В.В. ВОЕВОДИН

ВЫЧИСЛИТЕЛЬНЫЕ ОСНОВЫ ЛИНЕЙНОЙ АЛГЕБРЫ

В. В. ВОЕВОДИН

ВЫЧИСЛИТЕЛЬНЫЕ ОСНОВЫ ЛИНЕЙНОЙ АЛГЕБРЫ

Допущено Министерством
высшего и среднего специального образования СССР
в качестве учебного пособия
для студентов вузов,
обучающихся по специальности
«Прикладная математика»

ИЗДАТЕЛЬСТВО «НАУКА»
ГЛАВНАЯ РЕДАКЦИЯ
ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ
Москва 1977

518
В 63
УДК 519.95

Вычислительные основы линейной алгебры. В. В. Воеводин. Главная редакция физико-математической литературы изд-ва «Наука», М., 1977.

В книге последовательно изучаются ошибки округленця элементарных арифметических операций, их происхождение, свойства и влияние на вычислительные процессы, впервые рассматриваются вероятностные свойства ошибок округления. Описываются основные численные методы, связанные с решением систем, вычислением определителей, решением полной и частичной проблем собственных значений. Особое внимание уделяется изучению устойчивости и оценкам влияния ошибок округления.

Книга рассчитана на студентов вузов, изучающих прикладную математику, и будет полезна всем лицам, решающим задачи алгебры на ЭВМ.

Библиографических названий 9.

В $\frac{20204-122}{053(02)-77}$ 5-77

ОГЛАВЛЕНИЕ

Глава IV

Прямое разложение матрицы на множители

Глава V

Решение систем линейных алгебраических уравнений

Глава VI

Решение проблемы собственных значений

«Там, где большинству алгорифмистов-любителей кажется, что алгорифм готов для публикации, профессионал понимает, что тяжелая и утомительная работа только начинается»

Дж. Форсайт

ПРЕДИСЛОВИЕ

Сложная ли наука линейная алгебра?

Начало пятидесятых годов. Первые выпуски студентов-вычислителей, перед которыми открывается увлекательный мир решения новых еще никем неизведанных проблем. И вдруг вместо заманчивой перспективы неожиданное предложение — заняться созданием программного обеспечения электронных вычислительных машин (ЭВМ) для решения задач линейной алгебры. Особого восторга оно не вызвало.

Легко понять, почему это произошло. Мы были воспитаны в духе классических курсов, читаемых на математическом факультете. Линейная алгебра была преподнесена нам столь четко и ясно, что не оставалось никаких сомнений в том, что все основные задачи, рассматриваемые этой областью математики, полностью решены.

В самом деле, теория определителей исчерпывающе отвечала на вопрос о том, когда существует решение системы линейных алгебраических уравнений, а правило Крамера указывало его явный вид. Все спектральные задачи сводились в основном к двум задачам — определению корней алгебраического многочлена и решению систем уравнений. Более того, в нашем арсенале были такие «эффективные» численные методы, как метод Гаусса, метод Данилевского и др. Эти методы вроде бы позволяли решать соответствующие задачи линейной алгебры во всей их полноте. Поэтому порученная нам работа первона-

чально воспринималась как чисто механический процесс по переводу огромного количества известных к тому времени вычислительных алгорифмов с общепринятого языка математических формул на язык команд ЭВМ.

Действительность оказалась значительно сложнее. Лишь после многих неудач и ошибок мы стали понимать, что рядом с классической линейной алгеброй не только существует, но и успешно развивается совсем «другая» линейная алгебра, о которой почти ничего не говорилось ни в основных, ни даже в специальных курсах. Эта линейная алгебра была тесно связана со многими областями математики, уходила своими корнями в самые разнообразные приложения, заставляла учитывать особенности ЭВМ и языков программирования, требовала решения новых системных задач и никак не согласовывалась с широко распространенным мнением о всемогуществе ЭВМ. Называлась она «вычислительной», хотя данный термин далеко не полностью отражал содержание этой «другой» линейной алгебры и нередко низводил ее до уровня жонглирования математическими преобразованиями.

Сравнительная простота теории линейной алгебры и кажущаяся эффективность существовавших численных методов долгое время держали нас в своем плену. К сожалению, многие математики и сейчас попадают под их успокоительное обаяние, не замечая всей сложности, которая характерна для задач алгебры.

Причина подобного положения кроется, на наш взгляд, в основах обучения этой науке, в методике преподавания, в содержании обязательных и специальных курсов линейной алгебры, читаемых в вузах. Вычислительная алгебра сделала за последние пятнадцать лет громадный скачок вперед и является одним из самых развитых направлений численного анализа. Содержание же лекций, как правило, слабо отражает достигнутый прогресс и по-прежнему ведется в духе изложения различных фак-

тов типа теорем существования без учета проблем вычислений.

Знакомство с линейной алгеброй в вузе начинается для студентов-вычислителей с первых же лекций. Поэтому от того, что и как читается в теоретической части этого курса, во многом зависит формирование основы будущего восприятия всей вычислительной математики.

Нельзя не признавать красоту и изящество теории, построенной на понятиях линейной зависимости, базиса, определителя и т. п. Но все практические вычисления, связанные с ними, весьма неустойчивы. Поэтому методы исследования, используемые в теоретической части курса, оказываются не очень полезными при непосредственном их применении для конструирования численных методов и часто приводят просто к неправильному пониманию вычислительной стороны дела.

Однако наличие подобных фактов в действительности предоставляет лектору исключительно благоприятную возможность формирования научных взглядов тех студентов, для которых вычислительная математика как предмет должна занять существенное место в образовании. Конечно, реализация этой возможности требует изменения всего теоретического курса линейной алгебры, но выигрыш от такого изменения может быть очень большим. Численные методы алгебры станут естественной частью общего курса, не нужно будет дополнительно тратить драгоценные часы занятий на изложение их основ и, что самое главное, можно будет легко показать студенту громадную практическую значимость курса линейной алгебры во всей ее полноте.

Отсутствие органического единства в методике изложения теоретической и практической частей курса линейной алгебры не позволяет достичь должного эффекта в обучении студентов-вычислителей. В полной мере мы ощутили это на себе в первые годы работы. С тех пор

прошло много лет. Но описанная выше ситуация с удивительным постоянством повторяется снова и снова. К сожалению, и сейчас молодой специалист в области вычислительной математики нередко оказывается совершенно беспомощным, встретившись с необходимостью грамотно решить систему линейных алгебраических уравнений, не говоря уже о проблеме собственных значений.

Все эти причины побудили нас предпринять попытку подготовить ряд взаимосвязанных и построенных на единой основе учебных пособий по линейной алгебре, содержащих необходимый минимум теоретических знаний, без которого невозможно воспринять огромное вычислительное богатство, и отражающих современные проблемы в области конструирования численных методов.

Настоящее учебное пособие представляет собой третью книгу из этой серии после теоретического курса [1] и задачника [4]. Посвящено оно вычислительным основам линейной алгебры.

Основной замысел книги был навеян теми трудностями, с которыми приходится сталкиваться, изучая численные методы линейной алгебры. Уже при первом знакомстве с существующей литературой, например, по библиографическому указателю [8], возникает недоуменный вопрос: «Почему решению в общем-то небольшого числа различных задач линейной алгебры посвящено такое огромное количество работ?» Нельзя на него дать однозначный ответ, ибо это явление обусловлено многими причинами.

Для линейной алгебры характерна исключительная широта ее приложений. Учет конкретных особенностей задачи приводит к появлению новых модификаций численных методов, а желание решить данную задачу как можно лучше значительно увеличивает их число. И вот здесь, на наш взгляд, кроется одна из основных причин обилия публикаций.

Что значит решить задачу лучше? Если задача решается на ЭВМ, то типичной ситуацией для линейной алгебры является использование стандартных программ. По крайней мере, так должно быть. Но для пользователя ЭВМ безразлично, какой из численных методов заложен в основу той или иной стандартной программы. Его интересуют, как правило, лишь три ее характеристики: время счета, объем требуемой памяти ЭВМ и точность.

Относительно легко сравнить методы по первым двум характеристикам. Что же касается точности, то это уже сделать значительно сложнее. Трудно даже ответить на вопрос, как сравнивать методы по точности. Именно этим обстоятельством и объясняется наличие большого числа работ, в которых либо ничего не говорится о точности численных методов, либо доказательство преимущества одних из них сводится к неубедительным эмпирическим доводам и эмоциональным рассуждениям.

Мы уже отмечали обманчивость простоты формулировок задач линейной алгебры. Однако во всей полноте это постигается лишь тогда, когда проводится анализ влияния ошибок округления и возмущения входных данных на точность решения.

Существенный прогресс в исследовании устойчивости численных методов произошел сравнительно недавно и связан с возникновением так называемого обратного анализа ошибок. Основная идея этого анализа заключается в том, что реально вычисленное решение рассматривается как точное для той же задачи, но с возмущенными входными данными. При этом само возмущение выбирается так, чтобы его действие оказалось эквивалентным совокупному влиянию всех ошибок округления.

Обратный анализ предложил идею, но не инструмент для изучения ошибок округления. Даже для самых простых алгорифмов исследование эквивалентных возмущений остается тяжелой и утомительной работой, сопро-

вождающейся выполнением большого числа весьма тонких выкладок. Тем не менее обратный анализ позволил оценить совместное влияние ошибок округления и ошибок входных данных на точность результатов и провести на этой основе сравнение численных методов между собой.

Исследование лучших численных методов линейной алгебры показало удивительную малость их эквивалентных возмущений. Бо́льшая часть этих методов, предназначенных для решения систем уравнений и проблемы собственных значений, представлена в настоящей книге. В ней же описаны и многие вспомогательные алгебраические алгорифмы, позволяющие конструировать новые численные методы.

Возможно, что после знакомства с этой книгой у читателя появится желание использовать для решения своей задачи какой-либо иной метод. Прежде чем реализовать такое желание, стоит выполнить для нового метода столь же тщательный анализ ошибок, какой проделан здесь для всех численных методов.

Что же касается ответа на вопрос, поставленный в начале предисловия, то линейная алгебра действительно простая наука, если оставаться в пределах классических формулировок обязательного курса и не замечать тех сложных проблем, которые в ней существуют. А каково ваше мнение?

В. В. Воеводин

ГЛАВА I

МАТЕМАТИЧЕСКИЕ ОСОБЕННОСТИ МАШИННОЙ АРИФМЕТИКИ

Современная вычислительная техника стала необходимым звеном выполнения самых различных научных исследований. Она позволяет автоматизировать сложнейшие вычислительные процессы и получать достаточно быстро и в нужной форме решение многих задач. Однако за всем этим в действительности скрыто преобразование огромной информации. Это преобразование может быть весьма сложным или совсем простым, но в конечном счете оно всегда сводится к выполнению последовательности простейших операций, описанных системой команд электронной вычислительной машины.

Общение с вычислительной техникой на уровне системы команд не эффективно для подавляющего большинства пользователей ЭВМ. Поэтому оно осуществляется на уровне каких-либо специальных машинно-независимых языков типа алгола, фортрана и других. Такие языки содержат многие математические символы, с помощью которых принято описывать арифметические операции над числовыми данными. Однако это не означает, что арифметические операции на ЭВМ обладают теми же свойствами, что и математические операции.

Машинная арифметика имеет свои характерные особенности. Правильно учитывая их, можно достичь высокой эффективности в решении задач на ЭВМ. Невнимание же к этим особенностям нередко приводит к ошибочным результатам.

§ 1. Позиционные системы счисления

Общий эффект от решения задачи и даже возможность ее решения во многом определяется тем, как в действительности выполняются операции над числами. А это

в свою очередь зависит от принятой системы записи чисел или, как говорят, *системы счисления*.

Наиболее совершенным принципом записи чисел является тот, на котором основана наша десятичная система счисления. Известно, что любое неотрицательное число x может быть представлено в виде степенного ряда

$$x = a_n \cdot 10^n + a_{n-1} \cdot 10^{n-1} + \ldots + a_0 + a_{-1} \cdot 10^{-1} + a_{-2} \cdot 10^{-2} + \ldots$$

где коэффициенты a_i могут принимать значения 0, 1, 2, ..., ..., 9. Перечислив подряд все коэффициенты, указав положение запятой и приписав числу некоторый знак, мы приходим к следующей системе записи:

$$x = \pm a_n a_{n-1} \ldots a_0, \; a_{-1} a_{-2} \ldots$$

Несмотря на кажущуюся простоту, такая система явилась продуктом длительного исторического развития. Известный французский математик и физик Лаплас писал: «Мысль выражать все числа девятью знаками, придавая им, кроме значения по форме, еще значение по месту, настолько проста, что именно из-за этой простоты трудно понять, насколько она удивительна. Как нелегко прийти к этой методике, мы видим на примере величайших гениев греческой учености Архимеда и Аполлония, от которых эта мысль осталась скрытой».

Создание современной цифровой вычислительной техники не связано с какими-либо принципиально другими системами счисления. Запись чисел, с которыми оперирует ЭВМ, основана на той же идее, что и десятичная система. В математическом плане основные изменения невелики и заключаются в следующем.

Зафиксируем некоторое целое положительное число $p > 1$ и целые числа $\alpha_0, \alpha_1, \ldots, \alpha_{p-1}$. Пусть любое неотрицательное число x может быть представлено в виде ряда

$$x = b_n p^n + b_{n-1} p^{n-1} + \ldots + b_0 + b_{-1} p^{-1} + b_{-2} p^{-2} + \ldots, \quad (1.1)$$

где каждый из коэффициентов b_i может принимать одно из значений $\alpha_0, \alpha_1, \ldots, \alpha_{p-1}$. Снова перечислив подряд все коэффициенты, указав положение запятой и приписав числу некоторый знак, мы получим аналогичную запись:

$$x = \pm b_n b_{n-1} \ldots b_0, b_{-1} b_{-2} \ldots \quad (1.2)$$

Описанные системы счисления называются *позиционными*. Их название связано с тем, что роль, которую играет каждое число в записи (1.2), зависит от занимаемой им позиции. Отсчет позиции определяется положением запятой или, что то же самое, положением коэффициента b_0.

В литературе, связанной с вычислительной математикой, слово «позиция» чаще всего заменяется словом «разряд». Нумерация разрядов устанавливается в убывающем порядке подряд слева направо, причем первый разряд слева от запятой имеет нулевой номер. Различаются разряды числа до запятой и разряды после запятой. Число p называется *основанием* системы счисления, числа $\alpha_0, \alpha_1, \ldots, \alpha_{p-1}$ — *базисными*. Если используется система счисления с основанием p, то правую часть (1.2) называют p-ичной дробью. Дробь называется бесконечной, если в ее записи (1.2) имеется бесконечно много ненулевых коэффициентов, и конечной в противном случае. Обычно в записи дроби (1.2) опускаются все первые и последние нулевые коэффициенты. Опускается и запятая, если все коэффициенты после нее являются нулевыми.

Выбор базисных чисел $\alpha_0, \alpha_1, \ldots, \alpha_{p-1}$ определяется в основном требованиями удобства работы с вещественными числами в данной системе счисления. Не видно каких-либо особых преимуществ, которые дало бы использование базисных чисел, превосходящих по модулю основание системы счисления. Поэтому мы будем считать, что

$$|\alpha_k| < p \quad (1.3)$$

для всех k. В современной цифровой вычислительной технике чаще всего используются системы счисления с базисными числами $\alpha_k = k$.

Теорема 1.1. *Если базисные числа образуют совокупность* $0, 1, \ldots, p-1$, *то любое вещественное число может быть представлено в виде* p*-ичной дроби* (1.2).

Доказательство. Покажем, что любое число x может быть представлено в виде ряда (1.1). Очевидно, что достаточно рассмотреть лишь положительные числа x.

Существует целое число n_1 такое, что выполняются соотношения

$$p^{n_1} \leqslant x < p^{n_1+1},$$

в свою очередь зависит от принятой системы записи чисел или, как говорят, *системы счисления*.

Наиболее совершенным принципом записи чисел является тот, на котором основана наша десятичная система счисления. Известно, что любое неотрицательное число x может быть представлено в виде степенного ряда

$$x = a_n \cdot 10^n + a_{n-1} \cdot 10^{n-1} + \ldots + a_0 + a_{-1} \cdot 10^{-1} + a_{-2} \cdot 10^{-2} + \ldots$$

где коэффициенты a_i могут принимать значения 0, 1, 2, ..., ..., 9. Перечислив подряд все коэффициенты, указав положение запятой и приписав числу некоторый знак, мы приходим к следующей системе записи:

$$x = \pm a_n a_{n-1} \ldots a_0, \; a_{-1} a_{-2} \ldots$$

Несмотря на кажущуюся простоту, такая система явилась продуктом длительного исторического развития. Известный французский математик и физик Лаплас писал: «Мысль выражать все числа девятью знаками, придавая им, кроме значения по форме, еще значение по месту, настолько проста, что именно из-за этой простоты трудно понять, насколько она удивительна. Как нелегко прийти к этой методике, мы видим на примере величайших гениев греческой учености Архимеда и Аполлония, от которых эта мысль осталась скрытой».

Создание современной цифровой вычислительной техники не связано с какими-либо принципиально другими системами счисления. Запись чисел, с которыми оперирует ЭВМ, основана на той же идее, что и десятичная система. В математическом плане основные изменения невелики и заключаются в следующем.

Зафиксируем некоторое целое положительное число $p > 1$ и целые числа $\alpha_0, \alpha_1, \ldots, \alpha_{p-1}$. Пусть любое неотрицательное число x может быть представлено в виде ряда

$$x = b_n p^n + b_{n-1} p^{n-1} + \ldots + b_0 + b_{-1} p^{-1} + b_{-2} p^{-2} + \ldots, \quad (1.1)$$

где каждый из коэффициентов b_i может принимать одно из значений $\alpha_0, \alpha_1, \ldots, \alpha_{p-1}$. Снова перечислив подряд все коэффициенты, указав положение запятой и приписав числу некоторый знак, мы получим аналогичную запись:

$$x = \pm b_n b_{n-1} \ldots b_0, b_{-1} b_{-2} \ldots \quad (1.2)$$

Описанные системы счисления называются *позиционными*. Их название связано с тем, что роль, которую играет каждое число в записи (1.2), зависит от занимаемой им позиции. Отсчет позиции определяется положением запятой или, что то же самое, положением коэффициента b_0.

В литературе, связанной с вычислительной математикой, слово «позиция» чаще всего заменяется словом «разряд». Нумерация разрядов устанавливается в убывающем порядке подряд слева направо, причем первый разряд слева от запятой имеет нулевой номер. Различаются разряды числа до запятой и разряды после запятой. Число p называется *основанием* системы счисления, числа $\alpha_0, \alpha_1, \ldots, \alpha_{p-1}$ — *базисными*. Если используется система счисления с основанием p, то правую часть (1.2) называют p-ичной дробью. Дробь называется бесконечной, если в ее записи (1.2) имеется бесконечно много ненулевых коэффициентов, и конечной в противном случае. Обычно в записи дроби (1.2) опускаются все первые и последние нулевые коэффициенты. Опускается и запятая, если все коэффициенты после нее являются нулевыми.

Выбор базисных чисел $\alpha_0, \alpha_1, \ldots, \alpha_{p-1}$ определяется в основном требованиями удобства работы с вещественными числами в данной системе счисления. Не видно каких-либо особых преимуществ, которые дало бы использование базисных чисел, превосходящих по модулю основание системы счисления. Поэтому мы будем считать, что

$$|\alpha_k| < p \quad (1.3)$$

для всех k. В современной цифровой вычислительной технике чаще всего используются системы счисления с базисными числами $\alpha_k = k$.

Теорема 1.1. *Если базисные числа образуют совокупность* $0, 1, \ldots, p-1$, *то любое вещественное число может быть представлено в виде* p*-ичной дроби* (1.2).

Доказательство. Покажем, что любое число x может быть представлено в виде ряда (1.1). Очевидно, что достаточно рассмотреть лишь положительные числа x.

Существует целое число n_1 такое, что выполняются соотношения

$$p^{n_1} \leqslant x < p^{n_1+1}.$$

Из совокупности 1, 2, ..., $p-1$ выбираем наибольшее число b_{n_1}, для которого

$$b_{n_1}p^{n_1} \leqslant x < (b_{n_1}+1)\,p^{n_1}. \tag{1.4}$$

Если

$$x - b_{n_1}p^{n_1} = 0,$$

то ряд (1.1) получен. Предположим поэтому, что

$$x - b_{n_1}p^{n_1} > 0.$$

Находим далее целое число n_2 такое, что

$$p^{n_2} \leqslant x - b_{n_1}p^{n_1} < p^{n_2+1},$$

и затем из совокупности 1, 2, ..., $p-1$ выбираем число b_{n_2}, для которого

$$b_{n_2}p^{n_2} \leqslant x - b_{n_1}p^{n_1} < (b_{n_2}+1)\,p^{n_2}.$$

В силу соотношений (1.4) заключаем, что $n_1 > n_2$. Если

$$x - b_{n_1}p^{n_1} - b_{n_2}p^{n_2} = 0,$$

то получение ряда (1.1) закончено. Поэтому снова рассматриваем случай

$$x - b_{n_1}p^{n_1} - b_{n_2}p^{n_2} > 0.$$

Продолжая этот процесс, получаем последовательность целых чисел $n_1 > n_2 > n_3 > \ldots$ и чисел b_{n_1}, b_{n_2}, b_{n_3}, ..., выбираемых из совокупности 1, 2, ..., $p-1$. При этом, либо при некотором k

$$x - \sum_{i=1}^{k} b_{n_i}p^{n_i} = 0, \tag{1.5}$$

либо для всех k

$$p^{n_{k+1}} \leqslant x - \sum_{i=1}^{k} b_{n_i}p^{n_i} < p^{n_{k+1}+1}. \tag{1.6}$$

В силу полноты пространства вещественных чисел, соотношения (1.5), (1.6) означают, что

$$x = \sum_{i=1}^{\infty} b_{n_i}p^{n_i}.$$

Числа b_{n_1}, b_{n_2}, b_{n_3}, ... образуют последовательность ненулевых коэффициентов искомой p-ичной дроби.

Арифметические операции над числами, заданными в любой позиционной системе счисления, производятся по таким же правилам, что и в десятичной системе. Это объясняется тем, что все операции основаны на правилах выполнения действий над соответствующими полиномами. При этом нужно пользоваться таблицами сложения и умножения не десятичной системы, а системы с основанием p. Для каждой конкретной системы такие таблицы составляются весьма просто.

Позиционные системы счисления широко применяются для представления чисел в современной вычислительной технике. Наиболее часто применяется простейшая из них — *двоичная* система счисления. Использование именно позиционных систем объясняется возможностью реализации в них достаточно простых алгорифмов выполнения арифметических операций над числами.

УПРАЖНЕНИЯ

Всюду предполагается, что в качестве базисных взяты числа 0, 1, ..., $p-1$.

1. Написать p-ичную дробь числа p.
2. Составить таблицы умножения и сложения для двоичной системы счисления.
3. Как по p-ичной дроби числа найти его p^n-ичную дробь, где n — целое положительное число?
4. Будет ли дробь, конечная в системе счисления с одним основанием, конечной во всех других системах?
5. Какие из рациональных чисел могут быть точно представлены конечными p-ичными дробями?
6. Какую часть числа изображают разряды, стоящие слева (справа) от запятой?
7. Указать какой-нибудь алгорифм выполнения операции деления конечных p-ичных дробей.
8. Какой смысл можно приписать выражению $p \ln p$? Для какого p значение этого выражения минимально?

§ 2. Округление чисел

Никакие технические средства не позволяют выполнять арифметические операции над числами, заданными бесконечными дробями. Поэтому замена любого числа конечной дробью является необходимой операцией.

Округлением числа до s разрядов в заданной системе счисления называется операция замены этого числа таким числом, все разряды которого в той же системе счисле-

ния, начиная с $s-1$-го, являются нулевыми. Разность между округленным и округляемым числами называется *ошибкой округления*.

Заметим, что в данном определении ничего не говорится ни о способе выполнения операции округления, ни о том, насколько округленное число близко к округляемому. Это не случайно. В практике конструирования ЭВМ операции округления реализуются самыми различными способами. Единственное, что их объединяет — это малость ошибки округления по крайней мере для большинства чисел. Прежде чем говорить о тех требованиях, которые будут накладываться в дальнейшем на операцию округления, мы проведем некоторые исследования.

Один из простейших способов округления заключается в следующем. Пусть задана p-ичная дробь $x=b_n\ldots b_s b_{s-1}b_{s-2}\ldots$, которую для простоты будем считать неотрицательной. В качестве результата выполнения операции округления числа x до s разрядов берется число $x_s=b_n\ldots b_s$. Основное достоинство данного способа округления — простота реализации. Однако сразу же видны и некоторые недостатки.

Пусть в качестве базисных берутся числа $0, 1, \ldots, p-1$. Тогда для ошибки округления справедливо соотношение

$$|x_s-x|\leqslant p^s.$$

Равенство достигается в том случае, когда во всех разрядах числа x, начиная с $s-1$-го, стоят числа $p-1$. Уже сравнение с общепринятым правилом округления в десятичной системе показывает, что в рассмотренном способе округления оценка ошибки вдвое больше. Но более важным является то, что независимо от своей величины ошибка округления всегда имеет один и тот же знак, противоположный знаку округляемого числа. Это явление нежелательно, так как оно, по-видимому, должно приводить к быстрому накоплению ошибок в вычислительных процессах. В дальнейшем мы неоднократно подтвердим данное предположение.

Хотя описанный способ округления чисел и не является лучшим, тем не менее именно с ним тесно связаны все другие способы округления. В самом деле, как бы ни выполнялась операция округления, ее результатом

будет число, все разряды которого, начиная с $s-1$-го, являются нулевыми. Следовательно, операцию округления всегда можно трактовать как отбрасывание всех разрядов, начиная с $s-1$-го, и последующее добавление или вычитание некоторого числа, кратного p^s. Для того чтобы ошибка округления была малой, необходимо, чтобы было малым и это число.

Числа, имеющие нулевые разряды, начиная с $s-1$-го, образуют на вещественной оси равномерную сетку с шагом p^s. Среди этих чисел есть число x_s^*, наиболее близкое к x. Ясно, что

$$|x_s^*-x|\leqslant\frac{1}{2}p^s. \tag{2.1}$$

Из геометрических соображений следует, что наилучшее приближение x_s^* к x будет единственным, если в соотношении (2.1) имеет место строгое неравенство, и таких приближений будет два, если имеет место равенство. Легко проверить, что

$$x_s^*=\begin{cases} x_s, & \text{если } |x-x_s|<\frac{1}{2}p^s,\\ x_s+p^s, & \text{если } |x-x_s|>\frac{1}{2}p^s,\\ \text{либо } x_s, \text{ либо } x_s+p^s, & \text{если } |x-x_s|=\frac{1}{2}p^s. \end{cases} \tag{2.2}$$

Замена числа x числом x_s^* является операцией округления, причем лучшей во многих отношениях. Однако по сравнению с описанной ранее операцией она имеет два существенных недостатка. Как следует из (2.2), для ее реализации необходимо выполнять операцию сложения примерно в половине случаев. Так как округление числа осуществляется после каждой арифметической операции, его реализация согласно (2.2) фактически приводит к замедлению работы арифметических устройств ЭВМ. Кроме этого, есть некоторая неоднозначность в выполнении данного варианта операции округления, связанная с третьим случаем из (2.2). Мы увидим в дальнейшем, что она не так уж безобидна.

Естественным является желание объединить достоинства обоих способов округления. Покажем, что этого

можно добиться путем использования специальных систем счисления.

До сих пор мы предполагали, что в качестве базисных чисел p-ичной системы счисления используются числа $0, 1, \ldots, p-1$. При этом оказалось, что лучший по точности способ округления чисел в таких системах не является самым простым в реализации и приводит к замедлению выполнения арифметических операций. Рассмотрим теперь p-ичные позиционные системы счисления с другими наборами базисных чисел. Прежде чем исследовать возможность представления чисел в таких системах, посмотрим, чего можно достичь выбором базисных чисел.

Пусть числа задаются в p-ичной системе с базисными числами α_k. Наилучшее приближение x_s^* для x и в этой системе дает оценку ошибки округления на классе вещественных чисел не лучше, чем (2.1). Поэтому, если мы найдем базисные числа такие, чтобы для всех чисел x выполнялось неравенство

$$|x_s - x| \leqslant \frac{1}{2} p^s, \tag{2.3}$$

то эта система счисления во многих отношениях будет наилучшей. Именно, операция округления, заключающаяся в отбрасывании всех разрядов, начиная с $s-1$-го, будет не только самой простой, но и будет иметь наименьшую ошибку округления.

Для выполнения условия (2.3) необходимо и достаточно, чтобы все числа вида

$$0 \ldots 0\, b_{s-1} b_{s-2} \ldots$$

были по модулю не более $(1/2)\, p^s$. Для этого в свою очередь необходимо и достаточно выполнение условия

$$\max_{0 \leqslant k \leqslant p-1} |\alpha_k| \leqslant \frac{p-1}{2}. \tag{2.4}$$

Любая p-ичная система счисления должна иметь p различных базисных чисел. Но неравенство (2.4) имеет p различных целочисленных решений относительно α_k лишь при нечетном p, при этом сами α_k определяются однозначно. Именно,

$$\alpha_k = (1 + 2k - p)/2.$$

Таким образом, если искомая система счисления существует, то она должна иметь нечетное основание и базисные

числа $-(1/2)(p-1), \ldots, +(1/2)(p-1)$. Заметим, что в таких системах не нужно дополнительного изображения для знака числа, так как он учитывается в его цифровой части. Подобные системы счисления называются *сокращенными*.

Теорема 2.1. *Если p — нечетное и базисные числа образуют совокупность $-(1/2)(p-1), \ldots, +(1/2)(p-1)$, то любое вещественное число x может быть представлено в виде ряда* (1.1).

Доказательство. Будем снова считать для определенности, что число x положительное. Доказательство этой теоремы также основано на последовательном определении всех коэффициентов ряда (1.1) и в идейном плане почти не отличается от доказательства теоремы 1.1. Единственное отличие состоит в следующем. При получении ряда (1.1) в теореме 1.1 конечные дроби всегда приближали число x с недостатком. Теперь же коэффициенты ряда (1.1) находятся из условия, чтобы соответствующая конечная дробь наилучшим образом приближала число x, т. е. допускается как приближение с недостатком, так и с избытком. Мы не будем останавливаться на доказательстве более подробно, оставляя проведение его читателю в качестве упражнения.

Среди сокращенных позиционных систем счисления простейшей является *троичная* сокращенная система. Как уже отмечалось, в современной вычислительной технике наиболее широко используется двоичная система. С точки зрения округления чисел этот выбор не является лучшим, так как, например:

Троичная сокращенная сцстема счисления обладает более простым способом наилучшего округления чисел и не приводит к замедлению выполнения арифметических операций.

В дальнейшем мы покажем и другие преимущества сокращенных систем счисления с точки зрения влияния ошибок округления на вычислительный процесс.

УПРАЖНЕНИЯ

Всюду рассматривается сокращенная система счисления.

1. Написать p-ичную дробь числа p.
2. Доказать, что знак числа совпадает со знаком первого разряда.
3. Доказать, что $x_s > x$, если первый из ненулевых отброшенных разрядов отрицательный, и $x_s < x$, если он положительный.
4. Доказать, что число $(1/2)\, p^s$ не может быть задано конечной дробью.

5. Доказать, что при округлении конечных дробей случай неоднозначности в (2.2) не имеет места.

6. Какое множество чисел представляется неединственным образом?

7. Какую часть числа изображают разряды, стоящие слева (справа) от запятой?

§ 3. Фиксированная и плавающая запятая

Запоминание цифровой информации в современных вычислительных машинах основано на использовании достаточно простых однотипных элементов. Каждый из таких элементов представляет собой некоторое физическое устройство, имеющее p устойчивых физических состояний, где $p > 1$. При этом само устройство допускает возможность перевода любого своего состояния в любое другое. Эти элементы называются базисными и служат для моделирования одного числового разряда p-ичной системы счисления.

ЭВМ не может содержать бесконечно много базисных элементов. Поэтому она всегда имеет возможность оперировать лишь с конечным числом конечных p-ичных дробей. Это важный вывод, из которого вытекают все основные особенности машинной арифметики.

Требование унифицированного выполнения арифметических операций над числами приводит к необходимости унифицированного изображения в ЭВМ всех конечных дробей. Будем для простоты рассматривать изображение дробей без знака, считая, что их знак либо учитывается в самой системе счисления, либо изображается каким-то иным способом.

Пусть на изображение каждой дроби отводится одно и то же число τ базисных элементов. Ясно, что на τ элементах можно изображать не более τ разрядов любого числа. Чтобы это изображение можно было снова прочитать, необходимо установить взаимно однозначное соответствие между базисными элементами, отведенными для изображения каждого числа, и положением разрядов в числе относительно запятой. В зависимости от того, является ли это соответствие одним и тем же для всех изображаемых чисел или зависит от самого числа, различают два основных способа представления чисел в ЭВМ, называемых соответственно представлением с фиксированной и плавающей запятой.

Предположим, что каждые τ базисных элементов служат для изображения τ последовательных разрядов чисел, причем положение этих разрядов относительно запятой фиксировано и является одним и тем же для всех дробей. Будем считать, что на изображение разрядов, стоящих слева от запятой, отводится r элементов, где $r \geqslant 0$. Такой способ представления чисел и называется представлением с *фиксированной запятой*.

С помощью этого способа можно точно запоминать любую из конечных p-ичных дробей, имеющих не более r ненулевых разрядов слева от запятой и не более $\tau - r$ ненулевых разрядов справа от запятой. Все эти дроби x лежат в диапазоне

$$-p^r < x < p^r.$$

Один из недостатков представления чисел с фиксированной запятой виден сразу. Если p-ичная дробь много меньше p^r по модулю, то большая часть из отведенных τ базисных элементов изображает старшие нулевые разряды и фактически не используется. Поэтому аппроксимация числа такой дробью связана с большой относительной ошибкой. Однако для чисел, близких к p^r по модулю, для представления старших ненулевых разрядов используются все τ базисных элементов. В этом случае относительная ошибка аппроксимации чисел является минимальной. Абсолютная ошибка представления чисел с фиксированной запятой всегда лежит в одних и тех же пределах независимо от величины чисел.

Представление чисел с *плавающей запятой* заключается в следующем. Всякое ненулевое число x можно записать в виде

$$x = a\ p^b, \tag{3.1}$$

где b — целое число и

$$1/p \leqslant |a| < 1. \tag{3.2}$$

Число a называется *мантиссой* числа x, число b — его *порядком*. Пусть на изображение порядка без знака отводится r базисных элементов, на изображение мантиссы без знака $\tau - r$ элементов. Если теперь порядок и мантисса представлены как дроби с фиксированной запятой, то это и будет представлением числа x с плавающей запятой.

Заметим, что порядок всегда представляется точно, так как он является целым числом. Мантисса же будет представлена точно лишь для тех p-ичных дробей, которые имеют не более $t-r$ ненулевых старших разрядов. Именно эти дроби x из примерного диапазона

$$-p^{p^r} < x < p^{p^r}$$

и могут быть точно представлены описанным способом. Число нуль обычно изображается числом с нулевой мантиссой. Порядок этого числа не определен и в разных ЭВМ его величина может быть различной.

В современных ЭВМ применяются оба способа представления чисел. Выбор того или иного способа зависит от типа решаемых задач. На ЭВМ широкого назначения нередко допускаются обе формы представления чисел.

Операции над числами с фиксированной запятой выполняются быстрее, чем над числами с плавающей запятой. Это связано с тем, что при реализации операций в режиме плавающей запятой, по существу, приходится выполнять все действия с парами чисел с фиксированной запятой. Поэтому при решении тех задач, где положение запятой в числовых данных более или менее определено, использование представления чисел с фиксированной запятой позволяет получить ощутимый выигрыш во времени. К таким задачам относятся, например, финансовые расчеты, задачи количественного учета, многие задачи управления и т. п.

При решении научно-технических задач более удобно представление чисел с плавающей запятой. Это связано с тем, что в таких задачах, как правило, приходится иметь дело с числовыми данными из очень широкого диапазона.

В общем случае фиксированная запятая позволяет представлять числа в ЭВМ с одинаковой абсолютной точностью, плавающая запятая — с одинаковой относительной точностью. Однако заметим, что при одном и том же количестве базисных элементов, отведенных на представление числа, фиксированная запятая позволяет получить для некоторых чисел большую относительную точность, чем плавающая запятая. Это относится в основном к числам, близким к максимальным по модулю.

Умелое использование фиксированной запятой иногда позволяет добиться большей скорости и большей точности решения задачи, чем использование плавающей запятой. Еще большего эффекта можно достичь путем разумного сочетания вычислений с фиксированной и плавающей запятой. Однако мы не будем сколько-нибудь подробно останавливаться на этих вопросах.

Все дальнейшие исследования будут касаться лишь вычислений с плавающей запятой. Это связано с нашей ориентацией на изучение численных методов линейной алгебры. Линейная алгебра находит свое применение в основном в научно-технических задачах, а такие задачи чаще всего решаются на ЭВМ в режиме плавающей запятой.

УПРАЖНЕНИЯ

1. Привести примеры физических устройств, имеющих p устойчивых состояний.
2. Построить различные модели арифметических устройств для реализации операций сложения, вычитания и умножения на основе каких-либо конкретных базисных элементов.
3. На основе арифметических устройств, построенных в предыдущем упражнении, исследовать различие между выполнением операций в режимах фиксированной и плавающей запятой.
4. Предложить какой-либо способ изображения чисел в ЭВМ, отличный от фиксированной и плавающей запятой. Как при этом выполняются операции над числами?
5. Имеет ли место неоднозначность представления каких-либо чисел на конкретных ЭВМ в фиксированной и плавающей запятой?

§ 4. Особенности представления чисел на ЭВМ

Как уже отмечалось, современные вычислительные машины оперируют лишь с конечными p-ичными дробями. Далеко не всегда результат выполнения арифметической операции над конечными дробями является конечной дробью. Примерами могут служить операции деления, извлечения корня и т. п. Но даже если результат и будет конечной дробью, его чаще всего нельзя точно представить в том виде, который принят на ЭВМ. Это относится к форме записи чисел как с фиксированной, так и с плавающей запятой; примером может служить операция умножения чисел. Таким образом, при запоминании почти всех чисел в ЭВМ в сами числа вносится некоторая ошибка, связанная с их округлением. Ясно, что

Появление ошибок округления чисел неизбежно в любой современной ЭВМ. Величина этих ошибок зависит от конкретной реализации операции округления и от принятой формы представления чисел.

Конечно, за счет различных технических решений можно в какой-то мере влиять на ошибки округления. Однако есть некоторые принципиальные ограничения, связанные с их величиной. Эти ограничения нельзя преодолеть никакими техническими средствами, если оставаться в рамках существующих идей представления чисел на ЭВМ.

Обозначим через $\mathrm{fi}(x)$ конечную дробь, которая получается после округления числа x с фиксированной запятой до t-го разряда после запятой. Имеем

$$\mathrm{fi}(x) = x + \nu, \tag{4.1}$$

где ν есть ошибка округления. Каким бы образом ни было реализовано округление, согласно (2.1) на классе вещественных чисел нельзя получить оценки лучшей, чем

$$|\nu| \leqslant \frac{1}{2} p^{-t}. \tag{4.2}$$

Обозначим, далее, через $\mathrm{fl}(x)$ конечную дробь, которая получается после округления мантиссы числа x с плавающей запятой до t-го разряда после запятой. Тождество, аналогичное (4.1), теперь удобнее записать в таком виде:

$$\mathrm{fl}(x) = x(1+\varepsilon). \tag{4.3}$$

Величина ε не является ошибкой округления числа x, хотя связана с ней. Если $\mathrm{fl}(x) \neq 0$, то для мантиссы справедливы оценки (3.2), а для ее ошибки округления не может быть оценки лучшей, чем (4.2). Отсюда вытекает, что на классе вещественных чисел нельзя получить для ε оценки лучшей, чем

$$|\varepsilon| \leqslant \frac{1}{2} p^{-t+1}. \tag{4.4}$$

При больших t правая часть неравенства может быть сделана сколь угодно малой. Однако важно подчеркнуть, что

На любой ЭВМ для некоторого множества чисел будем иметь $\varepsilon = -1$ независимо от числа разрядов, отведенных на представление мантиссы.

В самом деле, сколько бы ни отводилось базисных элементов под изображение порядка чисел, их всегда будет конечное число. Пусть под порядок без знака отводится r элементов. Предположим для определенности, что в выбранной системе счисления базисные числа равны $0, 1, \ldots, p-1$. Тогда на r элементах можно изобразить целые числа, не превосходящие по модулю $p^r - 1$. Следовательно, минимальное положительное число, которое можно изобразить в ЭВМ с плавающей запятой, равно

$$\omega = p^{-p^r}.$$

Итак, с формальной точки зрения почти все ненулевые числа x, принадлежащие интервалу

$$-\omega < x < \omega, \tag{4.5}$$

заведомо нельзя представить в ЭВМ с плавающей запятой, соблюдая соотношение (4.4). Такие числа могут появиться в процессе вычислений. Например, они появляются после перемножения любых двух чисел x, y, удовлетворяющих соотношениям

$$\omega < x, y < \omega^{1/2}.$$

При этом сами числа x, y не принадлежат интервалу (4.5).

Числа из (4.5) необходимо заменять какими-то числами, представимыми в ЭВМ. Выход из создавшегося положения, по существу, единственный. Так как ω является «малым» числом, то все числа x из интервала (4.5) заменяются в ЭВМ нулем. Для этих чисел и получаем $\varepsilon = -1$. Исключение составляет число нуль, для которого можно считать $\varepsilon = 0$. Таким образом, на классе вещественных чисел можно получить лишь оценки следующего вида:

$$\begin{aligned} &\varepsilon \leqslant \frac{1}{2} p^{-t+1}, \quad \text{если } \mathrm{fl}(x) \neq 0, \\ &\varepsilon = -1, \quad \text{если } \mathrm{fl}(x) = 0, \text{ но } x \neq 0. \end{aligned} \tag{4.6}$$

Несмотря на «малость» интервала (4.5), с подобными числами приходится проводить вычисления значительно чаще, чем может показаться на первый взгляд. Как будет установлено в дальнейшем, именно к таким вычислениям приводят многие важнейшие численные методы линейной алгебры. Тот факт, что числа (4.5) нельзя представить

в ЭВМ с приемлемой относительной точностью, заставляет преодолевать немало трудностей при реализации методов на ЭВМ.

Для современных ЭВМ величина p^{-t} обычно находится в пределах $10^{-10}-10^{-15}$, величина ω — в пределах $10^{-15}-10^{-40}$. Вообще говоря, p^{-t} и ω не связаны между собой. Однако на всех ЭВМ, за исключением ЭВМ с переменной длиной мантиссы, выполняется соотношение

$$\omega < p^{-1-t}. \tag{4.7}$$

На ЭВМ, работающей с фиксированной запятой, обычно допускаются числа, не превосходящие по модулю единицы, на ЭВМ с плавающей запятой, — не превосходящие ω^{-1}. Если в процессе вычислений появляются числа, выходящие за эти границы, то в большинстве случаев вычислительный процесс останавливается. Это явление принято называть *переполнением*. Конечно, его надо учитывать при реализации алгорифмов на ЭВМ.

Отмеченные особенности представления чисел не могут быть устранены какими-либо техническими средствами. Можно сконструировать ЭВМ со сколь угодно малым числом ω. Но все равно оно будет отлично от нуля. Можно построить ЭВМ, у которой операция округления будет реализована самым лучшим способом, но оценки (4.2), (4.6) при этом все равно не будут улучшены. Нельзя избежать, по существу, и переполнения.

Чтобы не заниматься излишними деталями, связанными с особенностями представления чисел на конкретных ЭВМ, всюду в дальнейшем будет предполагаться выполнение оценок (4.2), (4.6). Операцию округления с такими оценками будем называть *правильной*.

Правильное округление обладает существенными достоинствами. Тем не менее на многих, если не большинстве, современных ЭВМ по тем или иным причинам округление не реализуется правильно. При этом в оценках типа (4.2), (4.4) справа появляется дополнительный множитель $\alpha > 1$. Если относительно ошибок округления необходимо знать лишь их мажорантные оценки, то подобная реализация операции округления равносильна потере $\log_p \alpha$ разрядов в представлении чисел на ЭВМ, так как

$$\alpha = p^{\log_p \alpha},$$

В действительности последствия неправильной реализации операции округления значительно серьезнее, чем просто потеря некоторого числа разрядов.

УПРАЖНЕНИЯ

1. Почему числа из интервала $(-\omega, \omega)$ заменяются нулем, а не каким-либо другим числом?
2. Можно ли на ЭВМ с плавающей запятой вычислить число ω^{-1}?
3. Построить график величины ν из (4.1) как функции от x.
4. Построить график величины ε из (4.3) как функции от x.
5. Насколько точно правые части неравенств (4.2), (4.4) оценивают левые части тех же неравенств?
6. Привести какие-либо аргументы, подтверждающие необходимость выполнения на ЭВМ соотношения (4.7)?

§ 5. Арифметические операции

В математическом описании большинства вычислительных алгорифмов допускается некоторая неоднозначность, связанная с коммутативностью, ассоциативностью, дистрибутивностью арифметических операций. При реализации на ЭВМ различных форм записи алгорифмов возникают неодинаковые эффекты. Это связано с тем, что арифметические операции на ЭВМ обладают иными свойствами, чем точные операции.

На разных ЭВМ одни и те же арифметические операции могут отличаться в деталях своего выполнения весьма значительно. Однако эти различия для нас не представляют особого интереса, так как математическое значение имеет лишь результат операции. Чтобы не заниматься деталями, связанными с конкретными ЭВМ, мы будем считать в дальнейшем, что:

Результат выполнения любой арифметической операции на ЭВМ совпадает с правильно округленным результатом точного выполнения той же операции; операция с нулевым результатом выполняется точно.

Одной из простейших операций является сложение (вычитание) чисел с фиксированной запятой. Предположим, что складываются (вычитаются) числа в p-ичной системе счисления и единица последнего представленного разряда равна p^{-t}. Если не происходит переполнение, то результат будет конечной p-ичной дробью, имеющей не более t разрядов после запятой и поэтому может быть

представлен в ЭВМ точно. Отсюда вытекает, что операция сложения (вычитания) чисел с фиксированной запятой может быть осуществлена без ошибок округления. В дальнейшем будем считать, что

$$\mathrm{fl}\,(x \pm y) = x \pm y \tag{5.1}$$

для всех чисел x, y, представленных в ЭВМ, если при этом сама операция выполнима.

Все остальные наиболее распространенные арифметические операции над числами с фиксированной запятой не обладают таким свойством. Это связано с тем, что при реализации других операций появляются числа, имеющие более t ненулевых разрядов после запятой. В этом случае ошибки округления неизбежны. Поэтому для всех чисел x, y, представленных в ЭВМ,

$$\mathrm{fl}\left(x \overset{\times}{\div} y\right) = x \overset{\times}{\div} y + \nu, \tag{5.2}$$

где

$$|\nu| \leqslant \frac{1}{2}\, p^{-t}. \tag{5.3}$$

Предполагается, конечно, что сами операции выполнимы.

Выполнение арифметических операций над числами с плавающей запятой приводит к появлению ошибок округления уже почти во всех случаях. Предположим, что числа заданы в p-ичной системе счисления и для представления мантисс отводится t разрядов. Тогда при выполнимости операции будем иметь

$$\mathrm{fl}\left(x \overset{+}{\underset{\div}{\times}} y\right) = \left(x \overset{+}{\underset{\div}{\times}} y\right)(1+\varepsilon), \tag{5.4}$$

где согласно (4.6)

$$\begin{gathered} |\varepsilon| \leqslant \frac{1}{2}\, p^{-t+1}, \quad \text{если } \mathrm{fl}\left(x \overset{+}{\underset{\div}{\times}} y\right) \neq 0, \\ \varepsilon = -1, \quad \text{если } \mathrm{fl}\left(x \overset{+}{\underset{\div}{\times}} y\right) = 0, \text{ но } \left(x \overset{+}{\underset{\div}{\times}} y\right) \neq 0. \end{gathered} \tag{5.5}$$

Формулы (5.5) нельзя существенно улучшить на множестве всех чисел, представленных в ЭВМ. Однако для некоторых практически важных случаев можно гарантировать, что $\varepsilon \neq -1$. Нетрудно проверить, что это будет, например, если складываются числа одинаковых знаков

или вычитаются числа разных знаков, если один из сомножителей в произведении по модулю не меньше единицы или знаменатель дроби по модулю не больше единицы.

Несмотря на то, что в случае операции сложения (вычитания) ошибка округления появляется, она все же имеет некоторую особенность, на которую стоит обратить внимание. Пусть

$$x = a_1 p^{b_1}, \quad y = a_2 p^{b_2},$$

где a_1, a_2 — мантиссы, b_1, b_2 — порядки чисел x, y. Будем считать, для определенности, что $b_1 \geqslant b_2$, и представим $x \pm y$ в таком виде:

$$x \pm y = (a_1 \pm a_2)\, p^{b_2 - b_1} p^{b_1}.$$

Мантиссы a_1, a_2 удовлетворяют условиям (3.2). Поэтому число

$$z = a_1 \pm a_2 p^{b_2 - b_1}$$

имеет не более $t + b_1 - b_2$ ненулевых разрядов после запятой и не более одного ненулевого разряда перед запятой. Этот разряд появляется лишь в том случае, когда $|z| \geqslant 1$. Следовательно, мантиссы чисел $x \pm y$ имеют не более $t + b_1 - b_2 + 1$ ненулевых разрядов после запятой.

В частности, при сложении чисел одного порядка с разными знаками (вычитании чисел одного порядка с одинаковыми знаками) мантисса результата имеет не более t ненулевых разрядов после запятой. Естественно считать, что в этом случае ошибка округления не должна появляться. В общем же случае:

Операция сложения (вычитания) чисел с плавающей запятой имеет ошибку округления, содержащую конечное число ненулевых разрядов. Число этих разрядов определяется величиной чисел, участвующих в операции, и не зависит от числа разрядов, отведенных для представления мантиссы.

В этом отношении операция сложения (вычитания) отличается от всех других операций. Так, например, в операциях деления, извлечения корня и т. п. ошибка округления содержит, как правило, бесконечное число разрядов. В операции умножения ошибка округления хотя и содержит конечное число разрядов, но все же их имеется порядка t.

Отмеченная особенность ошибки округления операции сложения (вычитания) чисел с плавающей запятой оказывает очень большое влияние на общее распределение ошибок округления в вычислительных процессах.

