.: Наука, 1976.

35. *Крылов В.И., Бобков В.В,, Монастырный П.И.* Начала теории вычислительных методов. Линейная алгебра и нелинейные уравнения. – Минск: Наука и техника, 1985.

36. *Ланс Дж.Н.* Численные методы для быстродействующих вычислительных машин. – М.: ИЛ, 1962.

37. *Лесин В.В., Лисовец Ю.П.* Основы методов оптимизации. – М.: Изд. МАИ, 1995.

38. *Мак-Кракен Д., Дорн У.* Численные методы и программирование на Фортране. – М.: Мир, 1977.

39. Под ред. *Монастырного П.И.* Сборник задач по методам вычислений. – М.: Наука, 1994.

40. *Мысовских И.П.* Лекции по методам вычислений. – М.: Физматгиз, 1962.

41. *Ортега Дж., Пул У.* Введение в численные методы решения дифференциальных уравнений. – М.: Наука, 1986.

42. *Ортега Дж., Рейнболдт В.* Итерационные методы решения нелинейных систем уравнений со многими неизвестными. – М.: Мир, 1975.

43. *Островский А.М.* Решение уравнений и систем уравнений. – М.: ИЛ, 1963.

44. *Парлетт Б.* Симметричная проблема собственных значений. – М.: Мир, 1983.

45. *Пугачев В.С.* Лекции по функциональному анализу. – М.: Изд. МАИ, 1996.

46. *Рябенький В.С.* Введение в вычислительную математику. – М.: Наука, 1994.

47. *Самарский А.А., Гулин А.В.* Численные методы. – М.: Наука, 1989.

48. *Самарский А.А., Николаев Е.С.* Методы решения сеточных уравнений. – М.: Наука , 1978.

49. *Тихонов А.Н., Арсенин В.Я.* Методы решения некорректных задач. – М.: Наука, 1985.

50. *Трауб Дж.* Итерационные методы решения уравнений. – М.: Мир, 1985.

51. *Треногин В.А.* Функциональный анализ. – М.: Наука, 1980.

52. *Турчак Л.И.* Основы численных методов. – М.: Наука, 1987.

53. *Уилкинсон Дж.Х.* Алгебраическая проблема собственных значений. – М.: Наука, 1970.

54. *Фаддеев Д.К., Фаддеева В.Н.* Вычислительные методы линейной алгебры. – М.: Физматгиз, 1960.

55. *Форсайт Дж., Молер К.* Численное решение систем линейных алгебраических уравнений. – М.: Мир, 1969.

56. *Хейгеман Л., Янг Д.* Прикладные итерационные методы. – М.: Мир, 1986.

57. *Хемминг Р.В.* Численные методы. – М.: Наука, 1968.

58. *Шарковский А.Н., Майстренко Ю.Л., Романенко Е.Ю.* Разностные уравнения и их приложения. – Киев: Наукова думка, 1986.

59. *Altman M.* An optimum cubically convergent iterative method of inverting a linear bounded operator in hilbert space. «Pacific J.Math.», 10, 1960, № 4, 1107-1113.

60. *Diaconu A.* Sur quelques méthodes itératives combinées. «Matematica» (RSR), 22(45), 1980, № 2, c.247-261.

61. *Schulz G.* Iterative Berechnung der reziproken Matrix. ZAMM, 13, 1933, c.57-59.

Предметный указатель

Учебное издание

Вержбицкий Валентин Михайлович

ЧИСЛЕННЫЕ МЕТОДЫ
(линейная алгебра и нелинейные уравнения)

Редактор *Ж.И. Яковлева*
Художественный редактор *Ю.Э. Иванова*
Технический редактор *Л.А. Овчинникова*
Оригинал-макет выполнен *С.В. Высоцким*

ЛР № 010146 от 25.12.96. Изд. № ФМ-192. Сдано в набор 28.09.99.
Подп. в печать 22.10.99. Формат 60x88$^1/_{16}$. Бум. газетн. Гарнитура «Таймс»
Печать офсетная. Объем 16,66 усл. печ. л. 16,91 усл. кр.-отт.
Тираж 8000 экз. Зак. № 9759.

ГУП издательство «Высшая школа», 101430, Москва, ГСП-4,
Неглинная ул., д. 29/14.

Отпечатано с оригинал-макета на Государственном унитарном
предприятии Смоленский полиграфический комбинат
Министерства Российской Федерации по делам печати,
телерадиовещания и средств массовых коммуникаций.
214020, Смоленск, ул. Смольянинова, 1.

Вержбицкий В.М.

В31 Численные методы (линейная алгебра и нелинейные уравнения): Учеб. пособие для вузов. — М.: Высш. шк., 2000. — 266 с.: ил.

ISBN 5-06-003654-5

В книге последовательно излагаются численные методы решения линейных алгебраических систем, обращения матриц, вычисления собственных чисел и собственных векторов матриц, а также методы решения нелинейных скалярных уравнений и систем таких уравнений. Показываются идеи, выводы и взаимосвязь методов, обсуждается их сравнительная эффективность, обосновывается сходимость, приводятся алгоритмы, численные примеры, задания для упражнений и лабораторных работ. Наряду с методическими погрешностями изучаемых процессов, уделяется внимание и погрешностям, связанным с их компьютерными реализациями.

Пособие предназначено для студентов математических и инженерных специальностей вузов и может быть полезно всем, кто связан с изучением и применением вычислительной математики.