При реализации некоторых алгорифмов возникает необходимость выполнять отдельные промежуточные вычисления с существенно бо́льшей точностью, чем допускается принятой системой представления чисел. На большинстве современных ЭВМ такие вычисления организуются следующим образом. Во-первых, имеется техническая возможность получать результат выполнения основных арифметических операций над t-разрядными числами не с t разрядами, а с $2t$ разрядами. При этом ошибка округления обычно искажает лишь последние из этих $2t$ разрядов. Во-вторых, имеется возможность программного доступа как к старшим t разрядам результата, так и к младшим его t разрядам. Используя эту возможность, можно программным путем осуществлять любые вычисления со сколь угодно большой точностью.

Несмотря на исключительную важность операций с удвоенным числом разрядов, возможность их реализации имеется не на всех ЭВМ. Это следует считать серьезной конструкторской недоработкой. Система команд и арифметические устройства ЭВМ должны быть такими, чтобы операции с удвоенным числом разрядов могли выполняться с наибольшей эффективностью.

В задачах линейной алгебры особое значение имеет максимально точное и быстрое вычисление выражений вида

$$\frac{\alpha+\sum_{i=1}^{n}\alpha_i\beta_i}{\beta}, \tag{5.6}$$

где все написанные числа являются t-разрядными. Если промежуточные вычисления осуществляются с удвоенным числом разрядов, то эти выражения вычисляются, как правило, с такой же относительной точностью, что и результат выполнения одной арифметической операции. Подобный режим вычислений мы будем называть *накоплением* и обозначать соответственно символами fi_2, fl_2. С технической точки зрения его реализация не вызывает принципиальных трудностей. При хорошо продуманной системе команд ЭВМ, относящихся к вычислениям с удвоенной точностью, выражения вида (5.6) в режиме накопления вычисляются почти за то же время, что и в режиме одинарной точности. Будем считать в дальнейшем, что

$$\mathrm{fi}_2\left(\frac{\alpha+\sum_{i=1}^{n}\alpha_i\beta_i}{\beta}\right)=\frac{\alpha+\sum_{i=1}^{n}\alpha_i\beta_i}{\beta}+\nu,$$

$$\mathrm{fl}_2\left(\frac{\alpha+\sum_{i=1}^{n}\alpha_i\beta_i}{\beta}\right)=\left(\frac{\alpha+\sum_{i=1}^{n}\alpha_i\beta_i}{\beta}\right)(1+\varepsilon);$$

при этом

$$|\nu|\leqslant\frac{1}{2}\,p^{t},\quad |\varepsilon|\leqslant\frac{1}{2}\,p^{t+1},\ \text{либо}\ \varepsilon=-1,$$

как и в других аналогичных случаях.

Символы fl, fi и fl_2, fi_2 будут использоваться в дальнейшем довольно часто. Тем не менее им не приписывается строгий математический смысл. В общем случае они будут указывать лишь на применяемый режим вычислений. Точное описание вычислительного процесса и его особенностей мы будем задавать указанием всех возникающих ошибок и их оценок.

В заключение обратим внимание на следующее обстоятельство. Какими бы малыми ни были ошибки округления, возникающие при выполнении арифметических операций, их появление существенно меняет математические свойства самих операций. Точные операции умножения и сложения *являются* коммутативными, ассоциативными и связаны между собой законом дистрибутивности. Операции умножения и сложения на ЭВМ уже *не являются* таковыми.

Коммутативность операций на ЭВМ гарантируется лишь тогда, когда ошибка округления однозначно определяется результатом точного выполнения операций. В частности, коммутативной будет операция сложения чисел с фиксированной запятой, так как ошибка округления здесь вообще отсутствует. По этой же причине данная операция будет и ассоциативной. Любые операции, коммутативные при точном выполнении, *можно сделать* коммутативными и на ЭВМ, если имеет место правильное округление. Как уже отмечалось, такое округление реализовано далеко не на всех современных ЭВМ.

Что же касается законов ассоциативности и дистрибутивности, то они *не выполняются* ни на одной из существующих ЭВМ. Соответствующие соотношения имеют место только для некоторых наборов чисел.

УПРАЖНЕНИЯ

1. Будет ли операция умножения на ЭВМ ассоциативной при каком-либо способе округления?
2. Будет ли операция сложения на ЭВМ с плавающей запятой ассоциативной при каком-либо способе округления?
3. Можно ли выбором способа округления добиться выполнения закона дистрибутивности на ЭВМ?
4. Привести примеры возможных пар алгебраических операций на ЭВМ, для которых выполняются законы коммутативности, ассоциативности и дистрибутивности.
5. Для каких чисел на ЭВМ с плавающей запятой операции сложения и умножения выполняются без ошибок округления?
6. Построить различные модели арифметических устройств для реализации операций сложения, вычитания и умножения с правильным округлением результата.
7. Построить систему команд ЭВМ, удобную для реализации вычислений с удвоенным числом разрядов. Насколько просто вычисляются выражения вида (5.6)?
8. Исследовать на какой-либо конкретной ЭВМ работу арифметического устройства при выполнении операций. Как реализована операция округления на этой ЭВМ?

§ 6. Порядок выполнения операций

Математическое описание процесса решения любой научно-технической задачи сводится в конечном счете к описанию некоторых арифметических выражений, связанных между собой логическими условиями. Современные ЭВМ имеют возможность выполнять лишь достаточно простые операции. Поэтому чтобы указать алгорифм вычисления арифметического выражения, необходимо определить порядок выполнения входящих в него операций, т. е. расставить нужным образом скобки.

С точки зрения точного выполнения операций расстановка скобок обычно не бывает однозначной. Это связано в основном с выполнением законов ассоциативности и дистрибутивности для операций. Однако для арифметических операций на ЭВМ указанные законы уже не имеют места. Следовательно, различные расстановки скобок в одном и том же арифметическом выражении будут теперь приводить к различным результатам.

Итак, *всякая задача при постановке на ЭВМ определяет в действительности целую совокупность вычислительных алгорифмов, отличающихся друг от друга порядком выполнения арифметических операций*. Несмотря на математическую эквивалентность всех этих модификаций в точном смысле, различие в вычислительном эффекте может быть огромным, в особенности с точки зрения численной устойчивости.

Рассмотрим простой, но довольно показательный пример. Пусть вычисляется сумма $z = \sum_{i=1}^{n} \alpha_i$, где все α_i — числа одного знака. Операция сложения коммутативна и ассоциативна. Поэтому точный результат не зависит от того, в каком порядке осуществляется суммирование.

Будем теперь вычислять эту сумму на ЭВМ в режиме плавающей запятой. Прибавим к первому слагаемому второе, к полученной сумме прибавим третье слагаемое и т. д. Согласно (5.4) имеем

$$\begin{aligned} \mathrm{fl}(z) &= (\ldots((\alpha_1+\alpha_2)(1+\varepsilon_1)+\alpha_3)(1+\varepsilon_2)+\ldots+\alpha_n)(1+\varepsilon_{n-1}) = \\ &= (\alpha_1+\alpha_2)\prod_{i=1}^{n-1}(1+\varepsilon_i)+ \\ &+\alpha_3\prod_{i=2}^{n-1}(1+\varepsilon_i)+\ldots+\alpha_n\prod_{i=n-1}^{n-1}(1+\varepsilon_i). \end{aligned} \tag{6.1}$$

Так как числа α_i одного знака, то $\varepsilon_i \neq 1$, поэтому

$$|\varepsilon_i| \leqslant \frac{1}{2}\beta^{-t+1}$$

для всех i. Перепишем теперь формулу (6.1) в таком виде:

$$\mathrm{fl}\left(\sum_{i=1}^{n}\alpha_i\right) = \sum_{i=1}^{n}(1+\mathrm{E}_i). \tag{6.2}$$

На современных ЭВМ величина β^{-t} очень мала. Следовательно, с точностью до $O(\beta^{-2t})$

$$\begin{aligned} |\mathrm{E}_1| &\lesssim \frac{(n-1)}{2}\beta^{-t+1}, \\ |\mathrm{E}_i| &\lesssim \frac{(n+1-i)}{2}\beta^{-t+1}, \quad 2 \leqslant i \leqslant n. \end{aligned} \tag{6.3}$$

Полученные формулы означают следующее. Как вытекает из (6.2), суммирование чисел на ЭВМ в режиме

плавающей запятой эквивалентно точному суммированию возмущенных чисел с относительным возмущением E_i в слагаемом α_i. Относительные возмущения неодинаковы. Согласно (6.3) они максимальны в первых слагаемых и минимальны в последних. Абсолютная ошибка Δ вычисленной суммы равна

$$\Delta = \sum_{i=1}^{n} \alpha_i E_i.$$

Так как оценки для E_i не зависят от α_i, то в общем случае ошибка Δ будет наименьшей, если числа суммировать в порядке возрастания их абсолютных значений, начиная с наименьшего.

Нетрудно понять причину неравноправности слагаемых с точки зрения вносимых в них возмущений. Формально каждое слагаемое участвует в процессе суммирования лишь один раз. Однако в образовании ошибок каждое слагаемое участвует столько раз, сколько раз суммируются какие-либо частичные суммы, зависящие от этого слагаемого.

Чтобы устранить подобное неравноправие слагаемых, поступим следующим образом. Будем осуществлять суммирование по этапам. При этом постараемся добиться того, чтобы на каждом из этапов во все слагаемые вносились относительные возмущения одного порядка.

На первом этапе разобьем слагаемые α_i на пары и сложим каждую из пар. Ясно, что в каждое слагаемое α_i, вошедшее в пары, будет внесено относительное возмущение одного порядка. Если n четное, то каждое из α_i входит в одну из пар, если n нечетное, то одно слагаемое, не вошедшее ни в одну из пар, на первом этапе в суммировании не участвует.

Рассмотрим теперь совокупность чисел, составленную из полученных на первом этапе сумм и того числа, которое в случае нечетного n не суммировалось. С этой совокупностью повторим ту же процедуру, что и на первом этапе со слагаемыми α_i. Повторяя затем процесс аналогичным образом, придем снова к формуле (6.2), но теперь для всех E_i будет иметь место такая оценка:

$$|E_i| \lesssim \frac{(1+\log_2 n)}{2} p^{t+1}.$$

Таким образом, только изменением порядка суммирования чисел можно добиться уменьшения оценки ошибки примерно в $n/\log_2 n$ раз.

Как уже отмечалось, всякая задача при постановке на ЭВМ определяет совокупность вычислительных алгоритмов, отличающихся друг от друга порядком выполнения арифметических операций. Это означает, что точному решению задачи соответствует множество приближенных решений, которые могут быть получены с помощью таких алгоритмов. Среди элементов этого множества могут быть как близкие к точному решению, так и очень далекие от него. Разброс приближенных решений говорит о степени неустойчивости задачи к порядку выполнения операций. Рассмотренный пример суммирования чисел показывает, что даже в простейших задачах этот разброс бывает значительным. В более сложных случаях он может быть таким, что при некоторых порядках выполнения операций решение задачи на ЭВМ даже не может быть получено.

Каким бы ни было множество приближенных решений, среди его элементов заведомо находится элемент, наиболее близкий к точному решению. Найти его или не очень далекий от него элемент — трудная задача, особенно в сложных вычислительных алгоритмах. Тем не менее, о принципиальной возможности оптимизации алгоритмов в отношении точности по порядку выполнения операций забывать не следует, особенно в сложных задачах, где влияние ошибок округления сказывается в наибольшей степени.

УПРАЖНЕНИЯ

1. Может ли множество приближенных решений задачи, определяемое порядком выполнения операций, содержать бесконечно много различных элементов?

2. Зависит ли состав множества приближенных решений задачи, определяемого порядком выполнения операций, от выбранного способа округления чисел?

3. Является ли описанный выше алгоритм суммирования чисел с плавающей запятой по парам оптимальным по точности?

4. Рассмотреть различные способы суммирования элементов прямоугольной матрицы. Сравните полученные оценки ошибок между собой.

5. Рассмотреть различные способы вычисления произведения многих чисел в режиме фиксированной запятой. Сравните полученные оценки ошибок между собой.

§ 7. Запись на машинно-независимых языках

При решении научно-технических задач на современных ЭВМ широко используются алгорифмические языки типа алгола, фортрана. На этих языках имеются многочисленные публикации алгорифмов, создаются библиотечные фонды, организуется обмен информацией по математическому обеспечению ЭВМ.

Отличительной чертой указанных языков является их *машинная независимость*, т. е. отсутствие в них каких-либо ссылок на конкретные особенности ЭВМ. В частности, в них нет никаких ссылок на число разрядов в представлении чисел, не указываются граница переполнения и граница понижения точности вблизи машинного нуля, не описывается способ округления чисел.

На уровне алгорифмических языков описание отличий ЭВМ друг от друга возможно лишь через входные параметры задачи. Однако эти параметры связаны с ЭВМ только смысловым содержанием. Поэтому в любом случае описание задачи на языках типа алгола, фортрана не зависит от конкретных особенностей ЭВМ.

Вычислительные машины не воспринимают запись непосредственно на универсальном языке. Прежде чем начать решение задачи, необходимо осуществить перевод ее записи с алгорифмического языка на язык системы команд конкретной ЭВМ. Следовательно, несмотря на машинную независимость исходного описания задачи, на ее решение в конечном счете влияют все особенности машинной арифметики конкретной ЭВМ, в частности, особенности представления чисел.

Таким образом, *любое машинно-независимое описание задачи на алгорифмическом языке порождает в действительности множество машинно-зависимых описаний и, соответственно, множество приближенных решений той же задачи, определяемое множеством конкретных ЭВМ.* Разброс множества приближенных решений характеризует степень устойчивости данного описания задачи на алгорифмическом языке к частным особенностям ЭВМ.

Ранее отмечалось, что за счет выбора порядка выполнения операций можно изменить вычислительные свойства алгорифма таким сбразом, что влияние ошибок округления будет минимальным. Сказанное в полной мере

относится и к выбору формы записи задачи на алгорифмическом языке. Однако в этом направлении можно пойти несколько дальше. Заменяя отдельные формулы в алгорифме на эквивалентные, можно добиваться и повышения устойчивости к особенностям конкретных ЭВМ. Мы будем интересоваться лишь такими особенностями, как граница переполнения и граница понижения точности вблизи машинного нуля, поскольку их проявление *неизбежно* на каждой ЭВМ.

Рассмотрим некоторые примеры. Пусть в режиме плавающей запятой вычисляется евклидова норма вещественного вектора $x=(x_1, x_2, \ldots, x_n)$. Согласно определению имеем

$$\|x\|_E = \sqrt{\sum_{i=1}^{n} x_i^2}. \tag{7.1}$$

Обозначим через ω минимальное положительное число, представимое на ЭВМ, и предположим, что $|x_i| < \omega^{1/2}$ для всех i. Так как в этом случае

$$\Pi(x_i^2) = 0,$$

то и евклидова норма вектора x окажется равной нулю.

Ясно, что при малых значениях координат вектора x относительная погрешность вычисления его евклидовой нормы по формуле (7.1) будет очень бэльшой. Это явление нельзя считать оправданным, так как норма вектора в данном случае может даже превосходить $\omega^{1/2}$, а числа порядка $\omega^{1/2}$ ни на одной ЭВМ не являются исключительными.

Заменим теперь формулу (7.1), переписав ее в следующем эквивалентном виде:

$$\|x\|_E = \begin{cases} 0, & \alpha = 0, \\ \alpha \sqrt{\sum_{i=1}^{n} (x_i/\alpha)^2}, & \alpha \neq 0, \end{cases} \tag{7.2}$$

где

$$\alpha = \max_{1 \leqslant i \leqslant n} |x_i|.$$

Проводя последовательно все вычисления и учитывая возникающие ошибки, находим

$$y_i = \mathrm{fl}\left(\left(\frac{x_i}{\alpha}\right)^2\right) = \left(\frac{x_i}{\alpha}(1+\varepsilon_i')\right)^2(1+\varepsilon_i''),$$

$$z = \mathrm{fl}\left(\sum_{i=1}^{n} y_i\right) = \sum_{i=1}^{n} y_i(1+\eta_i),$$

$$u = \mathrm{fl}\left(\sqrt{z}\right) = \sqrt{z}\,(1+\varepsilon),$$

$$v = \mathrm{fl}\,(\alpha u) = \alpha u\,(1+\eta)$$

и окончательно будем иметь

$$\mathrm{fl}\,(\|x\|_E) = $$

$$= (1+\eta)(1+\varepsilon)\sqrt{\sum_{i=1}^{n} x_i^2(1+\varepsilon_i')^2(1+\varepsilon_i'')(1+\eta_i)}. \qquad (7.3)$$

Пусть для определенности $\alpha = x_1$. Может случиться, что для некоторых i либо ε_i', либо ε_i'' равно -1. Предположим, что это имеет место для $i = 2, \ldots, k$. Согласно предположению относительно α величину y_1 можно не вычислять, а положить ее равной единице. Следовательно, $\varepsilon_1' = \varepsilon_1'' = 0$. Но тогда

$$\sum_{i=1}^{n} x_i^2(1+\varepsilon_i')^2(1+\varepsilon_i'')(1+\eta_i) = x_1^2\left(1+\eta_1-\sum_{i=2}^{k}\left(\frac{x_i}{x_1}\right)^2\right) +$$

$$+\sum_{i=2}^{k} x_i^2 + \sum_{i=k+1}^{n} x_i^2(1+\varepsilon_i')^2(1+\varepsilon_i'')(1+\eta_i).$$

Тот факт, что либо ε_i', либо ε_i'' равно -1 для $i = 2, \ldots, k$, означает выполнение неравенства $(x_i/x_1)^2 < \omega$, поэтому

$$\sum_{i=2}^{k}(x_i/x_1)^2 < (k-1)\,\omega.$$

Все y_i неотрицательны. Как уже отмечалось, в этом случае, независимо от способа суммирования, $\eta_i \neq -1$. Величина z ограничена снизу единицей и $\sqrt{z}$ можно вычислить с высокой относительной точностью, т. е. $\varepsilon \neq -1$. По этой же причине $\eta \neq -1$. Если вычисления ведутся в p-ичной системе счисления с правильным округлением, то ε и η не превосходят по модулю $(1/2)\,p^{-t+1}$. Таковыми же будут и ε_i', ε_i'' для $i > k$. Для величин η_i заведомо

$$|\eta_i| \leqslant \frac{n-k}{2}\,p^{-t+1}.$$

Если формулу (7.3) переписать следующим образом:

$$\mathrm{fl}\,(\|x\|_E) = \sqrt{\sum_{i=1}^{n}(x_i(1+\mathrm{E}_i))^2}, \qquad (7.4)$$

то теперь для E_i можно получить достаточно простые оценки. С точностью до величин второго порядка малости будем иметь

$$|\mathrm{E}_i| \lesssim \begin{cases} \dfrac{(n-k+4)\,p^{-t+1}+2(k-1)\,\omega}{4}, & i=1, \\ p^{-t+1}, & 2 \leqslant i \leqslant k, \\ \dfrac{n-k+7}{4}\,p^{-t+1}, & i > k. \end{cases}$$

Принимая во внимание, что $2 \leqslant p$, и учитывая (4.7), находим

$$|\mathrm{E}_i| \lesssim \frac{(n+6)}{4}\,p^{-t+1} \qquad (7.5)$$

для всех i, причем эти оценки не зависят от применяемого способа суммирования.

Таким образом, вычисление на ЭВМ евклидовой нормы вектора по формуле (7.2) в режиме плавающей запятой эквивалентно точному вычислению евклидовой нормы возмущенного вектора с относительными возмущениями E_i в координатах x_i. В оценку E_i не входит ω, поэтому формула (7.2) определяет алгорифм, устойчивый к понижению точности вблизи машинного нуля для всех конкретных ЭВМ. Этим свойством не обладает формула (7.1). Из (7.4), (7.5) вытекает, что

$$\mathrm{fl}\,(\|x\|_E) = \|x\|_E\,(1+\mathrm{E}),$$

где для E снова имеет место оценка (7.5). Следовательно, вычисление по формуле (7.2) всегда гарантирует высокую относительную точность результата.

Рассмотренный пример показывает возможность преодоления трудностей, связанных с понижением точности вблизи машинного нуля. В данном примере не возникло

каких-либо серьезных проблем с переполнением, хотя счет по формуле (7.2) более благоприятен и в этом отношении. В общем же случае трудности с переполнением могут возникать и в самых простых алгорифмах.

Пусть перемножаются n чисел $x_1, x_2, \ldots, x_n$, среди которых есть как большие по модулю, чем единица, так и меньшие, чем единица. Не безразлично, в каком порядке перемножать эти числа. Если перемножать их начиная с наименьших по модулю, то частичное произведение может стать меньше ω, т. е. машинным нулем. Но тогда и все произведение будет нулем, независимо от того, чему равно точное произведение. Если же перемножать числа, начиная с наибольшего, то очень быстро может наступить переполнение, несмотря на то, что все произведение является машинно-допустимым числом.

Следующий машинно-независимый алгорифм свободен от обоих отмеченных недостатков, если только само произведение всех чисел $x_1, x_2, \ldots, x_n$ не превосходит верхней границы представления чисел в ЭВМ.

Предположим, что заданные числа упорядочены в порядке неубывания их модулей, т. е. $|x_1| \leqslant |x_2| \leqslant \ldots \leqslant |x_n|$. Возьмем число x_1 и будем его последовательно умножать на $x_n, x_{n-1}, \ldots$ до тех пор, пока частичное произведение впервые не станет по модулю больше единицы. Затем полученное частичное произведение будем последовательно умножать на $x_2, x_3, \ldots$ до тех пор, пока новое частичное произведение не станет впервые по модулю меньше единицы. И так процесс повторяется. Когда среди сомножителей, не вошедших в частичное произведение, останутся только не большие по модулю, чем единица, или только не меньшие, чем единица, все они последовательно умножаются на полученное частичное произведение. На этом вычисление произведения n чисел заканчивается. Очевидно, что

$$\mathrm{fl}\left(\prod_{i=1}^{n} x_i\right) = \prod_{i=1}^{n} x_i (1+\varepsilon_i), \tag{7.6}$$

где $|\varepsilon_i| \leqslant \frac{1}{2} \beta^{-t+1}$.

Снова видим, что вычисление произведения чисел в режиме плавающей запятой эквивалентно точному вычислению произведения возмущенных чисел с относительными возмущениями ε_i, удовлетворяющими (7.6).

Итак, изменяя машинно-независимую форму записи алгорифмов в обоих примерах; удалось достичь устойчивости по отношению к рассмотренным особенностям представления чисел конкретных ЭВМ. Конечно, эти примеры очень просты, однако и они довольно часто встречаются на практике. Очень точное вычисление евклидовой нормы необходимо при реализации решения систем уравнений с помощью ортогональных преобразований, вычисление произведения чисел указанным способом встречается при нахождении определителей методом исключения, при решении систем уравнений некоторыми итерационными методами и т. п. Рассматривая эти примеры и проводя их обсуждение, мы хотим подчеркнуть следующее.

Чтобы машинно-независимые алгорифмические языки стали основой для накопления и обмена информацией по математическому обеспечению ЭВМ, любой алгорифм, записанный на любом из этих языков, должен быть устойчив к особенностям конкретных ЭВМ.

Конечно, преобразование алгорифмов ведет к увеличению времени счета. Однако в целом увеличение не столь велико, как может показаться. К тому же современный уровень развития вычислительной техники уже таков, что некоторое замедление счета при повышении его надежности стало не только оправданным, но и возможным.

Рассмотренные проблемы со всей остротой встали в линейной алгебре. Объясняется это тем, что именно задачи линейной алгебры довольно часто решаются на ЭВМ и входят как составные части во многие другие задачи. Поэтому разработка устойчивых вычислительных алгорифмов в этой области весьма актуальна. Следует отметить еще, что в последнее время в линейной алгебре стали появляться весьма эффективные и очень тонкие численные методы, работающие на грани представления чисел в ЭВМ. Описывать такие методы на машинно-независимых языках особенно трудно. Но об этом мы будем говорить позднее.

УПРАЖНЕНИЯ

Исследуйте на устойчивость к особенностям конкретных ЭВМ различные формы записи следующих формул:

1. Вычисление $\cos x$ через $\sin x$ при малых x.
2. Вычисление $\sin x$ через $\operatorname{tg} x$ при больших x.
3. Вычисление величины z, где $z = (2n)!!/(2n+1)!!$.

4. Вычисление величины v для многочлена $P(x)$, где $v = x - P(x)/P'(x)$.
5. Вычисление среднего геометрического большого количества чисел.
6. Вычисление среднего арифметического большого количества чисел разных знаков.

§ 8. Суммарный эффект влияния ошибок округления.

За исключением редких случаев, ошибка округления появляется в каждой арифметической операции. Поэтому при реализации на ЭВМ сложного вычислительного алгорифма на его окончательный результат будет оказывать влияние очень большое число ошибок округления результатов промежуточных вычислений.

Общий эффект влияния ошибок обычно учитывается следующим образом. Обозначим через A входные данные задачи, через B — результат их обработки по некоторому точному алгорифму φ и запишем, что

$$B = \varphi(A).$$

Предположим, что алгорифм φ включает в себя лишь те операции, которые имеются в списках команд ЭВМ. При реализации этого алгорифма на ЭВМ он будет заменен другим, вообще говоря, «близким» алгорифмом φ_t в силу неизбежных отличий машинной арифметики от точной. Следовательно, вместо B будет получен результат B_t, где

$$B_t = \varphi_t(A).$$

На множестве входных данных и множестве решений задачи могут быть введены операции сложения и вычитания элементов, умножения элемента на число и т. п. В этом случае

$$\Pi = B_t - B$$

есть ошибка вычисления на ЭВМ элемента B. Введя на множестве решений подходящим образом метрику, можно пытаться оценить величину Π, т. е. получить количественную оценку ошибки вычисленного решения задачи. Такой подход к оценке суммарного влияния ошибок округления получил название *прямого анализа ошибок*.

В настоящее время получил широкое распространение и другой подход к оценке влияния ошибок. Во многих задачах реально вычисленное решение B_t можно рассматривать как результат обработки некоторых возмущенных входных данных A_t по точному алгорифму φ, т. е.

$$B_t = \varphi(A_t). \tag{8.1}$$

В этом случае ошибку вычисленного решения характеризует и элемент $E = A_t - A$, который принято называть *эквивалентным возмущением*. Если формулу (8.1) переписать в виде

$$B_t = \varphi(A + E),$$

то реально вычисленное решение B_t задачи φ можно трактовать как точное решение той же задачи, но соответствующее возмущенным входным данным с возмущением E. Этим фактом и определяется название возмущения E как эквивалентного. Количественную оценку влияния ошибок округления можно получить, введя на множестве входных данных подходящим образом метрику и оценивая величину E. Такой подход к оценке суммарного влияния ошибок округления получил название *обратного анализа ошибок*.

По существу, мы уже встречались с обратным анализом. Исследуя общее влияние ошибок округления на сложение и умножение чисел, на вычисление евклидовой нормы вектора, нам удалось показать, что результат реальных вычислений в этих случаях можно трактовать как точное применение соответствующих алгорифмов к возмущенным входным данным. При этом были получены оценки эквивалентных возмущений.

В практических задачах входные данные редко бывают заданы точно. Обычно они получаются из каких-либо измерений либо предварительных расчетов и почти всегда содержат определенные ошибки. Обратный анализ показывает, что влияние ошибок округления при последующих вычислениях равносильно дополнительному внесению ошибок во входные данные. Сравнение величин первоначальных ошибок и эквивалентного возмущения при решении задачи позволяет правильно соизмерять точность входных данных с точностью самих вычислений.

Даже в случае математически точного задания входных данных ошибки в них почти неизбежно появляются за счет округления чисел при вводе в ЭВМ. Это минимальные из возможных ошибок.

Как будет показано в дальнейшем, для многих численных методов линейной алгебры имеет место весьма примечательный факт. Именно, *при правильной реализации эквивалентное возмущение оказывается соизмеримым по величине с ошибками округления входных данных.* Однако заметим, что столь высокая устойчивость методов достигается далеко *не при всякой реализации* и не сразу видно, как следует организовать вычисления, чтобы добиться устойчивости. Мы уже видели это на простом примере вычисления евклидовой нормы вектора.

Значительная часть обратного анализа ошибок в линейной алгебре выполняется по типичной схеме, которую можно показать на следующем примере. Пусть A — прямоугольная матрица, которая преобразуется в процессе реализации численного метода. Предположим, что математический процесс сводится к построению последовательности $A_0 = A, A_1, \ldots, A_N$, где

$$A_i = L_i A_{i-1}, \quad i = 1, 2, \ldots, N,$$

и матрицы L_i невырожденные. Если $L = L_N \ldots L_1$, то

$$A_N = LA. \tag{8.2}$$

Следовательно, матрица A_N получается в результате точного умножения матрицы A на матрицу L.

Реальный вычислительный процесс приводит в общем случае к построению такой последовательности:

$$\bar{A}_i = \bar{L}_i \bar{A}_{i-1} + \mu_{i-1}, \quad i = 1, 2, \ldots, N. \tag{8.3}$$

Здесь $\bar{L}_i$ — матрицы, реально получаемые в процессе вычислений, μ_{i-1} — матрица ошибок от умножения $\bar{A}_{i-1}$ на $\bar{L}_i$. Имеем

$$\bar{A}_N = \bar{L}_N \bar{A}_{N-1} + \mu_{N-1} = \ldots = \bar{L}_N \bar{L}_{N-1} \ldots \bar{L}_1 (A + \bar{L}_1^{-1}\mu_0 + \bar{L}_1^{-1}\bar{L}_2^{-1}\mu_1 + \ldots + \bar{L}_1^{-1}\bar{L}_2^{-1} \ldots \bar{L}_N^{-1}\mu_{N-1}). \tag{8.4}$$

Обозначим $\bar{L} = \bar{L}_N \ldots \bar{L}_1$ и, кроме этого, положим

$$E_N = \bar{L}_1^{-1}\mu_0 + \ldots + \bar{L}_1^{-1}\bar{L}_2^{-1} \ldots \bar{L}_N^{-1}\mu_{N-1}; \tag{8.5}$$

тогда

$$\bar{A}_N = \bar{L}(A + E_N). \tag{8.6}$$

Сравнивая (8.2), (8.6), заключаем, что реально вычисленная матрица $\bar{A}_N$ может рассматриваться как полученная в результате точного умножения возмущенной матрицы $A + E_N$ на матрицу $\bar{L}$; при этом для эквивалентного возмущения E_N имеется явная формула (8.5). Если вычисленные матрицы $\bar{L}_i$ асимптотически близки к унитарным, то

$$\|E_N\| \lesssim \sum_{k=0}^{N-1} \|\mu_k\|$$

для 2-нормы или евклидовой нормы.

В наших исследованиях будет в основном использоваться обратный анализ ошибок, значительно реже — прямой анализ. В отдельных вспомогательных задачах может возникнуть необходимость в использовании своих методик оценки суммарного влияния ошибок округления.

УПРАЖНЕНИЯ

Выполнить прямой и обратный анализ ошибок в упражнениях предыдущего параграфа. Везде ли может быть осуществлен обратный анализ?

ГЛАВА II

ТЕОРИЯ ВОЗМУЩЕНИЙ В ЛИНЕЙНОЙ АЛГЕБРЕ

Прежде чем переходить к детальному изучению численных методов, исследуем влияние малого возмущения входных данных на решение алгебраических задач. Предметом наших рассмотрений будут именно малые возмущения. Это вызвано тем, что суммарный эффект влияния ошибок округления в алгебраических процессах сводится в основном к малым возмущениям. К ним же сводятся и многие численные методы уточнения решений.

Относительно возмущений, возникающих из-за ошибок округления, можно сказать совсем немного. Они не только не являются гладкими функциями каких-либо параметров, связанных с входными данными, но даже не непрерывны.

Наиболее широкое применение в теории возмущений находят евклидова и спектральная нормы, что объясняется их инвариантностью к унитарным преобразованиям.

С целью упрощения записи мы будем довольно часто использовать знаки приближенных равенств и неравенств. Если они связывают какие-либо выражения, то предполагается, что написанные соотношения асимптотически верны с точностью до указанных в тексте членов.

§ 9. Сведение к простым матрицам

Решение систем линейных алгебраических уравнений и проблемы собственных значений являются основными задачами линейной алгебры. Покажем, что с точки зрения теории возмущения их исследование сводится к изучению аналогичных задач с простыми матрицами — диагональными и клеточно-диагональными.

Пусть для матрицы A находится сингулярное разложение. Это означает, что необходимо определить ненулевые векторы x, y и неотрицательные числа ρ, связанные с матрицей A такими соотношениями:

$$\begin{aligned} Ax &= \rho y, \\ A^* y &= \rho x. \end{aligned} \tag{9.1}$$

Известно [1], что векторы x, y, удовлетворяющие (9.1), не только существуют, но и образуют ортонормированные системы. Эти векторы называются соответственно *правыми* и *левыми сингулярными векторами*, числа ρ — *сингулярными числами*.

Предположим, что векторы x образуют столбцы унитарной матрицы X, векторы y — столбцы унитарной матрицы Y, числа ρ — диагональную матрицу P. Будем считать, что соответствующие (9.1) векторы x, y и числа ρ расположены в столбцах матриц с одинаковыми номерами. Тогда уравнения (9.1) эквивалентны двум матричным равенствам:

$$\begin{aligned} AX &= YP, \\ A^* Y &= XP \end{aligned}$$

или одному матричному разложению

$$A = YPX^*. \tag{9.2}$$

Оно и называется сингулярным разложением матрицы A.

Рассмотрим возмущенную матрицу $A + E$ и напишем для нее систему уравнений, аналогичную (9.1). Имеем

$$\begin{aligned} (A + E)\,\hat{x} &= \hat{\rho}\hat{y}, \\ (A + E)^*\,\hat{y} &= \hat{\rho}\hat{x}. \end{aligned}$$

Подставив вместо A ее сингулярное разложение (9.2) и сделав замену

$$X^*\hat{x} = \hat{u}, \quad Y^*\hat{y} = \hat{v}, \tag{9.3}$$

получим, что

$$\begin{aligned} (P + \Omega)\,\hat{u} &= \hat{\rho}\hat{v}, \\ (P + \Omega)^*\,\hat{v} &= \hat{\rho}\hat{u}, \end{aligned}$$

где $\Omega = Y^* E X$. В случае точной матрицы A мы имели бы такую систему:

$$\begin{aligned} Pu &= \rho v, \\ P^* v &= \rho u, \end{aligned}$$

ГЛАВА II

ТЕОРИЯ ВОЗМУЩЕНИЙ В ЛИНЕЙНОЙ АЛГЕБРЕ

Прежде чем переходить к детальному изучению численных методов, исследуем влияние малого возмущения входных данных на решение алгебраических задач. Предметом наших рассмотрений будут именно малые возмущения. Это вызвано тем, что суммарный эффект влияния ошибок округления в алгебраических процессах сводится в основном к малым возмущениям. К ним же сводятся и многие численные методы уточнения решений.

Относительно возмущений, возникающих из-за ошибок округления, можно сказать совсем немного. Они не только не являются гладкими функциями каких-либо параметров, связанных с входными данными, но даже не непрерывны.

Наиболее широкое применение в теории возмущений находят евклидова и спектральная нормы, что объясняется их инвариантностью к унитарным преобразованиям.

С целью упрощения записи мы будем довольно часто использовать знаки приближенных равенств и неравенств. Если они связывают какие-либо выражения, то предполагается, что написанные соотношения асимптотически верны с точностью до указанных в тексте членов.

§ 9. Сведение к простым матрицам

Решение систем линейных алгебраических уравнений и проблемы собственных значений являются основными задачами линейной алгебры. Покажем, что с точки зрения теории возмущения их исследование сводится к изучению аналогичных задач с простыми матрицами — диагональными и клеточно-диагональными.

Пусть для матрицы A находится сингулярное разложение. Это означает, что необходимо определить ненулевые векторы x, y и неотрицательные числа ρ, связанные с матрицей A такими соотношениями:

$$\begin{aligned} Ax &= \rho y, \\ A^* y &= \rho x. \end{aligned} \tag{9.1}$$

Известно [1], что векторы x, y, удовлетворяющие (9.1), не только существуют, но и образуют ортонормированные системы. Эти векторы называются соответственно *правыми* и *левыми сингулярными векторами*, числа ρ — *сингулярными числами*.

Предположим, что векторы x образуют столбцы унитарной матрицы X, векторы y — столбцы унитарной матрицы Y, числа ρ — диагональную матрицу P. Будем считать, что соответствующие (9.1) векторы x, y и числа ρ расположены в столбцах матриц с одинаковыми номерами. Тогда уравнения (9.1) эквивалентны двум матричным равенствам:

$$\begin{aligned} AX &= YP, \\ A^* Y &= XP \end{aligned}$$

или одному матричному разложению

$$A = YPX^*. \tag{9.2}$$

Оно и называется сингулярным разложением матрицы A.

Рассмотрим возмущенную матрицу $A + E$ и напишем для нее систему уравнений, аналогичную (9.1). Имеем

$$\begin{aligned} (A + E)\,\hat{x} &= \hat{\rho}\hat{y}, \\ (A + E)^*\,\hat{y} &= \hat{\rho}\hat{x}. \end{aligned}$$

Подставив вместо A ее сингулярное разложение (9.2) и сделав замену

$$X^*\hat{x} = \hat{u}, \quad Y^*\hat{y} = \hat{v}, \tag{9.3}$$

получим, что

$$\begin{aligned} (P + \Omega)\,\hat{u} &= \hat{\rho}\hat{v}, \\ (P + \Omega)^*\,\hat{v} &= \hat{\rho}\hat{u}, \end{aligned}$$

где $\Omega = Y^*EX$. В случае точной матрицы A мы имели бы такую систему:

$$\begin{aligned} Pu &= \rho v, \\ P^* v &= \rho u, \end{aligned}$$

если, конечно,

$$X^*x = u, \quad Y^*y = v. \tag{9.4}$$

Очевидно, что точные векторы u, v являются единичными. Поэтому есть некоторое основание предполагать, что при достаточно малом возмущении Ω векторы $\hat{u}$, $\hat{v}$ могут быть взяты близкими к единичным. Если используется евклидова норма векторов, то в силу ее инвариантности к унитарным преобразованиям из (9.3), (9.4) вытекает, что

$$\|x - \hat{x}\|_E = \|u - \hat{u}\|_E, \quad \|y - \hat{y}\|_E = \|v - \hat{v}\|_E.$$

Матрица P — диагональная. Следовательно, при изучении влияния возмущения на сингулярное разложение можно ограничиться рассмотрением лишь возмущения диагональной матрицы. Важно отметить, что для евклидовой и спектральной норм величины возмущений матриц A и P одинаковы, т. е. в этом случае

$$\|E\|_{E,2} = \|\Omega\|_{E,2}. \tag{9.5}$$

Сингулярное разложение матрицы позволяет исследовать влияние возмущения на решение системы линейных алгебраических уравнений. Пусть даны точная система

$$Ax = b \tag{9.6}$$

и возмущенная система

$$(A + E)\hat{x} = b + e. \tag{9.7}$$

Система (9.6) может быть совместной либо несовместной. Однако известно [1], что отыскание ее решения или нормального псевдорешения сводится к нахождению наименьшего по длине вектора x, минимизирующего *функционал невязки*

$$\Phi_0(x) = \|Ax - b\|_E^2.$$

Подставив вместо матрицы A ее сингулярное разложение (9.2) и сделав замену

$$X^*x = u, \quad Y^*b = d,$$

мы приходим к задаче минимизации функционала невязки

$$\Phi_0(u) = \|Pu - d\|_E^2$$

с диагональной матрицей P. Евклидовы нормы векторов

x и u совпадают, поэтому решение системы (9.6) однозначно определяется решением системы

$$Pu = d.$$

Аналогичные рассуждения показывают, что возмущенная система (9.7) будет эквивалентна системе

$$(P + \Omega)\hat{u} = d + \omega,$$

если положить

$$X^*\hat{x} = \hat{u}, \quad Y^*b = d,$$
$$Y^*EX = \Omega, \quad Y^*e = \omega.$$

При этом снова имеет место (9.5) и, конечно,

$$\|x - \hat{x}\|_E = \|u - \hat{u}\|_E, \quad \|e\|_E = \|\omega\|_E.$$

Сингулярное разложение позволяет свести к системе с диагональной матрицей не только систему (9.6), но и некоторые другие системы, матрицы которых определенным образом связаны с матрицей A. Рассмотрим, например, системы

$$(A^*A)^2 y = A^*b, \quad (A^*A)(A^*A)^{1/2} z = A^*b. \tag{9.8}$$

Подставив вместо матрицы A ее сингулярное разложение (9.2) и сделав замену

$$X^*y = v, \quad X^*z = w, \quad Y^*b = d,$$

мы приходим к системам

$$(P^*P)^2 v = P^*d, \quad (P^*P)(P^*P)^{1/2} w = P^*d. \tag{9.9}$$

Очевидно, что

$$\|y\|_E = \|v\|_E, \quad \|z\|_E = \|w\|_E.$$

Из диагонального вида матрицы P заключаем, что системы (9.9), а следовательно, и системы (9.8) всегда совместны. Отношение норм их нормальных решений к норме нормального псевдорешения системы (9.6) говорит о степени *согласованности* матрицы A и правой части b в (9.6).

Итак, при исследовании влияния возмущения на решение системы линейных алгебраических уравнений можно ограничиться изучением возмущения системы с диагональной матрицей.

Сингулярное разложение матрицы позволяет упростить исследование влияния возмущения матрицы на ее определитель. Легко проверить, что

$$|\det A - \det(A+\mathrm{E})| = |\det \mathrm{P} - \det(\mathrm{P}+\Omega)|.$$

Рассмотрим теперь задачу определения собственных значений, а также собственных и корневых векторов матрицы. Известно [1], что она связана с решением уравнений вида

$$(A-\lambda E)^p x = 0 \tag{9.10}$$

относительно чисел λ и векторов x при целых положительных p, не превосходящих кратности λ как корня характеристического многочлена. В возмущенном уравнении

$$(A+\mathrm{E}-\hat{\lambda}E)^{\hat{p}}\hat{x} = 0 \tag{9.11}$$

мы допускаем отличие $\hat{p}$ от p, так как могут быть различными размерности циклических подпространств матриц $A+\mathrm{E}$ и A.

Пусть матрица A подобна клеточно-диагональной матрице Λ. В частности, Λ может быть диагональной матрицей, если A имеет простую структуру, или, в общем случае, — канонической матрицей Жордана. Предположим, что преобразование Q приводит A к Λ. Тогда

$$A = Q\Lambda Q^{-1}.$$

Подставив это разложение в (9.10), (9.11) и сделав замену

$$Q^{-1}x = u, \quad Q^{-1}\hat{x} = \hat{u},$$

приходим к таким уравнениям:

$$\begin{aligned}(\Lambda-\lambda E)^p u &= 0,\\ (\Lambda+\Omega-\hat{\lambda}E)^{\hat{p}}\hat{u} &= 0,\end{aligned} \tag{9.12}$$

где

$$x-\hat{x} = Q(u-\hat{u}), \quad \Omega = Q^{-1}\mathrm{E}Q.$$

Сейчас нельзя утверждать, что в общем случае хотя бы для какой-нибудь нормы величины возмущений исходной и приведенной задач будут совпадать. Однако для

нормальной матрицы это снова имеет место. Нормальная матрица имеет полную систему ортонормированных собственных векторов [1], поэтому можно считать, что матрица Q унитарная и

$$\|x-\hat{x}\|_E = \|u-\hat{u}\|_E, \quad \|\mathrm{E}\|_{E,2} = \|\Omega\|_{E,2}.$$

Если матрица A нормальная, то $p=1$. Предположим, что возмущение E таково, что будет нормальной и матрица $A+\mathrm{E}$, тогда $\hat{p}=1$. Следовательно, вместо уравнений (9.12) можно рассматривать уравнения

$$\Lambda u = \lambda u, \quad (\Lambda+\Omega)\hat{u} = \hat{\lambda}\hat{u}.$$

С подобной ситуацией мы заведомо встретимся, изучая влияние эрмитова возмущения эрмитовой матрицы.

Таким образом, исследование основных задач линейной алгебры с точки зрения теории возмущений действительно сводится к исследованию аналогичных задач с простыми матрицами. В дальнейшем мы ограничимся в основном лишь рассмотрением этих случаев.

УПРАЖНЕНИЯ

1. Как связаны между собой сингулярные векторы матриц A, $A+\mathrm{E}$, $\mathrm{P}+\Omega$? Сравнить коэффициенты разложения этих векторов в каком-либо ортонормированном базисе.

2. Пусть решается система (9.6). Разложить правую часть по левым сингулярным векторам матрицы A, решение — по правым сингулярным векторам. Как связаны коэффициенты этих разложений между собой?

3. Как меняется решение системы (9.6) при малом изменении сингулярных чисел матрицы A?

4. Пусть $\|A\|_E = 1$. Доказать, что евклидовы нормы нормальных решений систем (9.8) не меньше евклидовой нормы нормального псевдорешения системы (9.6). В каком случае эти нормы равны?

5. Рассмотрим матрицу A простой структуры, но с кратными собственными значениями. Доказать, что при любой сколь угодно малой величине нормы возмущения E существует матрица $A+\mathrm{E}$, не имеющая простой структуры.

6. Пусть матрица A не имеет простой структуры. Доказать, что при любой сколь угодно малой величине нормы возмущения E существует матрица $A+\mathrm{E}$, которая не только имеет простую структуру, но и все ее собственные значения попарно различны.

7. Как меняется размерность циклических подпространств матрицы $A+\mathrm{E}$ при изменении возмущения E?

8. Можно ли вообще ставить вопрос об исследовании зависимости корневого базиса матрицы $A+\mathrm{E}$ от возмущения E?

§ 10. Невырожденные матрицы

Исследование возмущения невырожденной матрицы тесно связано с матрицами вида $E+H$, где элементы H достаточно малы, E — единичная матрица. Известно [1], что матрица $E+H$ невырожденная, если $\|H\|<1$ для какой-либо нормы. При этом

$$(E+H)^{-1}=E+\sum_{k=1}^{\infty}(-H)^k. \tag{10.1}$$

Пусть A — невырожденная матрица. Рассмотрим возмущенную матрицу $A+E$, где для величины E справедливо неравенство

$$\|E\|<\|A^{-1}\|^{-1}. \tag{10.2}$$

Тогда из (10.1) следуют такие разложения:

$$(A+E)^{-1}=(E+A^{-1}E)^{-1}A^{-1}=A^{-1}(E+EA^{-1})^{-1}=$$

$$=\left(E+\sum_{k=1}^{\infty}(-A^{-1}E)^k\right)A^{-1}=A^{-1}\left(E+\sum_{k=1}^{\infty}(-EA^{-1})^k\right), \tag{10.3}$$

откуда получаем, что

$$\|(A+E)^{-1}-A^{-1}\|\leqslant\|A^{-1}\|\sum_{k=1}^{\infty}\|E\|^k\|A^{-1}\|^k=\frac{\|E\|\|A^{-1}\|^2}{1-\|E\|\|A^{-1}\|}. \tag{10.4}$$

Введем относительные величины возмущений матриц A, A^{-1}. Именно,

$$\delta A=\frac{\|E\|}{\|A\|},\quad \delta A^{-1}=\frac{\|(A+E)^{-1}-A^{-1}\|}{\|A^{-1}\|}. \tag{10.5}$$

В этих обозначениях соотношение (10.4) означает, что

$$\delta A^{-1}\leqslant\frac{\nu_A\delta A}{1-\nu_A\delta A}, \tag{10.6}$$

где

$$\nu_A=\|A^{-1}\|\|A\|. \tag{10.7}$$

Предположим теперь, что решаются точная система линейных алгебраических уравнений

$$Ax=b \tag{10.8}$$

с невырожденной матрицей A и возмущенная система

$$(A+E)\hat{x}=b+e.$$

Если возмущение E удовлетворяет условию (10.2), то матрица $A+E$ будет невырожденной и обе системы имеют единственные решения.

Введем дополнительно к (10.5) относительные величины возмущений векторов x, b, т. е.

$$\delta x=\frac{\|\hat{x}-x\|}{\|x\|},\quad \delta b=\frac{\|e\|}{\|b\|}. \tag{10.9}$$

Ясно, что

$$\hat{x}=(A+E)^{-1}(b+e)=(E+A^{-1}E)^{-1}x+(A+E)^{-1}e,$$

поэтому

$$\hat{x}-x=\sum_{k=1}^{\infty}(-A^{-1}E)^k x+(A+E)^{-1}e,$$

и далее для любых согласованных норм имеем

$$\|\hat{x}-x\|\leqslant\left\|\sum_{k=1}^{\infty}(-A^{-1}E)^k x\right\|+\left\|A^{-1}\left(e+\sum_{k=1}^{\infty}(-EA^{-1})^k e\right)\right\|\leqslant$$

$$\leqslant\left(\sum_{k=1}^{\infty}\|A^{-1}\|^k\|E\|^k\right)\|x\|+\|A^{-1}\|\left(\|e\|+\|e\|\sum_{k=1}^{\infty}\|A^{-1}\|^k\|E\|^k\right)=$$

$$=\frac{\|A^{-1}\|\|E\|\|x\|}{1-\|E\|\|A^{-1}\|}+\frac{\|A^{-1}\|\|e\|}{1-\|E\|\|A^{-1}\|}.$$

Принимая во внимание неравенство $\|b\|\leqslant\|A\|\|x\|$, находим окончательно, что

$$\frac{\|\hat{x}-x\|}{\|x\|}\leqslant\frac{\|A^{-1}\|\|E\|+\|A^{-1}\|\frac{\|e\|}{\|x\|}}{1-\|E\|\|A^{-1}\|}\leqslant$$

$$\leqslant\frac{\|A^{-1}\|\|A\|}{1-\|A^{-1}\|\|A\|\frac{\|E\|}{\|A\|}}\left(\frac{\|E\|}{\|A\|}+\frac{\|e\|}{\|b\|}\right)$$

или, в обозначениях (10.5), (10.7), (10.9),

$$\delta x\leqslant\frac{\nu_A}{1-\nu_A\delta A}(\delta A+\delta b). \tag{10.10}$$

Полученные формулы (10.6), (10.10) дают количественные оценки возмущения обратной матрицы и решения системы линейных алгебраических уравнений при изме-

нении матрицы и правой части системы. Из них вытекает, что в окрестности любой невырожденной матрицы обратная матрица и решение системы являются непрерывными функциями входных данных. При этом соотношение (10.2) определяет окрестность, в которой гарантируется непрерывность по матрице. Непрерывность решения по правой части имеет место всюду.

В заключение остановимся на одном следствии из разложения (10.1). Пусть матрица $E+H$ унитарная. Это будет тогда и только тогда, когда выполняется равенство

$$(E+H)^* = (E+H)^{-1}.$$

Но согласно (10.3) асимптотически при малых H оно эквивалентно соотношению

$$(E+H)^* \cong E-H,$$

откуда следует, что

$$H \cong -H^*.$$

Итак, для того, чтобы матрица $E+H$ асимптотически была унитарной, необходимо и достаточно, чтобы матрица H асимптотически была косоэрмитовой.

УПРАЖНЕНИЯ

1. Доказать, что элементы обратной матрицы и решения системы с невырожденной матрицей являются дифференцируемыми функциями входных данных.

2. Доказать, что для элементов обратной матрицы справедливо соотношение

$$\frac{\partial \{A^{-1}\}_{kl}}{\partial \{A\}_{ij}} = -\{A^{-1}\}_{ki}\{A^{-1}\}_{jl}.$$

3. Доказать, что для элементов решения системы (10.8) имеют место следующие равенства:

$$\frac{\partial \{x\}_l}{\partial \{A\}_{ij}} = -\{A^{-1}\}_{li}\{x\}_j, \quad \frac{\partial \{x\}_l}{\partial \{b\}_k} = \{A^{-1}\}_{lk}.$$

4. На основе полученных выше соотношений вывести формулы для главных членов возмущений обратной матрицы и решения системы. Сравнить эти формулы с (10.6), (10.10).

5. Вывести формулу для главного члена возмущения определителя матрицы $P+\Omega$, где P — невырожденная диагональная матрица.

6. Пусть матрица $E+H$ унитарная и $\|H\|<1$. Доказать, что все собственные значения η матрицы H удовлетворяют уравнению

$$\eta = -\eta(\eta+1).$$

§ 11. Непрерывность корней алгебраического многочлена

Для некоторых величин в линейной алгебре существуют явные формулы, связывающие их с другими величинами. Например, есть формулы, выражающие определитель матрицы через ее элементы, компоненты решения системы линейных уравнений через определители и т. д. Характер зависимости таких величин исследуется относительно просто, по крайней мере в теоретическом плане.

Однако нельзя получить явные формулы, выражающие корни многочлена выше четвертой степени через его коэффициенты. Следовательно, не могут быть непосредственно исследованы зависимости собственных значений и корневых векторов от элементов матрицы. Ввиду важности решения этих вопросов мы проведем сейчас некоторые исследования. Всюду будем предполагать, что многочлены имеют старшие коэффициенты, равные единице.

Рассмотрим произвольный многочлен $P(z)$ степени n с комплексными коэффициентами a_i, где

$$P(z) = z^n + a_{n-1}z^{n-1} + \ldots + a_1 z + a_0.$$

Пусть последовательность многочленов

$$P_s(z) = z^n + a_{n-1,s}z^{n-1} + \ldots + a_{1,s}z + a_{0,s}$$

с комплексными коэффициентами $a_{i,s}$ сходится к $P(z)$, т. е.

$$\lim_{s\to\infty} a_{i,s} = a_i$$

для всех i. Эти соотношения мы будем отождествлять в дальнейшем с равенством

$$\lim_{s\to\infty} P_s(z) = P(z). \tag{11.1}$$

Многочлены $P(z)$ и $P_s(z)$ имеют по n корней, считая каждый из них столько раз, какова его кратность. Но сразу нельзя сказать, как относятся корни многочлена $P_s(z)$ к корням $P(z)$ при больших s.

Лемма 11.1. *Для любого многочлена $P(z)$ степени n и любого комплексного числа z_0 по крайней мере один корень $P(z)$ находится в круге*

$$|z-z_0| \leqslant \sqrt[n]{|P(z_0)|}.$$

Доказательство. Из формул Вьета, связывающих корни многочлена с его коэффициентами, следует, что с точностью до знака произведение всех корней многочлена $P(z)$ равно a_0. Поэтому один из корней заведомо находится в круге

$$|z| \leqslant \sqrt[n]{|a_0|}. \tag{11.2}$$

Разложим далее многочлен $P(z)$ по степеням $z - z_0$. При этом старший коэффициент останется без изменения, а свободный член будет равен $P(z_0)$. Утверждение леммы теперь является следствием неравенства (11.2).

Обозначим через $z_1, z_2, \ldots, z_r$ попарно различные корни многочлена $P(z)$. Согласно лемме 11.1 в каждом из кругов

$$|z - z_i| \leqslant \sqrt[n]{|P_s(z_i)|}, \tag{11.3}$$

где $1 \leqslant i \leqslant r$, находится по крайней мере один корень $z_{i,s}$ многочлена $P_s(z)$. Значение многочлена в любой точке непрерывно зависит от своих коэффициентов. Следовательно, из (11.1) вытекают такие равенства:

$$\lim_{s \to \infty} P_s(z_i) = 0.$$

Для всех достаточно больших s круги (11.3) не имеют общих точек и корни $z_{1,s}, z_{2,s}, \ldots, z_{r,s}$ являются попарно различными. Но тогда

$$\lim_{s \to \infty} z_{i,s} = z_i \tag{11.4}$$

для $1 \leqslant i \leqslant r$. Если многочлен $P(z)$ имеет лишь простые корни, то соотношения (11.4) означают непрерывную зависимость всех его корней от коэффициентов.

Пусть теперь $z_1, z_2, \ldots, z_n$ и $z_{1,s}, z_{2,s}, \ldots, z_{n,s}$ представляют полные наборы корней многочленов $P(z)$ и $P_s(z)$. Среди них могут быть равные, однако мы не предполагаем какой-либо связи между кратностями корней $P(z)$ и $P_s(z)$.

Теорема 11.1. *Корни многочленов $P_s(z)$ можно перенумеровать таким образом, что будут выполняться соотношения*

$$\lim_{s \to \infty} z_{i,s} = z_i \tag{11.5}$$

для $1 \leqslant i \leqslant n$.

Доказательство будем проводить методом индукции. Утверждение теоремы справедливо для многочленов первой и второй степени, в чем можно убедиться, исследуя явные формулы, выражающие корни этих многочленов через коэффициенты. Предположим поэтому, что оно справедливо для многочленов степени не выше $n-1$.

Все корни многочлена $P(z)$ могут быть равны между собой. Однако из (11.4) следует, что среди корней каждого многочлена $P_s(z)$ можно найти такой корень, например, $z_{1,s}$, что соотношение (11.5) будет выполняться для $i = 1$.

Обозначим через $R(z)$, $R_s(z)$ частные от деления $P(z)$, $P_s(z)$ соответственно на $z - z_1$, $z - z_{1,s}$. Пусть

$$R(z) = z^{n-1} + b_{n-2}z^{n-2} + \ldots + b_1 z + b_0,$$
$$R_s(z) = z^{n-1} + b_{n-2,s}z^{n-2} + \ldots + b_{1,s}z + b_{0,s}.$$

Из тождества $P(z) \equiv (z - z_1)R(z)$ находим, что

$$\begin{aligned} b_{n-2} &= a_{n-1} + z_1, \\ b_{n-3} &= a_{n-2} + z_1 b_{n-2}, \\ &\cdots\cdots\cdots\cdots \\ b_0 &= a_1 + z_1 b_1. \end{aligned} \tag{11.6}$$

Аналогично определяются и коэффициенты многочлена $R_s(z)$. Именно,

$$\begin{aligned} b_{n-2,s} &= a_{n-1,s} + z_{1,s}, \\ b_{n-3,s} &= a_{n-2,s} + z_{1,s}b_{n-2,s}, \\ &\cdots\cdots\cdots\cdots \\ b_{0,s} &= a_{1,s} + z_{1,s}b_{1,s}. \end{aligned}$$

Переходя к пределу в правых и левых частях последних соотношений и сравнивая их с (11.6), заключаем, что

$$\lim_{s \to \infty} R_s(z) = R(z).$$

Корнями $R(z)$ являются числа $z_2, \ldots, z_n$, корнями $R_s(z)$ — числа $z_{2,s}, \ldots, z_{n,s}$. Возможность их упорядочивания согласно утверждению теоремы вытекает из индукционного предположения.

Таким образом, *корни алгебраического многочлена являются непрерывными функциями коэффициентов в любой области их изменения.*

Доказанная теорема позволяет утверждать, что при малом возмущении коэффициентов многочлена его корни изменяются мало. Однако на существенную малость этого изменения в общем случае рассчитывать нельзя. Действительно, корни всех многочленов $P(z)$ и $P_s(z)$ ограничены сверху по модулю некоторым числом $\alpha > 1$. Выберем число $\varepsilon > 0$ и пусть для $s > s_0$

$$\max_{0 \leqslant i \leqslant n-1} |a_i - a_{i,s}| \leqslant \varepsilon.$$

Так как

$$P(z) - P_s(z) = \sum_{i=0}^{n-1} (a_i - a_{i,s}) z^i,$$

то

$$|P_s(z_l)| = \prod_{j=1}^{n} |z_l - z_{j,s}| \leqslant \varepsilon \frac{\alpha^n - 1}{\alpha - 1}$$

для всех z_l. Отсюда следует, что существует такой корень z_ν, что при всех достаточно малых ε будем иметь

$$|z_\nu - z_{\nu,s}| \leqslant \left(\varepsilon \frac{\alpha^n - 1}{\alpha - 1}\right)^{1/n}.$$

Порядок зависимости от ε в этом неравенстве достигается. Рассмотрим, например, многочлен $P(z) = z^n$. Он имеет n-кратный корень $z_0 = 0$. Многочлен же $P_s(z) = z^n - \varepsilon_s$ имеет корни $z_{l,s}$, совпадающие с корнями n-й степени из ε_s. Очевидно, что

$$|z_0 - z_{l,s}| = |\varepsilon_s|^{1/n}.$$

Итак, *возмущение коэффициентов многочлена на величины порядка ε может привести к изменению его корней на величины порядка $\varepsilon^{1/n}$*. Это явление связано исключительно с наличием кратных корней.

Снова рассмотрим последовательность многочленов $P_s(z)$, сходящуюся к многочлену $P(z)$. Пусть многочлен $P(z)$ представлен в виде произведения $P(z) = Q(z) R(z)$, где $Q(z)$ и $R(z)$ взаимно простые. Представим каждый из многочленов $P_s(z)$ в виде произведения $P_s(z) = Q_s(z) \times R_s(z)$ таким образом, чтобы выполнялись предельные соотношения

$$\lim_{s \to \infty} Q_s(z) = Q(z),$$
$$\lim_{s \to \infty} R_s(z) = R(z).$$

В этих условиях справедлива

Лемма 11.2. *Скорость сходимости последовательностей многочленов $Q_s(z)$, $R_s(z)$ не меньше, чем скорость сходимости последовательности многочленов $P_s(z)$.*

Доказательство. Пусть $z_1, z_2, \ldots, z_p$ — попарно различные корни многочлена $Q(z)$ и их кратности равны $m_1, m_2, \ldots, m_p$. Так как $Q(z)$ и $R(z)$ взаимно простые, то при всех s, бо́льших некоторого s_0,

$$|R_s(z_j)| \geqslant \delta > 0 \tag{11.7}$$

для $1 \leqslant j \leqslant p$. Далее имеем

$$Q_s(z) R_s(z) \equiv Q(z) R(z) + \varepsilon_s(z), \tag{11.8}$$

где $\varepsilon_s(z)$ есть некоторый многочлен степени не выше $n-1$. Условие сходимости последовательности многочленов $P_s(z)$ к $P(z)$ означает, что

$$\lim_{s \to \infty} \varepsilon_s(z) = 0.$$

При этом скорость сходимости определяется скоростью убывания коэффициентов $\varepsilon_s(z)$.

Выберем произвольное число $\varepsilon > 0$. Найдется такое число $s_1 > s_0$, что для $s > s_1$ все коэффициенты многочлена $\varepsilon_s(z)$ будут по модулю меньше ε. Дифференцируя тождество (11.8) и учитывая (11.7), легко получить, что при $s > s_1$ выполняются неравенства

$$|Q_s^{(h)}(z_j)| < N\varepsilon \tag{11.9}$$

для всех корней z_j и $0 \leqslant h < m_j$. Здесь число N не зависит от ε.

Последовательность многочленов $Q_s(z)$ сходится к $Q(z)$. Скорость сходимости определяется скоростью убывания коэффициентов многочлена $\tau_s(z) = Q_s(z) - Q(z)$. Этот многочлен имеет степень не выше $m_1 + \ldots + m_p - 1$, причем в точках z_j можно оценить $m_1 + \ldots + m_p$ его значений и значений его производных, так как

$$\tau_s^{(h)}(z_j) = Q_s^{(h)}(z_j). \tag{11.10}$$

Будем трактовать соотношения (11.10) как систему линейных алгебраических уравнений относительно коэффициентов многочлена $\tau_s(z)$. Согласно (11.9) правые части системы являются величинами порядка ε, матрица

системы полностью определяется корнями z_j. Следовательно, существует такое число M, не зависящее от ε, что все коэффициенты многочлена $\tau_s(z)$ будут по модулю меньше $M\varepsilon$ для $s > s_1$.

Таким образом, последовательность многочленов $Q_s(z)$ сходится с такой же скоростью, что и последовательность многочленов $P_s(z)$. Аналогичное утверждение справедливо, конечно, и для последовательности многочленов $R_s(z)$.

Следствие. *Если коэффициенты многочлена возмущаются на величины порядка ε, то любой его корень кратности m может измениться на величину порядка $\varepsilon^{1/m}$. Все простые корни меняются на величины порядка ε.*

УПРАЖНЕНИЯ

1. Пусть последовательность многочленов $P_s(z)$ сходится к $P(z)$. Предположим, что z_l есть простой корень $P(z)$ и последовательность корней $z_{l,s}$ многочленов $P_s(z)$ сходится к z_l. Обозначим через $R(z)$ частное от деления $P(z)$ на $z-z_l$. Доказать, что при больших s имеет место асимптотическое равенство

$$z_{l,s}-z_l \cong (P_s(z_l)-P(z_l))/R(z_l).$$

2. Доказать, что простые корни являются дифференцируемыми функциями коэффициентов многочлена; при этом

$$\frac{\partial z_l}{\partial a_p} = -\frac{z_l^p}{R(z_l)}.$$

3. Доказать, что собственные значения матрицы являются непрерывными функциями ее элементов.

4. Доказать, что простые собственные значения и соответствующие им собственные векторы являются дифференцируемыми функциями элементов матрицы.

5. Пусть λ — простое собственное значение матрицы A. Доказать, что

$$\frac{\partial \lambda}{\partial \{A\}_{ij}} = \frac{\{x\}_j \{y\}_i}{(x, y)},$$

где x, y являются собственными векторами матриц A, A^*, соответствующими собственным значениям λ, $\bar{\lambda}$.

6. Пусть, в дополнение к условиям предыдущего упражнения, z является собственным вектором матрицы A^* и соответствует собственному значению μ, где $\mu \neq \lambda$. Доказать, что

$$\sum_k \frac{\partial \{x\}_k}{\partial \{A\}_{ij}} \{z\}_k = \frac{\{x\}_j \{z\}_i}{\bar{\lambda}-\bar{\mu}}.$$

7. Пусть $P(z)$ — вещественный многочлен, имеющий простой вещественный корень z_0. Доказать, что при достаточно малом вещественном возмущении коэффициентов многочлена корень, ближайший к z_0, будет вещественным.

§ 12. Локализация собственных значений

Различные задачи линейной алгебры связаны с собственными значениями матрицы. Исследование таких задач нередко приводит к необходимости локализовать собственные значения, т. е. определить те области комплексной плоскости, в которых они находятся. Конечно, локализация собственных значений по элементам матрицы должна осуществляться достаточно простыми средствами. Во всяком случае эти средства должны быть существенно проще, чем численные методы определения собственных значений.

В курсе линейной алгебры [1] доказывается ряд утверждений, с помощью которых можно решить некоторые задачи локализации.

Пусть исследуются собственные значения $\lambda_1, \lambda_2, \ldots, \lambda_n$ матрицы A порядка n с комплексными элементами a_{ij}. Согласно (11.2) по крайней мере одно собственное значение находится в круге $|\lambda| \leqslant |\det A|^{1/n}$. Используя неравенство Адамара [1] для определителей матриц A и A^*, заключаем, что по крайней мере одно собственное значение находится в каждом из кругов

$$|\lambda| \leqslant \max_{1\leqslant i\leqslant n} \left(\sum_{j=1}^{n} |a_{ij}|^2\right)^{1/2}, \quad |\lambda| \leqslant \max_{1\leqslant j\leqslant n} \left(\sum_{i=1}^{n} |a_{ij}|^2\right)^{1/2}$$

и, следовательно, в меньшем из этих кругов.

Некоторые неравенства получаются с помощью матричных норм. Известно [1], что все собственные значения матрицы A находятся в каждом из кругов

$$|\lambda| \leqslant \|A\|, \quad |\lambda| \leqslant \|A^*\|$$

для любой согласованной нормы. Для 1- или ∞-нормы эти неравенства принимают такой вид:

$$|\lambda| \leqslant \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |a_{ij}|, \quad |\lambda| \leqslant \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{ij}|.$$

Евклидова норма дает слабую оценку, так как в действительности [1] для собственных значений матрицы A имеет место неравенство

$$\sum_{i=1}^{n} |\lambda_i|^2 \leqslant \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|^2.$$

Обозначим через ρ_1 и ρ_n соответственно максимальное и минимальное сингулярные числа матрицы A. Спектральная норма матрицы равна ρ_1. Поэтому все собственные значения матрицы A находятся в круге $|\lambda| \leqslant \rho_1$. Аналогичное рассуждение для обратной матрицы приводит к неравенству $\rho_n \leqslant |\lambda|$. В силу непрерывной зависимости собственных и сингулярных чисел от элементов матрицы это неравенство справедливо и для вырожденной матрицы A. Итак, все собственные значения матрицы A находятся в кольце $\rho_n \leqslant |\lambda| \leqslant \rho_1$.

Довольно общий принцип построения областей, локализующих собственные значения, основан на следующей идее. Пусть A — произвольная матрица и $B(A)$ — некоторое арифметическое условие, выполнение которого достаточно для невырожденности матрицы A. Если λ является собственным значением, то матрица $A - \lambda E$ вырожденная. Поэтому для того, чтобы λ было собственным значением матрицы A, необходимо невыполнение условия $B(A - \lambda E)$. Это и определяет некоторую область, в которой должны находиться все собственные значения.

Лемма 12.1. *Для того чтобы матрица A была невырожденной, достаточно выполнение неравенств*

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|$$

для $i = 1, 2, \ldots, n$.

Доказательство. Предположим, что матрица A вырожденная. Тогда однородная система линейных алгебраических уравнений

$$\sum_{j=1}^{n} a_{ij} x_j = 0, \qquad i = 1, 2, \ldots, n,$$

имеет ненулевое решение. Пусть x_k — наибольшая по модулю координата этого решения. Запишем k-е уравнение системы в таком виде:

$$a_{kk} x_k = -\sum_{j \neq k} a_{kj} x_j,$$

откуда следует, что

$$|a_{kk}||x_k| \leqslant \sum_{j \neq k} |a_{kj}||x_j|,$$

и, окончательно,

$$|a_{kk}| \leqslant \sum_{j \neq k} |a_{kj}| \left|\frac{x_j}{x_k}\right| \leqslant \sum_{j \neq k} |a_{kj}|.$$

Это соотношение противоречит условиям леммы.

Следствие. *Для того чтобы λ было собственным значением матрицы A, необходимо выполнение неравенства*

$$|\lambda - a_{ii}| \leqslant \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \tag{12.1}$$

хотя бы для одного значения i, где $1 \leqslant i \leqslant n$, или, другими словами:

Любое собственное значение матрицы A лежит по крайней мере в одном из кругов с центрами a_{ii} и радиусами $\sum_{j \neq i} |a_{ij}|$, где $1 \leqslant i \leqslant n$.

Области (12.1) называются *кругами Гершгорина*. Они широко используются в самых различных исследованиях, связанных с собственными значениями. Покажем, что имеет место

Теорема 12.1. *Если s кругов Гершгорина образуют область G, изолированную от остальных кругов, то в G находится ровно s собственных значений матрицы A.*

Доказательство основано на непрерывной зависимости собственных значений матрицы от ее элементов. Представим матрицу A в виде суммы $A = B + C$, где B — диагональная матрица с элементами a_{ii}, C — матрица с нулевой диагональю. Рассмотрим теперь семейство матриц $A_\varepsilon = B + \varepsilon C$, где $0 \leqslant \varepsilon \leqslant 1$. Сравнивая круги Гершгорина с одинаковыми центрами a_{ii} для матриц A и A_ε, замечаем, что их радиусы отличаются множителем ε.

Обозначим через G_ε замкнутую область, составленную из кругов Гершгорина для матрицы A_ε, центры которых

принадлежат G. Через F_ε обозначим замкнутую область, составленную из остальных кругов. Ясно, что

$$G_\varepsilon \subset G, \quad G_\varepsilon \cap F_\varepsilon = 0 \tag{12.2}$$

для $0 \leqslant \varepsilon \leqslant 1$. При $\varepsilon = 0$ область G_0 содержит ровно s собственных значений матрицы A_0. Эти собственные значения будут непрерывно меняться при изменении ε. Предположим, что при некотором ε одно из них вышло из области G_ε. Тогда в силу непрерывности и второго условия из (12.2) найдется такое ε, при котором одно из собственных значений матрицы A_ε не будет принадлежать ни G_ε, ни F_ε. Это невозможно, поэтому при всех допустимых значениях ε область G_ε содержит ровно s собственных значений матрицы A_ε. Но при $\varepsilon = 1$ область G_ε совпадает с G, а матрица A_ε — с матрицей A.

Следствие. *Если какой-либо круг Гершгорина изолирован, то он содержит точно одно собственное значение.*

Следствие. *Если при некотором i для всех $k \neq i$ выполняются неравенства*

$$|a_{kk} - a_{ii}| > \sum_{j \neq k} |a_{kj}| + \sum_{j \neq k} |a_{ij}|, \tag{12.3}$$

то круг Гершгорина

$$|\lambda - a_{ii}| \leqslant \sum_{j \neq k} |a_{ij}| \tag{12.4}$$

содержит точно одно собственное значение.

Для доказательства этого утверждения достаточно заметить, что выполнение условий (12.3) гарантирует изолированность круга (12.4) от остальных кругов.

Как уже отмечалось, локализация собственных значений должна осуществляться достаточно простыми средствами. Но круги Гершгорина определяются столь просто, что их можно явно написать и исследовать для любой из матриц вида

$$B = DAD^{-1}, \tag{12.5}$$

где D — диагональная матрица. Если $d_1, d_2, \ldots, d_n$ — элементы матрицы D, то любое собственное значение λ матрицы A будет находиться в одном из кругов

$$|\lambda - a_{ii}| \leqslant \sum_{j \neq i} |a_{ij}| \left| \frac{d_i}{d_j} \right|.$$

Выбирая подходящим образом матрицу D, можно изменять радиусы кругов Гершгорина, делать отдельные круги или группы кругов изолированными и т. д. Использование преобразования (12.5) позволяет во многих случаях существенно точнее локализовать собственные значения матрицы A.

В исследованиях, связанных с кругами Гершгорина, всюду рассматривалась матрица A. Однако аналогичные утверждения справедливы и для матрицы A'. При этом области (12.1) заменяются на такие:

$$|\lambda - a_{ii}| \leqslant \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ji}|. \tag{12.6}$$

Они также называются кругами Гершгорина.

УПРАЖНЕНИЯ

1. Можно ли утверждать, что все собственные значения матрицы A лежат в любых n кругах из $2n$ кругов (12.1), (12.6)?

2. Пусть элемент a_{ii} и коэффициенты характеристического многочлена матрицы A вещественные. Доказать, что при выполнении условий (12.3) собственное значение, расположенное в круге (12.4), — вещественное.

3. Предположим, что матрица A перестановкой строк и столбцов не может быть приведена к клеточно-треугольному виду. Доказать, что все ее собственные значения лежат внутри объединения кругов Гершгорина, за исключением того случая, когда собственное значение является общей граничной точкой всех n кругов.

4. Пусть λ — собственное значение матрицы A и дефект матрицы $A - \lambda E$ равен m. Доказать, что λ лежит по крайней мере в m кругах Гершгорина.

5. Доказать, что каждое собственное значение матрицы A лежит по крайней мере в одной из областей

$$|\lambda - a_{ii}|\,|\lambda - a_{jj}| \leqslant \sum_{s \neq i} |a_{is}| \sum_{s \neq j} |a_{js}|,$$

где $1 \leqslant i, j \leqslant n$ и $i \neq j$. Эти области называются овалами Кассини.

6. Используя неравенство Гёльдера [1], доказать, что каждое собственное значение матрицы A лежит по крайней мере в одном из кругов

$$|\lambda - a_{ii}| \leqslant \left(\sum_{j \neq i} |a_{ij}| \right)^{\alpha} \left(\sum_{j \neq i} |a_{ji}| \right)^{1-\alpha},$$

где $0 \leqslant \alpha \leqslant 1$.

7. Пусть Λ — диагональная матрица. Исследовать круги Гершгорина для матрицы $\Lambda + \Omega$. Как они меняются при уменьшении элементов матрицы Ω?

8. Предположим, что элементы матрицы Ω являются величинами порядка ω. Рассмотрим простое собственное значение λ матрицы Λ. Используя преобразование (12.5), доказать, что соответствующий диагональный элемент матрицы $\Lambda+\Omega$ отличается от ее собственного значения на величину порядка ω^2.

9. Что дает использование преобразования (12.5) для исследования случая кратного собственного значения λ матрицы Λ?

10. Используя преобразование (12.5), локализовать собственные значения матрицы $\Lambda+\Omega$, где Λ — каноническая матрица Жордана.

11. Сравнить величины возмущений собственных значений, полученные в упражнениях 7—10, с величинами возмущений корней алгебраического многочлена. В чем причина их различий даже по порядку малости?

§ 13. Клеточно-диагональные матрицы

Исследование клеточно-диагональных матриц связано в основном с полной проблемой собственных значений. Известно [1], что любая квадратная матрица подобна клеточно-диагональной матрице, у которой собственные значения различных клеток различные. В частности, такой клеточно-диагональной матрицей является каноническая матрица Жордана. Как уже отмечалось, исследование возмущения матриц общего вида сводится к изучению возмущения клеточно-диагональных матриц.

Пусть даны матрицы A, B, C размеров соответственно $n\times n$, $m\times m$, $n\times m$. Рассмотрим матричное уравнение

$$AZ-ZB=C, \tag{13.1}$$

где Z — искомая матрица размеров $n\times m$. Приравнивая элементы правой и левой частей этого уравнения, заключаем, что оно эквивалентно системе из nm линейных алгебраических уравнений относительно nm элементов матрицы Z.

Теорема 13.1. *Уравнение* (13.1) *имеет единственное решение тогда и только тогда, когда матрицы A и B не имеют общих собственных значений.*

Доказательство. Для того чтобы уравнение (13.1) имело единственное решение, необходимо и достаточно, чтобы однородное уравнение

$$AZ-ZB=0 \tag{13.2}$$

имело лишь нулевое решение. Поэтому, не уменьшая общности, можно ограничиться доказательством теоремы для уравнения (13.2).

Необходимость. Пусть уравнение (13.2) имеет только нулевое решение. Предположим, что при этом λ является общим собственным значением матриц A, B. Обозначим через x, y собственные векторы матриц A, B', соответствующие λ, и рассмотрим матрицу $Z_0=xy'$ ранга единица. Очевидно, что $Z_0\neq 0$, но

$$AZ_0-Z_0B=(Ax)\,y'-x\,(B'y)'=(\lambda x)\,y'-x\,(\lambda y)'=0.$$

Полученное противоречие доказывает, что матрицы A и B не могут иметь общих собственных значений.

Достаточность. Пусть матрицы A и B не имеют общих собственных значений, но уравнение (13.2) имеет ненулевое решение Z_0. Обозначим через r ранг матрицы Z_0. Ясно, что $r\geqslant 1$. Матрица Z_0 эквивалентна [1] матрице E_r, где E_r — диагональная матрица, у которой первые r диагональных элементов равны единице, а остальные — нулю. Следовательно, существуют такие невырожденные матрицы R, S, что

$$Z_0=RE_rS. \tag{13.3}$$

Подставив теперь Z_0 из (13.3) в уравнение (13.2), получаем, что $(R^{-1}AR)E_r=E_r(SBS^{-1})$. Сравнение элементов правой и левой частей этого соотношения показывает, что матрицы $R^{-1}AR$ и SBS^{-1} являются клеточно-треугольными, причем диагональные клетки, стоящие в левом верхнем углу, равны и имеют порядок r. Поэтому характеристические многочлены матриц $R^{-1}AR$ и SBS^{-1} или, что то же самое, матриц A и B имеют общий делитель степени r. Это противоречит условию, что матрицы A и B не имеют общих собственных значений. Следовательно, уравнение (13.2) не может иметь ненулевого решения.

Рассмотрим клеточно-диагональную матрицу Λ, клетки $\Lambda_1, \Lambda_2, \ldots, \Lambda_r$ которой не имеют общих собственных значений. Пусть $\Lambda+\Omega$ — возмущенная матрица. Разобьем матрицу Ω на прямоугольные клетки так, чтобы ее диагональные клетки имели те же размеры, что и соответствующие клетки матрицы Λ. Обозначим

$$\Lambda=\begin{bmatrix}\Lambda_1 & & 0\\ & \Lambda_2 & \\ & & \ddots\\ 0 & & \Lambda_r\end{bmatrix},\qquad \Omega=\begin{bmatrix}\Omega_{11} & \Omega_{12} & \ldots & \Omega_{1r}\\ \Omega_{21} & \Omega_{22} & \ldots & \Omega_{2r}\\ \ldots & \ldots & \ldots & \ldots\\ \Omega_{r1} & \Omega_{r2} & \ldots & \Omega_{rr}\end{bmatrix}.$$

Будем приводить матрицу $\Lambda+\Omega$ подобным преобразованием к клеточно-диагональному виду. Это означает, что нужно найти невырожденную матрицу $\hat{X}$ и клеточно-диагональную матрицу $\hat{\Lambda}$, для которых

$$\hat{X}^{-1}(\Lambda+\Omega)\hat{X}=\hat{\Lambda}.$$

Конечно, диагональные клетки матрицы $\hat{\Lambda}$ должны иметь те же размеры, что и клетки матрицы Λ.

Если Ω — нулевая матрица, то $\hat{X}$ — единичная. Поэтому при малых Ω будем искать матрицу $\hat{X}$ в виде $\hat{X}=E+\mathrm{H}$, где H — малая матрица. Разобьем матрицу H на клетки H_{ij} аналогично Ω. Принимая во внимание (10.1), находим

$$\hat{X}^{-1}(\Lambda+\Omega)\hat{X}\cong(E-\mathrm{H})(\Lambda+\Omega)(E+\mathrm{H})\cong$$
$$\cong\Lambda+\Omega-\mathrm{H}\Lambda+\Lambda\mathrm{H}.$$

Подберем теперь H так, чтобы с точностью до малых второго порядка правая часть полученного соотношения была бы клеточно-диагональной матрицей, аналогичной Λ. Для этого положим

$$\mathrm{H}_{kk}=0 \tag{13.4}$$

для всех k, а внедиагональные клетки H_{lk} определим из уравнений

$$\mathrm{H}_{lk}\Lambda_k-\Lambda_l\mathrm{H}_{lk}=\Omega_{lk}. \tag{13.5}$$

Согласно условию матрицы Λ_k и Λ_l не имеют общих собственных значений при $k\neq l$, следовательно, уравнения (13.5) разрешимы. Пусть элементы Ω малы по сравнению с расстояниями между множествами собственных значений матриц Λ_k и Λ_l при $k\neq l$. В этом случае матрицы H_{lk} будут иметь тот же порядок малости, что и Ω_{lk}. Обозначив через $\hat{\Omega}$ клеточно-диагональную матрицу, составленную из диагональных клеток матрицы Ω, получим, что $\hat{\Lambda}\cong\Lambda+\hat{\Omega}$; при этом

$$\hat{\Lambda}_k\cong\Lambda_k+\Omega_{kk} \tag{13.6}$$

для всех k.

Формула (13.6) определяет главные члены возмущений собственных значений. Однако аналогичным способом можно получить и более точное соотношение. Пусть матрица H

вычисляется согласно (13.4), (13.5). Это означает, что она удовлетворяет уравнению

$$\Omega-\hat{\Omega}=\mathrm{H}\Lambda-\Lambda\mathrm{H}$$

и имеет тот же порядок малости, что и матрица Ω. Далее находим, что с точностью до членов третьего порядка малости

$$(E+\mathrm{H})^{-1}(\Lambda+\Omega)(E+\mathrm{H})=$$
$$=(E-\mathrm{H}+\mathrm{H}^2-\ldots)(\Lambda+\Omega)(E+\mathrm{H})\cong$$
$$\cong\Lambda+\Omega-\mathrm{H}\Lambda+\Lambda\mathrm{H}-\mathrm{H}(\Omega-\mathrm{H}\Lambda+\Lambda\mathrm{H})+\Omega\mathrm{H}=$$
$$=\Lambda+\hat{\Omega}+\Omega\mathrm{H}-\mathrm{H}\hat{\Omega}. \tag{13.7}$$

Итак, с точностью до членов третьего порядка малости матрица $\Lambda+\Omega$ подобна матрице $\Lambda+\hat{\Omega}$ с возмущением $\Omega\mathrm{H}-\mathrm{H}\hat{\Omega}$.

Для нахождения клеточно-диагональной матрицы, которой подобна матрица в левой части (13.7), снова воспользуемся асимптотической формулой (13.6), заменяя матрицы Λ и Ω соответственно матрицами $\Lambda+\hat{\Omega}$ и $\Omega\mathrm{H}-\mathrm{H}\hat{\Omega}$. В силу непрерывной зависимости собственных значений от элементов матрицы, клетки $\Lambda_k+\Omega_{kk}$ при малых Ω_{kk} не будут иметь общих собственных значений. Поэтому

$$\hat{\Lambda}_k\cong\Lambda_k+\Omega_{kk}+\sum_{\substack{l=1\\ l\neq k}}^{r}\Omega_{kl}\mathrm{H}_{lk}. \tag{13.8}$$

Это равенство уже верно с точностью до членов третьего порядка малости.

Исследование осуществляется более эффективно, если решение уравнений (13.5) можно написать в явном виде. Рассмотрим один из важнейших случаев, когда матрица Λ диагональная. Не ограничивая общности, можно считать, что каждая из клеток Λ_k является скалярной матрицей.

Обозначим через λ_i и $\hat{\lambda}_i$, где $1\leqslant i\leqslant n$, собственные значения матриц Λ и $\hat{\Lambda}$, через ω_{ij} — элементы матрицы Ω. С точностью до величин второго порядка малости $\hat{\lambda}_i$ совпадают с собственными значениями клеток (13.6), т. е. получаются путем сдвига собственных значений клеток Ω_{kk} на диагональные элементы клеток Λ_k. Известно [1], что

сумма квадратов модулей собственных значений матрицы не превосходит квадрата ее евклидовой нормы, поэтому

$$\sum_{i=1}^{n} |\lambda_i - \hat{\lambda}_i|^2 \leqslant \sum_{k=1}^{n} \|\Omega_{kk}\|_E^2 \qquad (13.9)$$

и заведомо

$$\sum_{i=1}^{n} |\lambda_i - \hat{\lambda}_i|^2 \leqslant \|\Omega\|_E^2.$$

Полученное соотношение асимптотически верно для любой матрицы Ω. Если же матрицы Λ и Ω эрмитовы, то оно оказывается верным независимо от величины Ω [5]. Для нормальной матрицы Ω асимптотическое неравенство (13.9) переходит в асимптотическое равенство.

В случае диагональной матрицы Λ уравнения (13.5) легко решаются. Совместно с (13.4) получаем следующие выражения для элементов η_{ij} матрицы H:

$$\eta_{ij} \cong \begin{cases} 0, & \lambda_i = \lambda_j, \\ \omega_{ij}/(\lambda_j - \lambda_i), & \lambda_i \neq \lambda_j. \end{cases} \qquad (13.10)$$

Таким образом, *при возмущении элементов диагональной матрицы Λ на величины порядка ω все ее собственные значения согласно* (13.9) *также меняются на величины порядка ω. Как показывает* (13.10), *корневой базис матрицы $\Lambda+\Omega$ может быть выбран и упорядочен так, что каждый из его векторов отличается от соответствующего собственного вектора матрицы Λ снова на величину порядка ω.* Заметим, что нельзя говорить о сравнении базисов из собственных векторов матриц Λ и $\Lambda+\Omega$, так как матрица $\Lambda+\Omega$ может его не иметь.

Для простых собственных значений эти выводы уточняются. Пусть собственное значение λ_p матрицы Λ является простым; тогда соответствующие клетки матриц Λ и $\hat{\Lambda}$ будут иметь первый порядок. Теперь из (13.6) следует асимптотическое равенство $\hat{\lambda}_p \cong \lambda_p + \omega_{pp}$ для собственного значения $\hat{\lambda}_p$ возмущенной матрицы. Равенство, верное с точностью до членов третьего порядка, вытекает из (13.8), (13.10). Именно,

$$\hat{\lambda}_p \cong \lambda_p + \omega_{pp} + \sum_{i \neq p} \omega_{pi}\omega_{ip}/(\lambda_p - \lambda_i). \qquad (13.11)$$

Итак, *собственное значение $\hat{\lambda}_p$ матрицы $\Lambda+\Omega$, соответствующее простому собственному значению λ_p матрицы Λ, отличается от ее диагонального элемента лишь на величину порядка ω^2.* В этом случае формулы (13.10) при $j=p$, $i \neq p$ дают асимптотические выражения для координат нормированного собственного вектора матрицы $\Lambda+\Omega$, соответствующего $\hat{\lambda}_p$.

УПРАЖНЕНИЯ

1. Найти явный вид решения уравнения (13.1), если матрицы A и B являются каноническими ящиками либо каноническими матрицами Жордана.
2. Доказать, что множество собственных значений оператора $AZ-ZB$ с учетом их кратности совпадает с множеством чисел вида $\lambda-\mu$, где λ, μ — собственные значения матриц A, B.
3. Что представляют собственные «векторы» оператора $AZ-ZB$? Как они связаны с собственными векторами матриц A и B?
4. Написать на порядок более точные выражения, чем (13.10), (13.11).
5. Получить точные оценки погрешностей формул (13.10), (13.11).
6. Пусть матрицы Λ и Ω эрмитовы. Доказать, что матрица H, определяемая формулами (13.4), (13.5), будет косоэрмитовой.
7. Доказать, что характеристические многочлены матриц Λ и $\Lambda+\Omega$ совпадают с точностью до членов второго порядка малости.
8. Пусть A — эрмитова матрица. Доказать, что базис из ее собственных векторов можно выбрать так, что он будет непрерывно зависеть от элементов матрицы.
9. Как меняются формулы, если не требовать выполнения условия (13.4)?
10. Пусть каким-либо способом найдены корневые векторы матрицы $\Lambda+\Omega$. Как они выражаются через корневые векторы, определяемые соотношениями (13.4), (13.5)?

§ 14. Матрицы общей структуры

Корни многочлена являются непрерывными функциями коэффициентов. При этом возмущение простых корней имеет тот же порядок малости, что и возмущение самих коэффициентов. Однако кратные корни могут меняться весьма существенно. Собственные значения матрицы совпадают с корнями характеристического многочлена. Следовательно, есть основания предполагать, что кратные собственные значения по сравнению с простыми будут также изменяться значительно.

Полученные результаты пока этого не подтверждают. Более того, оказалось, что все собственные значения

матрицы простой структуры, независимо от их кратности, имеют тот же порядок возмущения, что и матрица.

Согласно (13.6) асимптотическое исследование влияния возмущения матрицы общей структуры сводится к аналогичной задаче для матрицы с одинаковыми собственными значениями. Но такая матрица заведомо подобна матрице вида

$$\Lambda = \begin{vmatrix} \lambda_0 & \alpha_1 & & & \\ & \lambda_0 & \alpha_2 & & \\ & & \ddots & & \\ & & & \lambda_0 & \alpha_{n-1} \\ & & & & \lambda_0 \end{vmatrix}. \qquad (14.1)$$

Поэтому можно попытаться получить дополнительные сведения, изучая возмущение матриц (14.1).

Влияние возмущения зависит не только от его величины, но и от того, где оно сосредоточено. Предположим, что возмущается лишь один элемент в позиции (i, j) на величину ω_{ij}. Если $i < j$, то все собственные значения остаются без изменения; если $i = j$, то на ω_{ii} меняется лишь одно собственное значение; если же $i > j$, то меняются $i - j + 1$ собственных значений.

Вычисляя характеристический многочлен возмущенной матрицы (14.1), нетрудно установить, что изменившиеся собственные значения являются корнями многочлена

$$(\lambda_0 - \lambda)^{i-j+1} - (-1)^{i-j+1}\omega_{ij}\alpha_j\alpha_{j+1}\ldots\alpha_{i-1}.$$

Отсюда вытекает, что в общем случае наибольшее влияние на собственные значения оказывает возмущение, находящееся в позиции $(n, 1)$. По порядку зависимости оно такое же, как и для кратных корней многочлена. При этом матрица (14.1) должна представлять собой канонический ящик Жордана.

Таким образом, большие по порядку возмущения собственных значений матрицы действительно могут иметь место. Но это связано только с наличием клеток Жордана в структуре матрицы.

УПРАЖНЕНИЯ

1. Целесообразно ли с точки зрения точности вычислять собственные значения матрицы как корни характеристического многочлена?

2. Пусть матрица A получена путем малого возмущения матрицы простой структуры и имеет жордановы клетки. Возможны ли большие по порядку возмущения собственных значений матрицы A при малом ее возмущении?

3. Есть ли основания опасаться потери точности собственных значений из-за жордановых клеток, появившихся в результате влияния ошибок округления?

4. Подобным преобразованием ненулевые числа $\alpha_1, \ldots, \alpha_{n-1}$ матрицы (14.1) можно сделать как угодно малыми и как угодно большими. Влечет ли за собой такое преобразование изменение зависимости собственных значений от возмущения матрицы?

5. Исследовать влияние возмущения двухдиагональной матрицы с неодинаковыми собственными значениями.

6. Могут ли появиться жордановы клетки при эрмитовом (неэрмитовом) возмущении эрмитовой матрицы?

7. Пусть λ является корнем кратности p характеристического многочлена матрицы A. Предположим, далее, что дефект матрицы $A - \lambda E$ равен m. Доказать, что существует такое возмущение нормы ε для матрицы A, при котором все собственные значения, соответствующие λ, изменятся на величины порядка $\varepsilon^{1/(p-m+1)}$.

§ 15. Сингулярное разложение

Продолжим исследование возмущения клеточно-диагональной матрицы, однако на этот раз в связи с сингулярным разложением. Новые результаты будут иметь много общего с полученными ранее. Существенное различие заключается лишь в том, что теперь мы рассматриваем только унитарные преобразования матрицы.

Пусть даны матрицы A, B, C, D размеров соответственно $n \times n$, $m \times m$, $n \times m$, $n \times m$ и система матричных уравнений

$$AU - VB = C, \quad UB^* - A^*V = D, \qquad (15.1)$$

где U, V — искомые матрицы размеров $n \times m$.

Теорема 15.1. *Система* (15.1) *имеет единственное решение тогда и только тогда, когда матрицы A и B не имеют общих сингулярных чисел.*

Доказательство. Преобразуем систему (15.1) в эквивалентную, но более простого вида. Для A и B существуют сингулярные разложения

$$A = Q\Lambda R, \quad B = FMG, \qquad (15.2)$$

где Λ, M — диагональные матрицы с неотрицательными элементами λ_i, μ_i, а остальные матрицы — унитарные. Подставив разложения (15.2) в (15.1) и выполнив несложные преобразования, приходим к системе

$$\Lambda\hat{U} - \hat{V}M = \hat{C}, \quad \hat{U}M - \Lambda\hat{V} = \hat{D}. \qquad (15.3)$$

Здесь

$$\hat{U} = RUG^*, \quad \hat{V} = Q^*VF,$$
$$\hat{C} = Q^*CG^*, \quad \hat{D} = RDF.$$

Очевидно, что достаточно исследовать систему (15.3). Но сравнивая элементы ее правых и левых частей, заключаем, что она распадается на системы второго порядка относительно элементов матриц $\hat{U}$, $\hat{V}$. Определители этих систем отличны от нуля тогда и только тогда, когда $\lambda_i \neq \mu_j$ для всех i, j. Отсюда и вытекает утверждение теоремы.

Рассмотрим квадратную клеточно-диагональную матрицу P, клетки $P_1, P_2, \ldots, P_r$ которой не имеют общих сингулярных чисел. Пусть $P + \Omega$ — возмущенная матрица. Разобьем матрицу Ω на прямоугольные клетки Ω_{ij} так, чтобы ее диагональные клетки имели те же размеры, что и соответствующие клетки матрицы P. Будем приводить матрицу $P + \Omega$ к клеточно-диагональному виду с помощью унитарных преобразований. Это означает, что нужно найти унитарные матрицы $\hat{X}$, $\hat{Y}$ и клеточно-диагональную матрицу $\hat{P}$, для которых

$$\hat{Y}^* (P + \Omega) \hat{X} = \hat{P}.$$

При этом предполагается, что клетки $\hat{P}_k$ матрицы $\hat{P}$ имеют те же размеры, что и клетки матрицы P.

Будем опять искать матрицы $\hat{X}$, $\hat{Y}$ как возмущенные единичные матрицы, т. е. в виде сумм $\hat{X} = E + H$, $\hat{Y} = E + T$, где H и T — малые матрицы. Так как $\hat{X}$ и $\hat{Y}$ должны быть унитарными, то асимптотически H и T будут косоэрмитовыми. Эти матрицы мы разобьем на клетки H_{ij} и T_{ij} по тому же принципу, что и Ω. Имеем

$$\hat{Y}^* (P + \Omega) \hat{X} \cong (E - T)(P + \Omega)(E + H) \cong$$
$$\cong P + \Omega - TP + PH.$$

Подберем матрицы H и T так, чтобы с точностью до малых второго порядка малости правая часть полученного соотношения была бы клеточно-диагональной матрицей. Для этого положим

$$H_{kk} = T_{kk} = 0 \tag{15.4}$$

для всех k, а внедиагональные клетки H_{lk}, T_{lk} определим из систем

$$T_{kl}P_l - P_kH_{kl} = \Omega_{kl},$$
$$T_{lk}P_k - P_lH_{lk} = \Omega_{lk}.$$

В силу условий на матрицы H и T,

$$T_{kl} = -T_{lk}^*, \quad H_{kl} = -H_{lk}^*,$$

поэтому в действительности будем иметь системы

$$P_kH_{kl} - T_{kl}P_l = -\Omega_{kl},$$
$$H_{kl}P_l^* - P_k^*T_{kl} = \Omega_{lk}^*. \tag{15.5}$$

Матрицы P_k и P_l не имеют общих сингулярных чисел при $k \neq l$, следовательно, системы (15.5) разрешимы. Предположим, что элементы Ω малы по сравнению с расстояниями между множествами сингулярных чисел матриц P_k и P_l при $k \neq l$. В этом случае матрицы H_{kl}, T_{kl} будут иметь тот же порядок малости, что и Ω.

Пусть $\hat{\Omega}$ — клеточно-диагональная матрица, составленная из диагональных клеток Ω. Матрицы H и T удовлетворяют уравнению $\Omega - \hat{\Omega} = TP - PH$, поэтому $\hat{P} \cong P + \hat{\Omega}$, при этом, конечно,

$$\hat{P}_k = P_k + \Omega_{kk} \tag{15.6}$$

для всех k.

Формула (15.6) определяет главные члены возмущений сингулярных чисел, а решения систем (15.5) — главные члены возмущений сингулярных векторов. Снова исследование осуществляется более эффективно, если решение систем (15.5) можно написать в явном виде.

Предположим, что матрица P — диагональная с неотрицательными элементами, расположенными в порядке невозрастания. В этом случае все ее клетки P_k являются скалярными матрицами. Напомним, что с помощью унитарных преобразований к такому виду можно привести любую матрицу. Обозначим через ρ_i, $\hat{\rho}_i$ сингулярные числа матриц P, $\hat{P}$, через ω_{ij} — элементы матрицы Ω. Теперь системы (15.5) легко решаются. Совместно с (15.4)

получаем следующие выражения для элементов η_{ij}, τ_{ij} матриц H, T:

$$\eta_{ij} \cong \begin{cases} 0, & \text{если } \rho_i = \rho_j, \\ \dfrac{\omega_{ij}\rho_i + \bar{\omega}_{ji}\rho_j}{\rho_j^2 - \rho_i^2}, & \rho_i \neq \rho_j, \end{cases} \tag{15.7}$$

$$\tau_{ij} \cong \begin{cases} 0, & \text{если } \rho_i = \rho_j, \\ \dfrac{\omega_{ij}\rho_j + \bar{\omega}_{ji}\rho_i}{\rho_j^2 - \rho_i^2}, & \rho_i \neq \rho_j. \end{cases} \tag{15.8}$$

Оценим возмущение сингулярных чисел матрицы P или, что то же самое, отклонение сингулярных чисел матриц $\hat{P}_k$ от диагональных элементов матриц P_k. Известно [1], что сингулярные числа любой квадратной матрицы A совпадают с собственными значениями матрицы $(AA^*)^{1/2}$. Матрицы P_k, по предположению, скалярные, поэтому для $P_k \neq 0$ имеем

$$\hat{P}_k\hat{P}_k^* \cong (P_k + \Omega_{kk})(P_k + \Omega_{kk})^* \cong$$
$$\cong P_k^2 + P_k\Omega_{kk}^* + \Omega_{kk}P_k \cong \left(P + \frac{\Omega_{kk} + \Omega_{kk}^*}{2}\right)^2.$$

Если же $P_k = 0$, то

$$\hat{P}_k\hat{P}_k^* \cong \Omega_{kk}\Omega_{kk}^*. \tag{15.9}$$

Матрица

$$P_k + \frac{\Omega_{kk} + \Omega_{kk}^*}{2}$$

получена путем эрмитова возмущения диагональной матрицы P_k. Для оценки возмущений ее собственных значений можно воспользоваться соотношением (13.9). В случае (15.9) мы примем во внимание то, что сумма квадратов всех сингулярных чисел равна квадрату евклидовой нормы матрицы. Учитывая сказанное, получаем

$$\sum_{i=1}^{n} |\rho_i - \hat{\rho}_i|^2 \cong \sum_{k=1}^{r} \|\Omega_{kk}\|_E^2 \tag{15.10}$$

и, конечно,

$$\sum_{i=1}^{n} |\rho_i - \hat{\rho}_i|^2 \lessapprox \|\Omega\|_E^2. \tag{15.11}$$

Итак, при возмущении элементов матрицы P на величины порядка ω все ее сингулярные числа согласно (15.10)

также меняются на величины порядка ω. Как показывают (15.7), (15.8), сингулярные векторы матрицы $P + \Omega$ могут быть выбраны и упорядочены так, что они отличаются от соответствующих сингулярных векторов матрицы P снова на величины порядка ω. Правильность соотношения (15.11) была установлена лишь при малых Ω. В действительности оно выполняется независимо от величины возмущения.