УДК 519.6
ББК 22.19

В издательстве «ВЫСШАЯ ШКОЛА»

ИМЕЮТСЯ В НАЛИЧИИ КНИГИ ИЗ СЕРИИ «ВЫСШАЯ МАТЕМАТИКА»

Архипов Г. И., Садовничий В. А., Чубариков В. Н. **Лекции по математическому анализу.** Учебник. — 695 с.

Книга является учебником по курсу математического анализа и посвящена дифференциальному и интегральному исчислениям функций одной и нескольких переменных. В ее основу положены лекции, прочитанные авторами на механико-математическом факультете МГУ им. М. В. Ломоносова. В учебнике предложен новый подход к изложению ряда основных понятий и теорем анализа, а также и к самому содержанию курса. Она доступна широкому кругу читателей, а первая ее часть может быть использована при изучении ряда тем по алгебре и началам математического анализа в математических школах.

Виноградов И. М. **Элементы высшей математики. (Аналитическая геометрия. Дифференциальное исчисление. Основы теории чисел).** Учебник. — 511 с.

Книга написана на основе лекций по высшей математике, которые автор читал в Ленинградском политехническом институте, а также его широко известного учебника по теории чисел. Математика воспринимается автором этой книги как мощное орудие для решения проблем естествознания. Поэтому он стремится возможно быстрее дать в руки студентов это орудие, на большом количестве простых, но принципиально важных примеров научить пользоваться ими. При этом он не делает различия, будет ли впоследствии оно применяться в инженерной работе или в абстрактных математических исследованиях. Часть, посвященная теории чисел, является классическим учебником, который стал настольной книгой специалистов в этой области.

Нечаев В. И. **Элементы криптографии (Основы теории защиты информации).** Учебное пособие. — 109 с.

Книга является первым учебным пособием по теории защиты информации, фундаментом которой является прикладная теория чисел. В основу книги положены лекции, читавшиеся автором на математическом факультете Московского педагогического государственного университета. В пособии рассматриваются современные методы шифрования по открытому ключу и электронная подпись. Книга отличается широтой охвата материала в области защиты информации. В ней также содержится интересный исторический очерк развития криптографии. Она доступна широкому кругу читателей, начиная с учащихся старших классов математических школ.

Привалов И. И. **Введение в теорию функций комплексного переменного.** Учебник. — Изд. 14-е. — 432 с.

Книга написана известным автором и является одним из наиболее апробированных и хорошо себя зарекомендовавших учебников по теории функций комплексного переменного. Она отличается строгостью выводов и простотой изложения материала.

Садовничий В. А. **Теория операторов.** Учебник. — Изд. 3-е. — 368 с.

Учебник соответствует программе курсов «Функциональный анализ», «Теория операторов», «Анализ III», которые читаются в университетах и педагогических вузах. В книге приведены основные теоретико-множественные понятия, представлена общая теория метрических, топологических, линейных топологических и нормированных пространств, общая теория меры, измеримых функций и интеграла Лебега. Подробно рассмотрены теория операторов в гильбертовом пространстве, спектральная теория самосопряженных операторов, применения методов теории аналитических функций в спектральной теории несамосопряженных операторов, теория преобразования Фурье и обобщенные функции.

КНИГИ, НАХОДЯЩИЕСЯ В ПЕЧАТИ

Бахвалов Н. С., Лапин А. В., Чижонков Е. В. **Численные методы в задачах и упражнениях.** Учебное пособие. — 26 л.

Учебное пособие содержит элементы теории, примеры решений задач и упражнения для самостоятельной работы. Представленные задачи разбиты по рекомендуемым темам семинарских занятий, а их подбор призван способствовать закреплению материала, излагаемого в теоретическом курсе. Типовые задачи снабжены решениями, которые могут быть использованы студентами для самостоятельного изучения предмета и овладения общими принципами применения вычислительных методов. Ответы и указания помогут преподавателям в выборе полезных и интересных задач в соответствии со спецификой вуза.

Виноградова И. А., Олехник С. Н., Садовничий В. А. **Задачи и упражнения по математическому анализу. В 2-х ч.** Учебное пособие. — 80 л.

Учебное пособие соответствует программе курса математического анализа для студентов механико-математических и математических факультетов университетов, пединститутов и технических вузов. Задачник отражает современное развитие математики и должен заменить известное пособие Б. Д. Демидовича. В отличие от задачника Б. Д. Демидовича большая часть задач в данном пособии приводится с решениями, в связи с чем оно может быть полезно при самостоятельном изучении предмета.

Гашков Г. И., Чубариков В. Н. **Арифметика. Алгоритмы. Сложность вычислений.** Учебное пособие. — 25 л.

Книга состоит из 17 параграфов, тесно связанных целью, поставленной авторами — ознакомить читателя с материалом, который скорейшим путем приводит к современным проблемам теории чисел и теории сложности арифметических алгоритмов. Задачи в каждом параграфе сгруппированы по единству идей и содержания, иногда их объединяют связанные между собой сходные методы решения или просто схожесть формулировки. Часто группы задач заканчиваются красивой и очень трудной задачей. Но если прорешать все задачи подряд, то и она не покажется трудной. Каждый параграф сопровождается указаниями и решениями. В книге имеется достаточное количество и несложных задач.

ISBN 5-06-003654-5