В этих исследованиях предполагалось, что матрица P квадратная. Если P — прямоугольная, то изменения невелики. Действительно, дополним матрицы P и Ω нулевыми столбцами (строками) до квадратной. Из формул (15.7), (15.8) вытекает, что элементы матрицы H (матрицы T), появившиеся за счет такого расширения, будут равны нулю. Следовательно, формулы (15.7), (15.8), (15.10) имеют место и в случае прямоугольной матрицы P, если «несуществующие» элементы матриц P, Ω считать нулевыми.

Единственное отличие заключается в несколько ином виде матрицы $\hat{P}$. В ней теперь останутся все элементы ω_{ij} ниже (правее) диагональных элементов, соответствующих нулевым сингулярным числам матрицы P.

УПРАЖНЕНИЯ

1. Пусть модуль разности сингулярных чисел матриц A и B не меньше δ. Доказать, что для решений системы (15.1) справедливо соотношение

$$\max\{\|U\|_E, \|V\|_E\} \leqslant \frac{1}{\delta}(\|C\|_E^2 + \|D\|_E^2)^{1/2}.$$

2. Рассмотрим диагональную матрицу P с неотрицательными элементами и пусть элементы Ω являются величинами порядка ω. Доказать, что сингулярное число матрицы $P + \Omega$, соответствующее простому сингулярному числу матрицы P, отличается от вещественной части диагонального элемента матрицы $P + \Omega$ на величину порядка ω^2.

3. Доказать, что сингулярные числа непрерывно зависят от элементов матрицы.

4. Доказать, что базисы из сингулярных векторов можно выбрать так, что они будут непрерывно зависеть от элементов матрицы.

5. Можно ли таким же приемом, как при выводе формулы (13.8), получить аналогичную формулу в случае сингулярного разложения?

6. Как меняются формулы (15.7)—(15.11), если не требовать выполнения условий (15.4)?

7. Пусть каким-либо способом найдены ортонормированные сингулярные векторы матрицы $P + \Omega$. Как они выражаются через сингулярные векторы, определяемые соотношениями (15.4), (15.5)?

§ 16. Проекции псевдорешения

Любое псевдорешение неустойчиво [1] к возмущению оператора, если дефект оператора отличен от нуля. Это связано с тем, что образ возмущенного оператора может значительно отличаться от образа точного оператора и даже иметь другую размерность.

Однако во всяком псевдорешении можно выделить его устойчивую часть. *Важно подчеркнуть, что эта часть может быть найдена численным способом по приближенно заданной информации. Неустойчивую же часть псевдорешения нельзя определить по приближенной информации и для ее оценки следует привлекать дополнительные сведения.*

Как уже отмечалось, при исследовании влияния возмущения на псевдорешение можно ограничиться рассмотрением возмущения системы линейных алгебраических уравнений с диагональной матрицей из сингулярных чисел. Пусть

$$Pu = d \tag{16.1}$$

— точная система. Обозначим через $\rho_1, \ldots, \rho_n$ диагональные элементы матрицы P и будем считать, что $\rho_1 \geqslant \rho_2 \geqslant \ldots \geqslant \rho_n \geqslant \rho_{n+1} = 0$. Предположим далее, что

$$(P + \Omega)\,\hat{u} = d + \omega$$

— возмущенная система.

Если матрица P ненулевая, то среди $\rho_1, \ldots, \rho_n, \rho_{n+1}$ есть хотя бы одна пара не равных между собой соседних чисел. Пусть $\rho_k - \rho_{k+1} \neq 0$ и элементы матрицы Ω и вектора ω достаточно малы по сравнению с $\rho_k - \rho_{k+1}$. Обозначим через X_k, $\hat{X}_k$ подпространства, натянутые на первые k правых сингулярных векторов матриц P и $P+\Omega$. Проведенные ранее исследования позволяют утверждать, что эти подпространства мало отличаются друг от друга. Поэтому можно ожидать, что будут мало отличаться и проекции u_k, $\hat{u}_k$ псевдорешений u, $\hat{u}$ точной и возмущенной систем на X_k, $\hat{X}_k$.

Разобьем каждую из рассматриваемых матриц на четыре прямоугольные клетки, считая клетку в левом верхнем углу квадратной порядка k. Если

$$P = \begin{bmatrix} P_{11} & 0 \\ 0 & P_{22} \end{bmatrix},$$

то u_k совпадает с нормальным псевдорешением системы

$$\begin{bmatrix} P_{11} & 0 \\ 0 & 0 \end{bmatrix} u = d,$$

и, следовательно,

$$u_k = \begin{bmatrix} P_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} d. \tag{16.2}$$

Пусть далее

$$P + \Omega = \begin{bmatrix} P_{11} + \Omega_{11} & \Omega_{12} \\ \Omega_{21} & P_{22} + \Omega_{22} \end{bmatrix}.$$

При исследовании сингулярного разложения мы установили существование матриц

$$\begin{bmatrix} E & H_{12} \\ -H_{12}^* & E \end{bmatrix}, \quad \begin{bmatrix} E & T_{12} \\ -T_{12}^* & E \end{bmatrix},$$

таких, что

$$\begin{bmatrix} E & -T_{12} \\ T_{12}^* & E \end{bmatrix} \begin{bmatrix} P_{11} + \Omega_{11} & \Omega_{12} \\ \Omega_{21} & P_{22} + \Omega_{22} \end{bmatrix} \begin{bmatrix} E & H_{12} \\ -H_{12}^* & E \end{bmatrix} \simeq$$

$$\simeq \begin{bmatrix} P_{11} + \Omega_{11} & 0 \\ 0 & P_{22} + \Omega_{22} \end{bmatrix}.$$

Поэтому $\hat{u}_k$ асимптотически совпадает с нормальным псевдорешением системы

$$\begin{bmatrix} E & T_{12} \\ -T_{12}^* & E \end{bmatrix} \begin{bmatrix} P_{11} + \Omega_{11} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} E & -H_{12} \\ H_{12}^* & E \end{bmatrix} \hat{u} = d + \omega$$

и, следовательно,

$$\hat{u}_k \cong \begin{bmatrix} E & H_{12} \\ -H_{12}^* & E \end{bmatrix} \begin{bmatrix} (P_{11} + \Omega_{11})^{-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} E & -T_{12} \\ T_{12}^* & E \end{bmatrix} (d + \omega). \tag{16.3}$$

Теперь легко получить асимптотическое выражение для ошибки $u_k - \hat{u}_k$. Представив векторы d, ω в виде сумм

$$d = \begin{bmatrix} d_1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ d_2 \end{bmatrix}, \quad \omega = \begin{bmatrix} \omega_1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ \omega_2 \end{bmatrix},$$

где d_1, ω_1 имеют размерность k, находим согласно (16.2), (16.3), что

$$u_k - \hat{u}_k \cong \begin{bmatrix} P_{11}^{-1}\Omega_{11}P_{11}^{-1}d_1 - P_{11}^{-1}\omega_1 + P_{11}^{-1}T_{12}d_2 \\ H_{12}^* P_{11}^{-1} d_1 \end{bmatrix}. \tag{16.4}$$

Пусть известно, что точная система совместна. Выберем в качестве ρ_k наименьшее ненулевое сингулярное

число. Тогда $d_2 = 0$, проекция u_k совпадает с нормальным решением u_0 системы (16.1), а из (15.5) вытекает, что

$$T_{12} = -P_{11}^{-1}\Omega_{21}^*, \quad H_{12}^* = -\Omega_{12}^* P_{11}^{-1}. \tag{16.5}$$

Если обозначить

$$\delta P = \frac{\|\Omega\|_E}{\|P\|_E}, \quad \delta d = \frac{\|w\|_E}{\|d\|_E}, \quad \delta u_0 = \frac{\|u_0 - \tilde{u}_k\|_E}{\|u_0\|_E},$$

$$\nu_P^+ = \|P^+\|_E \|P\|_E,$$

то совместно с (16.5) соотношение (16.4) приводит к оценке

$$\delta u_0 \lesssim \nu_P^+ (\delta P + \delta d), \tag{16.6}$$

асимптотическая связь которой с (10.10) очевидна.

Оценка (16.6) сохраняется и для «почти совместной» системы (16.1), т. е. при достаточно малом, хотя и отличном от нуля векторе d_2. Проекция u_k в этом случае будет совпадать с нормальным псевдорешением системы (16.1).

Таким образом, *если точная система линейных алгебраических уравнений совместна или почти совместна и возмущение мало по сравнению с минимальным ненулевым сингулярным числом точной матрицы, то нормальное псевдорешение можно определить по возмущенной системе с такой же точностью, как и для системы с невырожденной матрицей.*

В случае несовместности точной системы влияние возмущения матрицы становится более заметным. Если снова предположить, что оно достаточно мало по сравнению с минимальным ненулевым сингулярным числом, то при введенных выше обозначениях будем иметь

$$\delta u_0 \lesssim \nu_P^+ (\delta P + \delta d) + \nu_P^+ (\nu_P^+ \delta P + \delta d)\, (\|d_2\|_E / \|d_1\|_E). \tag{16.7}$$

Здесь u_0 уже является нормальным псевдорешением системы (16.1). Его точность в значительной мере зависит от отношения $\|d_2\|_E$ к $\|d_1\|_E$, т. е. от *степени согласования матрицы и правой части исходной системы.*

Оценки (16.6), (16.7) получены для возмущений, достаточно малых по сравнению с минимальным ненулевым сингулярным числом. Они позволяют высказать предположение о том, что при подходящем выборе номера k проекция $\tilde{u}_k$ будет достаточно хорошо приближать нормальное псевдорешение u_0 точной системы (16.1) и в самом

общем случае. Мы проведем сейчас необходимое обоснование этого предположения.

Обозначим через Δ_k полную ошибку $u_0 - \tilde{u}_k$ и представим ее в следующем виде:

$$u_0 - \tilde{u}_k = (u_0 - u_k) + (u_k - \tilde{u}_k).$$

Разность $u_k - \tilde{u}_k$ легко оценивается с учетом (16.4), разность же $u_0 - u_k$ неустойчива к возмущению и для ее оценки необходимо привлечь дополнительные сведения о точной задаче. Рассмотрим поэтому наряду с системой (16.1) связанную с ней согласно (9.8), (9.9) систему

$$(P^*P)^2 v = P^* d.$$

Если v_0 — ее нормальное решение, то прямое сравнение этой системы с (16.1) показывает, что

$$\|u_0 - u_k\|_E \leqslant \rho_{k+1}^2 \|v_0\|_E. \tag{16.8}$$

В случае совместности системы (16.1) имеем

$$\|d_2\|_E \leqslant \rho_{k+1}^2 \|v_0\|_E. \tag{16.9}$$

Предположим для простоты, что $\rho_1 = 1$, и пусть входные данные системы (16.1) заданы с малой абсолютной ошибкой порядка ε. Тогда из (16.4), (16.8), (16.9) следует, что в случае совместности точной системы норма главного члена полной ошибки Δ_k с точностью до констант будет ограничена сверху функцией вида

$$\|\Delta_k\| \lesssim \frac{\varepsilon}{\rho_k - \rho_{k+1}} + \rho_{k+1}^2 \tag{16.10}$$

и функцией вида

$$\|\Delta_k\| \lesssim \frac{\varepsilon}{(\rho_k - \rho_{k+1})\,\rho_k} + \rho_{k+1}^2 \tag{16.11}$$

в противном случае. Минимизируя правые части (16.10), (16.11) с помощью выбора соответствующего номера k, можно определить ту проекцию $\tilde{u}_k$, которая лучше всего приближает нормальное псевдорешение u_0. О достижимой при этом точности говорит

Лемма 16.1. *Для $\varepsilon \leqslant (4n)^{-1}$ и любого набора чисел ρ_k, где $1 = \rho_1 \geqslant \rho_2 \geqslant \ldots \geqslant \rho_n \geqslant \rho_{n+1} = 0$, имеют место соотношения*

$$\min_{1 \leqslant k \leqslant n} \left(\frac{\varepsilon}{\rho_k - \rho_{k+1}} + \rho_{k+1}^2 \right) \leqslant 3\,(n\varepsilon)^{2/3}, \tag{16.12}$$

$$\min_{1 \leqslant k \leqslant n} \left(\frac{\varepsilon}{(\rho_k - \rho_{k+1})\,\rho_k} + \rho_{k+1}^2 \right) \leqslant 4\,(n\varepsilon)^{1/2}.$$

Доказательство. Рассмотрим сегмент $[c/2, c]$, где $0<c\leqslant 1$. На нем находится не более n чисел из $\rho_1, \rho_2, \ldots, \rho_n$, поэтому существует другой сегмент $[\alpha, \beta]$ при $c/2\leqslant\alpha<\beta\leqslant c$ и $\beta-\alpha\geqslant c/2n$, внутри которого нет сингулярных чисел. Выберем в качестве ρ_k ближайшее к β и не меньшее его сингулярное число. Тогда

$$\rho_k-\rho_{k+1}\geqslant c/2n, \quad \rho_k\leqslant c, \quad \rho_{k+1}\geqslant c/2$$

и, следовательно,

$$\min_{1\leqslant k\leqslant n}\left(\frac{n}{\rho_k-\rho_{k+1}}+\rho_{k+1}^2\right)\leqslant\frac{2n\varepsilon}{c}+c^2.$$

Пусть c таково, что правая часть полученного неравенства достигает минимума. Это дает $c=(n\varepsilon)^{1/3}\leqslant 1$, и первое соотношение леммы доказано. Имеем далее

$$\min_{1\leqslant k\leqslant n}\left(\frac{n}{(\rho_k-\rho_{k+1})\rho_k}+\rho_{k-1}^2\right)\leqslant\frac{4n\varepsilon}{c^2}+c^2.$$

Правая часть достигает минимума при $c=(4n\varepsilon)^{1/4}\leqslant 1$. Величина этого минимума подтверждает справедливость второго соотношения леммы.

Итак, *если входные данные системы заданы с точностью порядка* ε, *то одна из проекций* $\tilde{u}_k$ *приближает нормальное псевдорешение* u_0 *с точностью порядка* $(n\varepsilon)^{\alpha}$. *Если исходная система совместна, то* $\alpha\geqslant 2/3$ *и* $\alpha\geqslant 1/2$ — *в противном случае.*

Нетрудно построить примеры систем с таким распределением сингулярных чисел, при которых достигаются наименьшие порядки точности. Пусть $\rho_1=1$ и есть некоторое количество нулевых сингулярных чисел. Для совместных систем достигается порядок $(n\varepsilon)^{2/3}$, если все остальные сингулярные числа расположены равномерно между $(n\varepsilon)^{1/3}$ и нулем. Для несовместных систем достигается порядок $(n\varepsilon)^{1/2}$, если все остальные сингулярные числа расположены равномерно между $(n\varepsilon)^{1/4}$ и нулем.

Заметим, что наличие малых сингулярных чисел матрицы системы не обязательно свидетельствует о невозможности вычислить псевдорешение с достаточно хорошей точностью. Если матрица имеет группу больших сингулярных чисел, а остальные сингулярные числа соизмеримы с точностью входных данных или меньше; то из (16.10), (16.11) следует, что одна из проекций $\tilde{u}_k$ приближает

нормальное псевдорешение u_0 с точностью порядка ε как для совместной, так и для несовместной системы. Этот факт имеет исключительное значение для обоснования большинства численных методов решения систем линейных алгебраических уравнений с вырожденной матрицей.

Если матрица системы квадратная и невырожденная, то норма ошибки Δ решения возмущенной системы имеет вид $\|\Delta\|=\varepsilon/\rho$, где ρ — того же порядка, что и минимальное сингулярное число матрицы системы. Как вытекает из первого соотношения (16.12), норму ошибки Δ можно представить в таком же виде и в случае произвольной совместной системы, при этом ρ по порядку зависимости от сингулярных чисел и точности входных данных удовлетворяет неравенству

$$\max\{\rho_n, (n\varepsilon)^{1/3}\}\leqslant\rho.$$

Из второго соотношения (16.12) вытекает, что для несовместной системы $\|\Delta\|=\varepsilon/\rho^2$; при этом по порядку зависимости

$$\max\{\rho_n, (n\varepsilon)^{1/4}\}\leqslant\rho.$$

УПРАЖНЕНИЯ

1. Можно ли утверждать, что проекция $\tilde{u}_k$, для которой ошибка $u_k-\tilde{u}_k$ будет минимальной, обеспечивает асимптотическую близость $\tilde{u}_k$ к нормальному псевдорешению u_0?

2. Вывести формулу, аналогичную (16.6), для несовместной системы (16.1).

3. Что означает малое (большое) значение нормы нормального псевдорешения системы (9.9) по отношению к системе (16.1)?

4. Оценить минимальные по k значения $\|\Delta_k\|$, вычисляемые согласно (16.10), (16.11), для сингулярных чисел $\rho_k=k^{-1}$ и $\rho_k=k^{-2}$, $1\leqslant k\leqslant n$. Сравнить полученные результаты с (16.12).

5. Доказать, что для матриц T_{12}, H_{12} из (16.4) справедливо асимптотическое неравенство

$$\|T_{12}\|_E, \|H_{12}\|_E\leqslant\frac{(\|\Omega_{12}\|_E^2+\|\Omega_{21}\|_E^2)^{1/2}}{\rho_k-\rho_{k+1}}.$$

6. Учитывая (16.4), (16.8), (16.9), получить точную оценку для нормы полной ошибки Δ_k, содержащую все члены.

7. Есть ли основания опасаться потери точности решения системы из-за несовместности, появившейся в результате влияния ошибок округления?

8. Оценить нормы невязок

$$r_k=P\tilde{u}_k-d, \quad \tilde{r}_k=(P+\Omega)\tilde{u}_k-(d+\omega).$$

9. Пусть точная система совместна. Можно ли с помощью соответствующего выбора проекции $\hat{u}_k$ добиться одновременно малости невязок и устойчивости вычисления проекции?

10. Что можно получить с помощью выбора проекции $\hat{u}_k$ в отношении величины невязок в случае несовместности точной системы?

§ 17. Нормальное псевдорешение

Необходимость определения проекций псевдорешений и подпространств сингулярных векторов возникает далеко не во всех задачах, связанных с системой линейных алгебраических уравнений (16.1). Значительно чаще требуется лишь вычислить с приемлемой точностью *нормальное псевдорешение*. С точки зрения теоретического исследования и практической реализации эта задача нередко сводится к минимизации *регуляризирующего функционала*

$$\Phi_\alpha(u) = \alpha \|u\|_E^2 + \|Pu - d\|_E^2, \qquad (17.1)$$

где число $\alpha \geqslant 0$. Снова, не ограничивая существенно общности, можно считать, что P является диагональной матрицей из сингулярных чисел.

Обозначим через $e_1, e_2, \ldots$ — координатные векторы, через $\beta_1, \beta_2, \ldots$ — координаты вектора d и пусть сингулярные числа $\rho_1, \ldots, \rho_t$ отличны от нуля, а остальные — равны нулю. Если

$$u = \sum_k \alpha_k e_k,$$

то

$$\Phi_\alpha(u) = \sum_{k=1}^{t} (\alpha |\alpha_k|^2 + |\rho_k \alpha_k - \beta_k|^2) + \alpha \sum_{k>t} |\alpha_k|^2 + \sum_{p>t} |\beta_p|^2.$$

Отсюда следует, что минимум $\Phi_\alpha(u)$ достигается в том случае, когда последние координаты $\alpha_{t+1}, \alpha_{t+2}, \ldots$ — нулевые и для каждого $k \leqslant t$ выражение

$$\alpha |\alpha_k|^2 + |\rho_k \alpha_k - \beta_k|^2$$

принимает минимальное значение. Это дает для $k \leqslant t$

$$\alpha_k = \frac{\rho_k \beta_k}{\alpha + \rho_k^2}.$$

Таким образом, при каждом $\alpha > 0$ минимум регуляризирующего функционала (17.1) достигается на единственном векторе

$$u_\alpha = \sum_{k=1}^{t} \frac{\rho_k \beta_k}{\alpha + \rho_k^2} e_k. \qquad (17.2)$$

При $\alpha = 0$ регуляризирующий функционал (17.1) совпадает с функционалом невязки

$$\Phi_0(u) = \|Pu - d\|_E^2.$$

Его минимальное значение достигается на псевдорешениях системы (16.1) и для нормального псевдорешения u_0 справедлива формула

$$u_0 = \sum_{k=1}^{t} \frac{\beta_k}{\rho_k} e_k. \qquad (17.3)$$

Сравнение (17.2), (17.3) позволяет установить некоторые соотношения, связывающие u_α и u_0. Имеем

$$u_0 - u_\alpha = \alpha \sum_{k=1}^{t} \frac{\beta_k}{\rho_k (\alpha + \rho_k^2)} e_k.$$

Для любого $\alpha > 0$

$$\frac{\sqrt{2}\rho_k}{\alpha + \rho_k^2} \leqslant \frac{1}{\sqrt{\alpha}}, \qquad (17.4)$$

поэтому

$$\|u_0 - u_\alpha\|_E \leqslant \alpha\gamma, \quad \|u_0 - u_\alpha\|_E \leqslant \frac{\sqrt{\alpha}\eta}{\sqrt{2}}, \qquad (17.5)$$

где

$$\gamma^2 = \sum_{k=1}^{t} \frac{|\beta_k|^2}{\rho_k^2}, \qquad \eta^2 = \sum_{k=1}^{t} \frac{|\beta_k|^2}{\rho_k^4}. \qquad (17.6)$$

Очевидно, далее, что

$$\|u_\alpha\|_E \leqslant \|u_0\|_E. \qquad (17.7)$$

Таким образом, *при малых значениях α вектор u_α может служить приближением снизу к нормальному псевдорешению u_0*. Неравенства (17.5) определяют при этом величину ошибки.

Непосредственной проверкой легко убедиться, что вектор u_α удовлетворяет системе уравнений

$$(P^*P+\alpha E)\,u_\alpha = P^*d. \tag{17.8}$$

При $\alpha>0$ матрица системы является положительно определенной. Следовательно,

$$u_\alpha = (P^*P+\alpha E)^{-1}P^*d. \tag{17.9}$$

Учитывая (17.2), (17.4), находим

$$\|u_\alpha\|_E \leqslant \frac{\|d\|_F}{\sqrt{2\alpha}}.$$

Вместе с (17.9) это означает справедливость неравенства

$$\|(P^*P+\alpha E)^{-1}P^*d\|_E \leqslant \frac{\|d\|_F}{\sqrt{2\alpha}} \tag{17.10}$$

для любых матриц P и векторов d при $\alpha>0$. Невязки векторов u_α и u_0 связаны между собой таким соотношением

$$\|Pu_\alpha - d\|_E = \left(\sum_{k=1}^{l}\left(\frac{\rho_k^2}{\alpha+\rho_k^2}-1\right)^2|\beta_k|^2+\sum_{p>l}|\beta_p|^2\right)^{1/2} =$$

$$=\left(\alpha^2\sum_{k=1}^{l}\frac{|\beta_k|^2}{(\alpha+\rho_k^2)^2}+\sum_{p>l}|\beta_p|^2\right)^{1/2} \leqslant$$

$$\leqslant \alpha\eta + \|Pu_0-d\|_E. \tag{17.11}$$

Рассмотрим возмущенную систему линейных алгебраических уравнений с матрицей $\hat{P}$ и правой частью $\hat{d}$, где

$$\hat{P}=P+\Omega, \quad \hat{d}=d+\omega. \tag{17.12}$$

Определение приближенного псевдорешения $\hat{u}_\alpha$ по возмущенным $\hat{P}$ и $\hat{d}$ приводит к системе уравнений

$$(\hat{P}^*\hat{P}+\alpha E)\,\hat{u}_\alpha = \hat{P}^*\hat{d}. \tag{17.13}$$

Из (17.9), (17.12), (17.13) получаем

$$(\hat{P}^*\hat{P}+\alpha E)(\hat{u}_\alpha-u_\alpha) = P^*(Pu_\alpha-d)-\hat{P}^*(\hat{P}^*u_\alpha-\hat{d}) =$$

$$= -\Omega(Pu_\alpha-d)-\hat{P}^*(\Omega u_\alpha-\omega).$$

Следовательно,

$$\hat{u}_\alpha-u_\alpha = (\hat{P}^*\hat{P}+\alpha E)^{-1}\delta+(\hat{P}^*\hat{P}+\alpha E)^{-1}\hat{P}^*\nu,$$

где $\delta=-\Omega^*(Pu_\alpha-d)$, $\nu=-(\Omega u_\alpha-\omega)$.

Для положительно определенной матрицы спектральная норма совпадает с максимальным собственным значением, поэтому $\|(\hat{P}^*\hat{P}+\alpha E)^{-1}\|_2 \leqslant \alpha^{-1}$. Учитывая (17.11), будем иметь

$$\|(\hat{P}^*\hat{P}+\alpha E)^{-1}\delta\|_E \leqslant \|(\hat{P}^*\hat{P}+\alpha E)^{-1}\|_2\,\|\delta\|_E \leqslant$$

$$\leqslant \frac{\|\Omega\|_2}{\alpha}\|Pu_\alpha-d\|_E \leqslant \|\Omega\|_2\,\eta+\frac{\|\Omega\|_2}{\alpha}\|Pu_0-d\|_E.$$

Воспользовавшись формулами (17.7), (17.10), находим

$$\|(\hat{P}^*\hat{P}+\alpha E)^{-1}\hat{P}^*\nu\|_E \leqslant \frac{\|\nu\|_F}{\sqrt{2\alpha}} \leqslant \frac{\|\Omega\|_2\|u_0\|_E+\|\omega\|_F}{\sqrt{2\alpha}}.$$

Теперь можно оценить отклонение $\hat{u}_\alpha$ от u_0

$$\|u_0-\hat{u}_\alpha\|_E \leqslant \|u_0-u_\alpha\|_E+\|\hat{u}_\alpha-u_\alpha\|_E \leqslant \alpha\gamma+$$

$$+\|\Omega\|_2\,\eta+\frac{\|\Omega\|_2}{\alpha}\|Pu_0-d\|_E+\frac{\|\Omega\|_2\|u_0\|_E+\|\omega\|_F}{\sqrt{2\alpha}}. \tag{17.14}$$

Правая часть неравенства при некотором α достигает своего минимума. Это значение α будет обеспечивать почти наилучшее приближение $\hat{u}_\alpha$ к точному нормальному псевдорешению u_0.

Предположим, что входные данные системы заданы с малой абсолютной ошибкой порядка ε. Если точная система (16.1) совместна, то $Pu_0-d=0$. В этом случае правая часть (17.14) по характеру зависимости от α и ε есть функция вида $\alpha+\varepsilon+\varepsilon/\alpha^{1/2}$. При $\alpha=\varepsilon^{2/3}$ она принимает значение порядка $\varepsilon^{2/3}$. Если же точная система не имеет ни одного решения, то $Pu_0-d\neq 0$ и правая часть (17.14) есть функция вида $\alpha+\varepsilon/\alpha+\varepsilon\,\alpha^{1/2}$. При $\alpha=\varepsilon^{1/2}$ она принимает значение порядка $\varepsilon^{1/2}$.

Таким образом, *если входные данные системы заданы с точностью порядка ε, то при некотором значении α вектор $\hat{u}_\alpha$ приближает нормальное псевдорешение u_0 с точностью порядка $\varepsilon^{2/3}$ в случае разрешимости исходной системы и с точностью порядка $\varepsilon^{1/2}$ в противном случае.*

Параметр α, обеспечивающий необходимое приближение $\hat{u}_\alpha$, не может быть найден лишь по возмущенной системе. Для его определения требуется привлечение дополнительных сведений о точной задаче.

УПРАЖНЕНИЯ

1. Доказать, что для любой матрицы A и $\alpha > 0$ справедливы неравенства

$$\|(A^*A+\alpha E)^{-1}A^*\|_{2,E} \leqslant \|A^+\|_{2,E},$$

$$\|(A^*A+\alpha E)^{-1}A^*\|_{2,E} \leqslant \frac{\|E\|_{2,E}}{2\alpha}.$$

2. Доказать, что величины γ, η в (17.6) совпадают с евклидовыми нормами нормальных решений систем (9.9).

3. Доказать, что разность $u_\alpha - u_\beta$ удовлетворяет уравнению

$$(P^*P+\alpha E)(P^*P+\beta E)(u_\alpha - u_\beta) = (\beta-\alpha)P^*d.$$

4. Пусть матрица системы имеет группу больших сингулярных чисел, а остальные сингулярные числа соизмеримы с точностью входных данных или меньше. Какую точность может обеспечить в этом случае выбор вектора $\hat{u}_\alpha$? Сравнить полученные результаты с соответствующими результатами предыдущего параграфа.

5. Рассмотрим класс систем с ограниченными величинами γ, η в (17.6). Какую точность в случае самой «плохой» системы из этого класса может обеспечить выбор вектора $\hat{u}_\alpha$? Сравнить полученные результаты с соответствующими результатами предыдущего параграфа.

6. Достигаются ли на каких-либо системах порядки точности $\varepsilon^{2/3}$ и $\varepsilon^{1/2}$?

7. Есть ли основания опасаться потери точности нормального псевдорешения системы из-за несовместности, появившейся в результате влияния ошибок округления?

8. Получить независимо от (17.14) оценку нормы возмущения нормального псевдорешения при возмущении лишь правой части системы. Сравнить эту оценку с (17.14).

9. Зависит ли точность приближения $\hat{u}_\alpha$ к u_0 от распределения сингулярных чисел матрицы системы?

ГЛАВА III

ВСПОМОГАТЕЛЬНЫЕ АЛГЕБРАИЧЕСКИЕ ОПЕРАЦИИ

Современные численные методы линейной алгебры весьма разнообразны по своим вычислительным схемам. Несмотря на это, большинство из них основано на последовательном выполнении ряда простых алгебраических операций, общее число которых относительно невелико. Это в первую очередь линейные преобразования векторов, двухсторонние преобразования матриц, вычисление матриц преобразований и т. п. Поэтому мы начнем детальное исследование влияния ошибок округления в численных методах с изучения именно таких операций.

Как правило, вычислительные схемы исследуемых алгорифмов будем выбирать такими, чтобы они были устойчивыми как в случае вещественных, так и в случае комплексных вычислений. Однако анализ ошибок будем проводить только для вещественного случая. Схемы для комплексных вычислений исследуются аналогично, при этом в оценках меняются лишь числовые коэффициенты.

§ 18. Преобразование вращения

Пусть на плоскости Oxy задан вектор a своими прямоугольными координатами u, v. Построим вектор a', повернув вектор a вокруг точки O на угол α против часовой стрелки. Обозначим через u', v' координаты вектора a'. Из курса аналитической геометрии известно, что координаты u', v' связаны с координатами u, v такими соотношениями:

$$u' = u\cos\alpha - v\sin\alpha, \quad v' = u\sin\alpha + v\cos\alpha. \tag{18.1}$$

Если через T обозначим матрицу второго порядка

$$T = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix}, \tag{18.2}$$

то соотношения (18.1) в матричной записи означают, что

$$a' = Ta. \tag{18.3}$$

Матрица (18.2) называется *матрицей вращения*, а преобразования вида (18.3) — *преобразованиями вращения*; угол α называется *углом поворота*.

Матрица вращения является ортогональной при любом α. Но среди чисел $\cos\alpha$, $\sin\alpha$ одно или оба числа, как правило, будут иррациональными. Поэтому в общем случае матрицу вращения нельзя точно представить в ЭВМ, даже если для $\cos\alpha$, $\sin\alpha$ имеются явные формулы. Следовательно, в реальных условиях будем иметь дело с матрицами $\tilde{T}$ вида

Рис. 18.1.

$$\tilde{T} = \begin{bmatrix} \tilde{c} & -\tilde{s} \\ \tilde{s} & \tilde{c} \end{bmatrix}, \tag{18.4}$$

где $\tilde{c}$, $\tilde{s}$ — некоторые действительные числа, полученные тем или иным способом. Заметим, что матрица $\tilde{T}$ лишь множителем τ отличается от ортогональной матрицы, где $\tau = (\tilde{c}^2 + \tilde{s}^2)^{1/2}$, и будет ортогональной тогда и только тогда, когда $\tau = 1$.

Таким образом, в реальных условиях вместо выражений (18.1) придется вычислять выражения

$$u' = u\tilde{c} - v\tilde{s}, \quad v' = u\tilde{s} + v\tilde{c}. \tag{18.5}$$

Этот процесс реализуется на ЭВМ вполне устойчиво. Обозначим

$$\mathrm{fl}(\tilde{T}a) - \tilde{T}a = f \tag{18.6}$$

и пусть f_1, f_2 — координаты вектора f. Имеем

$$\begin{aligned}
p &= \mathrm{fl}(u\tilde{c}) = u\tilde{c}(1+\varepsilon_1),\\
q &= \mathrm{fl}(v\tilde{s}) = v\tilde{s}(1+\varepsilon_2),\\
r &= \mathrm{fl}(p-q) = (u\tilde{c}(1+\varepsilon_1) - v\tilde{s}(1+\varepsilon_2))(1+\varepsilon_3),\\
m &= \mathrm{fl}(u\tilde{s}) = u\tilde{s}(1+\varepsilon_4),\\
n &= \mathrm{fl}(v\tilde{c}) = v\tilde{c}(1+\varepsilon_5),\\
l &= \mathrm{fl}(m+n) = (u\tilde{s}(1+\varepsilon_4) + v\tilde{c}(1+\varepsilon_5))(1+\varepsilon_6).
\end{aligned} \tag{18.7}$$

Предположим сначала, что среди величин ε_i нет равных -1, т. е. справедливы оценки $|\varepsilon_i| \leqslant (1/2)\,p^{t-1}$ для всех i. Тогда получаем

$$|f_1| \lesssim (|u\tilde{c}| + |v\tilde{s}|)\,p^{t+1}, \quad |f_2| \lesssim (|u\tilde{s}| + |v\tilde{c}|)\,p^{-t+1},$$

откуда следует, что

$$\|f\|_E \lesssim \sqrt{2}\,\tau p^{-t+1}\|a\|_E.$$

Если среди ε_i есть равные -1, то это означает, что соответствующие вычисляемые величины по модулю не превосходят минимального положительного числа ω, которое можно представить в ЭВМ. Чисел ε_i, равных -1, может быть в (18.7) не более четырех. Поэтому окончательно оценка для $\|f\|_E$ будет такой:

$$\|f\|_E \lesssim \sqrt{2}\,\tau p^{-t+1}\|a\|_E + 2\sqrt{2}\,\omega. \tag{18.8}$$

Эта формула показывает применение прямого анализа ошибок к исследованию процесса вычисления выражений (18.5). Полученный результат можно истолковать и с точки зрения обратного анализа ошибок. Из (18.6) вытекает, что

$$\mathrm{fl}(\tilde{T}a) = \tilde{T}(a+\varepsilon), \tag{18.9}$$

где $\varepsilon = \tilde{T}^{-1}f$. Принимая во внимание вид матрицы $\tilde{T}$ и оценку (18.8), находим, что

$$\|\varepsilon\|_E \lesssim \sqrt{2}\,p^{-t+1}\|a\|_E + 2\sqrt{2}\,\tau^{-1}\omega. \tag{18.10}$$

Итак, вектор, реально вычисленный по формулам (18.5), совпадает с вектором, точно вычисленным по тем же формулам, но исходя из возмущенного вектора $a+\varepsilon$, где эквивалентное возмущение ε удовлетворяет условию (18.10).

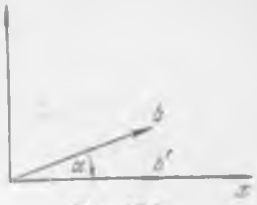

Рис. 18.2.

Преобразования вращения встречаются в самых различных вычислительных алгорифмах линейной алгебры. При этом один из важнейших случаев определения угла α связан со следующей задачей. Пусть на плоскости Oxy задан ненулевой вектор b своими прямоугольными координатами x, y. Повернем его вокруг точки O на такой угол α, чтобы по-

лученный вектор b' оказался лежащим на координатной оси Ox. Согласно (18.1) это означает, что для угла α должно выполняться соотношение $x\sin\alpha + y\cos\alpha = 0$. Поэтому можно, например, считать, что

$$\cos\alpha = \frac{x}{(x^2+y^2)^{1/2}}, \quad \sin\alpha = -\frac{y}{(x^2+y^2)^{1/2}}. \tag{18.11}$$

Полученные формулы определяют лишь одно из возможных значений для $\cos\alpha$, $\sin\alpha$. Однако во всех известных вычислительных задачах этих значений бывает достаточно.

Прямое вычисление $\cos\alpha$, $\sin\alpha$ по формулам (18.11) невозможно, если вектор b нулевой, и заведомо неустойчиво, если вектор b достаточно мал. Поэтому реальные вычисления будем выполнять по измененным формулам. Обозначим $z = \max\{|x|, |y|\}$. Если $z = 0$, то положим

$$\cos\alpha = 1, \quad \sin\alpha = 0. \tag{18.12}$$

Если же $z \neq 0$, то вычисляем $x_1 = x/z$, $y_1 = y/z$ и далее

$$\cos\alpha = \frac{x_1}{(x_1^2+y_1^2)^{1/2}}, \quad \sin\alpha = -\frac{y_1}{(x_1^2+y_1^2)^{1/2}}.$$

Оценим теперь влияние ошибок округления на вычисления по этим формулам. Пусть $z \neq 0$ и предположим для определенности, что $z = |x|$. Тогда

$$\cos\alpha = \operatorname{sign} x_1 \frac{1}{(1+y_1^2)^{1/2}}, \quad \sin\alpha = -\frac{y_1}{(1+y_1^2)^{1/2}}, \tag{18.13}$$

при этом очевидно, что $|y_1| \leqslant 1$. Имеем

$$\begin{aligned}
&l = \mathrm{fl}\left(\frac{y}{z}\right) = \frac{y}{z}(1+\varepsilon_1),\\
&p = \mathrm{fl}(l^2) = l^2(1+\varepsilon_2),\\
&q = \mathrm{fl}(1+p) = (1+l^2)(1+\varepsilon_3),\\
&r = \mathrm{fl}(q^{1/2}) = q^{1/2}(1+\varepsilon_4),\\
&m = \mathrm{fl}\left(\frac{1}{r}\right) = \frac{1}{r}(1+\varepsilon_5),\\
&n = \mathrm{fl}\left(\frac{l}{r}\right) = \left(\frac{l}{r}\right)(1+\varepsilon_6),\\
&\tilde{c} = \operatorname{sign} x \cdot m,\\
&\tilde{s} = -n,
\end{aligned} \tag{18.14}$$

вычисляя тем самым элементы матрицы (18.4).

Независимо от исходного вектора b, полученная матрица $\tilde{T}$ будет близка к ортогональной. В самом деле, пусть $|l| < \omega^{1/2}$; тогда $p = 0$ и, не ограничивая общности, можно считать, что $\varepsilon_i = 0$ для $i \geqslant 3$. В этом случае

$$\tilde{c}^2 = 1, \quad \tilde{s} = -l.$$

Если обозначить

$$\tau = 1 + \nu, \tag{18.15}$$

то для ν получаем оценку

$$|\nu| \lesssim \omega/2. \tag{18.16}$$

Если же $|l| \geqslant \omega^{1/2}$, то для всех $i \geqslant 2$

$$|\varepsilon_i| \lesssim (1/2)\, p^{i-1}.$$

Легко проверить, что теперь

$$|\nu| \lesssim (5/4)\, p. \tag{18.17}$$

С учетом (4.7), эта оценка включает в себя оценку (18.16), поэтому соотношение (18.17) имеет место всегда, если, конечно, матрица вращения вычисляется согласно (18.12), (18.13).

Если все вычисления осуществляются точно, то образом вектора b является вектор, имеющий координаты $(x^2+y^2)^{1/2}$ и 0. В качестве первой координаты вычисленного вектора мы возьмем

$$\mathrm{fl}_2\left(zr\left(\tilde{c}^2+\tilde{s}^2\right)\right) = zr\left(\tilde{c}^2+\tilde{s}^2\right)(1+\varepsilon_7), \tag{18.18}$$

а вторую положим равной нулю. Легко проверить, что такой вектор является образом вектора

$$zr(1+\varepsilon_7)\begin{bmatrix}\tilde{c}\\ \tilde{s}\end{bmatrix} = (1+\varepsilon_7)\begin{bmatrix}x(1+\varepsilon_3)\\ y(1+\varepsilon_1)(1+\varepsilon_6)\end{bmatrix} \tag{18.19}$$

при точном линейном преобразовании с матрицей, вычисленной согласно (18.14). Как уже отмечалось выше, величины ε_5, ε_6 не равны -1, величина же ε_7 не равна -1, так как $r \geqslant 1$, $\tilde{c}^2+\tilde{s}^2 \cong 1$. Если $\varepsilon_1 \neq -1$, то вектор (18.19) отличается от исходного вектора b на вектор e, где

$$\|e\|_E \lesssim \frac{\sqrt{2}+\sqrt{5}}{1-p}\, p^{t+1}\, \|b\|_E.$$

Если же $\varepsilon_1 = -1$, то это означает, что $|y| < |x|\omega$, и поэтому, с учетом (4.7), теперь имеем

$$\|e\|_E \leqslant \frac{\sqrt{5}}{2} p^{-t+1} \|b\|_E. \tag{18.20}$$

Формула (18.18) несколько сложна. Однако заметим, что в численных методах она встречается относительно редко. Общий уровень ошибок практически не изменится, если в качестве первой координаты вычисленного вектора взять $\mathrm{fl}(zr)$ вместо (18.18).

Подведем итоги выполненных исследований. Итак, пусть заданы два вектора a, b, и по вектору b вычисляется матрица вращения T согласно формулам (18.12), (18.13). Если вычисления выполняются по алгорифму (18.12), (18.14), то реально полученная матрица $\tilde{T}$ имеет вид (18.4) и будет отличаться от ортогональной матрицы множителем $\tau = 1 + \nu$, где для ν справедлива оценка (18.17). Преобразование вектора a с помощью матрицы $\tilde{T}$ согласно формулам (18.5) по алгорифму (18.7) или преобразование вектора b по тем же формулам с вычислением единственной ненулевой координаты согласно (18.18) эквивалентно точному преобразованию по формулам (18.5) возмущенных векторов. Эквивалентное возмущение e для вектора a удовлетворяет неравенству (18.10), для вектора b — неравенству (18.20) и, следовательно, для любого вектора c, включая тот, по которому строилась матрица $\tilde{T}$, — неравенству

$$\|e\|_E \leqslant \sqrt{2}\, p^{-t+1} \|c\|_E + 2\sqrt{2}\omega. \tag{18.21}$$

Аналогичные результаты могут быть получены и для комплексных векторов. В этом случае вместо матрицы вращения в преобразовании (18.3) берется унитарная матрица T вида

$$T = \begin{bmatrix} \bar{c} & -\bar{s} \\ s & c \end{bmatrix},$$

где c, s — комплексные числа такие, что $|c|^2 + |s|^2 = 1$, а черта означает комплексное сопряжение. Условие обращения в нуль второй координаты вектора $b' = Tb$ снова приводит к уравнению $xs + yc = 0$, откуда заключаем, что в качестве чисел c, s можно, например, взять

$$c = \frac{x}{(|x|^2 + |y|^2)^{1/2}}, \quad s = -\frac{y}{(|x|^2 + |y|^2)^{1/2}}.$$

Все дальнейшие исследования, по существу, повторяют выполненные исследования для вещественного случая. В аналогичных оценках меняются лишь числовые константы.

Преобразования с матрицами вращения второго порядка редко используются в численных методах, однако весьма часто применяются преобразования с матрицами вида

$$T_{ij} = \begin{bmatrix} & & i_{\text{стб}} & & j_{\text{стб}} & \\ 1 & & & & & 0 \\ & \ddots & & & & \\ & & \cos\alpha & \cdots & -\sin\alpha & \cdots & i_{\text{стр}} \\ & & \vdots & & \vdots & & \\ & & \sin\alpha & \cdots & \cos\alpha & \cdots & j_{\text{стр}} \\ & & & & & \ddots & \\ 0 & & & & & & 1 \end{bmatrix}, \tag{18.22}$$

которые отличаются от единичной матрицы лишь четырьмя элементами, расположенными на пересечении строк и столбцов с номерами i, j. Конечно, все проведенные выше исследования остаются в силе и для этих матриц. Матрицы (18.22) также называются *матрицами вращения*, а соответствующие преобразования — *преобразованиями вращения*. Изменения для комплексного случая очевидны.

УПРАЖНЕНИЯ

1. Доказать, что для реально вычисленной матрицы вращения $\tilde{T}$ второго порядка матрица $\tilde{T}\tilde{T}^*$ является скалярной.

2. Доказать, что реально вычисленная матрица вращения $\tilde{T}$ второго порядка удовлетворяет соотношению

$$\|\tilde{T}\tilde{T}^* - E\|_2 \leqslant \frac{5}{2} p^{-t+1}.$$

3. Доказать, что реально вычисленная матрица вращения $\tilde{T}$ асимптотически близка к точно вычисленной матрице T.

4. Получить оценку для отклонения $\|\tilde{T} - T\|$.

5. Пусть A — произвольная матрица и $\tilde{T}_{ij}$ — реально вычисленная матрица вращения вида (18.22). Исследовать вид матриц эквивалентных возмущений при вычислении произведений $\tilde{T}_{ij}A$ и $A\tilde{T}_{ij}$.

6. Пусть в условиях упражнения 5 имеем

$$\mathrm{fl}(\tilde{T}_{ij}A) = \tilde{T}_{ij}(A + M), \quad \mathrm{fl}(A\tilde{T}_{ij}) = (A + N)\tilde{T}_{ij}.$$

Доказать, что с точностью до членов порядка ω

$$\|M\|_E \leqslant p^{-t+1}\left(2\sum_k (a_{ik}^2+a_{jk}^2)\right)^{1/2},$$

$$\|N\|_E \leqslant p^{-t+1}\left(2\sum_k (a_{ki}^2+a_{kj}^2)\right)^{1/2},$$

где a_{kl} — элементы матрицы A.

§ 19. Последовательность преобразований вращения

Оценим теперь эквивалентное возмущение при выполнении последовательности преобразований вращения. Рассмотрим n-мерный вектор z с координатами $z_1, \ldots, z_n$ и последовательность $T_{i_1 j_1}, \ldots, T_{i_N j_N}$ матриц вращения. Пусть $\tilde{T}_{i_1 j_1}, \ldots, \tilde{T}_{i_N j_N}$ реально заданные матрицы. Мы не будем интересоваться способом их вычисления, но предположим, что все они удовлетворяют условию (18.17). Обозначим

$$z_k = fl\left(\tilde{T}_{i_k j_k} z_{k-1}\right), \quad z_0 = z,$$

для $1 \leqslant k \leqslant n$. Ясно, что

$$z_k = \tilde{T}_{i_k j_k}(z_{k-1} + \varepsilon_{k-1}), \tag{19.1}$$

где ε_{k-1} — эквивалентное возмущение преобразования вращения с матрицей $\tilde{T}_{i_k j_k}$. Последовательно используя соотношение (19.1), находим

$$z_N = \left(\tilde{T}_{i_N j_N} \ldots \tilde{T}_{i_1 j_1}\right)(z + E).$$

Таким образом, вектор z_N, реально полученный после выполнения N последовательных преобразований вращения с матрицами $\tilde{T}_{i_1 j_1}, \ldots, \tilde{T}_{i_N j_N}$, можно рассматривать как вектор, полученный после точного выполнения тех же преобразований над возмущенным вектором $z+E$. Оценка для $\|E\|_E$ определяется многими факторами, однако уже сразу можно предположить, что она будет зависеть от вида последовательности пар индексов матриц вращения.

Одна из оценок выводится совсем просто. Пусть $z_{k1}, \ldots, z_{kn}$ — координаты вектора z_k. Согласно (18.21) будем иметь

$$\|\varepsilon_k\|_E \leqslant \sqrt{2}\, p^{-t+1}\left(z_{ki_k}^2 + z_{kj_k}^2\right)^{1/2} + 2\sqrt{2}\,\omega.$$

Так как матрицы $\tilde{T}_{i_k j_k}$ близки к ортогональным, то

$$\|E\|_E \lesssim \sum_{k=0}^{N-1} \|e_k\|_E, \tag{19.2}$$

поэтому всегда выполняется неравенство

$$\|E\|_E \lesssim \sqrt{2}\, p^{-t+1} \sum_{k=0}^{N-1} \left(z_{ki_k}^2 + z_{kj_k}^2\right)^{1/2} + 2\sqrt{2}\, N\omega. \tag{19.3}$$

К тому же, $\|z_k\|_E \cong \|z\|_E$, следовательно,

$$\|E\|_E \lesssim \sqrt{2}\, N p^{-t+1} \|z\|_E + 2\sqrt{2}\, N\omega. \tag{19.4}$$

Оценка (19.4) справедлива для любой последовательности из матриц вращения и для некоторых последовательностей она почти достигается. Предположим, например, что вектор z имеет лишь одну ненулевую координату в позиции с номером i_1 и все матрицы вращения близки к единичным. Тогда для любой последовательности индексов вида

$$i_1, j_1; i_1, j_2; \ldots; i_1, j_N$$

евклидова норма эквивалентного возмущения E будет совпадать с точностью до констант с правой частью (19.4). Оценка (19.4) почти достигается и для последовательностей

$$i_1, j_1; j_1, j_2; \ldots; j_{N-1}, j_N,$$

если только все матрицы вращения близки к матрицам перестановок. Эти примеры показывают, что оценка (19.4) практически неулучшаема для всех последовательностей матриц вращения, у которых любые две соседние матрицы имеют хотя бы один общий индекс. Такие последовательности мы будем называть *сильно связанными*.

В одном весьма важном случае удается получить оценку эквивалентного возмущения E, зависящую от N очень слабо. Назовем последовательность матриц вращения *несвязанной*, если все индексы матриц различны. Конечно, последовательность может быть несвязанной лишь тогда, когда $N \leqslant n/2$. Для несвязанных последовательностей ошибки, возникшие после выполнения любого преобразования, не будут меняться при всех последующих преобразованиях. Более того, *результат выполнения несвязанной последовательности преобразований, включая всю*

совокупность ошибок округления, не зависит вообще от порядка выполнения самих преобразований. Теперь вместо неравенства (19.2) в действительности будет иметь место асимптотическое равенство

$$\|E\|_E \cong \sum_{k=0}^{N-1} \|e_k\|_E,$$

а вместо соотношения (19.3) — неравенство

$$\|E\|_E \lesssim 2p^{-2(t-1)} \sum_{k=0}^{N-1} (z_{i_k}^2 + z_{l_k}^2) + 8p^{t+1}\omega \sum_{k=0}^{N-1} (z_{i_k}^2 + z_{l_k}^2)^{1/2} + 8N\omega^2.$$

Так как индексы у матриц вращения не совпадают, а сами матрицы близки к ортогональным, то

$$z_{i_k}^2 + z_{l_k}^2 \cong z_{i_k}^2 + z_{l_k}^2,$$

поэтому

$$\sum_{k=0}^{N-1} (z_{i_k}^2 + z_{l_k}^2) \leqslant \|z\|_E^2.$$

В соответствии с неравенством Коши — Буняковского

$$\sum_{k=0}^{N-1} (z_{i_k}^2 + z_{l_k}^2)^{1/2} \leqslant \sqrt{N} \left(\sum_{k=0}^{N-1} (z_{i_k}^2 + z_{l_k}^2) \right)^{1/2},$$

и окончательно находим, что теперь

$$\|E\|_E \lesssim \sqrt{2}\, p^{-t+1} \|z\|_E + 2\sqrt{2N}\,\omega. \qquad (19.5)$$

Если последовательность матриц вращения можно разбить на k групп таких, что внутри каждой группы матрицы вращения не имеют одинаковые индексы, то из (19.5) вытекает, что

$$\|E\|_E \lesssim \sqrt{2k}\, p^{-t+1} \|z\|_E + 2\sqrt{2kN}\,\omega. \qquad (19.6)$$

Заметим, что выполнение любой последовательности из N преобразований вращения можно трактовать как выполнение N несвязанных последовательностей, каждая из которых содержит лишь одно преобразование. При этом для суммарного эквивалентного возмущения E справедливы как неравенство (19.4), так и неравенство (19.6). В данном случае обе оценки совпадают и, как мы уже отмечали, они практически неулучшаемы для сильно связанных последовательностей.

Однако в общем случае выполнение преобразований вращения можно сводить к несвязанным последовательностям не единственным способом, что видно на примере самой несвязанной последовательности. Чтобы на основе формулы (19.6) получить наилучшую оценку эквивалентного возмущения для любой последовательности преобразований вращения, необходимо определить минимальное число несвязанных последовательностей, к которым сводится исходная последовательность.

Предположим, что в последовательности матриц вращения две соседние матрицы не имеют общих индексов. Тогда результат преобразований не изменится, если эти матрицы поменять местами. Если одну последовательность матриц вращения можно получить из другой с помощью перестановок соседних матриц, не имеющих общих индексов, то такие последовательности будем называть *эквивалентными*. Ясно, что результат выполнения эквивалентных последовательностей преобразований будет одним и тем же, включая всю совокупность ошибок округления.

Среди эквивалентных между собой последовательностей существует такая, которая распадается на наименьшее число несвязанных последовательностей. Это наименьшее число мы назовем *индексом эквивалентности*. Очевидно, что знание индекса позволяет получить на основе формулы (19.6) наилучшую оценку возмущения E для всего множества эквивалентных последовательностей. Так как правая часть (19.6) не зависит от углов поворота, то в действительности индекс эквивалентности дает возможность оценить возмущение для всех последовательностей матриц вращения, эквивалентных с точностью до выбора углов поворота.

Вычислять индекс и даже сравнивать различные последовательности можно с помощью процесса преобразования самих последовательностей к некоторой *канонической форме*.

Пусть последовательность матриц вращения расположена в строке слева направо. Выберем в этой последовательности все матрицы, каждая из которых не имеет

слева ни одной матрицы с общими индексами. Предположим, что таких матриц оказалось s_1. Ясно, что они образуют несвязанную последовательность и с помощью перестановок с соседними матрицами все их можно поставить в любом порядке первыми в заданной последовательности, сохранив при этом относительное расположение остальных матриц. Выберем далее среди оставшихся матриц все матрицы, не имеющие слева ни одной матрицы с общими индексами, кроме матриц из выбранной группы. Пусть таких матриц оказалось s_2. Переставим их в каком-либо порядке вслед за матрицами первой группы. Продолжая этот процесс, приведем исходную последовательность к эквивалентной, распадающейся на несвязанные последовательности из s_1, s_2, ..., s_k матриц вращения.

Число k равно индексу эквивалентности. В самом деле, рассмотрим какую-нибудь эквивалентную последовательность с минимальным числом несвязанных последовательностей. Применив описанное выше преобразование, мы не увеличим в ней их число. Пусть несвязанные последовательности состоят теперь из r_1, r_2, ..., r_l матриц вращения и при этом $l<k$. Если какая-либо матрица не имеет слева ни одной матрицы с общими индексами, то такие матрицы не появятся слева ни при каких эквивалентных перестановках. Но отсюда вытекает, что $r_1=s_1$, следовательно, $r_2=s_2$ и т. д. Поэтому $k=l$.

Исследуем две последовательности матриц вращения, наиболее часто используемые в численных методах. Обе последовательности называются *циклическими* и описываются одной и той же совокупностью пар индексов

$$
\begin{array}{llll}
1,\ 2; & & & \\
1,\ 3; & 2,\ 3; & & \\
1,\ 4; & 2,\ 4; & 3,\ 4; & \\
\dots\dots\dots\dots\dots\dots\dots\dots & & & \\
1,\ m+1; & 2,\ m+1; & 3,\ m+1; & \dots;\ m,\ m+1; \\
\dots\dots\dots\dots\dots\dots\dots\dots & & & \\
1,\ n; & 2,\ n; & 3,\ n; & \dots;\ m,\ n,
\end{array}
$$

где $m<n$. В первом случае совокупность пар индексов упорядочивается по строкам, во втором — по столбцам, причем сами строки и столбцы упорядочиваются сверху вниз и слева направо.

Глядя на эти последовательности, трудно обнаружить существенную связь между ними. Однако приведя их к канонической форме, замечаем, что упорядоченность пар индексов становится одинаковой и имеет вид

$$
\begin{array}{llll}
1,\ 2; & & & \\
1,\ 3; & & & \\
1,\ 4; & 2,\ 3; & & \\
1,\ 5; & 2,\ 4; & & \\
1,\ 6; & 2,\ 5; & 3,\ 4; & \\
\dots\dots\dots\dots & & & \\
1,\ n; & 2,\ n-1; & 3,\ n-2; & \dots;\ m,\ n-m+1; \\
 & 2,\ n; & 3,\ n-1; & \dots;\ m,\ n-m+2; \\
 & & 3,\ n; & \dots;\ m,\ n-m+3; \\
 & & & \dots\dots\dots \\
 & & & m,\ n;
\end{array}
\tag{19.7}
$$

где каждая строка соответствует несвязанной последовательности.

Таким образом, *с точностью до выбора углов поворота циклические последовательности эквивалентны*. Как вытекает из (19.7), их индекс равен $m+n-2$. Поэтому согласно (19.6) для эквивалентных возмущений этих последовательностей справедливо одно и то же неравенство

$$
\|\mathrm{E}\|_E \leqslant \{2(m+n-2)p^{-t+1}\|z\|_E + \\
+2(m(m+n-2)(2n-m+1))^{1/2}\omega. \tag{19.8}
$$

В дальнейшем мы неоднократно воспользуемся полученными оценками для изучения самых различных численных методов. При этом практически всегда будет выполняться соотношение $\|z\|_E > \omega^{1/2}$. В этих условиях оценки упрощаются. Пусть выполняется N произвольных преобразований вращения. Как следует из (19.4)

$$
\|\mathrm{E}\|_E \leqslant \{2Np^{-t+1}\|z\|_E. \tag{19.9}
$$

Предположим, что последовательность матриц вращения состоит из k несвязанных групп. Тогда из (19.6) получаем, что

$$
\|\mathrm{E}\|_E \leqslant \{2kp^{-t+1}\|z\|_E. \tag{19.10}
$$

Для обеих циклических последовательностей согласно (19.8) имеем

$$
\|\mathrm{E}\|_E \leqslant \{2(m+n-2)p^{-t+1}\|z\|_E, \tag{19.11}
$$

если $m < n$, и

$$\|E\|_E \lesssim \sqrt{2}\,(2n-3)\,p^{-t+1}\|z\|_E, \qquad (19.12)$$

если $m = n-1$.

Полученные оценки подтвердили наше предположение о том, что общий эффект влияния ошибок округления зависит не только от числа выполненных преобразований вращения, но и от того, в какой последовательности преобразуются элементы. В некоторых задачах мы сможем в известной мере выбирать эту последовательность и, следовательно, строить лучшие по точности методы.

Рассмотрим в качестве примера одну простую задачу. Пусть заданы n-мерный вектор s и координатный вектор $e = (1, 0, \ldots, 0)'$. Постараемся подобрать такую последовательность матриц вращения, чтобы умножение на их произведение U переводило вектор s в вектор, коллинеарный e, т. е.

$$Us = \alpha e. \qquad (19.13)$$

Предположим, что какой-либо вектор имеет ненулевые координаты в позициях i и j. Выбрав соответствующим образом угол поворота, можно перевести этот вектор путем умножения на матрицу вращения T_{ij} в такой вектор, у которого одна из координат в позициях i или j будет исключена, т. е. равна нулевой. Искомая совокупность матриц вращения, определяющая преобразование (19.13), может быть построена на основе последовательного исключения всех координат вектора s, кроме первой.

Порядок исключения не является однозначным и, выбирая его подходящим образом, можно уменьшить общее влияние ошибок округления. В вычислительной практике наиболее часто координаты исключаются подряд, начиная со второй, путем умножения на матрицы $T_{12}, T_{13}, \ldots, T_{1n}$. Эта последовательность является сильно связанной и согласно оценке (19.9) эквивалентное возмущение E будет удовлетворять неравенству

$$\|E\|_E \lesssim \sqrt{2}\,(n-1)\,p^{-t+1}\|s\|_E. \qquad (19.14)$$

Теперь исключим те же координаты в другом порядке. Сначала, умножая на матрицы $T_{12}, T_{34}, T_{56}, \ldots$, исключим координаты с номерами 2, 4, 6, ... Затем, умножая на матрицы $T_{13}, T_{57}, T_{9,11}, \ldots$, исключим координаты

с номерами 3, 7, 11, ... Далее, умножая на матрицы $T_{15}, T_{9,13}, T_{17,21}, \ldots$, исключим координаты с номерами 5, 13, 21, ... Ясно, что внутри каждой такой группы матрицы вращения не имеют одинаковых индексов, а всего групп будет не более $\log_2(2n)$. Согласно оценке (19.10), в этом случае будем иметь

$$\|E\|_E \leqslant \sqrt{2}\log_2(2n)\,p^{-t+1}\|s\|_E. \qquad (19.15)$$

Эта оценка существенно лучше оценки (19.14)

УПРАЖНЕНИЯ

1. Рассмотрим вектор s размерности n. Предположим, что при каждом умножении на матрицу вращения исключается наименьшая по модулю его координата. Пусть вторая преобразуемая координата является наименьшей по модулю из оставшихся ненулевых координат. Доказать, что эквивалентное возмущение E удовлетворяет в этом случае соотношению

$$\|E\|_E \lesssim 2\sqrt{2(n-1)}\,p^{-t+1}\|s\|_E. \qquad (19.16)$$

2. Предположим, что при каждом умножении исключается наибольшая по модулю координата. Доказать, что теперь эквивалентное возмущение удовлетворяет соотношению

$$\|E\|_E \lesssim \sqrt{2}\,(n-1)\,p^{-t+1}\|s\|_E. \qquad (19.17)$$

3. Сравнить между собой оценки (19.14)—(19.17).

4. Пусть прямоугольная матрица умножается слева (справа) на последовательность матриц вращения, соответствующую любой из оценок (19.9)—(19.12). Доказать, что в этом случае для эквивалентного возмущения имеют место те же оценки (19.9)—(19.12) с заменой, конечно, нормы вектора нормой матрицы.

5. Справедливо ли утверждение упражнения 4 для оценок (19.16), (19.17)?

6. Пусть $\bar{R}$ есть точное произведение реально вычисленных матриц вращения, соответствующих любой из циклических последовательностей. Доказать, что существует такая ортогональная матрица R, что

$$\|R-\bar{R}\|_E \lesssim \frac{5}{4}(m+n-2)\,n\,p^{-t+1}. \qquad (19.18)$$

§ 20. Преобразование отражения

Предположим, что в пространстве $Oxyz$ задана плоскость π с единичным нормальным вектором w. Возьмем произвольный вектор z и преобразуем его по правилу отраже-

ния от плоскости π. Если z представить в виде суммы $z = x + y$, где x перпендикулярен w, а y — коллинеарен w, то отраженный вектор z' будет иметь такой вид: $z' = x - y$.

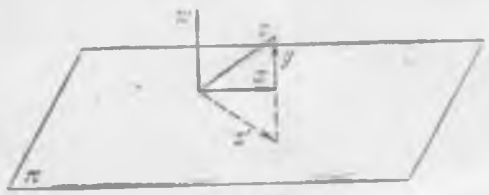

Рис. 20.1.

Преобразование $z \Rightarrow z'$ является линейным и для его матрицы U можно указать явный вид. Именно,

$$U = E - 2ww'.$$

В самом деле, если $(x, w) = 0$, а $y = \alpha w$, где α — число, то

$$Uz = z - 2ww'x - 2ww'y =$$
$$= z - 2(x, w)w - 2\alpha(w, w)w = x + y - 2\alpha w = x - y = z'.$$

Преобразование отражения имеет n-мерный аналог, причем не только вещественный, но и комплексный. Пусть w — единичный вектор, т. е. $(w, w) = 1$. Построим матрицу

$$U = E - 2ww^*$$

и рассмотрим преобразование

$$z' = Uz. \tag{20.1}$$

По аналогии с трехмерным случаем это преобразование называется *преобразованием отражения*, а его матрица — *матрицей отражения*.

В трехмерном вещественном случае преобразование отражения является ортогональным, так как, очевидно, оно сохраняет длины всех векторов. В общем случае матрица отражения не только унитарная, но и эрмитова. Действительно,

$$U^* = (E - 2ww^*)^* = E - 2w^{**}w^* = E - 2ww^* = U,$$
$$UU^* = (E - 2ww^*)^2 = E - 4ww^* + 4ww^*ww^* =$$
$$= E - 4ww^* + 4(w, w)ww^* = E.$$

Легко проверить, что преобразование (20.1) оставляет без

изменения все векторы, ортогональные w, и меняет на противоположные векторы, коллинеарные w.

Для запоминания матрицы отражения и выполнения преобразования (20.1) совсем не обязательно иметь элементы матрицы U в явном виде. Если преобразование отражения (20.1) выполнять по формуле

$$z' = z - 2(z, w)w, \tag{20.2}$$

то для его реализации достаточно знать лишь координаты вектора w.

Формула (20.2) показывает одно интересное свойство преобразования отражения. Именно, определяющий его вектор w коллинеарен разности образа и прообраза. Следовательно, он может быть восстановлен по этой разности с точностью до числового множителя, равного по модулю единице, если, конечно, сама разность не является нулевой. Заметим, что умножение вектора w на любое число, по модулю равное единице, не меняет преобразования отражения.

Один из важнейших способов построения матрицы отражения связан с ее восстановлением по образу и прообразу. Пусть заданы ненулевые векторы s и e, причем e — единичной длины. Подберем такой вектор w, чтобы соответствующее преобразование отражения переводило вектор s в вектор, коллинеарный e, т. е. $Us = \alpha e$. Искомое преобразование унитарное, поэтому $|\alpha| = \|s\|_2$. Как уже отмечалось, вектор w должен иметь вид

$$w = \frac{1}{\rho}(s - \alpha e),$$

где ρ — нормирующий множитель. Имеем

$$|\rho|^2 = (s - \alpha e, s - \alpha e) = 2((s, s) - \mathrm{Re}(s, \alpha e)).$$

Чтобы разность $s - \alpha e$ была заведомо отличной от нуля, выберем аргумент числа α так, чтобы скалярное произведение $(s, \alpha e)$ было отрицательным. Тогда $|\rho|^2 = 2(s, s - \alpha e)$. Но теперь проверяем, что

$$Us = s - 2(s, w)w = s - \frac{2}{\rho}(s, s - \alpha e)w = s - \rho w = \alpha e.$$

Наиболее часто в качестве вектора e берется один из координатных векторов. Предположим, например, что $e =$

$= (1, 0, \ldots, 0)'$. Обозначим через s_i, ω_i координаты векторов s, ω. В этом случае

$$\alpha = -|\alpha|\frac{s_1}{|s_1|}, \quad \rho^2 = 2(|\alpha|^2 + |\alpha|\cdot|s_1|)$$

и, далее,

$$\omega_1 = \frac{s_1 + |\alpha|\frac{s_1}{|s_1|}}{(2(|\alpha|^2 + |\alpha|\,|s_1|))^{1/2}}, \quad \omega_i = \frac{s_i}{(2(|\alpha|^2 + |\alpha|\,|s_1|))^{1/2}}, \quad i \geqslant 2.$$

Для повышения устойчивости реальные вычисления будем выполнять по следующей схеме. Пусть $|\alpha| \neq 0$. Определяем координаты u_i вектора

$$u = \frac{1}{|\alpha|}s$$

и положим

$$v_1 = \frac{u_1}{|u_1|}(1 + |u_1|), \quad v_i = u_i, \quad i \geqslant 2. \tag{20.3}$$

Если $u_1 = 0$, то будем считать, что $u_1/|u_1|$ есть какое-то число, по модулю равное единице. Теперь матрицу отражения можно представить в виде

$$U = E - \frac{1}{\gamma}vv^*, \tag{20.4}$$

где координаты вектора v определены согласно (20.3), а $\gamma = 1 + |u_1|$. Если $|\alpha| = 0$, то берем $v = e$, $\gamma = 1/2$. Ясно, что всегда $0{,}5 \leqslant \gamma \leqslant 2$, $1 \leqslant \|v\|_E \leqslant 2$.

Исследуем ошибки, возникающие при вычислении вектора v и числа γ. Будем считать, что $|\alpha|$ вычисляется с удвоенной точностью, т. е.

$$|\tilde{\alpha}| = \mathrm{fl}_2(|\alpha|) \equiv |\alpha|(1 + \varepsilon).$$

Если при вычислении $|\alpha|$ используется алгорифм, описанный в § 7, то

$$|\varepsilon| \lesssim \frac{1}{2}p^{-t+1}. \tag{20.5}$$

Пусть $|\tilde{\alpha}| \neq 0$; тогда находим

$$\begin{aligned} \tilde{u}_i &= \mathrm{fl}\left(\frac{s_i}{|\tilde{\alpha}|}\right) \equiv \frac{s_i}{|\tilde{\alpha}|}(1 + \eta_i), \quad i \geqslant 1, \\ \tilde{\gamma} &= \mathrm{fl}(1 + |\tilde{u}_1|) \equiv (1 + |\tilde{u}_1|)(1 + \mu), \\ \tilde{v}_1 &= \operatorname{sign} s_1 \tilde{\gamma}, \\ \tilde{v}_i &= \tilde{u}_i, \quad i \geqslant 2. \end{aligned} \tag{20.6}$$

Величина $\tilde{\gamma}$ не может быть малой, следовательно, μ удовлетворяет неравенству вида (20.5). Если же $\eta_i = -1$, то это означает, что

$$|s_i| \lesssim |\alpha|\omega_i. \tag{20.7}$$

Оценим отклонение вычисленной матрицы $\tilde{U}$ от унитарной. Предположим, что h величин η_i равны -1. Пусть

$$\eta_i' = \begin{cases} \eta_i, & \eta_i \neq -1, \\ 0, & \eta_i = -1. \end{cases}$$

Имеем

$$(\tilde{u}, \tilde{u}) = \frac{\sum_{i=1}^{n} s_i^2(1 + \eta_i)^2}{|\alpha|^2(1 + \varepsilon)^2} = \frac{\sum_{i=1}^{n} s_i^2(1 + \eta_i')^2 + \vartheta}{|\alpha|^2(1 + \varepsilon)^2}.$$

Согласно (20.7) $|\vartheta| \lesssim h|\alpha|^2\omega^2$. Далее,

$$\sum_{i=1}^{n} s_i^2(1 + \eta_i')^2 = (1 + \eta)^2 \sum_{i=1}^{n} s_i^2 = (1 + \eta)^2|\alpha|^2,$$

где η удовлетворяет неравенству вида (20.5). Поэтому

$$(\tilde{u}, \tilde{u}) = 1 + \nu, \quad |\nu| \lesssim 2p^{-t+1}.$$

Теперь находим

$$\frac{(\tilde{v}, \tilde{v})}{\tilde{\gamma}} = \frac{\tilde{\gamma}^2 + \sum_{i=2}^{n}\tilde{u}_i^2}{\tilde{\gamma}} = \tilde{\gamma} + \frac{1 - \tilde{u}_1^2 + \nu}{\tilde{\gamma}} =$$

$$= \tilde{\gamma} + \frac{\nu}{\tilde{\gamma}} + \frac{1 - |\tilde{u}_1|}{1 + \mu} \lesssim 2 + \left(2|\tilde{u}_1|\mu + \frac{\nu}{1 + |\tilde{u}_1|}\right).$$

Так как $0 \leqslant |\tilde{u}_1| \leqslant 1 + \nu$, то, принимая во внимание оценки для μ, ν, получаем

$$2|\tilde{u}_1|\mu + \frac{\nu}{1 + |\tilde{u}_1|} \lesssim \max\left\{\nu, 2\mu + \frac{\nu}{2}\right\} \lesssim 2p^{-t+1}.$$

Следовательно,

$$\frac{(\tilde{v}, \tilde{v})}{\tilde{\gamma}} = 2 + \delta, \quad |\delta| \leqslant 2p^{-t+1}. \tag{20.8}$$

Вычисленная матрица отражения

$$\tilde{U} = E - \frac{1}{\tilde{\gamma}} \tilde{v}\tilde{v}'$$

всегда эрмитова. Легко проверить, что ее собственными векторами являются вектор $\tilde{v}$ и любой вектор, ортогональный $\tilde{v}$, а собственными значениями число $-(1+\delta)$ и $n-1$ раз число $+1$. Обозначим через $\hat{U}$ унитарную матрицу, имеющую такие же собственные векторы, а собственные значения соответственно число -1 и $n-1$ раз число $+1$. Если $\tilde{U} = \hat{U} + \Delta$, то Δ есть эрмитова матрица ранга 1 и ее единственное ненулевое собственное значение равно $-\delta$. Учитывая оценку для δ, заключаем, что

$$\|\Delta\| \leqslant 2p^{-t+1} \tag{20.9}$$

как для 2-нормы, так и для евклидовой нормы. Отсюда, в частности, следует, что

$$1 - 2p^{-t+1} \leqslant \|\tilde{U}\|_2, \ \|\tilde{U}^{-1}\|_2 \leqslant 1 + 2p^{-t+1}. \tag{20.10}$$

Рассмотрим теперь влияние ошибок округления на процесс реализации преобразования отражения. Пусть это преобразование выполняется согласно формуле

$$Uz = z - \frac{(z, v)}{\gamma} v.$$

Будем считать, что при вычислении $(z, \tilde{v})/\tilde{\gamma}$ используется операция накопления. Имеем

$$\begin{aligned} r &= \mathrm{fl}_2((z, \tilde{v})/\tilde{\gamma}) = ((z, \tilde{v})/\tilde{\gamma})(1+\sigma), \\ k_i &= \mathrm{fl}(r\tilde{v}_i) = r\tilde{v}_i(1+\nu_i), \quad i \geqslant 1, \\ \tilde{z}_i' &= \mathrm{fl}(z_i - k_i) = (z_i - k_i)(1+\tau_i), \quad i \geqslant 1. \end{aligned} \tag{20.11}$$

Здесь z_i — координаты вектора z; z_i' — координаты вычисленного вектора $\tilde{U}z$. Обозначим

$$\mathrm{fl}(\tilde{U}z) - \tilde{U}z = f. \tag{20.12}$$

Если ни одна из ошибок в (20.11) не равна -1, то несложные вычисления, учитывающие (20.8), показывают, что

$$\|f\|_E \leqslant 2{,}5p^{-t+1}\|z\|_E. \tag{20.13}$$

В общем случае в правую часть оценки (20.13) добавляется слагаемое, зависящее от ω. Предположим, что $\sigma = -1$. Тогда все остальные ошибки оказываются равными нулю и мы получим

$$\|f\|_E \leqslant \sqrt{2}\,\omega. \tag{20.14}$$

Величины ν_i, τ_i не могут быть равны -1 одновременно. Если $\nu_i = -1$, то $k_i = 0$. Но отсюда вытекает, что $\tau_i = 0$. Таким образом, равенство минус единице некоторых величин ν_i, τ_i приводит к увеличению правой части (20.13) не более, чем на $\sqrt{n}\,\omega$. Принимая во внимание (20.14), заключаем, что всегда

$$\|f\|_E \leqslant 2{,}5p^{t+1}\|z\|_E + \sqrt{n}\,\omega. \tag{20.15}$$

Полученные соотношения позволяют выполнить обратный анализ ошибок. Из (20.12) следует, что

$$\mathrm{fl}(\tilde{U}z) = \tilde{U}(z+\tau), \tag{20.16}$$

где $\tau = \tilde{U}^{-1}f$. Но согласно (20.10), (20.15)

$$\|\tau\|_E \leqslant 2{,}5p^{t+1}\|z\|_E + \sqrt{n}\,\omega. \tag{20.17}$$

При точных вычислениях образом вектора s, по которому строилась матрица отражения, является вектор $s' = \alpha e$. В практических вычислениях вектор $\tilde{s}'$ находят не по формулам (20.11), а считают, что

$$\tilde{s}' = \mathrm{fl}(\tilde{U}s) = -\operatorname{sign} s_1 \cdot |\tilde{\alpha}|\, e. \tag{20.18}$$

Пусть

$$\mathrm{fl}(\tilde{U}s) = \tilde{U}(s + \hat{\rho}),$$

где $\hat{\rho}$ — эквивалентное возмущение. Оценим норму вектора $\hat{\rho}$.

Вектор $s + \hat{\rho}$ есть прообраз вектора $\tilde{s}'$ при преобразовании с матрицей $\tilde{U}$, поэтому $s + \hat{\rho} = \tilde{U}^{-1}\tilde{s}'$. Но легко проверить, что

$$\tilde{U}^{-1} = E - \frac{1}{(\tilde{v}, \tilde{v}) - \tilde{\gamma}} \tilde{v}\tilde{v}'$$

или, принимая во внимание (20.8),

$$\tilde{U}^{-1} = E - \frac{1}{\tilde{\gamma}(1+\delta)} \tilde{v}\tilde{v}^*.$$

Следовательно,

$$s + \hat{\rho} = -\operatorname{sign} s_1 |\tilde{\alpha}| \left(e - \frac{(e_1, \tilde{v})}{\tilde{\gamma}(1+\delta)} \tilde{v}\right). \qquad (20.19)$$

Обозначим через $s_i + \hat{\rho}_i$ координаты вектора $s + \hat{\rho}$. Из (20.6), (20.19) находим, что для $i \geqslant 2$

$$s_i + \hat{\rho}_i = \frac{\operatorname{sign} s_1 |\tilde{\alpha}| \tilde{v}_1 \tilde{v}_i}{\tilde{\gamma}(1+\delta)} = \frac{|\tilde{\alpha}| \tilde{u}_i}{1+\delta} = \frac{s_i(1+\eta_i)}{1+\delta}.$$

Далее,

$$s_1 + \hat{\rho}_1 =$$

$$= -\operatorname{sign} s_1 |\tilde{\alpha}| \left(1 - \frac{\tilde{v}_1^2}{\tilde{\gamma}(1+\delta)}\right) = -\operatorname{sign} s_1 |\tilde{\alpha}| \left(1 - \frac{\tilde{\gamma}}{1+\delta}\right) =$$

$$= \frac{s_1(1+\eta_1) + \operatorname{sign} s_1 \left((|\tilde{\alpha}| + |\tilde{s}_1|(1+\eta_1))\mu + |\tilde{\alpha}|\delta\right)}{1+\delta}.$$

Теперь получаем оценку

$$|\hat{\rho}_i| \leqslant \begin{cases} 2{,}5p^{i+1}|s_i|, & \text{если } \eta_i \neq -1, \\ |\alpha|\omega, & \text{если } \eta_i = -1, \end{cases}$$

для $i \geqslant 2$, и оценку

$$|\hat{\rho}_1| \leqslant \begin{cases} (2{,}5|\alpha| - |s_1|)p^{i+1}, & \text{если } \eta_1 \neq -1, \\ 2{,}5|\alpha|p^{i+1}, & \text{если } \eta_1 = -1, \end{cases}$$

для $i = 1$. Из этих оценок вытекает, что всегда

$$|\hat{\rho}|_E \leqslant 2{,}5\sqrt{2}\, p^{i+1}|s|_E. \qquad (20.20)$$

УПРАЖНЕНИЯ

1. Является ли единичная матрица матрицей отражения?
2. Какие из диагональных матриц являются матрицами отражения?
3. Доказать, что определитель любой матрицы отражения равен -1.
4. Доказать, что реально вычисленная матрица отражения $\tilde{U}$ удовлетворяет соотношению

$$|\tilde{U}\tilde{U}^* - E|_E \leqslant 4p^{i+1}.$$

5. Пусть v и γ — вектор и число, определяющие матрицу отражения (20.4). Доказать, что

$$\left|\frac{1}{\gamma} vv^*\right|_E = \left|\frac{1}{\gamma} vv^*\right|_E = 2.$$

6. Доказать, что собственные значения матрицы $E - \frac{1}{2\gamma} vv^*$ равны $+1$, кроме одного, которое равно нулю.
7. Доказать, что при умножении вектора z на матрицу $E - \frac{1}{2\gamma} vv^*$ сохраняется оценка (20.15)
8. Можно ли при умножении вектора на матрицу $E - \frac{1}{2\gamma} vv^*$ выполнить обратный анализ ошибок?

§ 21. Последовательность преобразований отражения

Выполнение последовательности преобразований отражения является составной частью многих численных методов линейной алгебры. При этом почти всегда преобразуются не все координаты вектора, а только часть из них.

Рассмотрим вектор z и построим преобразование отражения, изменяющее лишь координаты в позициях $i_1, i_2, \ldots, i_r$. Не ограничивая общности, можно считать, что это последние координаты вектора. В самом деле, пусть P — такая матрица перестановок, что все координаты, подлежащие преобразованию, являются последними для вектора Pz. Если U — искомая матрица отражения, то легко проверить, что

$$Uz = P\left(E - \frac{1}{\gamma}(Pv)(Pv)^*\right)(Pz).$$

Матрица в круглых скобках является также матрицей отражения и умножение на нее вектора Pz изменяет только последние r координат Pz.

Итак, пусть преобразование с матрицей U меняет последние r координат вектора z. Представим векторы z, v в блочном виде,

$$z = \begin{bmatrix} z' \\ z'' \end{bmatrix}, \quad v = \begin{bmatrix} v' \\ v'' \end{bmatrix},$$

где векторы z'', v'' имеют размерность r. Так как

$$Uz = z - \frac{(z, v)}{\gamma} v,$$

то для того, чтобы в векторе Uz изменялись лишь последние r координат, необходимо и достаточно, чтобы $v' = 0$. Но в этом случае матрица U будет иметь такое строение:

$$U = \left[\begin{array}{c|c} E & 0 \\ \hline 0 & E - \frac{1}{\gamma} v'' v''^* \end{array}\right]. \qquad (21.1)$$

Матрица, стоящая в нижнем правом углу, есть матрица отражения порядка r. Определяющие ее вектор v'' и число γ находятся только по изменяющимся координатам вектора z. Всюду в дальнейшем, говоря о преобразовании части координат, мы будем подразумевать в действительности умножение на матрицу отражения вида (21.1). При этом остаются в силе все полученные ранее оценки ошибок с заменой, конечно, нормы вектора z на норму вектора z'' и числа n на число r. Но так как $\|z''\|_E \leqslant \|z\|_E$ и $r \leqslant n$, то прежние оценки остаются в силе и в своем первоначальном виде.

Предположим теперь, что над вектором z выполняется N преобразований отражения с матрицами $\tilde{U}_1, \tilde{U}_2, \ldots, \tilde{U}_N$. Мы не будем сейчас интересоваться способом вычисления этих матриц, но будем считать, что выполняются оценки (20.8), (20.16), (20.20). Обозначим $z_k = fl(\tilde{U}_k z_{k-1})$ для всех k, причем $z_0 = z$. Ясно, что

$$z_k = \tilde{U}_k (z_{k-1} + \tau_{k-1}), \qquad (21.2)$$

где τ_{k-1} — эквивалентное возмущение преобразования отражения с матрицей $\tilde{U}_k$. Последовательно используя соотношение (21.2) и учитывая близость матриц $\tilde{U}_k$ к ортогональным, получаем

$$z_N = (\tilde{U}_N \ldots \tilde{U}_1)(z + \mathrm{T}),$$

где

$$\|\mathrm{T}\|_E \leqslant \sum_{k=0}^{N-1} \|\tau_k\|_E. \qquad (21.3)$$

Снова видим, что вектор z_N, реально полученный после выполнения N последовательных преобразований отражения, можно рассматривать как вектор, полученный после точного выполнения тех же преобразований над

возмущенным вектором $z + \mathrm{T}$, причем для эквивалентного возмущения T справедливо соотношение (21.3). Так как матрицы $\tilde{U}_k$ близки к ортогональным, то

$$\|z_k\|_E \cong \|z\|_E.$$

Если ни одна из матриц $\tilde{U}_1, \ldots, \tilde{U}_N$ не строилась по векторам $z_0, \ldots, z_{N-1}$, то согласно (20.16)

$$\|\mathrm{T}\|_E \lesssim 2{,}5Np^{-t+1}\|z\|_E + N\sqrt{n}\,\omega. \qquad (21.4)$$

Довольно часто мы будем иметь дело с последовательностью матриц $\tilde{U}_1, \ldots, \tilde{U}_N$, одна из которых строится по какому-то из векторов $z_0, \ldots, z_{N-1}$. В этом случае, учитывая (20.20), получаем

$$\|\mathrm{T}\|_E \lesssim (2{,}5\sqrt{2} + 2{,}5(N-1))\,p^{-t+1}\|z_E\| + (N-1)\sqrt{n}\,\omega. \qquad (21.5)$$

За исключением особых случаев, преобразуемый вектор z не будет малым. Поэтому можно считать, что выполняется соотношение $\|z\| > \omega^{1/3}$. В этом случае вместо оценки (21.4) будем иметь

$$\|\mathrm{T}\|_E \lesssim 2{,}5Np^{-t+1}\|z\|_E. \qquad (21.6)$$

Для $N \geqslant 3$ оценка (21.5) заменяется такой:

$$\|\mathrm{T}\|_E \lesssim \frac{2{,}5\sqrt{2}+5}{3} Np^{-t+1}\|z\|_E. \qquad (21.7)$$

УПРАЖНЕНИЯ

1. Пусть матрица A умножается слева на последовательность из N матриц отражения. Доказать, что эквивалентное возмущение M с точностью до членов порядка ω удовлетворяет соотношению

$$\|\mathrm{M}\|_E \lesssim 2{,}5Np^{-t+1}\|A\|_E. \qquad (21.8)$$

2. Доказать, что соотношение (21.8) остается в силе и в том случае, когда матрица A умножается на N матриц отражения как слева, так и справа.

3. Пусть $\bar{W}$ есть точное произведение N реально вычисленных матриц отражения. Доказать, что существует такая ортогональная матрица W, что

$$\|W - \bar{W}\|_E \lesssim 2Np^{-t+1}. \qquad (21.9)$$

§ 22. Сравнение точности преобразований вращения и отражения

Оценки (19.9), (21.7) могут создать впечатление, что преобразования вращения и отражения обладают похожими свойствами с точки зрения влияния ошибок округления на вычислительный процесс. Однако этот вывод был бы преждевременным.

Если матрица отражения имеет размерность, n, то преобразование с этой матрицей изменяет в общем случае n координат вектора. Преобразование же с матрицей вращения всегда изменяет лишь две координаты. Поэтому, как правило, преобразование с матрицей отражения является более содержательным и для решения одной и той же задачи требуется выполнить значительно меньше преобразований отражения, чем преобразований вращения.

Мы уже встречались с подобной ситуацией в задаче построения унитарного преобразования, переводящего заданный вектор s в вектор, коллинеарный координатному вектору e. Для решения этой задачи требуется выполнить одно преобразование отражения или $(n-1)$ преобразований вращения, если размерность векторов равна n. Такое соотношение между необходимым числом преобразований отражения и вращения является типичным.

Высказанные соображения не означают, что при решении одной и той же алгебраической задачи, связанной с большим числом преобразований вектора, отношение правой части оценки (19.9) к правой части оценки (21.7) будет всегда величиной порядка n. Исследуя последовательность преобразований вращения, мы видели, что на общую оценку ошибок влияет не только число выполненных преобразований вращения, но и выбранная последовательность индексов матриц вращения.

Вполне возможно, что для решения одной и той же задачи можно использовать различные последовательности матриц вращения. Поэтому, прежде чем сравнивать точность преобразований вращения и отражения, постараемся понять, каким может быть минимальный уровень ошибок в этих преобразованиях.

Рассмотрим следующий гипотетический пример. Предположим, что все матрицы вращения настолько близки к единичным, что каждое их действие на координаты

вектора равносильно лишь округлению координат. Пусть выполняется N преобразований вращения над вектором z размерности n. Обозначим через $z_1, \ldots, z_n$ его координаты, через $z_{N1}, \ldots, z_{Nn}$ координаты вектора, полученного после выполнения N преобразований. Тогда

$$z_{Ni} = z_i\left(1+\varepsilon_1^{(i)}\right)\ldots\left(1+\varepsilon_{N_i}^{(i)}\right).$$

Здесь N_i есть число преобразований, в которых участвовала координата, стоящая в позиции i. Согласно предположению $\left|\varepsilon_j^{(i)}\right| \leqslant (1/2)\,p^{t+1}$ для всех i, j. Кроме этого,

$$\sum_{i=1}^{n} N_i = 2N,$$

и мы имеем

$$\|E\|_E \lesssim \frac{1}{2}\left(\sum_{i=1}^{n} N_i^2 |z_i|^2\right)^{1/2} p^{t+1}. \tag{22.1}$$

Несмотря на то, что высказанные предположения относительно реализации преобразований вращения не совсем реальны, соотношение (22.1) для эквивалентного возмущения E отличается от реального лишь постоянным множителем в правой части.

Каковы бы ни были числа N_i, на классе векторов z с заданной величиной евклидовой нормы справедливо неравенство

$$\left(\sum_{i=1}^{n} N_i^2 |z_i|^2\right)^{1/2} \leqslant \max_i N_i \|z\|_E. \tag{22.2}$$

Очевидно, что правая часть имеет минимум в том случае, тогда $N_i = 2N/n$ для всех i. Так как неравенство (22.2) достигается, то оценки ошибок при различных порядках преобразования элементов не могут быть лучше, чем оценка

$$\|E\|_E \gtrsim \alpha \frac{N}{n} p^{t+1} \|z\|_E, \tag{22.3}$$

где α — некоторая константа.

Если оценка для какой-либо последовательности преобразований вращения отличается от (22.3), то это означает, что или она завышена, или в значительной мере зависит от углов поворота.

Рассмотрим с этой точки зрения результаты исследования влияния ошибок округления, полученные в § 19. Оценка (19.9) не является оценкой вида (22.3) и хуже ее примерно в n раз. Но как мы уже отмечали, эта оценка почти достигается для некоторых сильно связанных последовательностей. Следовательно, для таких последовательностей суммарное эквивалентное возмущение должно в значительной мере зависеть от углов поворота. С другой стороны, оценки (19.11), (19.12) для циклических последовательностей являются оценками вида (22.3) и поэтому исключительно эффективны.

Итак, при выполнении N преобразований вращения существуют такие порядки преобразования координат, при которых эквивалентное возмущение удовлетворяет соотношению (22.3). Конечно, остается открытым вопрос, обеспечивают ли такие последовательности решение соответствующих задач вычислительной алгебры. В дальнейшем мы дадим на него положительный ответ.

Исследование достижимости оценок ошибок для последовательности преобразований отражения осуществляется существенно проще. Снова рассмотрим гипотетический пример. Пусть матрицы отражения будут близки к диагональным, элементы которых равны либо $+1$, либо -1. Предположим, что действие каждой матрицы отражения на координаты вектора равносильно лишь округлению координат. Если выполняется N преобразований, то при тех же обозначениях, что были сделаны выше, мы будем иметь

$$z_{Ni} = \pm z_i (1 + \varepsilon_1^{(i)}) \ldots (1 + \varepsilon_N^{(i)}),$$

где

$$|\varepsilon_j^{(i)}| \leqslant (1/2)\, p^{-t+1}$$

для всех i, j. Отсюда вытекает, что

$$\| E \|_E \leqslant \frac{1}{2} N p^{-t+1} \| z \|_E. \tag{22.4}$$

Следовательно, оценки (21.6), (21.7) по существу неулучшаемы.

Как мы уже отмечали, одно преобразование отражения решает такую же задачу, как и $n-1$ преобразований вращения. Поэтому, сравнивая оценки (22.3), (22.4), можно заключить, что преобразования вращения не могут

гарантировать существенно большей точности, чем преобразования отражения. Однако этот вывод справедлив лишь тогда, когда используется операция накопления скалярных произведений. Если же все вычисления ведутся с одинарной точностью, то влияние ошибок округления в типичных последовательностях преобразований вращений будет в $\sqrt{n}$ или даже n раз по порядку меньше, чем в преобразованиях отражения, решающих ту же задачу.

УПРАЖНЕНИЯ

1. Доказать, что любая матрица отражения U может быть представлена в виде

$$U = RDR^*, \tag{22.5}$$

где R есть произведение матриц вращения, а матрица D отличается от единичной лишь тем, что один из ее диагональных элементов равен -1.

2. Пусть вектор v матрицы отражения U из (20.4) имеет l ненулевых координат. Доказать, что матрица R в разложении (22.5) состоит не меньше, чем из $l-1$ матриц вращения.

3. Выполнить анализ ошибок преобразований отражения для случая, когда все вычисления ведутся с одинарной точностью без накопления скалярных произведений.

§ 23. Двухсторонние унитарные преобразования

В огромном многообразии численных методов линейной алгебры имеется значительное количество алгорифмов, связанных с последовательным преобразованием заданной матрицы путем умножения на унитарные матрицы вращения или отражения. Если преобразования *односторонние*, например, левые (правые), то каждый столбец (строка) матрицы преобразуется независимо и для исследования влияния ошибок округления таких процессов могут быть использованы результаты, полученные ранее для преобразования векторов. *Двухсторонние* преобразования требуют более тщательного анализа.

Пусть матрица A умножается слева на последовательность матриц $\tilde{Q}_1, \tilde{Q}_2, \ldots, \tilde{Q}_s$ и справа на последовательность матриц $\tilde{R}_1, \tilde{R}_2, \ldots, \tilde{R}_p$. Будем считать, что все матрицы унитарные и имеют один и тот же тип, т. е. являются матрицами вращения либо матрицами отражения. Для нас сейчас безразлично, каким способом

вычислены эти матрицы. Важно лишь, чтобы выполнялись полученные ранее оценки ошибок округления.

Не возникает никаких проблем с анализом ошибок, если сначала выполняются все преобразования с одной стороны, а затем все преобразования с другой стороны. Действительно, пусть, например, первыми осуществляются преобразования слева. В качестве промежуточного результата мы получим матрицу

$$\tilde{B} = fl(\tilde{Q}_s \dots \tilde{Q}_1 A),$$

при этом согласно уже проведенным исследованиям

$$\tilde{B} = \tilde{Q}_s \dots \tilde{Q}_1 (A + T), \tag{23.1}$$

где $\|T\|_E \lesssim \alpha p^{t+1} \|A\|_E$. Величина α зависит от выбранной последовательности преобразований $\tilde{Q}_1, \dots, \tilde{Q}_s$, но не зависит от элементов матрицы A. Далее находим матрицу $\tilde{C} = fl(\tilde{B}\tilde{R}_1 \dots \tilde{R}_p)$; при этом аналогично (23.1)

$$\tilde{C} = (\tilde{B} + \Gamma)\tilde{R}_1 \dots \tilde{R}_p, \tag{23.2}$$

где $\|\Gamma\|_E \lesssim \beta p^{t+1} \|\tilde{B}\|_E$. Снова величина β зависит от выбранной последовательности преобразований $\tilde{R}_1, \dots, \tilde{R}_p$, но не зависит от элементов матрицы $\tilde{B}$. Согласно предположению все преобразования асимптотически близки к унитарным. Теперь, принимая во внимание инвариантность евклидовой нормы к унитарным преобразованиям, заключаем из (23.1), (23.2), что

$$\tilde{C} = \tilde{Q}_s \dots \tilde{Q}_1 (A + N)\tilde{R}_1 \dots \tilde{R}_p,$$

где $\|N\|_E \lesssim (\alpha + \beta) p^{t+1} \|A\|_E$. Такая же оценка имеет место и в том случае, когда сначала выполняются правые преобразования, а затем левые.

Однако большинство вычислительных алгорифмов с двухсторонними преобразованиями устроено иначе. Именно, преобразования с одной стороны обычно не выполняются подряд, а перед каждым или некоторыми из них выполняется одно или несколько преобразований с другой стороны.

При выводе оценок для норм эквивалентных возмущений, возникающих при выполнении последовательности односторонних преобразований вращения, были учтены весьма тонкие соотношения между результатами промежуточных вычислений. На первый взгляд кажется, что выполнение между двумя преобразованиями вращения каких-либо преобразований с другой стороны может нарушить эти соотношения. Но заметим, что в действительности, например, при левых преобразованиях в качестве промежуточных результатов появились лишь суммы квадратов модулей элементов строк матрицы. Поэтому, если между двумя левыми преобразованиями вращения включить любое правое унитарное преобразование или даже любую последовательность правых унитарных преобразований, то при этом все названные суммы останутся без изменения.

Следовательно, изменение очередности выполнения двухсторонних преобразований вращения справа и слева меняет общую мажорантную оценку евклидовой нормы эквивалентного возмущения лишь в членах второго порядка малости.

Аналогичные рассуждения показывают, что такой же вывод справедлив и в отношении изменения очередности выполнения двухсторонних преобразований отражения.

Напомним, что общий эффект от влияния ошибок округлений мы оценивали величиной нормы эквивалентного возмущения при преобразованиях с неточно вычисленными матрицами вращения или отражения. При этом отклонение вычисленных матриц от унитарных практически не играло существенной роли. С такой ситуацией мы будем встречаться почти всегда, изучая численные методы решения систем линейных алгебраических уравнений.

При исследовании подобных преобразований величина отклонения матриц преобразования от унитарных становится важной. Пусть над матрицей A совершается последовательность подобных преобразований с унитарными матрицами $Q_1, \dots, Q_s$. Так как для унитарных матриц обратная совпадает с сопряженной, то это означает, что находится некоторая матрица

$$B = Q_s^* \dots Q_1^* A Q_1 \dots Q_s.$$

Реально полученные матрицы $\tilde{Q}_1, \dots, \tilde{Q}_s$ уже не будут унитарными. Поэтому, строго говоря, мы должны были бы вычислять матрицу

$$\tilde{B} = \tilde{Q}_s^{-1} \dots \tilde{Q}_1^{-1} A \tilde{Q}_1 \dots \tilde{Q}_s.$$

Но $\tilde{Q}_1^{-1}, \ldots, \tilde{Q}_s^{-1}$ нельзя определить точно ни в случае матриц вращения, ни в случае матриц отражения. Это обстоятельство и заставляет нас ограничиться вычислением матрицы

$$\tilde{B} = \mathrm{fl}\,(\tilde{Q}_s^* \ldots \tilde{Q}_1^* A \tilde{Q}_1 \ldots \tilde{Q}_s).$$

Во всех рассмотренных преобразованиях

$$\mathrm{fl}\,(\tilde{Q}_s^* \ldots \tilde{Q}_1^* A \tilde{Q}_1 \ldots \tilde{Q}_s) \equiv \tilde{Q}_s^* \ldots \tilde{Q}_1^* (A + \mathrm{M})\, \tilde{Q}_1 \ldots \tilde{Q}_s$$

для некоторого эквивалентного возмущения M. Теперь покажем, что

$$\tilde{Q}_s^* \ldots \tilde{Q}_1^* (A + \mathrm{M})\, \tilde{Q}_1 \ldots \tilde{Q}_s = \tilde{Q}_s^{-1} \ldots \tilde{Q}_1^{-1} (A + \Delta)\, \tilde{Q}_1 \ldots \tilde{Q}_s \quad (23.3)$$

и дадим оценку нормы Δ.

Умножим обе части равенства (23.3) справа на матрицу $\tilde{Q}_s^{-1} \ldots \tilde{Q}_1^{-1}$. Это дает соотношение

$$\tilde{Q}_s^{-1} \ldots \tilde{Q}_1^{-1} \Delta = (\tilde{Q}_s^* \ldots \tilde{Q}_1^* - \tilde{Q}_s^{-1} \ldots \tilde{Q}_1^{-1})\, A + \tilde{Q}_s^* \ldots \tilde{Q}_1^* \mathrm{M}. \quad (23.4)$$

Так как матрицы преобразования близки к унитарным, то

$$\| \Delta \|_E \lesssim \| (\tilde{Q}_s^* \ldots \tilde{Q}_1^* - \tilde{Q}_s^{-1} \ldots \tilde{Q}_1^{-1})\, A \|_E + \| \mathrm{M} \|_E.$$

Поэтому оценка нормы Δ сводится, по существу, к оценке нормы первого слагаемого из (23.4).

Некоторую сложность представляет лишь исследование преобразований вращения. Пусть $\tilde{T}$ — вычисленная матрица вращения второго порядка, а c — вектор; тогда

$$\tilde{T}^{-1} c = \tilde{T}^* (c + (\tilde{T}^{*\,-1} \tilde{T}^{-1} - E)\, c).$$

Принимая во внимание (18.15), (18.17), заключаем, что

$$\| \tilde{T}^{*\,-1} \tilde{T}^{-1} - E \|_2 \lesssim \frac{5}{2}\, p^{t+1}.$$

Следовательно, если обозначить

$$\tilde{T}^{-1} c = \tilde{T}^* (c + \varepsilon), \quad (23.5)$$

то будем иметь

$$\| \varepsilon \|_E \leqslant \frac{5}{2}\, p^{t+1} \| c \|_E. \quad (23.6)$$

Предположим теперь, что $\tilde{Q}_1, \ldots, \tilde{Q}_s$ есть вычисленная последовательность матриц вращения, индексы которых меняются, например, в циклическом порядке. Рассмотрим более внимательно выражение $\tilde{Q}_s^{-1} \ldots \tilde{Q}_1^{-1} A$. Согласно (23.5) последовательное умножение матрицы A слева на матрицы $\tilde{Q}_1^{-1}, \ldots, \tilde{Q}_s^{-1}$ можно трактовать как умножение на матрицы $\tilde{Q}_1^*, \ldots, \tilde{Q}_s^*$ с внесением некоторых ошибок. Но эти ошибки с точностью до числовых коэффициентов образуются по тому же закону, что и ошибки округления при умножении на последовательность $\tilde{Q}_1^*, \ldots, \tilde{Q}_s^*$. Поэтому, воспользовавшись анализом ошибок, выполненным в § 19, находим, что

$$\tilde{Q}_s^{-1} \ldots \tilde{Q}_1^{-1} A = \tilde{Q}_s^* \ldots \tilde{Q}_1^* (A + \Gamma),$$

где согласно (19.12) с заменой (18.21) на (23.6) имеем

$$\| \Gamma \|_E \lesssim \frac{5}{2}\, (2n-3)\, p^{t+1} \| A \|_E.$$

Но это означает, что также

$$\| (\tilde{Q}_s^* \ldots \tilde{Q}_1^* - \tilde{Q}_s^{-1} \ldots \tilde{Q}_1^{-1})\, A \|_E \leqslant \frac{5}{2}\, (2n-3)\, p^{t+1} \| A \|_E. \quad (23.7)$$

Аналогично получаются оценки для других последовательностей матриц вращения.

Исследование преобразований отражения осуществляется по той же схеме. Если $\tilde{U}$ — вычисленная матрица отражения, а z — вектор, то $\tilde{U}^{-1} z = \tilde{U}^* (z + (\tilde{U}^{*-1} \tilde{U}^{-1} - E)\, z)$. Принимая во внимание (20.9), заключаем, что

$$\| \tilde{U}^{*\,-1} \tilde{U}^{-1} - E \|_2 \lesssim 4p^{t+1}.$$

Следовательно, если $\tilde{U}^{-1} z = \tilde{U}^* (z + \mu)$, то $\| \mu \|_E \leqslant 4p^{t+1} \| z \|_E$. Если выполняется $n-1$ преобразований отражения, то аналогичные рассуждения показывают, что

$$\| (\tilde{U}_{n-1}^* \ldots \tilde{U}_1^* - \tilde{U}_{n-1}^{-1} \ldots \tilde{U}_1^{-1})\, A \|_E \lesssim 4\,(n-1)\, p^{t+1} \| A \|_E.$$

Теперь уже нетрудно получить полную оценку для нормы Δ из (23.3), принимая во внимание соответствующие оценки для нормы M. Если в качестве матриц подобного преобразования берется последовательность матриц вращения, индексы которых меняются в циклическом порядке при $m = n-1$, то

$$\| \Delta \|_E \lesssim \frac{4 \cdot 2 + 5}{2}\, (2n-3)\, p^{t+1} \| A \|_E. \quad (23.8)$$

Если последовательность матриц вращения можно разбить на k групп, внутри которых матрицы вращения не имеют общих индексов, то

$$\|\Delta\|_E \leqslant \frac{4\sqrt{2}+5}{2} k p^{t+1} \|A\|_E. \tag{23.9}$$

И, наконец, если в качестве матриц подобного преобразования берется последовательность из $n-1$ матриц отражения, то в этом случае

$$\|\Delta\|_E \leqslant \frac{5\sqrt{2}+22}{3}(n-1)\, p^{t+1} \|A\|_E. \tag{23.10}$$

УПРАЖНЕНИЯ

1. Пусть $\bar{R}$ есть точное произведение реально вычисленных матриц вращения, соответствующих циклической последовательности. Используя соотношение (23.7), доказать, что

$$\|\bar{R}\bar{R}^* - E\|_E \leqslant (5/2\sqrt{2})(n+m-2)\, p^{-t+1}. \tag{23.11}$$

2. Доказать, что в условиях упражнения 1 существует ортогональная матрица R такая, что

$$\|R - \bar{R}\|_E \leqslant (5/4\sqrt{2})(n+m-2)\, p^{-t+1}. \tag{23.12}$$

Сравнить этот результат с (19.18), (21.9).

3. Получить оценку, аналогичную (23.12), для последовательности матриц отражения. Сравнить результат с (21.9).

§ 24. Неунитарные преобразования

Евклидова и 2-норма инвариантны к унитарным преобразованиям, поэтому не может происходить существенного увеличения в целом элементов пресбразуемых векторов и матриц. Это очень важно, так как на каждом шаге ошибки округления в основном пропорциональны величинам элементов. *Неунитарные преобразования не обладают естественной устойчивостью*, однако иногда вычисления можно организовать так, что в некоторой ограниченной форме устойчивость все же будет иметь место.

Рассмотрим *простейшие* неунитарные матрицы. Пусть a и b — векторы размерности n. По аналогии с матрицей отражения можно построить матрицы вида

$$V = E + ab^*. \tag{24.1}$$

Если s — пресбразуемый вектор, то теперь $Vs = s + (s, b)\, a$. Среди матриц (24.1) наиболее часто используются те, в которых либо вектор a, либо вектор b является координатным. Второй из векторов обычно определяется условиями задачи.

Предположим, что в качестве вектора b взят r-й координатный вектор e_r. Одна из важнейших вычислительных задач требует найти по заданному вектору s такой вектор a, чтобы первые r координат вектора $s + (s, e_r)\, a$ совпадали с соответствующими координатами вектора s, а остальные равнялись нулю. Обозначим через $s_1, \ldots, s_n$ и $a_1, \ldots, a_n$ координаты векторов s и a. Ясно, что $(s, e_r) = s_r$. Следовательно, искомый вектор a существует лишь в том случае, если $s_r \neq 0$. Но тогда очевидно, что

$$a_p = \begin{cases} 0, & p \leqslant r, \\ -s_p/s_r, & p > r. \end{cases} \tag{24.2}$$

Матрицы (24.1), в которых в качестве вектора b взят вектор e_r, а первые r координат вектора a нулевые, называются матрицами типа N_r. Они весьма часто используются в вычислительной практике. Эти матрицы отличаются от единичных лишь поддиагональными элементами в r-м столбце, т. е.

$$N_r = \begin{bmatrix} 1 & & & & 0 \\ & \ddots & & & \\ & & 1 & & \\ & & n_{r+1,\,r} & \ddots & \\ & & \vdots & & \\ & & n_{n,\,r} & & 1 \end{bmatrix}; \tag{24.3}$$

при этом

$$N_r^{-1} = \begin{bmatrix} 1 & & & & 0 \\ & \ddots & & & \\ & & 1 & & \\ & & -n_{r+1,\,r} & \ddots & \\ & & \vdots & & \\ & & -n_{n,\,r} & & 1 \end{bmatrix}.$$

Легко находится и произведение матриц $N_1^{-1} N_2^{-1} \ldots N_r^{-1}$. По существу, для его определения не нужно производить никаких вычислений, так как ненулевые элементы расположены лишь в первых r столбцах и они совпадают с элементами матриц $N_1^{-1}, N_2^{-1}, \ldots, N_r^{-1}$. Именно,

$$N_1^{-1} N_2^{-1} \ldots N_r^{-1} = \begin{bmatrix} 1 & & & & 0 \\ -n_{21} & 1 & & & \\ -n_{31} & -n_{32} & \ddots & & \\ \vdots & \vdots & & 1 & \\ -n_{n1} & -n_{n2} & \ldots & -n_{nr} \ldots & 1 \end{bmatrix}. \tag{24.4}$$

Конечно, по такому же принципу строится и произведение матриц $N_1N_2 \dots N_r$.

Пусть задан вектор z размерности n и находится последовательность векторов $z_1, \dots, z_r$, $r<n$, с помощью умножения на матрицы $N_1, N_2, \dots, N_r$, т. е.

$$z_k = N_k z_{k-1}, \quad z_0 = z, \quad k = 1, 2, \dots, r. \tag{24.5}$$

В реальных вычислениях вместо матриц $N_1, \dots, N_r$ будем иметь матрицы $\tilde{N}_1, \dots, \tilde{N}_r$. Кроме этого, будут возникать ошибки и при выполнении операций (24.5). Поэтому в действительности мы получим векторы $\tilde{z}_1, \dots, \tilde{z}_r$, где

$$\tilde{z}_k = \mathrm{fl}(\tilde{N}_k \tilde{z}_{k-1}) \equiv \tilde{N}_k z_{k-1} + \mu_{k-1}. \tag{24.6}$$

Здесь μ_{k-1} — вектор ошибок, возникающий из-за неточного вычисления произведений $\tilde{N}_k \tilde{z}_{k-1}$. Принимая во внимание (8.4), (8.5), можно записать, что

$$\tilde{z}_r = \tilde{N}_r \tilde{N}_{r-1} \dots \tilde{N}_1 (z + \mu),$$

где

$$\mu = \tilde{N}_1^{-1}\mu_0 + \tilde{N}_1^{-1}\tilde{N}_2^{-1}\mu_1 + \dots + \tilde{N}_1^{-1}\tilde{N}_2^{-1} \dots \tilde{N}_r^{-1}\mu_{r-1}.$$

Первые k строк матрицы $\tilde{N}_k$, $1 \leqslant k \leqslant r$, представляют собой строки единичной матрицы. Следовательно, первые k элементов векторов $\tilde{z}_{k-1}$ и $\tilde{z}_k$ совпадают. Но тогда первые k элементов вектора ошибок μ_{k-1} являются нулевыми. Согласно (24.4) для любого k в произведении $\tilde{N}_1^{-1} \dots \tilde{N}_k^{-1}$ лишь первые k столбцов отличны от столбцов единичной матрицы. Поэтому для любого k

$$\tilde{N}_1^{-1}\tilde{N}_2^{-1} \dots \tilde{N}_k^{-1}\mu_{k-1} = \mu_{k-1},$$

откуда вытекает, что

$$\mu = \sum_{k=1}^{r} \mu_{k-1}. \tag{24.7}$$

Таким образом, при выполнении последовательности преобразований с матрицами $\tilde{N}_1, \dots, \tilde{N}_r$ эквивалентное возмущение μ связано с ошибками μ_{k-1}, возникающими на отдельных шагах, простым соотношением (24.7).

Вычисление матриц N_r по формулам (24.2) возможно лишь при условии $s_r \neq 0$. Однако иногда оно может не выполняться. Существует значительное число алгорифмов, в которых для его обеспечения используются перестановки элементов. Обычно это означает, что в формуле (24.6) вместо матрицы $\tilde{N}_k$ стоит произведение $\tilde{N}_k P_{kk'}$, где $P_{kk'}$, $k' \geqslant k$ есть матрица перестановок столбцов с номерами k и k'. В этом случае для эквивалентного возмущения μ будем иметь

$$\mu = P_{11'}\tilde{N}_1^{-1}\mu_0 + P_{11'}\tilde{N}_1^{-1}P_{22'}\tilde{N}_2^{-1}\mu_1 + \dots$$
$$\dots + P_{11'}\tilde{N}_1^{-1}P_{22'}\tilde{N}_2^{-1} \dots P_{rr'}\tilde{N}_r^{-1}\mu_{r-1}.$$

Рассмотрим более подробно произведение $P_{11'}\tilde{N}_1^{-1} \dots P_{kk'}\tilde{N}_k^{-1}$. Докажем, что его последние $n-k$ столбцов совпадают с последними $n-k$ столбцами произведения $P_{11'}P_{22'} \dots P_{kk'}$. Для $k=1$ это утверждение очевидно, так как умножение матрицы $P_{11'}$ справа на $\tilde{N}_1^{-1}$ меняет только ее первый столбец. Пусть оно верно для $k-1$. Дополнительное умножение справа на $P_{kk'}$ эквивалентно перестановке столбцов с номерами k и k'. В силу условия $k' \geqslant k$ последние $n-k+1$ столбцов полученного произведения будут совпадать с последними $n-k+1$ столбцами произведения $P_{11'}P_{22'} \dots P_{kk'}$. Еще одно умножение справа на $\tilde{N}_k^{-1}$ меняет в этом произведении лишь столбец с номером k.

Итак, утверждение доказано. По-прежнему первые k элементов вектора ошибок μ_{k-1} являются нулевыми. Следовательно, для любого $k \geqslant 1$

$$P_{11'}\tilde{N}_1^{-1} \dots P_{kk'}\tilde{N}_k^{-1}\mu_{k-1} = P_{11'} \dots P_{kk'}\mu_{k-1}.$$

Окончательно заключаем, что теперь

$$\mu = \sum_{k=1}^{r} P_{11'} \dots P_{kk'}\mu_{k-1}. \tag{24.8}$$

Сравнение формул (24.7), (24.8) с (8.5) показывает, что эквивалентное возмущение в рассмотренных преобразованиях с матрицами вида N_r не зависит явно от этих матриц. Поэтому опасность неустойчивости может возникнуть лишь в том случае, если велики сами векторы ошибок μ_{k-1}. Мы уже отмечали, что ошибки, сделанные на отдельных шагах, в целом пропорциональны величинам координат векторов. Поэтому и важно, чтобы эти коор-

динаты были по возможности меньше. Такую задачу выполняет подходящий выбор матриц перестановок. Детальный анализ векторов μ_{k-1} мы проведем несколько позднее, когда будут описаны вычислительные процессы, приводящие к преобразованиям с матрицами типа N_r.

Рассмотрим еще один вид неунитарных преобразований. Пусть теперь координатный вектор $\bar{e}_r$ взят в качестве вектора a матрицы (24.1). В этом случае матрица (24.1) будет отличаться от единичной лишь одной r-й строкой. Наиболее важные в практическом отношении матрицы этого типа имеют вид

$$M_r = \begin{bmatrix} 1 & & & & & 0 \\ & \ddots & & & & \\ & & 1 & & & \\ m_{r1} & \dots & m_{r,r-1} & m_{rr} & m_{r,r+1} & \dots & m_{rn} \\ & & & & 1 & & \\ & & & & & \ddots & \\ 0 & & & & & & 1 \end{bmatrix}, \qquad (24.9)$$

при этом

$$M_r^{-1} = \begin{bmatrix} 1 & & & & & 0 \\ & \ddots & & & & \\ & & 1 & & & \\ -m_{r1} & \dots & -m_{r,r-1} & m_{rr}^{-1} & -m_{r,r+1} & \dots & -m_{rn} \\ & & & & 1 & & \\ & & & & & \ddots & \\ 0 & & & & & & 1 \end{bmatrix}.$$

Предположим, что снова задан вектор z размерности n и находится последовательность векторов $z_1, \dots, z_r$, $r \leqslant n$, с помощью умножения на матрицы $M_2, M_3, \dots, M_r$, т. е.

$$z_k = M_k z_{k-1}, \quad z_1 = z, \quad k = 2, 3, \dots, r. \qquad (24.10)$$

Обратим внимание на существенные отличия этого процесса от процесса (24.5). Если в (24.5) вектор z_k имеет k первых координат, совпадающих с соответствующими координатами вектора z_{k-1}, то в (24.10) вектор z_k отличается от z_{k-1} всего лишь одной k-й координатой. Эта координата представляет собой сумму попарных произведений некоторых чисел, поэтому для ее вычисления удобно применять режим накопления.

Если матрицы M_r определяются с помощью некоторых вычислений, то реальный вычислительный процесс будет описываться соотношениями

$$z_k = \mathrm{fl}_s(\tilde{M}_k \tilde{z}_{k-1}) = \tilde{M}_k \tilde{z}_{k-1} + \nu_{k-1}.$$

Однако у вектора ν_{k-1} при любых k будет отлична от нуля лишь одна k-я координата. Конечно, снова

$$\tilde{z}_k = \tilde{M}_r \tilde{M}_{r-1} \dots \tilde{M}_2 (z + \nu),$$

где

$$\nu = \tilde{M}_2^{-1}\nu_1 + \tilde{M}_2^{-1}\tilde{M}_3^{-1}\nu_2 + \dots + \tilde{M}_2^{-1}\tilde{M}_3^{-1}\dots\tilde{M}_r^{-1}\nu_{r-1}.$$

Произведение $M_2^{-1}\dots M_k^{-1}$ простым способом связано со своими сомножителями. Легко проверить, что аналогично (24.4) будем иметь

$$M_2^{-1}M_3^{-1}\dots M_k^{-1} = \begin{bmatrix} 1 & & & 0 \\ -m_{21} & 1 & & \\ \dots & \dots & \ddots & \\ -m_{k1} & -m_{k2} & \dots & 1 \end{bmatrix}.$$

По такому же принципу строится и произведение $M_2M_3\dots M_k$. Принимая во внимание вид векторов ν_{k-1}, получаем, что для любого k

$$\tilde{M}_2^{-1}\tilde{M}_3^{-1}\dots\tilde{M}_k^{-1}\nu_{k-1} = \nu_{k-1},$$

поэтому

$$\nu = \sum_{k=2}^{r} \nu_{k-1}. \qquad (24.11)$$

Несмотря на то, что формула (24.11) внешне похожа на (24.7), она должна представлять больший интерес, так как строение векторов ν_{k-1} значительно проще, чем μ_{k-1}. Это дает основание надеяться, что в вычислительных алгоритмах, использующих преобразования с матрицами типа M_r, можно достичь высокой точности.

УПРАЖНЕНИЯ

1. Доказать, что

$$(E + ab^*)^{-1} = E - \frac{1}{1 + (a, b)}\, ab^*.$$

2. Доказать, что

$$\left(1 + \frac{\|a\|_2 \|b\|_2}{|1 + (a, b)|}\right)^{-1} \leqslant \|E + ab^*\|_2 \leqslant 1 + \|a\|_2 \|b\|_2.$$

3. Найти собственные значения и собственные векторы матрицы $E + ab^*$.

4. Доказать, что для матриц вида (24.3) выполняются соотношения:

$$N_1N_2\dots N_k = N_1 + N_2 + \dots + N_k - (k-1)E,$$
$$N_1^{-1}N_2^{-1}\dots N_k^{-1} = (k+1)E - N_1 - N_2 - \dots - N_k. \qquad (24.12)$$

5. Будут ли выполняться соотношения (24.12), если в левых частях взять сомножители в другом порядке?

6. Пусть $i_1, i_2, \ldots, i_r$ — какая-либо перестановка из чисел $1, 2, \ldots, r$, отличающаяся от нормальной. Рассмотрим вместо преобразований (24.5) преобразования

$$z_k = N_{i_k} z_{k-1}, \quad z_0 = z, \quad k = 1, 2, \ldots, r.$$

Доказать, что в случае произвольного вектора z формула (24.7) уже не имеет места.

7. Рассмотреть аналоги упражнений 4—6 в отношении матриц вида (24.9).

§ 25. Ортогонализация

Процесс *ортогонализации* системы векторов является составной частью многих численных методов, поэтому мы более подробно остановимся на исследовании возникающих в нем ошибок.

Пусть задана линейно независимая система векторов $a_1, a_2, \ldots, a_n$. Будем строить такую систему ортонормированных векторов $b_1, b_2, \ldots, b_n$, что для всех $k = 1, 2, \ldots, n$ векторы $b_1, b_2, \ldots, b_k$ являются базисом подпространства P_k, натянутого на $a_1, a_2, \ldots, a_k$. Так как векторы b_k нормированы, то их можно представить в виде

$$b_k = \frac{1}{\|v_k\|_E} v_k, \tag{25.1}$$

где $v_1, v_2, \ldots, v_k$ — ортогональный, но не обязательно нормированный базис P_k. Для $k = 1$ положим $v_1 = a_1$ и будем искать v_{k+1} как линейную комбинацию векторов $a_{k+1}, b_1, \ldots, b_k$, т. е.

$$v_{k+1} = a_{k+1} + \sum_{i=1}^{k} c_i b_i. \tag{25.2}$$

Условие ортогональности вектора v_{k+1} к ортогональным векторам $v_1, \ldots, v_k$ или, что то же самое, векторам $b_1, \ldots, b_k$ дает

$$c_i = -(a_{k+1}, b_i), \tag{25.3}$$

поэтому окончательно находим

$$v_{k+1} = a_{k+1} - \sum_{i=1}^{k} (a_{k+1}, b_i) b_i. \tag{25.4}$$

Ошибки округления, возникающие при численной реализации процесса ортогонализации, изменяют свойства

получаемой системы векторов $b_1, \ldots, b_n$. Именно, эта система в точном смысле уже не будет эквивалентна системе $a_1, \ldots, a_n$ и не будет ортонормированной.

Оценим сначала степень неэквивалентности вычисленной системы $\tilde{b}_1, \ldots, \tilde{b}_n$ исходной системе векторов. Согласно общей идее обратного анализа постараемся показать, что система $\tilde{b}_1, \ldots, \tilde{b}_k$ будет для всех k эквивалентна возмущенной системе $a_1 + \varepsilon_1, \ldots, a_k + \varepsilon_k$ и определим величины норм эквивалентных возмущений $\varepsilon_1, \ldots, \varepsilon_k$.

Введем некоторые обозначения. Подпространства, натянутые на векторы $\tilde{b}_1, \ldots, \tilde{b}_k$, будем обозначать через $\tilde{P}_k$, координаты векторов a_i, b_k и других соответственно через a_{ji}, b_{jk} и т. д.

Рассмотрим процесс вычисления вектора $\tilde{b}_1$. Каким бы способом ни вычислялась длина вектора v_1 и какую бы ошибку она ни содержала, вектор $\tilde{b}_1$ будет коллинеарен вектору a_1, если только деление координат вектора v_1 на его вычисленную длину $\mathrm{fl}(\|v_1\|_E)$ осуществляется точно. Следовательно, вся неэквивалентность в данном случае возникает лишь при этом делении. Имеем для всех j

$$\tilde{b}_{j1} = \mathrm{fl}\left(\frac{v_{j1}}{\mathrm{fl}(\|v_1\|_E)}\right) = \frac{v_{j1}}{\mathrm{fl}(\|v_1\|_E)}(1 + \mu_{j1}),$$

где μ_{j1} удовлетворяет обычным соотношениям в зависимости от полученного результата вычислений. Отсюда уже нетрудно установить, что вектор $\tilde{b}_1$ коллинеарен вектору $a_1 + \varepsilon_1$, где

$$\|\varepsilon_1\|_E \leqslant \frac{1}{2} p^{1-t} \|a_1\|_E + n^{1/2} \Pi(\|v_1\|_E)\, \omega.$$

Предположим теперь, что вычислены векторы $\tilde{b}_1, \ldots, \tilde{b}_k$ и получены оценки норм для $\varepsilon_1, \ldots, \varepsilon_k$ при некотором $k \geqslant 1$. Пусть вектор b_{k+1} вычисляется согласно формулам (25.1), (25.4). Каким бы способом ни вычислялись коэффициенты c_i из (25.3) и какие бы ошибки они ни содержали, вектор $\tilde{v}_{k+1}$ будет принадлежать сумме подпространства $\tilde{P}_k$ и подпространства, натянутого на a_{k+1}, если лишь суммирование в формуле (25.2) реализуется точно. Поэтому дополнительная неэквивалентность на этом этапе может возникнуть только за счет неточного вычисления суммы в правой части (25.2).

Будем определять координаты вектора $\tilde{v}_{k+1}$, используя операцию вычисления скалярного произведения в режиме накопления. Тогда для всех j

$$\tilde{v}_{j,k+1} = \mathrm{fl}_k\left(a_{j,k+1} + \sum_{i=1}^{k} \tilde{c}_i \tilde{b}_{ji}\right) =$$

$$= \left(a_{j,k+1} + \sum_{i=1}^{k} \tilde{c}_i \tilde{b}_{ji}\right)(1+\nu_{j,k+1}),$$

где $\nu_{j,k+1}$ снова удовлетворяет обычным соотношениям. Эту формулу можно записать в таком виде:

$$\tilde{v}_{j,k+1} = (a_{j,k+1} + e'_{j,k+1}) + \sum_{i=1}^{k} \tilde{c}_i \tilde{b}_{ji}.$$

Здесь $|e'_{j,k+1}| < \omega$, если $\nu_{j,k+1} = -1$ и $|e'_{j,k+1}| \lesssim (1/2)\,|\tilde{v}_{j,k+1}|\,p^{-t+1}$, если $\nu_{j,k+1} \neq -1$. Нормировка вектора $\tilde{v}_{k+1}$ вносит дополнительное возмущение в координаты $a_{j,k+1}$, которое оценивается так же, как при вычислении вектора $\tilde{b}_1$. Окончательно получаем, что вектор $\tilde{b}_{k+1}$ принадлежит сумме подпространства $\tilde{P}_k$ и подпространства, натянутого на вектор $a_{k+1} + e_{k+1}$. При этом

$$\|e_{k+1}\|_E \lesssim p^{-t+1}\,\|\tilde{v}_{k+1}\|_E + n^{1/2}(1 + \mathrm{fl}\,(\|\tilde{v}_{k+1}\|))\,\omega,$$

где $\|\tilde{v}_{k+1}\|_E$ есть точное значение евклидовой нормы вектора $\tilde{v}_{k+1}$, $\mathrm{fl}\,(\|\tilde{v}_{k+1}\|_E)$ — вычисленное в процессе нормировки.

Если вычисления выполняются точно, то евклидова норма вектора v_{k+1} не превосходит евклидовой нормы вектора a_{k+1} для всех k, так как из (25.1), (25.4) вытекает, что

$$(v_{k+1},\ v_{k+1}) = (a_{k+1},\ a_{k+1}) - \sum_{i=1}^{k} (a_{k+1},\ b_i)^2. \qquad (25.5)$$

Поэтому, не ограничивая существенно общности, можно считать, что для всех $k\ \|\tilde{v}_{k+1}\|_E \lesssim \|a_{k+1}\|_E$. В практических задачах мы будем иметь дело лишь с такими векторами $a_1, \ldots, a_n$, для которых при всех k выполняются соотношения

$$\|a_{k+1}\|_E > \omega^{1/2}. \qquad (25.6)$$

Учитывая условие (4.7) и полученные выше оценки, можно сделать следующий вывод.

Если при реализации суммирования в правой части (25.2) *используется операция вычисления скалярного произведения в режиме накопления, то реально вычисленные векторы* $\tilde{b}_1, \ldots, \tilde{b}_n$ *обладают тем свойством, что для всех* $k \geqslant 1$ *векторы* $\tilde{b}_1, \ldots, \tilde{b}_k$ *принадлежат подпространству, натянутому на векторы* $a_1 + e_1, \ldots, a_k + e_k$; *при этом*

$$\|e_k\|_E \lesssim p^{-t+1}\,\|a_k\|_E. \qquad (25.7)$$

Если внимательно посмотреть на полученный результат, то он должен показаться удивительным. В самом деле, для реализации процесса ортогонализации n векторов размерности n надо выполнить порядка $2n^3$ арифметических операций. Однако правые части неравенств (25.7) совсем *не зависят* от n. Более того, они только в два раза превышают нормы эквивалентных возмущений, которые возникают лишь при округлении мантисс координат векторов a_k до t знаков. С такими замечательными результатами мы будем встречаться не часто.

Заметим, что наш вывод относится лишь к тому, насколько вычисленная система векторов $\tilde{b}_1, \ldots, \tilde{b}_n$ неэквивалентна исходной системе $a_1, \ldots, a_n$. Мы еще не говорили о степени близости системы $\tilde{b}_1, \ldots, \tilde{b}_n$ к ортонормированной.

Посмотрим, как сказываются ошибки в вычисленных векторах $\tilde{b}_1, \ldots, \tilde{b}_k$ на ортогональность вектора $\tilde{v}_{k+1}$ к этим векторам. Предположим, что

$$\max_{1 \leqslant i \leqslant k} \|b_i - \tilde{b}_i\|_E \leqslant \tau, \qquad (25.8)$$

где τ — достаточно малое число. С помощью несложных вычислений получаем

$$(\tilde{b}_i,\ \tilde{b}_j) = \delta_{ij} + \tau_{ij}.$$

Здесь δ_{ij} — символ Кронекера и все τ_{ij} суть величины порядка τ. Пусть по векторам $a_{k+1}, \tilde{b}_1, \ldots, \tilde{b}_k$ вектор $\tilde{v}_{k+1}$ вычисляется без ошибок. В силу (25.5), (25.8) имеем

$$(\tilde{v}_{k+1},\ \tilde{v}_{k+1})^{1/2} = \left((a_{k+1},\ a_{k+1}) - \sum_{i=1}^{k} (a_{k+1},\ b_i)^2\right)^{1/2} + O(\tau).$$

Отсюда для $j \leqslant k$ следует, что

$$\tau_{k+1,j} = \frac{(\tilde{v}_{k+1},\ \tilde{b}_j)}{(\tilde{v}_{k+1},\ \tilde{v}_{k+1})^{1/2}} =$$

$$= \frac{(a_{k+1},\ \tilde{b}_j) - \sum_{i=1}^{k} (a_{k+1},\ \tilde{b}_i)(\tilde{b}_i,\ \tilde{b}_j)}{(\tilde{v}_{k+1},\ \tilde{v}_{k+1})^{1/2}} = -\frac{\sum_{i=1}^{k} (a_{k+1},\ \tilde{b}_i)\,\tau_{ij}}{(\tilde{v}_{k+1},\ \tilde{v}_{k+1})^{1/2}} =$$

$$= -\frac{\sum_{i=1}^{k} (a_{k+1},\ b_i)\,\tau_{ij}}{\left((a_{k+1},\ a_{k+1}) - \sum_{i=1}^{k} (a_{k+1},\ b_i)^2\right)^{1/2}} + O(\tau^2).$$

Введем в рассмотрение угол $\{a_{k+1},\ P_k\}$ между вектором a_{k+1} и подпространством P_k согласно [1]. Несложные вычисления показывают, что

$$\operatorname{ctg}^2\{a_{k+1},\ P_k\} = \frac{\sum_{i=1}^{k} (a_{k+1},\ b_i)^2}{(a_{k+1},\ a_{k+1}) - \sum_{i=1}^{k} (a_{k+1},\ b_i)^2}. \qquad (25.9)$$

Если среди чисел $\operatorname{ctg}^2\{a_{k+1},\ P_k\}$, $k \geqslant 1$ есть большие по модулю, то большими по модулю будут и некоторые из отношений

$$\frac{(a_{k+1},\ b_j)}{\left((a_{k+1},\ a_{k+1}) - \sum_{i=1}^{k} (a_{k+1},\ b_i)^2\right)^{1/2}}. \qquad (25.10)$$

Следовательно, даже при точном выполнении всех арифметических операций на $(k+1)$-м шаге процесса величины ошибок $\tau_{k+1,j}$ могут стать значительными по сравнению с ошибками τ_{ij}, полученными на предыдущих шагах.

Если мы проследим распространение ошибок на несколько последующих шагов, то положение окажется еще более серьезным, так как первоначальные ошибки будут умножаться на произведение отношений вида (25.10). Это означает, что для того, чтобы в значительной мере нарушилась ортогональность системы векторов $\tilde{b}_1, \ldots, \tilde{b}_n$, совсем не обязательно, чтобы какое-либо из отношений (25.10) было большим. Достаточно, чтобы большим было произведение таких отношений из различных шагов.

Рассмотренная нами схема метода ортогонализации является наиболее распространенной на практике. Как показали наши исследования, она *обеспечивает очень высокую степень устойчивости в смысле малости эквивалентных возмущений и очень неустойчива в смысле сохранения ортогональности системы получаемых векторов* $\tilde{b}_1, \ldots, \tilde{b}_n$.

Для устранения отмеченной неустойчивости будем несколько иначе определять вектор v_{k+1}. Условие его ортогональности к вычисленным векторам $\tilde{b}_1, \ldots, \tilde{b}_k$ дает для определения коэффициентов c_i линейной комбинации (25.2) следующую систему линейных алгебраических уравнений:

$$\begin{aligned}
&c_1(\tilde{b}_1,\ \tilde{b}_1) + c_2(\tilde{b}_2,\ \tilde{b}_1) + \ldots + c_k(\tilde{b}_k,\ \tilde{b}_1) = -(a_{k+1},\ \tilde{b}_1),\\
&c_1(\tilde{b}_1,\ \tilde{b}_k) + c_2(\tilde{b}_2,\ \tilde{b}_2) + \ldots + c_k(\tilde{b}_k,\ \tilde{b}_2) = -(a_{k+1},\ \tilde{b}_2),\\
&\qquad\qquad \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \qquad (25.11)\\
&c_1(\tilde{b}_1,\ \tilde{b}_k) + c_2(\tilde{b}_2,\ \tilde{b}_k) + \ldots + c_k(\tilde{b}_k,\ \tilde{b}_k) = -(a_{k+1},\ \tilde{b}_k).
\end{aligned}$$

Положим $v_{k+1} = \lim\limits_{s\to\infty} v_{k+1}^{(s)}$, где

$$v_{k+1}^{(s)} = v_{k+1}^{(s-1)} - \sum_{i=1}^{k} (v_{k+1}^{(s-1)},\ \tilde{b}_i)\,\tilde{b}_i, \qquad v_{k+1}^{(0)} = a_{k+1}. \qquad (25.12)$$

Обозначим через B матрицу системы (25.11), а через $\omega^{(s)}$ — вектор

$$\omega^{(s)} = ((v_{k+1}^{(s)},\ \tilde{b}_1),\ (v_{k+1}^{(s)},\ \tilde{b}_2),\ \ldots,\ (v_{k+1}^{(s)},\ \tilde{b}_k))'.$$

Из рекуррентного соотношения (25.12) получаем, что векторы $\omega^{(s)}$ и $\omega^{(s-1)}$ связаны между собой равенством $\omega^{(s)} = (E - B)\,\omega^{(s-1)}$, откуда вытекает, что

$$|\omega^{(s)}|_E \leqslant |E - B|_E\,|\omega^{(s-1)}|_E \leqslant \ldots \leqslant |E - B|_E^s\,|\omega^{(0)}|_E.$$

Если векторы $\tilde{b}_1, \ldots, \tilde{b}_k$ близки к ортонормированным, то $|E - B|_E < 1$ и последовательность векторов $\omega^{(s)}$ сходится к нулю. Следовательно, последовательность векторов $v_{k+1}^{(s)}$ сходится к такому вектору v_{k+1}, который ортогонален векторам $\tilde{b}_1, \ldots, \tilde{b}_k$. По построению этот вектор является вектором вида (25.2).

Таким образом, при точной реализации итерационного процесса (25.12) ошибки от неортогональности системы

векторов $\tilde{b}_1, \ldots, \tilde{b}_k$ не оказывают влияние на ортогональность к этим векторам всех последующих векторов.

При реальных вычислениях правые части соотношений (25.12) не могут быть определены точно, поэтому в действительности найденный вектор $\tilde{v}_{k+1}$ все же не будет ортогонален векторам $\tilde{b}_1, \ldots, \tilde{b}_k$. Будем считать, что векторы $\tilde{b}_1, \ldots, \tilde{b}_k$ не слишком сильно отличаются от ортонормированных и выполняется условие (25.6). В этом случае для всех s

$$\|\tilde{v}_{k+1}^{(s)}\|_E \leqslant \|a_{k+1}\|_E. \tag{25.13}$$

Предположим далее, что при вычислении скалярных произведений используется режим накопления. Для всех j будем иметь

$$\tilde{v}_{j,\,k+1}^{(s)} = \left(\tilde{v}_{j,\,k+1}^{(s-1)} - \sum_{i=1}^{k} (\tilde{v}_{k+1}^{(s-1)}, \tilde{b}_i)(1+\eta_{ji}^{(s-1)})\tilde{b}_{ji}\right)(1+\zeta_j^{(s-1)}). \tag{25.14}$$

Здесь $\eta_{ji}^{(s-1)}$, $\zeta_j^{(s-1)}$ удовлетворяют обычным для ошибок соотношениям. Из (25.14) вытекает, что для $1 \leqslant l \leqslant k$

$$(\tilde{v}_{k+1}^{(s)}, \tilde{b}_l) = \left(\tilde{v}_{k+1}^{(s-1)} - \sum_{i=1}^{k} (\tilde{v}_{k+1}^{(s-1)}, \tilde{b}_i)\, b_i,\ \tilde{b}_l\right) - (\tilde{v}_{k+1}^{(s-1)}, \tilde{b}_l)\,\eta_{jl}^{(s-1)} + (\hat{v}_{k+1}^{(s-1)}, \tilde{b}_l). \tag{25.15}$$

При этом координаты $\hat{v}_{j,\,k+1}^{(s-1)}$ вектора $\hat{v}_{k+1}^{(s-1)}$ связаны с величинами в (25.14) такими соотношениями:

$$\hat{v}_{j,\,k+1}^{(s-1)} = \left(\tilde{v}_{j,\,k+1}^{(s-1)} - \sum_{i=1}^{k} (\tilde{v}_{k+1}^{(s-1)}, \tilde{b}_i)\, b_{ji}\right)\zeta_j^{(s-1)}.$$

Обозначим

$$\tilde{w}^{(s)} = ((\tilde{v}_{k+1}^{(s)}, \tilde{b}_1), (\tilde{v}_{k+1}^{(s)}, \tilde{b}_2), \ldots, (\tilde{v}_{k+1}^{(s)}, \tilde{b}_k))'.$$

Из (25.15) следует равенство

$$\tilde{w}^{(s)} = (E-B)\,\tilde{w}^{(s-1)} + \sigma^{(s-1)},$$

где координаты $\sigma_l^{(s-1)}$, $1 \leqslant l \leqslant k$, вектора $\sigma^{(s-1)}$ таковы:

$$\sigma_l^{(s-1)} = -(\tilde{v}_{k+1}^{(s-1)}, \tilde{b}_l)\,\eta_{jl}^{(s-1)} + (\hat{v}_{k+1}^{(s-1)}, b_l).$$

Принимая во внимание (25.6), (25.13), находим, что

$$\|\sigma^{(s-1)}\|_E \leqslant p^{-t+1}\|a_{k+1}\|_E.$$

Окончательно имеем

$$\|\tilde{w}^{(s)}\|_E \leqslant \|E-B\|_E\|\tilde{w}^{(s-1)}\|_E + \|\sigma^{(s-1)}\|_E \leqslant$$
$$\leqslant \|E-B\|_E\|\tilde{w}^{(s-1)}\|_E + p^{-t+1}\|a_{k+1}\|_E \leqslant$$
$$\leqslant \|E-B\|_E^{s}\|\tilde{w}^{(0)}\|_E + p^{-t+1}\|a_{k+1}\|_E \sum_{i=0}^{s-1}\|E-B\|_E^{i} \leqslant$$
$$\leqslant \|E-B\|_E^{s}\|\tilde{w}^{(0)}\|_E + \frac{p^{-t+1}\|a_{k+1}\|_E}{1-\|E-B\|_E}.$$

Полученное неравенство означает, что, начиная с некоторого s, вычисленный вектор $\tilde{v}_{k+1}^{(s)}$ будет почти ортогонален векторам $\tilde{b}_1, \ldots, \tilde{b}_k$. Именно, евклидова норма его проекции на подпространство $\bar{P}_k$ асимптотически не будет превосходить $p^{-t+1}\|a_{k+1}\|_E$. Снова мы получили замечательный результат в отношении точности метода ортогонализации. Ведь только округление мантисс координат вектора a_{k+1} до t знаков изменяет евклидову норму его проекции на подпространство $\bar{P}_k$ на величину порядка $(1/2)\,p^{-t+1}\|a_{k+1}\|_E$. Итерационный процесс (25.12) дает ошибку лишь вдвое больше.

Отметим, что при практической реализации метода ортогонализации почти всегда можно брать $s=2$. При этом оценки эквивалентных возмущений по сравнению с (25.7) увеличиваются не более чем вдвое.

Рассмотренный процесс исправления неортогональности вычисляемых векторов $\tilde{b}_1, \tilde{b}_2, \ldots, \tilde{b}_n$ называется *переортогонализацией*. Возможны и другие способы исправления неортогональности. Пусть, например, процесс ортогонализации проводится при $s=1$, т. е. без переортогонализации. К полученной системе векторов снова применяется процесс ортогонализации при $s=1$. И так далее до тех пор, пока система векторов не станет ортогональной с нужной точностью. Этот процесс называется *повторной ортогонализацией*. По эффективности он уступает процессу переортогонализации. Уже при первом выполнении процесса ортогонализации вычисляемые векторы могут стать столь неортогональными, что проводить процесс ортогонализации до конца оказывается бессмысленно.

УПРАЖНЕНИЯ

1. Что меняется в процессе ортогонализации, если исходная система векторов линейно зависима?

2. Доказать, что существует такое унитарное преобразование заданной системы векторов $a_1, a_2, \ldots, a_n$, при котором преобразованные векторы, расположенные по строкам, образуют левую треугольную матрицу.

3. Доказать, что величины (25.3), (25.9) не меняются при унитарном преобразовании системы векторов $a_1, a_2, \ldots, a_n$.

4. Пусть векторы $a_1, a_2, \ldots, a_n$ являются строками левой треугольной матрицы. Что представляет собой процесс ортогонализации для этих векторов?

5. В условиях упражнения 4 вычислить величины (25.3), (25.9).

6. В условиях упражнения 4 проследить возникновение неортогональности вычисленных векторов $\bar{b}_1, \bar{b}_2, \ldots, \bar{b}_n$.

7. В условиях упражнения 4 проследить выполнение процесса переортогонализации.

8. В условиях упражнения 4 проследить выполнение процесса повторной ортогонализации.

9. Предположим, что система векторов $a_1, a_2, \ldots, a_n$ преобразуется в эквивалентную систему $b_1, b_2, \ldots, b_n$ согласно (25.1), (25.2). Доказать, что на классе таких преобразований процесс ортогонализации имеет наименьшее эквивалентное возмущение.

ГЛАВА IV

ПРЯМОЕ РАЗЛОЖЕНИЕ МАТРИЦЫ НА МНОЖИТЕЛИ

Разложение произвольной матрицы на множители позволяет во многих случаях свести решение исходной алгебраической задачи к последовательному решению нескольких аналогичных задач, но с более простыми матрицами. В этой главе мы будем изучать *прямые* методы разложения матриц, т. е. такие методы, которые реализуются за конечное число арифметических операций.

§ 26. Матрицы специального вида

Разложение матриц, как правило, основано на последовательном их преобразовании к матрицам, имеющим значительное число нулевых элементов. Такие матрицы обладают целым рядом специфических свойств. Мы опишем сейчас некоторые из этих матриц.

Треугольные матрицы. Матрица A называется *правой* (*левой*) *треугольной*, если для ее элементов a_{ij} выполняются соотношения

$$a_{ij}=0, \quad i>j \ (j>i).$$

Треугольные матрицы имеют много замечательных свойств, в силу которых они широко используются в построении самых различных методов решения задач алгебры. Так, например, для квадратных матриц сумма и произведение одноименных треугольных матриц есть треугольная матрица того же наименования, определитель треугольной матрицы равен произведению диагональных элементов, собственные значения треугольной матрицы совпадают с ее диагональными элементами, треугольная матрица легко обращается и обратная к ней также будет треугольной.

Доказательство. Предположим, что разложение (27.1) существует. Используя формулу Бине — Коши, находим

$$A\begin{bmatrix}1 & 2 & \dots & k-1 & m\\ 1 & 2 & \dots & k-1 & k\end{bmatrix} = \sum_{1\leqslant\alpha_1<\alpha_2<\dots<\alpha_k\leqslant n} B\begin{bmatrix}1 & 2 & \dots & k-1 & m\\ \alpha_1 & \alpha_2 & \dots & \alpha_{k-1} & \alpha_k\end{bmatrix} C\begin{bmatrix}\alpha_1 & \alpha_2 & \dots & \alpha_k\\ 1 & 2 & \dots & k\end{bmatrix}.$$

Так как C — правая треугольная матрица, то первые k ее столбцов содержат только один отличный от нуля минор k-го порядка, а именно главный минор. Следовательно,

$$A\begin{bmatrix}1 & 2 & \dots & k-1 & m\\ 1 & 2 & \dots & k-1 & k\end{bmatrix} = B\begin{bmatrix}1 & 2 & \dots & k-1 & m\\ 1 & 2 & \dots & k-1 & k\end{bmatrix} C\begin{bmatrix}1 & 2 & \dots & k\\ 1 & 2 & \dots & k\end{bmatrix} = b_{11}b_{22}\dots b_{k-1,k-1}b_{mk}c_{11}c_{22}\dots c_{kk}.$$

Положив в этой формуле $m=k$, получим

$$A\begin{bmatrix}1 & 2 & \dots & k\\ 1 & 2 & \dots & k\end{bmatrix} = B\begin{bmatrix}1 & 2 & \dots & k\\ 1 & 2 & \dots & k\end{bmatrix} C\begin{bmatrix}1 & 2 & \dots & k\\ 1 & 2 & \dots & k\end{bmatrix} = b_{kk}c_{kk}B\begin{bmatrix}1 & 2 & \dots & k-1\\ 1 & 2 & \dots & k-1\end{bmatrix} C\begin{bmatrix}1 & 2 & \dots & k-1\\ 1 & 2 & \dots & k-1\end{bmatrix} = b_{kk}c_{kk}A\begin{bmatrix}1 & 2 & \dots & k-1\\ 1 & 2 & \dots & k-1\end{bmatrix},$$

откуда вытекает первая группа соотношений (27.2). С другой стороны,

$$A\begin{bmatrix}1 & 2 & \dots & k-1 & m\\ 1 & 2 & \dots & k-1 & k\end{bmatrix} = \frac{b_{11}b_{22}\dots b_{kk}c_{11}c_{22}\dots c_{kk}b_{mk}}{b_{kk}} = \frac{B\begin{bmatrix}1 & 2 & \dots & k\\ 1 & 2 & \dots & k\end{bmatrix} C\begin{bmatrix}1 & 2 & \dots & k\\ 1 & 2 & \dots & k\end{bmatrix}}{b_{kk}} b_{mk} = \frac{b_{mk}}{b_{kk}} A\begin{bmatrix}1 & 2 & \dots & k\\ 1 & 2 & \dots & k\end{bmatrix},$$

что доказывает справедливость формулы для коэффициентов b_{mk}. Справедливость формулы для коэффициентов c_{km} устанавливается аналогично.

Таким образом, если разложение (27.1) существует, то с точностью до определения диагональных элементов оно единственно и определяется формулами (27.2). Существование хотя бы одного разложения мы установим несколько позднее.

Следствие. *Если матрица A эрмитова и ее главные миноры положительны, то существует разложение*

$$A = C^*C,$$

где C — правая треугольная матрица.

Докажем, что в данном случае разложение (27.1) возможно при $B=C^*$. Так как матрица A имеет положительные главные миноры, то можно взять

$$c_{11} = b_{11} = \left(A\begin{bmatrix}1\\1\end{bmatrix}\right)^{1/2} e^{i\varphi_1}, \dots, = c_{nn} = b_{nn} = \left(\frac{A\begin{bmatrix}1 & 2 & \dots & n\\ 1 & 2 & \dots & n\end{bmatrix}}{A\begin{bmatrix}1 & 2 & \dots & n-1\\ 1 & 2 & \dots & n-1\end{bmatrix}}\right)^{1/2} e^{i\varphi_n},$$

где $\varphi_1, \dots, \varphi_n$ — произвольные вещественные числа. Но тогда

$$c_{km} = b_{kk}\frac{A\begin{bmatrix}1 & 2 & \dots & k-1 & k\\ 1 & 2 & \dots & k-1 & m\end{bmatrix}}{A\begin{bmatrix}1 & 2 & \dots & k\\ 1 & 2 & \dots & k\end{bmatrix}} = \overline{b_{kk}\frac{A\begin{bmatrix}1 & 2 & \dots & k-1 & m\\ 1 & 2 & \dots & k-1 & k\end{bmatrix}}{A\begin{bmatrix}1 & 2 & \dots & k\\ 1 & 2 & \dots & k\end{bmatrix}}} = \overline{b}_{mk}.$$

Ясно, что матрица C определяется с точностью до умножения слева на диагональную матрицу, все диагональные элементы которой по модулю равны единице.

Следствие. *Если для некоторого j (i) элементы матрицы A удовлетворяют условиям*

$$a_{ij}=0, \quad i=1, 2, \dots, p<j \quad (j=1, 2, \dots, r<i),$$

то будут равны нулю и элементы матрицы C (B) с соответствующими номерами.

В самом деле, рассмотрим, например, элементы матрицы C. Согласно второй группе соотношений (27.2)

$$c_{ij} = c_{ii}\frac{A\begin{bmatrix}1 & 2 & \dots & i-1 & i\\ 1 & 2 & \dots & i-1 & j\end{bmatrix}}{A\begin{bmatrix}1 & 2 & \dots & i\\ 1 & 2 & \dots & i\end{bmatrix}}.$$

Но в силу равенства нулю элементов j-го столбца матрицы A заключаем, что

$$A\begin{bmatrix}1 & 2 & \dots & i-1 & i\\ 1 & 2 & \dots & i-1 & j\end{bmatrix} = 0, \quad i\leqslant p,$$

откуда и вытекает справедливость высказанного утверждения. Равенство нулю соответствующих элементов матрицы B доказывается аналогично.

Во многих прикладных задачах приходиться иметь дело с «разреженными» матрицами, т. е. матрицами, имеющими много нулевых элементов. Установленное следствие позволяет описать целый класс «разреженных» матриц, треугольные сомножители которых сохраняют специфику «разреженности» исходной матрицы. Пусть матрица A удовлетворяет условию теоремы 27.1 и имеет вид

$$A = \qquad (27.3)$$

где все ее ненулевые элементы находятся в заштрихованной области. Граница этой области может быть произвольной. Требуется лишь, чтобы любая вертикальная (горизонтальная) прямая линия имела с правой (левой) частью границы односвязное множество общих точек. Как вытекает из второго следствия теоремы 27.1, треугольные сомножители B и C будут иметь аналогичный вид. Именно

$$B = \quad , \quad C = \quad . \qquad (27.4)$$

Все ненулевые элементы матриц B и C находятся в заштрихованных сбластях, границы которых такие же, как у матрицы A.

Рассмотренное свойство матриц (27.3) позволяет получить два важных следствия. Пусть матрица A — ленточная и $a_{ij}=0$, если $l \leqslant i-j$ или $j-i \leqslant m$ для некоторых целых неотрицательных чисел l, m. В этом случае

матрица B в разложении (27.1) будет левой ленточной, матрица C — правой ленточной. При этом $b_{ij}=0$, если $l \leqslant i-j$, $c_{ij}=0$, если $j-i \leqslant m$. Если матрица A правая (левая) почти треугольная, то в разложении (27.1) матрица B (C) будет левой (правой) двухдиагональной.

Теорема 27.2. *Всякую невырожденную квадратную матрицу A порядка n можно представить в виде произведения унитарной матрицы U на правую треугольную матрицу C, т. е.*

$$A = UC. \qquad (27.5)$$

При этом

$$c_1 = \left(A^*A\begin{bmatrix}1\\1\end{bmatrix}\right)^{1/2} e^{i\varphi_1}, \ \ldots, \ c_{kk} = \left(\frac{A^*A\begin{bmatrix}1 & 2 & \ldots & k\\ 1 & 2 & \ldots & k\end{bmatrix}}{A^*A\begin{bmatrix}1 & 2 & \ldots & k-1\\ 1 & 1 & \ldots & k-1\end{bmatrix}}\right)^{1/2} e^{i\varphi_k},$$

$$c_{km} = c_{kk}\frac{A^*A\begin{bmatrix}1 & 2 & \ldots & k-1 & k\\ 1 & 2 & \ldots & k-1 & m\end{bmatrix}}{A^*A\begin{bmatrix}1 & 2 & \ldots & k\\ 1 & 2 & \ldots & k\end{bmatrix}}, \qquad (27.6)$$

где $m=k+1, \ldots, n$, $k=1, 2, \ldots, n-1$, а $\varphi_1, \ldots, \varphi_n$ — произвольные вещественные числа.

Доказательство. Пусть разложение (27.5) существует; тогда матрица C должна удовлетворять уравнению

$$A^*A = C^*C. \qquad (27.7)$$

Матрица A^*A эрмитова и ее главные миноры положительны, поэтому согласно первому следствию теоремы 27.1 матрица C может быть определена из уравнения (27.7). Представим матрицу A в виде $A=(AC^{-1})C$. Принимая во внимание (27.7), имеем

$$(AC^{-1})(AC^{-1})^* = AC^{-1}C^{-1*}A^* = A(C^*C)^{-1}A^* =$$

$$= A(A^*A)^{-1}A^* = AA^{-1}A^{*-1}A^* = E.$$

Следовательно, матрица AC^{-1} унитарная и возможность разложения (27.5) доказана.

Матрица C (U) определяется с точностью до умножения слева (справа) на диагональную матрицу, все элементы которой по модулю равны единице. Формулы (27.6) получаются непосредственно из соответствующих формул теоремы 27.1 и ее следствия.

УПРАЖНЕНИЯ

1. Доказать, что любая невырожденная матрица приводится к матрице, удовлетворяющей условиям теоремы 27.1, путем перестановок строк или столбцов.

2. Обозначим через A_k, B_k, C_k матрицы главных миноров порядка k для матриц A, B, C в разложении (27.1). Доказать, что для всех k справедливо равенство $A_k = B_k C_k$.

3. Сформулировать условия, при которых матрица может быть разложена в произведение правой и левой треугольных матриц.

4. На какие другие треугольные множители можно разложить квадратную матрицу? Каковы условия существования этих разложений?

5. Доказать, что первые k векторов-столбцов матрицы U в разложении (27.5) для всех k представляют собой ортонормированный базис подпространства, натянутого на первые k векторов-столбцов матрицы A.

6. Доказать, что любую невырожденную матрицу можно представить в виде произведения левой треугольной и унитарной матриц.

7. Какие еще возможны варианты разложения невырожденной матрицы в произведение двух матриц, из которых одна унитарная, а вторая треугольная?

8. Как связаны между собой определитель матрицы A и определители треугольных матриц в разложениях (27.1), (27.5)?

§ 28. Разложение на треугольные множители

Пусть матрица A удовлетворяет условиям теоремы 27.1. Так как в этом случае $a_{11} \neq 0$, то по элементам первого столбца матрицы A можно построить матрицу N_1 вида (24.3) такую, что в произведении $A_1 = N_1 A$ все поддиагональные элементы первого столбца будут нулевыми. При этом диагональный элемент останется без изменения и, следовательно, будет отличен от нуля.

Будем считать, что уже вычислены матрицы $N_1, N_2 \ldots$ $\ldots, N_r$ вида (24.3) для некоторого $r \geqslant 1$ и построена последовательность матриц

$$A_k = N_k A_{k-1}, \quad A_0 = A \tag{28.1}$$

для $1 \leqslant k \leqslant r$. Обозначим элементы матрицы A_k через $a_{ij}^{(k)}$ и предположим, что

$$\begin{aligned} a_{ij}^{(k)} &= 0 \quad \text{для} \quad i > j, \ j \leqslant r, \\ a_{ii}^{(k)} &\neq 0 \quad \text{для} \quad i \leqslant r. \end{aligned} \tag{28.2}$$

Это предположение заведомо выполняется при $r = 1$.

Покажем, что отличие от нуля главных миноров матрицы A позволяет продолжить процесс (28.1). Из (28.1) находим

$$A = (N_1^{-1} \ldots N_r^{-1}) A_r. \tag{28.3}$$

Применяя формулу Бине — Коши [1], получаем

$$A \begin{bmatrix} 1 & 2 \ldots r+1 \\ 1 & 2 \ldots r+1 \end{bmatrix} = \sum_{\alpha_l} (N_1^{-1} \ldots N_r^{-1}) \begin{bmatrix} 1 & 2 & \ldots & r+1 \\ \alpha_1 & \alpha_2 & \ldots & \alpha_{r+1} \end{bmatrix} A_r \begin{bmatrix} \alpha_1 & \alpha_2 \ldots \alpha_{r+1} \\ 1 & 2 \ldots r+1 \end{bmatrix}.$$

Матрица $N_1^{-1} \ldots N_r^{-1}$ левая треугольная с диагональными элементами, равными единице. Поэтому среди миноров $(r+1)$-го порядка, расположенных в первых $r+1$ строках, лишь главный минор отличен от нуля и равен он единице. Но тогда

$$A \begin{bmatrix} 1 & 2 \ldots r+1 \\ 1 & 2 \ldots r+1 \end{bmatrix} = A_r \begin{bmatrix} 1 & 2 \ldots r+1 \\ 1 & 2 \ldots r+1 \end{bmatrix} = a_{11}^{(1)} \ldots a_{rr}^{(r)} a_{r+1, r+1}^{(r)}.$$

Левая часть этого соотношения отлична от нуля в силу условий на главные миноры матрицы A, элементы $a_{11}^{(1)}, \ldots, a_{rr}^{(r)}$ отличны от нуля в силу предположений (28.2). Следовательно, $a_{r+1, r+1}^{(r)} \neq 0$. Теперь мы можем по элементам $(r+1)$-го столбца матрицы A_r построить матрицу N_{r+1} вида (24.3) и матрицу $A_{r+1} = N_{r+1} A_r$. В этой матрице по сравнению с A_r не меняются элементы в первых r столбцах и первых $r+1$ строках. По построению матрицы N_{r+1} поддиагональные элементы в $(r+1)$-м столбце матрицы A_{r+1} будут нулевыми.

Таким образом, процесс (28.1) продолжен еще на один шаг и условия (28.2) выполняются при замене r на $r+1$. Поэтому процесс (28.1) можно продолжить до $r = n-1$. Согласно (28.3) получаем, что

$$A = (N_1^{-1} \ldots N_{n-1}^{-1}) A_{n-1}.$$

Как следует из (24.4), матрица $N_1^{-1} \ldots N_{n-1}^{-1}$ левая треугольная с единичными диагональными элементами. Матрица A_{n-1} — правая треугольная. Условия, наложенные на матрицу A, гарантируют отличие от нуля всех диагональных элементов матрицы A_{n-1}, кроме, может быть, последнего.

Описанный процесс не только доказывает возможность разложения матрицы на треугольные множители, но и может быть использован для численного нахождения этого разложения. Он получил название — *метод Гаусса* для треугольного разложения матрицы.

Конечно, из-за влияния ошибок округления все матрицы будут определены неточно. Перемножение реально найденных треугольных матриц уже не даст исходную матрицу A и в действительности

$$(\tilde{N}_1 \dots \tilde{N}_{n-1})\tilde{A}_{n-1} = A + M,$$

где M — эквивалентное возмущение.

Процесс (28.1) в терминах преобразования столбцов был описан и исследован в § 24. Теперь вместо (24.6) имеем

$$\tilde{A}_k = fl(\tilde{N}_k \tilde{A}_{k-1}) \equiv \tilde{N}_k \tilde{A}_{k-1} + M_{k-1};$$

формула (24.7) означает, что

$$M = \sum_{k=1}^{n-1} M_{k-1}. \tag{28.4}$$

Ясно, что для оценки элементов эквивалентного возмущения M необходимо оценить элементы матриц ошибок M_{k-1}, возникающих при выполнении отдельных шагов.

Обозначим через $\tilde{n}_{ij}^{(k)}$, $\tilde{a}_{ij}^{(k)}$, $p_{ij}^{(k)}$ элементы реально вычисленных матриц $\tilde{N}_k$, $\tilde{A}_k$, M_k. Элементы матрицы $\tilde{A}_k$ естественным образом разбиваются на три различные группы согласно способам их получения. Именно,

$$\tilde{a}_{ij}^{(k)} = \begin{cases} \tilde{a}_{ij}^{(k-1)}, \text{ если } i \leqslant k, \text{ или } j < k, \\ (\tilde{a}_{ij}^{(k-1)} + \tilde{n}_{ik}^{(k)} \tilde{a}_{kj}^{(k-1)}(1 + \varepsilon_{ij}^{(k-1)}))(1 + \tau_{ij}^{(k-1)}), \\ \qquad \text{если } i, j > k, \\ 0, \text{ если } i > k, \ j = k. \end{cases}$$

Здесь $\varepsilon_{ij}^{(k-1)}$, $\tau_{ij}^{(k-1)}$ принимают обычные значения для ошибок выполнения арифметических операций. Соответственно этим группам будем получать и оценки для элементов матрицы M_{k-1}.

Наиболее просто оцениваются ошибки для элементов первой группы. Так как эти элементы не изменяются, то мы можем считать, что

$$\mu_{ij}^{(k-1)} = 0, \tag{28.5}$$

если либо $i \leqslant k$, либо $j < k$.

Рассмотрим ошибки, возникающие при вычислении элементов второй группы. Имеем

$$\tilde{a}_{ij}^{(k)} = (\tilde{a}_{ij}^{(k-1)} + \tilde{n}_{ik}^{(k)} \tilde{a}_{kj}^{(k-1)}(1 + \varepsilon_{ij}^{(k-1)}))(1 + \tau_{ij}^{(k-1)}) = \\ = \tilde{a}_{ij}^{(k-1)} + \tilde{n}_{ik}^{(k)} \tilde{a}_{kj}^{(k-1)} + \mu_{ij}^{(k-1)}.$$

Предположим, что $\varepsilon_{ij}^{(k-1)} = -1$. Это означает, что к элементу $\tilde{a}_{ij}^{(k-1)}$ прибавляется нулевой элемент. Но тогда $\tau_{ij}^{(k-1)} = 0$. Из равенства $\varepsilon_{ij}^{(k-1)} = -1$ следует неравенство $|\tilde{n}_{ik}^{(k)} \tilde{a}_{kj}^{(k-1)}| < \omega$, поэтому

$$|\mu_{ij}^{(k-1)}| < \omega. \tag{28.6}$$

Если же $\varepsilon_{ij}^{(k-1)} \neq -1$, но $\tau_{ij}^{(k-1)} = -1$, то

$$\tilde{a}_{ij}^{(k)} = \tilde{a}_{ij}^{(k-1)} + \tilde{n}_{ik}^{(k)} \tilde{a}_{kj}^{(k-1)}(1 + \varepsilon_{ij}^{(k-1)}) + \sigma_{ij}^{(k-1)}, \tag{28.7}$$

где $|\sigma_{ij}^{(k-1)}| < \omega$. Отсюда вытекает, что

$$\mu_{ij}^{(k-1)} = \tilde{n}_{ik}^{(k)} \tilde{a}_{kj}^{(k-1)} \varepsilon_{ij}^{(k-1)} + \sigma_{ij}^{(k-1)}.$$

Принимая во внимание (28.7), получаем

$$|\tilde{n}_{ik}^{(k)} \tilde{a}_{kj}^{(k-1)}| \lessapprox |\tilde{a}_{ij}^{(k)}| + |\tilde{a}_{ij}^{(k-1)}| + |\sigma_{ij}^{(k-1)}|.$$

Следовательно, в данном случае

$$|\mu_{ij}^{(k-1)}| \lessapprox \frac{1}{2} p^{t+1}(|\tilde{a}_{ij}^{(k)}| + |\tilde{a}_{ij}^{(k-1)}|) + \omega.$$

Наконец, если $\varepsilon_{ij}^{(k-1)} \neq -1$, $\tau_{ij}^{(k-1)} \neq -1$, то

$$\mu_{ij}^{(k-1)} \approx \tilde{n}_{ik}^{(k)} \tilde{a}_{kj}^{(k-1)} \varepsilon_{ij}^{(k-1)} + \tilde{a}_{ij}^{(k-1)} \tau_{ij}^{(k-1)},$$

что дает

$$|\mu_{ij}^{(k-1)}| \leqslant \frac{1}{2} p^{t+1}(|\tilde{a}_{ij}^{(k)}| + 2|\tilde{a}_{ij}^{(k-1)}|). \tag{28.8}$$

Объединяя (28.6) — (28.8), заключаем, что

$$|\mu_{ij}^{(k-1)}| < \frac{1}{2} p^{t+1}(|\tilde{a}_{ij}^{(k)}| + 2|\tilde{a}_{ij}^{(k-1)}|) + \omega, \tag{28.9}$$

если $i, j > k$.

Несмотря на то, что все элементы третьей группы равны нулю, ошибки здесь все же возникают. Это объясняется тем, что в общем случае соответствующие элементы матрицы $\bar{N}_k\bar{A}_{k-1}$ не будут равны нулю из-за ошибок, возникших при вычислении элементов матрицы $\bar{N}_k$. Поэтому

$$\mu_{ik}^{(k-1)} = -\left(\bar{a}_{ik}^{(k-1)} + \bar{n}_{ik}^{(k)}\bar{a}_{kk}^{(k-1)}\right).$$

Но

$$\bar{n}_{ik}^{(k)} = \mathrm{fl}\left(-\frac{\bar{a}_{ik}^{(k-1)}}{\bar{a}_{kk}^{(k-1)}}\right) = -\frac{\bar{a}_{ik}^{(k-1)}}{\bar{a}_{kk}^{(k-1)}}\left(1+\nu_{ik}^{(k-1)}\right),$$

следовательно,

$$\mu_{ik}^{(k-1)} = \bar{a}_{ik}^{(k-1)}\nu_{ik}^{(k-1)}.$$

Если $\nu_{ik}^{(k-1)} \neq -1$, то это означает, что

$$|\mu_{ik}^{(k-1)}| \leqslant \frac{1}{2}\,p^{t+1}\,|\bar{a}_{ik}^{(k-1)}|. \tag{28.10}$$

В случае, когда $\nu_{ik}^{(k-1)} = -1$, мы имеем

$$\left|\frac{\bar{a}_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}\right| < \omega,$$

и тогда

$$|\mu_{ik}^{(k-1)}| < |\bar{a}_{kk}^{(k-1)}|\,\omega. \tag{28.11}$$

Из (28.10), (28.11) вытекает, что в любом случае

$$|\mu_{ik}^{(k-1)}| \leqslant \frac{1}{2}\,p^{t+1}\,|\bar{a}_{ik}^{(k-1)}| + |\bar{a}_{kk}^{(k-1)}|\,\omega, \tag{28.12}$$

если, конечно, $i > k$.

Полученные оценки ошибок позволяют оценить элементы μ_{ij} эквивалентного возмущения M. Принимая во внимание соотношение (28.4) и оценки (28.5), (28.9), (28.12), находим, что

$$|\mu_{ij}| \lesssim \begin{cases} 0, & i = 1, \\ p^{t+1}\left(\frac{1}{2}|\bar{a}_{ij}^{(0)}| + 1{,}5|\bar{a}_{ij}^{(1)}| + \ldots + 1{,}5|\bar{a}_{ij}^{(i-2)}| + \right. \\ \left. + 0{,}5\,|\bar{a}_{ij}^{(i-1)}|\right) + (i-1)\,\omega, & j \geqslant i, \\ p^{t+1}\left(\frac{1}{2}|a_{ij}^{(0)}| + 1{,}5|\bar{a}_{ij}^{(1)}| + \ldots + 1{,}5|\bar{a}_{ij}^{(j-2)}| + \right. \\ \left. + |\bar{a}_{ij}^{(j-1)}|\right) + (j-1)\,\omega + |\bar{a}_{jj}^{(j-1)}|\,\omega, & j < i. \end{cases} \tag{28.13}$$

Если обозначить

$$a_k = \max_{\substack{i \geqslant k \\ j \geqslant k}} |\bar{a}_{ij}^{(k)}|$$

и пренебречь членами с ω, то из (28.13) следует, что

$$|\mu_{ij}| \lesssim \begin{cases} 0, & i = 1, \\ p^{t+1}\,(a_0 + 1{,}5a_1 + \ldots + 1{,}5a_{i-2} + 0{,}5a_{i-1}), & j \geqslant i, \\ p^{t+1}\,(a_0 + 1{,}5a_1 + \ldots + 1{,}5a_{j-2} + a_{j-1}), & j < i. \end{cases} \tag{28.14}$$

Если же обозначить

$$a = \max_{i,\,j,\,k} |\bar{a}_{ij}^{(k)}|,$$

то будем иметь

$$|\mu_{ij}| \leqslant \begin{cases} 0, & i = 1, \\ 1{,}5\,(i-1)\,p^{t+1}a, & j \geqslant i, \\ (1{,}5j-1)\,p^{t+1}a, & j < i. \end{cases} \tag{28.15}$$

Оценки (28.13) — (28.15) получены *без каких-либо предположений* относительно величины главных миноров матрицы A. Они еще раз подтвердили высказанное ранее мнение о том, что существенным источником неустойчивости в процессах типа (28.1) может быть лишь *значительный рост* элементов промежуточных матриц A_k в процессе (28.1).

Если не менять принципиально общую схему вычислений, то единственной возможностью в какой-то мере регулировать рост элементов является использование *перестановок* при реализации процесса (28.1).

Выберем в исходной матрице A любой элемент $\bar{a}_{i_1j_1}^{(0)} \neq 0$, который назовем *ведущим* или *главным* элементом 1-го шага, и рассмотрим матрицу $P_{1i_1}AP_{1j_1}$. В этой матрице в позиции (1.1) находится ненулевой элемент $\bar{a}_{i_1j_1}^{(0)}$. Поэтому можно построить матрицу $\hat{A}_1 = \hat{N}_1(P_{1i_1}AP_{1j_1})$ с нулевыми поддиагональными элементами в первом столбце. Предположим, что аналогично матрицам $A_1, \ldots, A_{k-1}$ в (28.1) уже вычислены матрицы $\hat{A}_1, \ldots, \hat{A}_{k-1}$. Среди элементов $\bar{a}_{ij}^{(k-1)}$ матрицы $\hat{A}_{k-1}$, удовлетворяющих условиям $i, j \geqslant k$, выберем любой элемент $\bar{a}_{i_kj_k}^{(k-1)} \neq 0$,

который назовем *ведущим* или *главным* элементом k-го шага, и построим матрицу

$$\tilde{A}_k = \hat{N}_k (P_{ki_k} \tilde{A}_{k-1} P_{ki_k}). \qquad (28.16)$$

Так как в позиции (k, k) матрицы $P_{ki_k}\tilde{A}_{k-1}P_{ki_k}$ находится ненулевой элемент, то матрицу $\hat{N}_k$ можно выбрать так, что поддиагональные элементы k-го столбца матрицы $\tilde{A}_k$ будут нулевыми. При этом, очевидно, сохраняются все нулевые элементы, полученные ранее в матрице $\tilde{A}_{k-1}$. Продолжая данный процесс, мы приходим теперь к разложению

$$A = (P_{1i_1}\hat{N}_1^{-1} \dots P_{n-1,\, i_{n-1}}\hat{N}_{n-1}^{-1})(\tilde{A}_{n-1}P_{n-1,\, i_{n-1}} \dots P_{1i_1}). \qquad (28.17)$$

Матрицы, стоящие в круглых скобках (28.17), уже не являются треугольными. Поэтому может показаться, что анализ ошибок, выполненный для процесса (28.1), не пригоден для процесса (28.16). Однако в действительности между обоими процессами имеется очень тесная связь. В самом деле, преобразуем выражения в левых скобках (28.17) следующим образом:

$$P_{1i_1}\hat{N}_1^{-1}P_{2i_2}\hat{N}_2^{-1}P_{3i_3} \dots \hat{N}_{n-2}^{-1}P_{n-1,\, i_{n-1}}\hat{N}_{n-1}^{-1} =$$

$$= (P_{1i_1}P_{2i_2} \dots P_{n-1,\, i_{n-1}}) \times (P_{n-1,\, i_{n-1}} \dots P_{2i_2}\hat{N}_1^{-1}P_{2i_2} \dots$$

$$\dots P_{n-1,\, i_{n-1}}) \times (P_{n-1,\, i_{n-1}} \dots P_{3i_3}\hat{N}_2^{-1}P_{3i_3} \dots P_{n-1,\, i_{n-1}}) \times \dots$$

$$\dots \times (P_{n-1,\, i_{n-1}}\hat{N}_{n-2}^{-1}P_{n-1,\, i_{n-1}}) \times \hat{N}_{n-1}^{-1}.$$

Напомним, что $i_k \geqslant k$ для всех k. Если обозначить

$$P_{n-1,\, i_{n-1}} \dots P_{ki_k}\hat{N}_{k-1}^{-1}P_{ki_k} \dots P_{n-1,\, i_{n-1}} = \tilde{N}_{k-1}^{-1}, \qquad (28.18)$$

то $\tilde{N}_{k-1}^{-1}$ снова есть матрица вида (24.3) и отличается от матрицы $\hat{N}_{k-1}^{-1}$ лишь перестановкой поддиагональных элементов в $(k-1)$-м столбце. Чтобы получить элементы матрицы $\tilde{N}_{k-1}^{-1}$, необходимо последовательно переставить элементы матрицы $\hat{N}_{k-1}^{-1}$, находящиеся в $(k-1)$-м столбце и в строках с номерами $(k, i_k), \dots, (n-1, i_{n-1})$. Теперь из (28.17) вытекает, что

$$\tilde{A} = (\tilde{N}_1^{-1} \dots \tilde{N}_{n-1}^{-1})\, \tilde{A}_{n-1}, \qquad (28.19)$$

где

$$\tilde{A} = (P_{n-1,\, i_{n-1}} \dots P_{1i_1})\, A\, (P_{1i_1} \dots P_{n-1,\, i_{n-1}}). \qquad (28.20)$$

Соотношения (28.18) — (28.20) показывают, что процесс (28.16) определяет разложение на треугольные множители матрицы $\tilde{A}$, которая согласно (28.20) получается из матрицы A путем перестановок ее строк и столбцов. Анализ ошибок, выполненный для матрицы A и процесса (28.1), уже без изменения переносится на матрицу $\tilde{A}$ и процесс (28.16). Этот процесс получил название — *метод Гаусса с перестановками* для треугольного разложения матрицы.

Таким образом, рост элементов матриц $\tilde{A}_k$ процесса (28.16) и, следовательно, общий уровень ошибок полностью определяется стратегией выбора ведущих или главных элементов. Существуют три наиболее распространенных стратегии.

1. В качестве ведущего элемента k-го шага выбирается максимальный по модулю элемент $\tilde{a}_{ij}^{(k-1)}$ матрицы $\tilde{A}_{k-1}$ при условиях $i \geqslant k$, $j = k$. Если имеется несколько максимальных по модулю элементов, то ведущим берется тот из них, который находится в строке с наименьшим номером. Эта стратегия называется выбором ведущего элемента *по столбцу*.

2. В качестве ведущего элемента k-го шага выбирается максимальный по модулю элемент $\tilde{a}_{ij}^{(k-1)}$ матрицы $\tilde{A}_{k-1}$ при условиях $i = k$, $j \geqslant k$. Если имеется несколько максимальных по модулю элементов, то ведущим берется тот из них, который находится в столбце с наименьшим номером. Эта стратегия называется выбором ведущего элемента *по строке*.

3. В качестве ведущего элемента k-го шага выбирается максимальный по модулю элемент $\tilde{a}_{ij}^{(k-1)}$ матрицы $\tilde{A}_{k-1}$ при условиях $i \geqslant k$, $j \geqslant k$. Если имеется несколько максимальных по модулю элементов, то сначала оставляем элементы, находящиеся в столбце с наименьшим номером, а ведущим берется тот из них, который находится в строке с наименьшим номером. Эта стратегия называется выбором ведущего элемента *по всей матрице*.

Применение первой и третьей стратегий обеспечивает выполнение неравенств $|\hat{n}_{ij}^{(k)}| \leqslant 1$ для элементов матриц $\hat{N}_k$.

В этих случаях оценки (28.13) — (28.15) можно улучшить, но только лишь в 1,5 раза. По-прежнему главным фактором остается рост элементов матриц A_k по сравнению с элементами исходной матрицы A.

Условия $|\tilde{n}_{ij}^{(k)}| \leqslant 1$ позволяют получить верхние оценки, показывающие возможный рост элементов матриц A_k. Пусть

$$\alpha_k = \max_{i,j} |a_{ij}^{(k)}|.$$

Если перестановки не выполнялись, то очевидно, что

$$|a_{ij}^{(k)}| = |a_{ij}^{(k-1)} + \tilde{n}_{ik}^{(k)} a_{kj}^{(k-1)}| \leqslant |a_{ij}^{(k-1)}| + |a_{kj}^{(k-1)}|,$$

поэтому $\alpha_k \leqslant 2\alpha_{k-1}$. Перестановки не меняют этого соотношения, поэтому

$$\alpha_k \leqslant 2^k \alpha_0. \tag{28.21}$$

К сожалению, при применении стратегии выбора ведущего элемента по столбцу оценка (28.21) может достигаться. Например, она достигается для матриц A вида

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

Значительно лучший результат известен для стратегии выбора ведущего элемента по всей матрице. Доказано [5], что если $\alpha_{k-1} = f(k)\alpha_0$, то

$$f(k) \leqslant k^{1/2}(2^1 3^{1/2} 4^{1/3} \dots k^{1/(k-1)})^{1/2}. \tag{28.22}$$

Правая часть (28.22) много меньше, чем 2^{k-1}. Однако оценка (28.22), по-видимому, сильно завышена, так как до сих пор не найдено ни одной матрицы, для которой $f(k) > k$.

Мы уделили много внимания росту элементов в методе Гаусса. Но заметим, что в практических вычислениях это явление встречается очень редко. Гораздо чаще наблюдается существенное уменьшение элементов.

УПРАЖНЕНИЯ

1. Рассмотреть применение метода Гаусса без выбора ведущего элемента к почти треугольной и трехдиагональной матрицам.
2. Доказать, что метод Гаусса с выбором ведущего элемента можно применять к любой невырожденной матрице.
3. Рассмотреть применение метода Гаусса с выбором ведущего элемента по столбцу (строке) к правой (левой) почти треугольной матрице. Что существенно меняется в вычислительном алгорифме, если ведущий элемент выбирается по всей матрице?
4. Рассмотреть применение метода Гаусса с выбором ведущего элемента по столбцу или строке к трехдиагональной матрице. Что существенно меняется в вычислительном алгорифме, если ведущий элемент выбирается по всей матрице?
5. Что можно сказать о величине диагональных элементов треугольных сомножителей, полученных в результате применения метода Гаусса с выбором ведущего элемента по всей матрице?
6. Доказать, что метод Гаусса с выбором главного элемента по всей матрице приводит любую, в том числе и вырожденную матрицу к трапецевидной.
7. Можно ли использовать выбор ведущего элемента по столбцу или строке для приведения исходной матрицы к трапецевидной?
8. Сравнить метод Гаусса без выбора ведущего элемента для левой треугольной матрицы и процесс ортогонализации для векторов-строк той же матрицы.
9. Доказать, что при использовании метода Гаусса с выбором ведущего элемента по столбцу для трехдиагональной матрицы максимальный по модулю элемент в каждом столбце возрастает не более, чем в два раза.

§ 29. Компактная схема

Мы уже неоднократно подчеркивали, что наличие в формулах выражений типа скалярных произведений позволяет эффективно применять операции накопления и тем самым снизить общий уровень ошибок. Но в рассмотренном выше методе Гаусса, по существу, нет возможности для применения таких операций. Это связано лишь с выбором численного метода для получения разложения матрицы на треугольные множители.

Снова рассмотрим матрицу A, удовлетворяющую условиям теоремы 27.1. Так как разложение (27.1) существует, то, приравнивая между собой элементы матрицы A и произведения BC, получаем

$$a_{ij} = \sum_{p=1}^{\min(i,\,j)} b_{ip} c_{pj}. \tag{29.1}$$

Будем считать, что $b_{ii}=1$ для всех i. Тогда из (29.1) следует:

$$
\begin{aligned}
&c_{11}=a_{11},\\
&c_{1j}=a_{1j},\quad b_{j1}=\frac{a_{j1}}{c_{11}},\qquad j=2,\ 3,\ \ldots,\ n,\\
&c_{il}=a_{il}-\sum_{p=1}^{i-1}b_{ip}c_{pl},\qquad i=2,\ 3,\ \ldots,\ n,\\
&c_{ij}=a_{ij}-\sum_{p=1}^{i-1}b_{ip}c_{pj},\quad b_{ji}=\frac{a_{ji}-\sum\limits_{p=1}^{i-1}b_{jp}c_{pi}}{c_{ii}},\\
&i=2,\ 3,\ \ldots,\ n,\quad j=i+1,\ i+2,\ \ldots,\ n.
\end{aligned}
\tag{29.2}
$$

Если матрица A почти треугольная, то одна из матриц B, C в действительности будет двухдиагональной. В этом случае формулы (29.2) упрощаются. Например, для правой почти треугольной матрицы A

$$
\begin{aligned}
&c_{11}=a_{11},\\
&c_{1j}=a_{1j},\quad b_{21}=\frac{a_{21}}{c_{11}},\qquad j=2,\ 3,\ \ldots,\ n,\\
&c_{ii}=a_{ii}-b_{i,i-1}c_{i-1,i},\qquad i=2,\ 3,\ \ldots,\ n,\\
&c_{ij}=a_{ij}-b_{i,i-1}c_{i-1,j},\quad b_{i+1,i}=\frac{a_{i+1,i}}{c_{ii}},\\
&i=2,\ 3,\ \ldots,\ n,\quad j=i+1,\ i+2,\ \ldots,\ n.
\end{aligned}
$$

В частности, для трехдиагональной матрицы A имеем

$$
\begin{aligned}
&c_{11}=a_{11},\\
&c_{12}=a_{12},\quad b_{21}=\frac{a_{21}}{c_{11}},\\
&c_{ii}=a_{ii}-b_{i,i-1}c_{i-1,i},\\
&c_{i,i+1}=a_{i,i+1},\quad b_{i+1,i}=\frac{a_{i+1,i}}{c_{ii}},\\
&i=2,\ 3,\ \ldots,\ n.
\end{aligned}
$$

Полученные формулы можно использовать для численного определения треугольных сомножителей B и C матрицы A. Соответствующий алгоритм называется *компактной схемой* метода Гаусса. Теперь в общем случае применение операций накопления вполне оправдано. Конечно, ошибки округления и здесь будут оказывать свое влияние и вместо матриц B, C мы получим некоторые другие матрицы $\tilde B$, $\tilde C$ с элементами $\tilde b_{ij}$, $\tilde c_{ij}$.

Пусть находятся элементы первой строки матрицы $\tilde C$ и первого столбца матрицы $\tilde B$. Положим $\tilde c_{11}=a_{11}$, $\tilde c_{1j}=a_{1j}$ и вычисляем

$$
\tilde b_{j1}=\mathrm{fl}\left(\frac{a_{j1}}{\tilde c_{11}}\right)=\frac{a_{j1}}{\tilde c_{11}}(1+\varepsilon_{j1}).
$$

Реально вычисленные элементы $\tilde b_{j1}$ можно рассматривать как точно вычисленные, исходя из возмущенных элементов $a_{j1}+\mu_{j1}$. Если $\varepsilon_{j1}\neq -1$, то $|\mu_{j1}|\lessapprox (1/2)\,p^{t+1}|\tilde c_{11}||\tilde b_{j1}|$. Если же $\varepsilon_{j1}=-1$, то это означает, что $|a_{j1}|<|\tilde c_{11}|\omega$, поэтому $|\mu_{j1}|\leqslant|\tilde c_{11}|\omega$. Окончательно получаем, что

$$
|\mu_{j1}|\lessapprox\tfrac{1}{2}\,p^{t+1}|\tilde c_{11}||\tilde b_{j1}|+|\tilde c_{11}|\omega.
$$

Исследование процесса вычисления других элементов осуществляется аналогично. Предположим, что вычисляются элементы $\tilde c_{ij}$ для $i>1$, включая и диагональный элемент $\tilde c_{ii}$. Имеем

$$
\tilde c_{ij}=\mathrm{fl}_2\left(a_{ij}-\sum_{p=1}^{i-1}\tilde b_{ip}\tilde c_{pl}\right)=\left(a_{ij}-\sum_{p=1}^{i-1}\tilde b_{ip}\tilde c_{pj}\right)(1+\tau_{ij}).
$$

Отсюда заключаем, что реально вычисленные элементы $\tilde c_{ij}$ можно рассматривать как точно вычисленные, исходя из возмущенных элементов матрицы A. Важно подчеркнуть, что при нахождении элемента $\tilde c_{ij}$ дополнительное возмущение μ_{ij} вносится только в элемент a_{ij}. Если $\tau_{ij}\neq -1$, то

$$
a_{ij}-\sum_{p=1}^{i-1}\tilde b_{ip}\tilde c_{pj}=\frac{\tilde c_{ij}}{1+\tau_{ij}}\cong\tilde c_{ij}(1-\tau_{ij}),
\tag{29.3}
$$

поэтому

$$
|\mu_{ij}|\lessapprox\tfrac{1}{2}\,p^{t+1}|\tilde c_{ij}|.
$$

Если же $\tau_{ij}=-1$, то это означает, что выражение,

стоящее в (29.3) слева, не превосходит ω по модулю. Но тогда $|\mu_{ij}|<\omega$. Следовательно, всегда

$$|\mu_{ij}|\leqslant \frac{1}{2}\,p^{t+1}|\tilde{c}_{ij}|+\omega.$$

Аналогично мы получаем, что при вычислении элементов b_{ji} дополнительное возмущение μ_{ji} вносится лишь в элемент a_{ji}. При этом

$$|\mu_{ji}|\leqslant \frac{1}{2}\,p^{t+1}|\tilde{c}_{ii}|\,|\tilde{b}_{ij}|+|\tilde{c}_{ii}|\omega.$$

Таким образом, при нахождении треугольного разложения матрицы согласно формулам (29.2) влияние ошибок округления снова может быть учтено в форме обратного анализа. Если $BC=A+\mathrm{M}$, то, объединяя полученные оценки, заключаем, что для элементов μ_{ij} эквивалентного возмущения M справедливы неравенства

$$|\mu_{ij}|\leqslant\begin{cases}0, & i=1,\\ \frac{1}{2}\,p^{t+1}|\tilde{c}_{ij}|+\omega, & j\geqslant i,\\ \frac{1}{2}\,p^{t+1}|\tilde{c}_{ij}|\,|\tilde{b}_{ij}|+|c_{ij}|\omega, & j<i.\end{cases}\tag{29.4}$$

Как и в методе Гаусса, неустойчивость треугольного разложения в компактной схеме связана в основном только с возможным ростом элементов. Обычно мы будем иметь дело с матрицами, элементы которых много больше $\omega^{1/3}$. В этом случае оценки (29.4) упрощаются. Именно,

$$|\mu_{ij}|\lesssim\begin{cases}0, & i=1\\ \frac{1}{2}\,p^{t+1}|\tilde{c}_{ij}|, & j\geqslant i,\\ \frac{1}{2}\,p^{t+1}|\tilde{c}_{ij}|\,|\tilde{b}_{ij}|, & j<i.\end{cases}\tag{29.5}$$

Наиболее эффективно компактная схема используется для получения треугольного разложения положительно определенных матриц. Согласно следствию теоремы 27.1 в этом случае имеет место разложение

$$A=C^*C.\tag{29.6}$$

В полном соответствии с (29.2) теперь получаем

$$\begin{aligned}&c_{11}=a_{11}^{1/2}, \quad c_{1j}=\frac{a_{1j}}{c_{11}}, && j>1,\\ &c_{ii}=\left(a_{ii}-\sum_{p=1}^{i-1}|c_{pi}|^2\right)^{1/2}, && i>1,\\ &c_{ij}=\frac{a_{ij}-\sum_{p=1}^{i-1}\bar{c}_{pi}c_{pj}}{c_{ii}}, && j>i.\end{aligned}\tag{29.7}$$

В частности, если матрица A вещественная, то

$$\begin{aligned}&c_{11}=a_{11}^{1/2}, \quad c_{1j}=\frac{a_{1j}}{c_{11}}, && j>1,\\ &c_{ii}=\left(a_{ii}-\sum_{p=1}^{i-1}c_{pi}^2\right)^{1/2}, && i>1,\\ &c_{ij}=\frac{a_{ij}-\sum_{p=1}^{i-1}c_{pi}c_{pj}}{c_{ii}}, && j>i.\end{aligned}\tag{29.8}$$

Эти формулы используются для получения разложения (29.6) подобно тому, как формулы (29.2) — для разложения (27.1). Соответствующий алгорифм называется *методом квадратного корня*. Если применяются операции накопления, то для реально вычисленной матрицы $\tilde{C}$ имеем

$$\tilde{C}^*\tilde{C}=A+\mathrm{M}.$$

Элементы μ_{ij} эквивалентного возмущения M удовлетворяют соотношениям, аналогичным (29.5). Именно,

$$|\mu_{ij}|\lesssim\begin{cases}\frac{1}{2}\,p^{t+1}|\tilde{c}_{ij}|\,|\tilde{c}_{ii}|, & j>i,\\ \frac{1}{2}\,p^{t+1}|\tilde{c}_{ji}|\,|\tilde{c}_{jj}|, & j<i,\\ p^{t+1}|\tilde{c}_{ii}|^2, & j=i.\end{cases}$$

Мы не будем останавливаться подробно на их выводе, ибо они получаются почти так же, как и соотношения (29.5).

Заметим лишь, что теперь можно исключительно эффективно оценить $\|M\|_E$:

$$\|M\|_E \leqslant p^{t+1}\left(\frac{1}{2}\sum_{j>i}|\tilde{c}_{ii}\tilde{c}_{ij}|^2+\sum_{i=1}^{n}|\tilde{c}_{ii}|^4\right)^{1/2} \leqslant$$

$$\leqslant p^{t+1}\left(\sum_{i=1}^{n}|\tilde{c}_{ii}|^2\sum_{j\geqslant i}|\tilde{c}_{ij}|^2\right)^{1/2} \leqslant p^{t+1}\left(\sum_{i=1}^{n}\left(\sum_{j\geqslant i}|\tilde{c}_{ij}|^2\right)^2\right)^{1/2} =$$

$$= p^{t+1}\left(\sum_{i=1}^{n}|\{\tilde{C}\tilde{C}^*\}_{ii}|^2\right)^{1/2} \leqslant p^{t+1}\left(\sum_{i,j=1}^{n}|\{\tilde{C}\tilde{C}^*\}_{ij}|^2\right)^{1/2} =$$

$$= p^{t+1}\|\tilde{C}\tilde{C}^*\|_E = p^{t+1}\|\tilde{C}^*\tilde{C}\|_E = p^{t+1}\|A\|_E.$$

Итак,

$$\|M\|_E \leqslant p^{t+1}\|A\|_E.$$

УПРАЖНЕНИЯ

1. Сравнить формулы (28.14), (29.5).

2. Выполнить анализ ошибок в компактной схеме метода Гаусса без предположения об использовании операций накопления. Сравнить полученные результаты с (28.14).

3. Как использовать перестановки в компактной схеме метода Гаусса?

4. Доказать, что используя перестановки в методе квадратного корня, можно получить в разложении (29.6) нормализованную треугольную матрицу C.

5. Доказать, что для эрмитовой матрицы A с ненулевыми главными минорами существует разложение

$$A = S^*DS, \tag{29.9}$$

где S — правая треугольная матрица, а D — диагональная с элементами ± 1.

6. Для разложения (29.9) написать формулы, аналогичные (29.7).

7. Написать формулы для разложения эрмитовой трехдиагональной матрицы.

8. Доказать, что для любой матрицы C, удовлетворяющей равенству (29.6), справедливы соотношения

$$\|C\|_2 = \|A\|_2^{1/2}, \quad \|A\|_E^{1/2} \leqslant \|C\|_E \leqslant n^{1/4}\|A\|_E^{1/2}. \tag{29.10}$$

9. Доказать, что при реализации метода квадратного корня для положительно определенной матрицы не может быть увеличения максимального по модулю элемента.

10. Что можно сказать об аналогах упражнений 8, 9 для разложения (29.9)?

§ 30. Разложение на унитарный и треугольный множители

Мы уже отмечали, что квадратная матрица может быть разложена в произведение унитарной и треугольной матриц. Существует немало алгоритмов для численного отыскания этого разложения. Однако в основе большинства из них лежат рассмотренные ранее процессы исключения элементов с помощью унитарных преобразований или процессы ортогонализации.

Пусть A — квадратная матрица порядка n. Согласно формулам (18.11) построим матрицу вращения T_{12} так, чтобы в матрице $A_1 = T_{12}A$ элемент в позиции (2.1) был равен нулю. Затем построим матрицу вращения T_{13} из условия обращения в нуль элемента в позиции (3.1) матрицы $A_2 = T_{13}A_1$. Ясно, что при этом нулевой элемент, полученный на первом шаге, останется без изменения. Далее выбираем последовательность матриц вращения, например, циклически по столбцам при $m = n-1$. Сами матрицы строим так, чтобы при очередном умножении на матрицу вращения T_{ij} исключался элемент в позиции (j, i).

Очевидно, что при реализации каждого шага сохраняются все нулевые элементы, полученные на предыдущих шагах. После выполнения $N = n(n-1)/2$ шагов мы получим правую треугольную матрицу A_N, причем $A_N = R_N A$, где $R_N = T_{n-1,n}T_{n-2,n}\ldots T_{13}T_{12}$. Отсюда вытекает, что

$$A = R_N^* A_N. \tag{30.1}$$

Матрица R_N^* — унитарная и требуемое разложение построено.

Влияние ошибок округления приведет к тому, что реально будут вычислены некоторые другие матрицы $\tilde{A}_N$, $\tilde{T}_{ij}$. Но как показывают исследования, выполненные в § 19, мы будем иметь

$$\tilde{A}_N = \tilde{R}_N(A+M), \tag{30.2}$$

где $\tilde{R}_N = \tilde{T}_{n-1,n}\tilde{T}_{n-2,n}\ldots\tilde{T}_{13}\tilde{T}_{12}$, причем в соответствии с (19.12)

$$\|M\|_E \leqslant \sqrt{2}\,(2n-3)\,p^{t+1}\|A\|_E. \tag{30.3}$$

Конечно, элементы могут исключаться и в каком-либо другом порядке. В частности, если матрицы вращения выбираются циклически по строкам, то для соответствующего эквивалентного возмущения снова справедлива оценка (30.3).

Мы уже отмечали в § 19, что обе циклические последовательности эквивалентны с точностью до выбора углов поворота. В случае использования этих последовательностей для получения разложения матрицы на унитарный и треугольный множители можно сделать более точный вывод. Так как вычисление угла поворота для любой матрицы T_{ij} связано лишь с элементами i-ой и j-ой строк, то *циклические последовательности по строкам и столбцам полностью эквивалентны*. Следовательно, при разложении матрицы на унитарный и треугольный множители они будут давать один и тот же результат, включая всю совокупность ошибок округления.

Разложение матрицы на унитарную и треугольную можно осуществить и с помощью матриц отражения. Действительно, по первому столбцу матрицы A построим матрицу отражения U_1 так, чтобы в матрице $A_1 = U_1 A$ все поддиагональные элементы первого столбца были нулевыми. Затем по второму столбцу матрицы A_1 построим матрицу отражения U_2 так, чтобы в матрице $A_2 = U_2 A_1$ все поддиагональные элементы второго столбца были нулевыми, а элемент в первой строке остался без изменения. Принимая во внимание вид (21.1) матрицы A_2, заключаем, что первый столбец матрицы A_1 от умножения на матрицу U_2 не меняется. После выполнения $N = n - 1$ шагов мы получим правую треугольную матрицу A_N и разложение (30.1), где $R_N = U_{n-1} \dots U_1$. Реальные вычисления приведут к разложению (30.2). Согласно оценке (21.7) теперь имеем

$$\|M\|_E \leqslant \frac{2{,}5\sqrt{2}+5}{3}(n-1)\,p^{-t+1}\|A\|_E.$$

Процесс ортогонализации также может быть использован для численного отыскания разложения матрицы A на унитарную и треугольную. Будем рассматривать строки A как векторы, подлежащие ортогонализации. Нормировка первого вектора сводится к умножению A слева на диагональную матрицу D_1, у которой лишь первый элемент отличен от единицы. Это дает матрицу $Q_1 = D_1 A$.

Вычисление вектора v_2 согласно (25.2) означает определение некоторой матрицы M_2 вида (24.9) и умножение на нее слева матрицы Q_1. Нормировка второго вектора — это очередное умножение слева на диагональную матрицу D_2, у которой отличен от единицы только второй элемент, и т. д.

Весь процесс состоит из $N = n$ шагов. Матричная трактовка k-го шага означает нахождение некоторой матрицы M_k вида (24.9), умножение вычисленной ранее матрицы Q_{k-1} слева на M_k, определение диагональной матрицы D_k, имеющей лишь k-й элемент, отличный от единицы, и вычисление матрицы $Q_k = D_k M_k A_{k-1}$. После выполнения N шагов мы получим матрицу Q_N; при этом

$$Q_N = (D_N M_N \dots D_2 M_2 D_1) A,$$

откуда вытекает, что

$$A = (D_N M_N \dots D_2 M_2 D_1)^{-1} Q_N.$$

Матрица $(D_N M_N \dots D_2 M_2 D_1)^{-1}$ — левая треугольная. Обозначив ее через A_N, приходим к разложению

$$A = A_N Q_N,$$

в котором по сравнению с (30.1) унитарный множитель является правым, а треугольный — левым.

Ошибки округления и в этом процессе приведут к вычислению некоторых других матриц $\tilde{Q}_N$, $\tilde{D}_N$, $\tilde{M}_N \dots \tilde{D}_1$. Если к тому же согласно (25.11) применяется переортогонализация векторов, то вместо каждой матрицы $\tilde{M}_k$ мы в действительности будем иметь произведение $\tilde{M}_k^{(s_k)} \dots \tilde{M}_k^{(2)} \tilde{M}_k^{(1)}$ матриц типа M_k. Тем не менее, в соответствии с (25.7), реальное разложение будет таким:

$$\tilde{Q}_N = (\tilde{D}_N \tilde{M}_N^{(s_N)} \dots \tilde{D}_2 \tilde{M}_2^{(s_2)} \dots \tilde{M}_2^{(1)} \tilde{D}_1)(A + M),$$

где

$$\|M\|_E \leqslant \left(\max_{2 \leqslant k \leqslant n} s_k\right) p^{-t+1} \|A\|_E.$$

Напомним, что в практических вычислениях всегда $s_k = 2$.

УПРАЖНЕНИЯ

1. Пусть матрица A приводится к треугольному виду с помощью умножения справа или слева на циклическую последовательность матриц вращения по строкам. Как, используя этот процесс, вычислить главные миноры матрицы A?

2. Можно ли вычислить главные миноры матрицы, если для ее приведения к треугольному виду последовательность матриц вращения выбирается циклически по столбцам?

3. Доказать, что с помощью умножения слева на подходящим образом выбранную последовательность матриц вращения или отражения можно привести исходную матрицу не только к правой, но и к левой треугольной матрице.

4. К какому виду можно привести матрицу с помощью правосторонних унитарных преобразований?

5. Можно ли улучшить коэффициент $\sqrt{2}$ в оценке (30.3)?

6. Доказать, что с помощью процесса ортогонализации можно разложить матрицу в произведение унитарной и правой треугольной.

§ 31. Разложение прямоугольных матриц

Рассмотренные выше вычислительные алгорифмы разложения матрицы на множители были описаны в основном на примерах квадратных матриц. Однако многие из них без изменения могут быть применены и к общим прямоугольным матрицам. Тем не менее мы все же рассмотрим этот случай несколько подробнее, обратив особое внимание на алгорифмы, которые для разложения прямоугольных матриц используются наиболее часто.

Любое разложение матрицы на множители в конечном счете сводится к ее эквивалентному преобразованию. При этом для решения большинства алгебраических задач необходимо, чтобы по виду окончательной матрицы можно было бы легко установить положение базисного минора. По существу, именно этим требованием и определяется наше стремление привести исходную матрицу к треугольному виду.

Если матрица удовлетворяет условиям теоремы 27.1, то единственный нулевой диагональный элемент треугольной матрицы может находиться только в конце диагонали. Поэтому базисный минор будет расположен в тех строках и столбцах треугольной матрицы, которые содержат ненулевые диагональные элементы. Если же исходная матрица произвольная, то формальное применение рассмотренных ранее алгорифмов может дать такую треугольную матрицу,

в которой нулевые диагональные элементы могут находиться в самых различных местах и не обязательно подряд или в конце диагонали. В этом случае определение базисного минора затруднительно и требует дополнительных вычислений. Для преодоления данных трудностей снова оказывается полезным использование перестановок.

Пусть A — прямоугольная матрица размеров $m \times n$ и ранга r. Применим к этой матрице процесс метода Гаусса с выбором ведущего элемента по всей матрице. Тогда после r шагов мы получим правую трапецевидную матрицу

$$A_r = \left[\begin{array}{c|c} A_{11}^{(r)} & A_{12}^{(r)} \\ \hline 0 & 0 \end{array}\right],$$

где $A_{11}^{(r)}$ — правая треугольная матрица порядка r с ненулевыми диагональными элементами. Если $r = n$ или $r = m$, то в матрице A_r будет отсутствовать соответственно последний клеточный столбец или последняя клеточная строка. Исходная матрица A может быть приведена и к левой трапецевидной матрице

$$\left[\begin{array}{c|c} A_{11}^{(r)} & 0 \\ \hline \hat{A}_{21}^{(r)} & 0 \end{array}\right],$$

если процесс метода Гаусса с выбором ведущего элемента по всей матрице применить к транспонированной матрице A'. Напомним, что базисным минором трапецевидной матрицы заведомо является ее минор порядка r в верхнем левом углу.

Для приведения прямоугольной матрицы A к трапецевидной можно использовать и унитарные преобразования. Выберем в A столбец с максимальной суммой квадратов модулей элементов и поставим его на место первого столбца. Если таких столбцов окажется несколько, то среди них возьмем столбец с наименьшим возможным номером. Теперь с помощью умножения слева на подходящим образом выбранную матрицу отражения или последовательность матриц вращения исключим поддиагональные элементы нового первого столбца. Пусть уже получена матрица, у которой первые k столбцов, $k \geqslant 1$, являются столбцами трапецевидной матрицы. Среди всех ее столбцов, кроме первых k, выберем тот, который имеет максимальную сумму квадратов модулей элементов, не входящих в

первые k строк полученной матрицы. Если таких столбцов окажется несколько, то среди них возьмем столбец с наименьшим возможным номером. Переставим выбранный столбец на место $(k+1)$-го столбца и исключим его поддиагональные элементы с помощью умножения слева на подходящую матрицу отражения или последовательность матриц вращения. После выполнения r таких шагов мы получим правую трапецевидную матрицу. Более того, как видно из ее построения, она будет и нормализованной. Применение этого процесса к матрице A' позволяет привести матрицу A к левой нормализованной трапецевидной форме.

Прямоугольную матрицу можно приводить с помощью унитарных преобразований не только к трапецевидной, но и к двухдиагональной матрице. В самом деле, выберем матрицу отражения U_1 таким образом, чтобы в матрице $A_1 = U_1 A$ обратились в нули элементы первого столбца, лежащие ниже диагонали. Затем выберем матрицу отражения V_1 так, чтобы элементы первого столбца матрицы $A_2 = A_1 V_1$ остались бы без изменения, а элементы первой строки, лежащие правее элемента в позиции (1.2), стали нулевыми. Умножая поочередно слева и справа на матрицы отражения и исключая элементы в таком порядке:

$$\begin{matrix} (2.1),\ (3.1),\ (4.1),\ \ldots,\ (m.1), \\ (1.3),\ (1.4),\ \ldots,\ (1.n), \\ (3.2),\ (4.2),\ \ldots,\ (m.2), \\ (2.4),\ \ldots,\ (2.n), \\ \ldots\ldots\ldots\ldots\ldots\ldots \end{matrix}$$

мы придем к правой двухдиагональной матрице D. Начиная исключение с элементов первой строки, можно аналогичным способом привести исходную матрицу A и к левой двухдиагональной матрице. Двухдиагональная матрица D связана с исходной матрицей A соотношением

$$D = UAV,$$

где U, V — унитарные матрицы. Конечно, в этом процессе можно использовать и матрицы вращения.

Мы не будем проводить какие-либо исследования ошибок округления в рассмотренных алгорифмах, так как полученные ранее результаты охватывают и случай прямоугольных матриц.

УПРАЖНЕНИЯ

1. Рассмотреть применение процесса ортогонализации для разложения прямоугольной матрицы на множители.
2. Можно ли применять метод Гаусса с выбором ведущего элемента по столбцу или строке для преобразования общей прямоугольной матрицы к трапецевидной?
3. Найти какой-либо базис ядра правой и левой трапецевидных матриц.
4. Можно ли использовать перестановки строк и столбцов в процессе преобразования прямоугольной матрицы к двухдиагональной?

§ 32. Унитарно подобное разложение

Пусть задана квадратная матрица A порядка n и над ней совершается последовательность подобных преобразований с матрицами $Q_1, \ldots, Q_s$. Если в результате выполнения этих преобразований получается некоторая матрица B, то

$$B = (Q_s^{-1} \ldots Q_1^{-1}) A (Q_1 \ldots Q_s). \qquad (32.1)$$

Отсюда следует, что

$$A = (Q_1 \ldots Q_s) B (Q_s^{-1} \ldots Q_1^{-1}).$$

Итак, любое подобное преобразование матрицы A, по существу, приводит к разложению ее на множители.

Известно [1] существование преобразований подобия, при которых матрица B в (32.1) является треугольной, квазидиагональной или имеет вид канонической формы Жордана. Однако все эти преобразования косвенным образом связаны с отысканием корней алгебраических многочленов и поэтому не могут быть получены в общем случае за конечное число арифметических операций. Тем не менее можно построить подобное преобразование (32.1) с матрицей B, существенно более простой, чем исходная матрица A.

Рассмотрим сначала матрицу A произвольного вида и покажем, что ее можно привести подобными унитарными преобразованиями, например, к левой почти треугольной матрице. По элементам первой строки A построим матрицу отражения U_1 так, чтобы первый элемент этой строки остался без изменения, а все ее элементы, лежащие правее элемента в позиции (1.2), стали нулевыми. Далее получаем матрицу AU_1 и затем матрицу $A_1 = U^* A U$. Согласно построению в первой строке матрицы AU_1 лишь первые

два элемента могут быть отличны от нуля. Но при умножении слева на матрицу U_1^* первая строка не будет меняться. Следовательно, первая строка матрицы A_1 имеет вид первой строки левой почти треугольной матрицы.

Предположим, что уже построены матрицы отражения $U_1, \ldots, U_k$ такие, что первые k строк, $k \geqslant 1$, матрицы

$$A_k = (U_k^* \ldots U_1^*)\, A\, (U_1 \ldots U_k) \qquad (32.2)$$

имеют вид первых k строк левой почти треугольной матрицы. По элементам $(k+1)$-й строки матрицы A_k построим матрицу отражения U_{k+1} так, чтобы первые k элементов этой строки остались без изменения, а все ее элементы, лежащие правее элемента в позиции $(k+1, k+2)$, стали нулевыми. Ясно, что первые $k+1$ строк матрицы $A_k U_{k+1}$ имеют вид соответствующих строк левой почти треугольной матрицы, а умножение слева на матрицу U_{k+1}^* не меняет первых $k+1$ строк. Продолжая этот процесс, мы получим левую почти треугольную матрицу

$$B = (U_{n-2}^* \ldots U_1^*)\, A\, (U_1 \ldots U_{n-2}), \qquad (32.3)$$

унитарно подобную матрице A.

Конечно, для реализации преобразования (32.3) можно использовать не только матрицы отражения, но и матрицы вращения. Анализ ошибок округления, возникающих в реальных вычислительных процессах, полностью охватывается анализом, проведенным в § 23, с заменой в формулах (23.8), (23.10) числа n на $n-1$.

Остановимся более подробно на одном частном, но очень важном случае. Пусть матрица A эрмитова; тогда будут эрмитовы все матрицы A_k из (32.2) и матрица B из (32.3). Но почти треугольная эрмитова матрица в действительности является эрмитовой трехдиагональной. Следовательно, в случае эрмитовой матрицы A процесс подобного преобразования (32.2), (32.3) становится особенно эффективным.

Однако в реализации процесса с эрмитовыми матрицами имеется ряд особенностей. Влияние ошибок округления приводит к тому, что эрмитовость матриц A_k из (32.2) будет нарушаться. Для ее восстановления обычно вычисляют половину элементов матриц A_k, лежащих ниже или выше главной диагонали, а остальным элементам приписывают принудительные значения. Это несколько

изменяет распределение ошибок по сравнению с процессами, рассмотренными ранее, поэтому анализ ошибок для эрмитовых матриц требует особых исследований.

Имеются и некоторые организационные проблемы, которые также оказывают влияние на анализ ошибок. Эрмитова матрица может быть задана лишь половиной своих элементов. Стремление использовать память ЭВМ более экономично приводит к желанию задавать половиной своих элементов и все промежуточные эрмитовы матрицы. Но не каждый вычислительный алгорифм позволяет на всех этапах обойтись половиной элементов. Пусть, например, матрица A эрмитова, а U — матрица отражения. Ясно, что матрицы A и U^*AU можно задать половиной своих элементов. Но если матрицу U^*AU находить через промежуточное вычисление матрицы AU, то не сразу видно, как обойтись половиной элементов на всех этапах, так как матрица AU уже не будет эрмитовой. По-видимому, требуется некоторое изменение вычислительной схемы, что, скорее всего, приведет к иному распределению ошибок.

Необходимая модификация вычислительной схемы осуществляется довольно просто. Принимая во внимание вид (20.4) матрицы отражения, будем иметь

$$U^*AU = \left(E - \frac{1}{\gamma} vv^*\right) A \left(E - \frac{1}{\gamma} vv^*\right) =$$
$$= A - \left(\frac{1}{\gamma} vv^*A - \frac{1}{2\gamma^2} vv^*Avv^*\right) -$$
$$- \left(\frac{1}{\gamma} vv^*A - \frac{1}{2\gamma^2} vv^*Avv^*\right)^*. \qquad (32.4)$$

Далее,

$$\left(\frac{1}{\gamma} vv^*A - \frac{1}{2\gamma^2} vv^*Avv^*\right) = v\left(\left(E - \frac{1}{2} v\left(\frac{1}{\gamma} v\right)^*\right) A \left(\frac{1}{\gamma} v\right)\right)^*.$$

Если обозначить

$$r = \frac{1}{\gamma} v, \quad p = \left(E - \frac{1}{2} vr^*\right) Ar, \qquad (32.5)$$

то согласно (32.4)

$$U^*AU = A - vp^* - pv^*. \qquad (32.6)$$

Задание матрицы A половиной ее элементов позволяет теперь вычислить аналогичную половину элементов

матрицы U^*AU без существенного увеличения памяти ЭВМ для хранения результатов промежуточных вычислений. Новая вычислительная схема значительно отличается от старой, поэтому на нее нельзя без соответствующего обоснования перенести результаты ранее выполненного анализа ошибок.

Пусть $\tilde{v}$, $\tilde{\gamma}$ — элементы вычисленной матрицы отражения $\tilde{U}$. Через $\tilde{r}$, $\tilde{p}$ обозначим векторы, точно вычисленные согласно (32.5), но исходя из величин $\tilde{v}$, $\tilde{\gamma}$. Тогда, в соответствии с (32.6),

$$\tilde{U}^*A\tilde{U} = A - \tilde{v}\tilde{p}^* - \tilde{p}\tilde{v}^*. \qquad (32.7)$$

При вычислении правой части (32.7) появятся ошибки. Поэтому в действительности мы будем иметь

$$\mathrm{fl}\,(A - \tilde{v}\tilde{p}^* - \tilde{p}\tilde{v}^*) = A - \tilde{v}\tilde{p}^* - \tilde{p}\tilde{v}^* + N,$$

где N — матрица суммарных ошибок.

Не все элементы N образуются одинаково. Если $\tilde{U}$ есть матрица отражения, вычисленная на $(k+1)$-м шаге процесса приведения матрицы к трехдиагональному виду, то элементы N, находящиеся в первых k строках и k столбцах, а также элемент в позиции $(k+1, k+1)$, можно считать нулевыми. Внедиагональные элементы $(k+1)$-й строки и $(k+1)$-го столбца удовлетворяют соотношениям типа (20.20), остальные элементы матрицы N зависят от способа вычисления правой части (32.7).

Анализ ошибок, который необходимо выполнить, во многом повторяет исследования § 20. Мы не будем проводить его подробно, а ограничимся указанием основных результатов. Вычисляем

$$\tilde{r} = \mathrm{fl}\left(\frac{1}{\tilde{\gamma}}\tilde{v}\right) = r + e,$$

$$\tilde{u} = \mathrm{fl}_2(A\tilde{r}) = A\tilde{r} + \eta,$$

$$\tilde{p} = \mathrm{fl}_2\left(\left(E - \frac{1}{2}\tilde{v}\tilde{r}^*\right)\tilde{u}\right) = \tilde{p} + \nu,$$

$$\tilde{A} = \mathrm{fl}_3(A - \tilde{v}\tilde{p}^* - \tilde{p}\tilde{v}^*) = A - \tilde{v}\tilde{p}^* - \tilde{p}\tilde{v}^* + N.$$

Здесь

$$\|e\|_E \lesssim \frac{1}{(2\tilde{\gamma})^{1/2}}\,p^{t+1},$$

$$\|\eta\|_E \lesssim \left(\frac{2}{\tilde{\gamma}}\right)^{1/2} p^{t+1}\,\|A\|_E,$$

$$\|\nu\|_E \lesssim 3{,}5\left(\frac{2}{\tilde{\gamma}}\right)^{1/2} p^{t+1}\,\|A\|_E,$$

$$\|N\|_E \lesssim 14{,}5\,p^{t+1}\,\|A\|_E.$$

Теперь, используя результаты § 23, можно сделать следующий вывод. Если эрмитова матрица A порядка n с помощью вычисленных матриц отражения $\tilde{U}_1, \ldots, \tilde{U}_{n-2}$ преобразуется подобно к трехдиагональной эрмитовой матрице $\tilde{B}$, то

$$\tilde{B} = \tilde{U}_{n-2}^{-1}\ldots\tilde{U}_1^{-1}(\tilde{A} + \Delta)\,\tilde{U}_1\ldots\tilde{U}_{n-2}, \qquad (32.8)$$

где

$$\|\Delta\|_E \lesssim 18{,}5\,(n-2)\,p^{t+1}\,\|A\|_E. \qquad (32.9)$$

Эта оценка примерно вдвое больше соответствующей оценки в случае преобразования произвольной матрицы к почти треугольной.

УПРАЖНЕНИЯ

1. Доказать, что с помощью унитарно подобных преобразований можно привести матрицу к правой почти треугольной.

2. Доказать, что для векторов r, v, p из (32.5) выполняются соотношения

$$\|r\|_E = \left(\frac{2}{\gamma}\right)^{1/2}, \quad \|vp^*\|_E \leqslant 2\,\|A\|_E.$$

3. Написать вычислительную схему алгорифма подобного преобразования эрмитовой матрицы к трехдиагональной с помощью матриц вращения.

4. В условиях упражнения 3 выполнить анализ ошибок. Доказать, что для циклического исключения элементов по столбцам в обозначениях (32.8) справедлива оценка

$$\|\Delta\|_E \lesssim 13\,(n-1)\,p^{-t+1}\,\|A\|_E.$$

5. Написать вычислительную схему алгорифма определения главных миноров трехдиагональной эрмитовой матрицы.

6. Написать вычислительную схему алгорифма разложения почти треугольной матрицы в произведения унитарной и треугольной.

§ 33. Некоторые замечания

Любое разложение матрицы на множители с формальной точки зрения можно получить следующим способом. Будем считать элементы сомножителей неизвестными величинами. Перемножим сомножители и приравняем элементы произведения к элементам исходной матрицы. Это даст некоторую систему нелинейных уравнений относительно неизвестных элементов. Добавим к данной системе уравнения, определяющие вид сомножителей. Решая теперь полученную систему, найдем элементы искомого разложения.

Мы уже встречались с таким способом получения треугольного разложения, рассматривая компактную схему метода Гаусса. Можно отметить еще возможность приведения квадратной матрицы к почти треугольной с помощью подобного треугольного преобразования [5]. Однако в большинстве других случаев возникающие при этом системы нелинейных уравнений оказываются столь сложными, что прямое их решение становится неэффективным.

Наиболее эффективное разложение матрицы на множители связано с выполнением последовательности исключений элементов с помощью умножения на элементарные матрицы. Конечно, такие процессы в действительности эквивалентны решению нелинейных систем уравнений, описывающих разложение матрицы. Геометрическая и алгебраическая интерпретация преобразований исключения лишь подсказывают рациональный путь решения этих систем. Отметим, что все рассмотренные нами до сих пор разложения основаны именно на процессах исключения. Лишь разложение, получаемое с помощью процесса ортогонализации, построено на другой идее.

В вычислительной практике используется небольшое число различных типов *элементарных* матриц. В первую очередь следует назвать *элементарные унитарные* матрицы — матрицы отражения и матрицы вращения. Среди *неунитарных элементарных* матриц можно выделить матрицы, отличающиеся от единичной недиагональными элементами в одном столбце или в одной строке. Все остальные элементарные матрицы, как правило, являются частными случаями этих матриц. С некоторыми из них мы уже имели дело. Это матрицы (21.1), (24.3), (24.9).

Многие разложения на множители прямо или косвенно связаны с разложением на множители унитарной и треугольной матриц. Эти разложения осуществляются не единственным способом и, конечно, не все они равноценны с точки зрения влияния ошибок округления. Для исследования таких разложений полностью применим рассмотренный ранее общий анализ ошибок.

Остановимся на двух разложениях, использующих неунитарные элементарные матрицы. Эти разложения существуют для любых квадратных матриц с ненулевыми главными минорами и имеют много общего с треугольным разложением.

Обозначим через R_i матрицы, которые отличаются от единичной лишь недиагональными элементами в i-м столбце. Выберем R_1 так, чтобы в матрице $A_1 = R_1 A$ обратились в нуль все недиагональные элементы первого столбца. Построение матрицы R_1 осуществляется так же, как в методе Гаусса. Затем выберем R_2 из условия обращения в нуль недиагональных элементов второго столбца матрицы $A_2 = R_2 A_1$. Ясно, что при этом первый столбец A_1 останется без изменения. После выполнения n шагов придем к диагональной матрице A_n, где

$$A_n = R_n R_{n-1} \dots R_1 A.$$

Отсюда следует, что

$$A = R_1^{-1} R_2^{-1} \dots R_n^{-1} A_n, \quad A^{-1} = A_n^{-1} R_n \dots R_2 R_1. \quad (33.1)$$

Обращение матриц R_i и матрицы A_n осуществляется совсем просто и мы получаем разложение на множители не только матрицы A, но и матрицы A^{-1}.

Описанный процесс получил название — *метод Жордана* для разложения матрицы на множители [2]. Можно показать, что он эквивалентен последовательному выполнению разложения на треугольные множители по методу Гаусса и последующему разложению правой треугольной матрицы на элементарные неунитарные матрицы. Метод Жордана уступает по скорости выполнения и точности методу Гаусса. Наиболее часто он находит применение в задачах, связанных с обращением матриц.

Рассмотрим далее матрицы S_{ij}, которые отличаются от единичной лишь элементом s_{ij} в позиции (i, j). Если $i \neq j$ и для какого-нибудь k элемент a_{jk} матрицы A

не равен нулю, то, выбирая подходящим образом значение s_{ij}, можно исключить элемент матрицы $S_{ij}A$ в позиции (i, k). Исключая элементы в различном порядке, можно получить огромное количество разложений матрицы A в произведение матриц типа S_{ij}. Все рассмотренные ранее элементарные неунитарные матрицы раскладываются в произведение матриц S_{ij}, поэтому сюда включается и разложение по Гауссу и разложение по Жордану.

Особого внимания заслуживает то из разложений, которое получается при реализации *метода оптимального исключения* [2]. Оно выполняется за такое же число арифметических операций, что и треугольное разложение по методу Гаусса. При этом само разложение более содержательно, чем треугольное разложение. Данное разложение позволяет более эффективно использовать память ЭВМ при решении систем линейных алгебраических уравнений. Однако по точности оно уступает треугольному разложению. Процесс разложения состоит из последовательного умножения слева на матрицы

$$S_{21},\ S_{12},$$
$$S_{31},\ S_{32},\ S_{23},\ S_{13},$$
$$\dots\dots\dots\dots\dots\dots$$
$$S_{k1},\ \dots,\ S_{k,k-1},\ S_{k-1,k},\ \dots,\ S_{1k},$$
$$\dots\dots\dots\dots\dots\dots$$

причем умножение на каждую матрицу S_{ij} сопровождается исключением элемента в позиции (i, j).

Процессы исключения элементов можно использовать для подобного преобразования квадратной матрицы к более простому виду. Некоторые из таких процессов мы уже рассмотрели в § 32. Они составляют основу так называемых прямых методов решения полной проблемы собственных значений и позволяют определять коэффициенты характеристического многочлена матрицы. Значительное число этих методов описано в [6].

УПРАЖНЕНИЯ

1. Доказать, что любую невырожденную левую треугольную матрицу A порядка n можно разложить в произведение

$$A = N_1 N_2 \dots N_{n-1} D, \tag{33.2}$$

где $N_1, \dots, N_{n-1}$ — матрицы вида (24.3), а D — диагональная матрица, составленная из диагональных элементов A.

2. Доказать, что любую невырожденную левую треугольную матрицу A порядка n можно разложить в произведение

$$A = \check{N}_{n-1}\check{N}_{n-2} \dots \check{N}_1 D, \tag{33.3}$$

где $\check{N}_1, \dots, \check{N}_{n-1}$ — матрица вида (24.3), а D — диагональная матрица, составленная из диагональных элементов A.

3. Выполнить анализ ошибок для разложений (33.2), (33.3). Показать, что с точки зрения точности разложение (33.3) менее удовлетворительно, чем (33.2).

4. Выполнить аналоги упражнений 1 — 3 для невырожденной правой треугольной матрицы.

5. Исследовать различные разложения ортогональной матрицы с помощью матриц вращения и отражения. Выполнить анализ ошибок.

6. Доказать, что диагональные элементы матрицы A_{n-1} из (28.1) и матрицы A_n из (33.1) совпадают.

7. Доказать, что эквивалентное возмущение при разложении матрицы на множители методом Жордана совпадает с эквивалентным возмущением, полученным при треугольном разложении по методу Гаусса и последующем разложении правой треугольной матрицы.

8. Можно ли при разложении матрицы методом Жордана использовать перестановки?

9. Разложить матрицы (24.3), (24.9) в произведение матриц типа S_{ij}.

§ 34. Сравнительная характеристика разложений

Разложение матрицы на множители является основой построения большинства численных методов линейной алгебры. Чем эффективнее осуществляется разложение, тем лучшими характеристиками обычно обладает и метод. Нельзя дать однозначный ответ на вопрос «Какое из разложений лучше?», так как разные задачи предъявляют к разложениям различные требования. Тем не менее по некоторым характеристикам не только можно, но и нужно проводить сравнение. Для исследованных нами разложений такие характеристики приведены в табл. 34.1. Предполагается, что все матрицы квадратные и имеют один и тот же порядок, равный n.

Скорость. Общее время, затрачиваемое на получение разложения, по существу, определяется числом арифметических операций, которые необходимо при этом выполнить. В графе «Число операций» табл. 34.1 приведены главные члены числа арифметических операций для всех разложений. В случае использования преобразований вращения одну треть от общего числа операций составляют операции сложения, две трети — операции умножения. Для остальных разложений число операций сложения и

Таблица 34.1

Сравнительная характеристика разложений

Вид сомножителей, способ получения	Режим вычисл.	Число операций	Точность	Дополн. память
Треугольные				
исключение	fl	$(2/3)\,n^3$	αn	0
компактная схема	fl_2	$(2/3)\,n^3$	β	0
компактная схема, $A>0$	fl_2	$(1/3)\,n^3$	1,0	0
Треугольный, унитарный				
отражение	fl_2	$(4/3)\,n^3$	$2{,}9n$	$2n$
вращение (циклич.)	fl	$2n^3$	$2{,}9n$	$0{,}5n^2$
ортогонализация	fl_2	$2n^3$	1,0	$0{,}5n^2$
Треуг. нормализ., унитарный				
отражение	fl_2	$(4/3)\,n^3$	$1{,}8n$	$2n$
вращение (циклич.)	fl	$2n^3$	$1{,}8n$	$0{,}5n^2$
Двухдиаг., унитарные				
отражение	fl_2	$(8/3)\,n^3$	$5{,}9n$	$4n$
вращение (циклич.)	fl	$4n^3$	$5{,}8n$	$1{,}0n^2$
Почти треуг., унитарные				
отражение	fl_2	$(10/3)\,n^3$	$5{,}9n$	$2n$
вращение (циклич.)	fl	$5n^3$	$5{,}8n$	$0{,}5n^2$
Трехдиаг., унитарные				
отражение, $A=A^*$	fl_2	$(4/3)\,n^3$	$18{,}5n$	$2n$
вращение (циклич.), $A=A^*$	fl	$2n^3$	$8n$	$0{,}5n^2$

умножения примерно одинаково. Операции деления и извлечения квадратного корня главный член не определяют.

Самым быстрым разложением является треугольное. Есть некоторые основания предполагать, что ни для какого содержательного разложения, основанного на исключении элементов или процессах ортогонализации, общее число необходимых арифметических операций не может быть меньше, чем у треугольного разложения. Однако другие разложения, получаемые за такое же число операций, существуют. Мы уже отмечали разложение, определяемое методом оптимального исключения [2]. Самое медленное из рассмотренных разложений выполняется в 15 раз дольше, чем самое быстрое.

Объем памяти ЭВМ. Алгебраические задачи, особенно с матрицами большого порядка, требуют для своего решения значительных ресурсов памяти ЭВМ. Один из путей экономии памяти — размещение информации о сомножителях на месте исходной матрицы. Не всегда бывает достаточно этого места, нередко требуется дополнительная память. В графе «Дополнительная память» табл. 34.1 приведены главные члены числа полных слов памяти ЭВМ, которые необходимо добавить для размещения сомножителей. При этом предполагается, что вся память, отведенная для исходной матрицы, также используется для сомножителей.

Наиболее эффективно используется память ЭВМ в треугольных разложениях. Совсем немного дополнительной памяти требуется во всех разложениях, связанных с преобразованиями отражения. Однако в разложениях, основанных на процессе ортогонализации и преобразованиях вращения, дополнительная память весьма значительна. Особенно большая дополнительная память требуется для запоминания сомножителей в случае приведения исходной матрицы к двухдиагональному виду с помощью преобразований вращения.

Точность. Это одна из важнейших, а чаще всего решающая характеристика любого численного метода, в том числе и разложения матрицы на множители. Как показали наши исследования, эквивалентное возмущение M для любого из рассмотренных разложений матрицы A удовлетворяет неравенству

$$\|M\|_E \leqslant f(n)\,p^{-t+1}\|A\|_E,$$

где функция $f(n)$ зависит только от n и способа получения разложения.

В графе «Точность» табл. 34.1 приведены главные члены функции $f(n)$. Наилучшую точность в смысле малости $f(n)$ имеют разложение, полученное на основе процесса ортогонализации, и треугольное разложение для положительно определенной матрицы. В случае общей матрицы точность треугольного разложения зависит от роста элементов. Этот рост определяет значение параметров α, β. Во всех разложениях, связанных с преобразованиями отражения и вращения, точность вполне приемлема.

Режим вычислений. *Оценки точности, которые приведены в табл. 34.1, гарантируются лишь в том случае, когда разложения осуществляются по рассмотренным выше вычислительным схемам. Любое изменение вычислительной схемы должно обосновываться соответствующим анализом ошибок, так как иначе возможна катастрофическая потеря точности.*

Символ fl в графе «Режим» табл. 34.1 означает, что для достижения соответствующей точности можно ограничиться вычислениями с одинарной точностью. Символ fl_2 означает, что для достижения указанной точности использование операций накопления обязательно.

С точки зрения практического использования *весьма привлекательными являются все разложения, основанные на преобразованиях отражения. Они имеют много достоинств, среди которых выделим следующие.*

1. Для преобразований отражения существует эффективная мажорантная оценка точности, по порядку величины не более чем в n раз превышающая минимально возможную.

2. Преобразования отражения требуют выполнения объема вычислительной работы, лишь вдвое превышающего минимально возможный.

3. При выполнении преобразований отражения невозможен значительный рост величин элементов промежуточных вычислений.

4. Информация о сомножителях может быть практически размещена на месте исходной матрицы. Требуется лишь небольшая дополнительная память.

Реализация всех разложений, основанных на преобразованиях отражения, не вызывает каких-либо существенных трудностей. Поэтому наличие эффективных оценок точности является важным аргументом в пользу их широкого применения. Конечно, в каждом конкретном случае может оказаться полезным использование и других разложений. Однако при этом должны быть приведены веские доводы, так как в общем случае большого выигрыша по сравнению с разложениями, основанными на преобразованиях отражения, ожидать не приходится.

УПРАЖНЕНИЯ

1. Почему требуется большая дополнительная память для некоторых из разложений?

2. Каким способом можно существенно уменьшить дополнительную память для размещения сомножителей, полученных на основе преобразований вращения?

3. Будет ли устойчивым вычислительный алгорифм при других способах запоминания преобразований вращения?

4. Какую роль играет сложность вычислительных алгорифмов при сравнении разложений?

5. Какому из разложений вы отдаете предпочтение и почему?

ГЛАВА V

РЕШЕНИЕ СИСТЕМ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

Настоящая глава посвящена, в основном, исследованию численных методов решения систем линейных алгебраических уравнений. Эта задача является одной из важнейших в численном анализе и рассмотрению различных ее аспектов уделяется много внимания.

Теория решения линейных систем достаточно проста и давно известна, однако практическая реализация численных методов вызывает немало трудностей. Это связано прежде всего с тем, что многие методы весьма чувствительны к влиянию ошибок округления и возмущению входных данных. Реальная опасность потери точности заставляет нас считать исследование устойчивости неотъемлемой частью любого численного метода решения систем линейных алгебраических уравнений.

Мы рассмотрим широкий круг вопросов, относящихся к линейным системам. Будут изучены численные методы для решения систем с невырожденными матрицами и прямоугольными матрицами полного ранга. Особое внимание будет уделено исследованию особенностей неустойчивых систем и построению для таких систем численно устойчивых методов.

§ 35. Системы специального вида

Решение систем линейных алгебраических уравнений общего вида обычно сводится к последовательному решению одной или нескольких систем

$$Gu = l \tag{35.1}$$

со специальными матрицами G. Мы рассмотрим сейчас численные методы решения таких систем. При этом будем предполагать, что матрицы G не слишком близки к выро-

жденным. Точнее, они должны оставаться невырожденными в пределах изменения получаемых эквивалентных возмущений. Решение систем, матрицы которых меняют свой ранг в пределах уровня возмущений, мы рассмотрим позднее.

Согласно общей идее обратного анализа ошибок постараемся показать, что реально вычисленное решение $\tilde{u}$ системы (35.1) будет точным решением некоторой возмущенной системы

$$(G+\Delta)\,\tilde{u}=l+\nu. \tag{35.2}$$

Для каждого из численных методов мы проведем исследование соответствующих эквивалентных возмущений Δ, ν.

Система с треугольной матрицей. Одним из лучших методов решения систем с треугольной матрицей является так называемая *обратная подстановка*. Пусть, для определенности, матрица системы правая треугольная. Записав подробно все уравнения системы (35.1), будем иметь

$$\begin{aligned} g_{11}u_1+g_{12}u_2+\ldots+g_{1n}u_n&=l_1,\\ g_{22}u_2+\ldots+g_{2n}u_n&=l_2,\\ \ldots\ldots\ldots\ldots\ldots&\\ g_{nn}u_n&=l_n. \end{aligned} \tag{35.3}$$

Ясно, что $u_n=l_n/g_{nn}$. Предположим, что уже вычислены u_n, u_{n-1}, ..., u_{i+1} из последних $n-i$ уравнений (35.3). Из i-го уравнения находим

$$u_i=\frac{l_i-g_{i,\,i+1}u_{i+1}-\ldots-g_{i,\,n}u_n}{g_{ii}}.$$

Таким образом последовательно определяем все координаты u_n, u_{n-1}, ..., u_1 вектора u.

Полученные формулы удобны для применения операции накопления. Пусть $\tilde{u}_n$, $\tilde{u}_{n-1}$, ..., $\tilde{u}_{i+1}$ — реально вычисленные величины; тогда

$$\tilde{u}_i=\mathrm{fl}_2\left(\frac{l_i-g_{i,\,i+1}\tilde{u}_{i+1}-\ldots-g_{i,\,n}\tilde{u}_n}{g_{ii}}\right)=$$

$$=\frac{l_i-g_{i,\,i+1}\tilde{u}_{i+1}-\ldots-g_{i,\,n}\tilde{u}_n}{g_{ii}}(1+\varepsilon_i),$$

где ε_i принимает обычные для ошибок значения. Если $\varepsilon_i\neq-1$, то

$$\tilde{u}_i=\frac{l_i-g_{i,\,i+1}\tilde{u}_{i+1}-\ldots-g_{i,\,n}\tilde{u}_n}{g_{ii}+\delta_{ii}},$$

где

$$|\delta_{ii}|\leqslant\frac{1}{2}|g_{ii}|\beta^{-t+1}. \tag{35.4}$$

Если же $\varepsilon_i=-1$, то это означает, что

$$\left|\frac{l_i-g_{i,\,i+1}\tilde{u}_{i+1}-\ldots-g_{i,\,n}\tilde{u}_n}{g_{ii}}\right|<\omega$$

и тогда

$$\tilde{u}_i=\frac{(l_i+\nu_i)-g_{i,\,i+1}\tilde{u}_{i+1}-\ldots-g_{i,\,n}\tilde{u}_n}{g_{ii}},$$

где

$$|\nu_i|\lesssim|g_{ii}|\,\omega. \tag{35.5}$$

Итак, реально вычисленное решение $\tilde{u}$ системы (35.1) с треугольной матрицей G является точным решением возмущенной системы (35.2). При этом матрица эквивалентного возмущения Δ диагональная, а ее элементы удовлетворяют неравенствам (35.4). Элементы эквивалентного возмущения ν в правой части системы удовлетворяют неравенствам (35.5). Отметим, что всегда $\Delta\nu=0$.

Система с ортогональной матрицей. Если матрица G системы (35.1) ортогональная, то решение u находится весьма просто. Именно,

$$u=G'l, \tag{35.6}$$

откуда следует, что $u_r=\sum_{p=1}^{n}g_{pr}l_p$ для $1\leqslant r\leqslant n$. Эти формулы можно использовать и для численного определения решения. Вычисляем

$$\tilde{u}_r=\mathrm{fl}_2\left(\sum_{p=1}^{n}g_{pr}l_p\right)=(1+\mu_r)\sum_{p=1}^{n}g_{pr}l_p. \tag{35.7}$$

Здесь μ_r снова принимает обычные для ошибок значения. Отсюда следует, что $\tilde{u}=u+\delta$, где

$$\|\delta\|_E\lesssim\frac{1}{2}\beta^{-t+1}\|u\|_E+\sqrt{n}\,\omega.$$

Вектор $\tilde{u}$ будет точным решением системы вида (35.2), если положить $\Delta=0$, $\nu=-G\delta$. Так как матрица G ортогональная, то

$$\|\nu\|_E\lesssim\frac{1}{2}\beta^{-t+1}\|l\|_E+\sqrt{n}\,\omega. \tag{35.8}$$

Если матрица G не является ортогональной в точном смысле, но близка к таковой, то решение системы (35.1) в этом случае все равно находят согласно (35.6). По существу, это означает замену матрицы G матрицей G'^{-1}. В силу близости матрицы G к ортогональной, $\|G\|_E$ и $\|G'^{-1}\|_E$ близки к единице, поэтому

$$\|G'^{-1}-G\|_E=\|G'^{-1}(E-G'G)\|_E\leqslant$$
$$\leqslant\|G'^{-1}\|_E\|E-G'G\|_E\cong\|E-G'G\|_E.$$

Следовательно, вектор $\bar{u}$, координаты которого определяются по формулам (35.7), будет точным решением возмущенной системы (35.2). При этом эквивалентное возмущение ν в правой части удовлетворяет неравенству (35.8), а эквивалентное возмущение Δ матрицы системы — неравенству

$$\|\Delta\|_E\lessapprox\|E-G'G\|_E, \tag{35.9}$$

Как мы покажем в § 38, процесс решения системы с матрицей, близкой к ортогональной, можно организовать так, что возмущение Δ не будет влиять на точность. Для этого вычисления по формулам типа (35.7) придется проводить не один раз, а несколько.

Система с двухдиагональной матрицей. Рассмотрим, для определенности, систему с правой двухдиагональной матрицей. Имеем

$$\begin{aligned} g_{11}u_1+g_{12}u_2 &= l_1,\\ g_{22}u_2+g_{23}u_3 &= l_2,\\ \cdots\cdots\cdots\cdots&\\ g_{nn}u_n &= l_n. \end{aligned}$$

Эту систему снова решаем с помощью обратной подстановки, но теперь применение операций накопления не дает существенного выигрыша в точности. Предположим, что из последних $n-i$ уравнений уже определены $\bar{u}_n, \bar{u}_{n-1}, \ldots, \bar{u}_{i+1}$. Из i-го уравнения находим

$$\bar{u}_i=fl\left(\frac{l_i-g_{i,\,i+1}\bar{u}_{i+1}}{g_{ii}}\right)=\frac{(l_i-g_{i,\,i+1}\bar{u}_{i+1}(1+\varepsilon_i'))(1+\varepsilon_i'')(1+\varepsilon_i''')}{g_{ii}},$$

где ε_i', ε_i'', ε_i''' принимают обычные для ошибок значения.

Если все ошибки отличны от -1, то ε_i' определяет эквивалентное возмущение элемента $g_{i,\,i+1}$, а ε_i'', ε_i''' — эквивалентное возмущение элемента g_{ii}. Если же среди ошибок будут равные -1, то они учитываются эквивалентным возмущением правой части. Реально вычисленное решение $\bar{u}$ будет точным решением возмущенной системы (35.2). При этом матрица эквивалентного возмущения Δ правая двухдиагональная, а ее элементы δ_{ii}, $\delta_{i,i+1}$ удовлетворяют неравенствам

$$\begin{aligned} |\delta_{ii}| &\lessapprox |g_{ii}|\,p^t,\\ |\delta_{i,i+1}| &\lessapprox \tfrac{1}{2}|g_{i,i+1}|\beta^{-t+1}. \end{aligned} \tag{35.10}$$

Элементы ν_i эквивалентного возмущения ν удовлетворяют неравенствам

$$|\nu_i|\lessapprox(1+|g_{ii}|)\beta^{-t}. \tag{35.11}$$

Система с почти треугольной матрицей. Будем считать, для определенности, что матрица системы правая почти треугольная. Такую систему целесообразно решать следующим способом. Сначала, с помощью умножения слева на подходящим образом выбранную последовательность матриц вращения T_{12}, T_{23}, ..., $T_{n-1,n}$, приводим исходную систему к системе с правой треугольной матрицей. Затем решение полученной системы находится с помощью обратной подстановки. Рассмотренный ранее анализ ошибок полностью охватывает этот случай. Если предположить, что

$$\|G\|_E,\ \|l\|_E<\beta^{t-3}, \tag{35.12}$$

то нетрудно показать, что реально вычисленное решение $\bar{u}$ будет точным решением возмущенной системы (35.2); при этом

$$\begin{aligned} \|\Delta\|_E &\lessapprox \sqrt{2}\,np^{t+1}\|G\|_E,\\ \|\nu\|_E &\lessapprox \sqrt{2}\,np^{t+1}\|l\|_E. \end{aligned} \tag{35.13}$$

Система с трехдиагональной матрицей. Эту систему можно решать таким же способом, как и систему с почти треугольной матрицей. Однако теперь использование операций накопления при решении системы с треугольной матрицей не дает существенного выигрыша в точности. Если всюду использовать лишь прежние вычисления с

одинарной точностью, то в предположении (35.12) вместо (35.13) будем иметь

$$\begin{aligned} \|\Delta\|_E &\lesssim (3\sqrt{2}+1)\, p^{-t+1} \|G\|_E, \\ \|v\|_E &\lesssim \sqrt{2}\, np^{-t+1} \|l\|_E. \end{aligned} \tag{35.14}$$

УПРАЖНЕНИЯ

1. Выполнить анализ ошибок при решении систем с треугольной матрицей с помощью исключения элементов. Сравнить результаты, относящиеся к различным порядкам исключения, между собой.
2. Выполнить анализ ошибок для обратной подстановки в случае вычислений с одинарной точностью.
3. Можно ли систему с ортогональной матрицей решать с помощью исключения элементов?
4. Рассмотреть использование процесса исключения элементов с выбором главного элемента по столбцу (строке) для решения системы с правой (левой) почти треугольной матрицей. Выполнить анализ ошибок и сравнить результаты с (35.13).
5. Рассмотреть использование процесса исключения элементов с выбором главного элемента по столбцу или строке для решения системы с трехдиагональной матрицей. Выполнить анализ ошибок и сравнить результаты с (35.14).
6. Рассмотреть применение метода квадратного корня для решения системы с положительно определенной трехдиагональной матрицей. Выполнить анализ ошибок и сравнить результаты с (35.14).

§ 36. Решение систем с невырожденными матрицами

Значительная часть наиболее известных численных методов решения систем линейных алгебраических уравнений

$$Ax = b \tag{36.1}$$

основана на разложении матрицы A на множители. В зависимости от того, как связаны сомножители с матрицей A, различают две схемы построения методов.

В первой схеме предполагается, что явно известны сами сомножители, на которые разложена матрица A. Пусть

$$A = BC. \tag{36.2}$$

Решение системы (36.1) сводится к последовательному решению таких систем:

$$\begin{aligned} By &= b, \\ Cx &= y. \end{aligned} \tag{36.3}$$

Во второй схеме предполагается, что найдены матрицы L, S, G, для которых выполняется соотношение

$$LAS = G. \tag{36.4}$$

Тогда

$$x = Su, \tag{36.5}$$

где u есть решение системы

$$Gu = l \tag{36.6}$$

с матрицей G из (36.4) и правой частью

$$l = Lb. \tag{36.7}$$

Решение системы (36.1) сводится теперь к вычислению вектора l согласно (36.7), решению системы (36.6) и определению искомого вектора x по формуле (36.5). В данной схеме матрицы L и S обычно бывают представлены в виде произведения элементарных матриц.

Все формы разложения матрицы, которые рассмотрены нами, имеют вид либо (36.2), либо (36.4). Решение систем (36.3), (36.6), по существу, было изучено в § 35. Поэтому численные методы для систем линейных алгебраических уравнений (36.1) можно строить, вообще говоря, на основе любых исследованных ранее разложений.

Эти методы в отношении скорости и объема требуемой памяти ЭВМ будут обладать такими же характеристиками, как и соответствующие разложения матрицы. Главный член числа арифметических операций остается без изменения, так как для решения систем (36.1) при наличии разложений (36.2), (36.4) нужно выполнить на порядок меньше вычислительной работы, чем для получения самих разложений. При этом, по существу, не требуется никакой дополнительной памяти ЭВМ по сравнению с той, которая уже была использована при разложении матрицы. Поэтому выбор вида разложения матрицы для построения численного метода решения систем линейных алгебраических уравнений, обладающего нужными характеристиками скорости и объема памяти ЭВМ, можно осуществлять, используя табл. 34.1.

Связь точности решения систем с точностью разложения гораздо сложнее. В общем случае удается лишь показать, что реально вычисленное решение $\bar{x}$ близко к

некоторому вектору $\tilde{x}$, который является точным решением возмущенной системы

$$(A + E)\tilde{x} = b + e$$

с относительно малыми возмущениями E, e. Во многих случаях при этом оказывается возможным получить и оценки вида

$$\begin{aligned} \|E\|_E &\leqslant q(n)\, p^{t+1} \|A\|_E, \\ \|e\|_E &\leqslant \psi(n)\, p^{t+1} \|b\|_E, \\ \|\tilde{x} - x\|_E &\leqslant \theta(n)\, p^{t+1} \|\tilde{x}\|_E, \end{aligned} \qquad (36.8)$$

где функции $q(n)$, $\psi(n)$, $\theta(n)$ зависят только от порядка матрицы A и типа ее разложения. Все или некоторые из этих функций могут быть намного больше, чем соответствующие функции $f(n)$ из табл. 34.1.

Точность разложения матрицы на множители является одной из важнейших характеристик, определяющих общую погрешность решения системы (36.1). Как вытекает из формулы (10.10), (36.8), справедливо соотношение

$$\frac{\|\tilde{x} - x\|_E}{\|x\|_E} \lesssim \theta(n)\, p^{t+1} + \nu_A (\varphi(n) + \psi(n))\, p^{t+1}, \qquad (36.9)$$

где x — точное решение системы (36.1), а ν_A — спектральное или евклидово число обусловленности матрицы A. Отсюда следует, что точность решения системы, по существу, зависит лишь от максимальной из функций в (36.8). Но $\varphi(n)$ заведомо не может быть меньше, чем $f(n)$. Поэтому разложение, для которого максимальная из функций $\varphi(n)$, $\psi(n)$, $\theta(n)$ не слишком сильно превосходит $f(n)$, мы будем считать эффективным по точности. Как будет показано ниже, все разложения из табл. 34.1 являются таковыми.

По-видимому, нет никакой необходимости исследовать всевозможные комбинации допустимых видов матриц в разложениях (36.2), (36.4). Численные методы, использующие разложения любого типа, не могут обеспечить решение систем (36.1) с меньшими затратами времени и памяти ЭВМ, чем методы, основанные на разложениях из табл. 34.1. К тому же не известно ни одного разложения, более точного, чем эти. Сказанное выше позволяет нам ограничиться в дальнейшем исследованием точности тех

методов решения систем уравнений, которые основаны на использовании разложений из табл. 34.1.

Из всех разложений в табл. 34.1 лишь разложение на треугольные множители, выполненное по компактной схеме, имеет вид (36.2). Построим соответствующие численные методы решения систем линейных алгебраических уравнений (36.1). Тогда, решая вспомогательные системы (36.3) с помощью обратной подстановки, описанной в § 35, мы будем вносить дополнительные эквивалентные возмущения в диагональные элементы матриц B и C. Но ошибки такого порядка уже могли появиться во всех элементах матриц B и C при их вычислении. Поэтому для методов этой группы имеем

$$q(n) \leqslant 2f(n), \quad \psi(n) = \theta(n) = 0. \qquad (36.10)$$

Рассмотрим теперь разложения вида (36.4) из табл. 34.1. Пусть матрица G — треугольная и получена с помощью умножения матрицы A слева на последовательность матриц вращения, отражения или матриц типа (24.3). При решении системы (36.6) снова будут вноситься дополнительные эквивалентные возмущения в диагональные элементы матрицы G, но значительно большие ошибки уже могли появиться во всех элементах этой матрицы при ее вычислении. Эквивалентное возмущение в векторе b при вычислении правой части l образуется по тому же закону, как и при преобразовании матрицы A. Следовательно, для методов, основанных на этих разложениях,

$$q(n) \cong f(n), \quad \psi(n) = f(n), \quad \theta(n) = 0. \qquad (36.11)$$

Дополнительное включение перестановок строк и столбцов в преобразование матрицы A не меняет полученных соотношений.

Если с матрицей A выполняются аналогичные правосторонние преобразования, то в этом случае

$$q(n) \cong f(n), \quad \psi(n) = 0, \quad \theta(n) = f(n). \qquad (36.12)$$

Предположим теперь, что мы используем разложения (36.4), основанные на двухсторонних унитарных преобразованиях матрицы A. Матрица G может не быть треугольной. Однако в § 35 было показано, что какой бы вид она не имела, решение системы (36.6) дает эквивалентные возмущения в G и l, зависящие от n существенно слабее,

чем $f(n)$. Так как матрицы L, S унитарные, то эти эквивалентные возмущения легко переводятся в эквивалентные возмущения соответственно A и b. При вычислении векторов l, x эквивалентные возмущения образуются по такому же закону, как и при унитарном одностороннем преобразовании матрицы A. Поэтому для соответствующей группы численных методов решения систем линейных алгебраических уравнений имеем

$$\varphi(n) \cong f(n), \quad \psi(n) \lesssim \frac{1}{2} f(n), \quad \theta(n) \lesssim \frac{1}{2} f(n). \quad (36.13)$$

Наконец, рассмотрим использование процесса ортогонализации. Матрица L в разложении (36.4) представлена в виде произведения матриц типа (24.9) и диагональных матриц, у которых не более одного элемента отлично от единицы. Если при вычислении вектора l согласно (36.7) применять операции накопления, то реально вычисленный вектор будет иметь в каждой своей координате такие же ошибки, как и при правильном округлении этих координат. Но матрица G близка к унитарной. Следовательно, евклидова норма вектора ошибок в l может быть с таким же весом перенесена в решение u. Мы уже отмечали, что система (36.6) с матрицей G, близкой к унитарной, может быть решена настолько точно, что эквивалентное возмущение войдет лишь в правую часть и согласно (35.8) будет весьма малым. Принимая во внимание значение функции $f(n)$ для процесса ортогонализации, заключаем, что теперь

$$\varphi(n) = f(n), \quad \psi(n) = 0, \quad \theta(n) \leqslant f(n). \quad (36.14)$$

Полученные оценки для функций (36.8) позволяют сделать общий вывод о величине отклонения реально вычисленного вектора $\tilde{x}$ от точного решения x системы (36.1). Для этого воспользуемся неравенством (36.9) и заметим, что всегда $\nu_A \geqslant 1$. Из соотношений (36.9) — (36.14) вытекает, что для численных методов решения систем линейных алгебраических уравнений, основанных на рассмотренных разложениях из табл. 34.1, асимптотически будем иметь

$$\frac{\|\tilde{x} - x\|_E}{\|x\|_E} \lesssim 2\nu_A f(n)\, p^{t+1}. \quad (36.15)$$

Конечно, эта оценка справедлива лишь в том случае, когда вычислительный алгорифм на всех этапах реализуется по описанным выше процедурам.

Отметим некоторые из известных методов, характеристики которых определяются табл. 34.1 и формулой (36.15).

Метод Гаусса [6]. Основан на разложении (36.4). Матрица G — правая треугольная, матрица L представлена как произведение матриц типа (24.3). Если используется одна из стратегий выбора ведущего элемента, связанная с изменением порядка просмотра столбцов, то матрица S будет матрицей перестановок. При изменении порядка просмотра строк матрица L будет представлена как произведение матриц (24.3) и матриц перестановок.

Компактная схема метода Гаусса [6]. Основана на разложении (36.2). Матрица B — левая треугольная, матрица C — правая треугольная.

Метод квадратного корня [2]. Основан на разложении (36.2) для положительно определенной матрицы A. Матрица C — правая треугольная, $B = C^*$.

Метод отражений [2]. Основан на разложении (36.4). Матрица G — правая треугольная, матрица L представлена как произведение матриц отражения.

Нормализованный метод отражений [7]. Основан на разложении (36.4). Матрица G — нормализованная левая (правая) треугольная, матрица $S(L)$ представлена как произведение матриц отражения; матрица $L(S)$ является матрицей перестановок.

Двухсторонний метод отражений [3]. Основан на разложении (36.4). Матрица G — двухдиагональная, матрицы L и S представлены как произведения матриц отражения.

Симметричный метод отражений [7]. Основан на разложении (36.4) и применяется для эрмитовых матриц A. Матрица G — эрмитова трехдиагональная, матрица L представлена как произведение матриц отражения и совпадает с матрицей S^*.

Метод ортогонализации [2]. Основан на разложении (36.4). Матрица G — унитарная, матрица L представлена как произведение матриц типа (24.9) и диагональных матриц, у которых не более одного элемента отлично от единицы. Матрица S совпадает с единичной.

Все рассмотренные методы особенно удобны для решения систем линейных алгебраических уравнений с многими

правыми частями и одной и той же матрицей. В этом случае соответствующие разложения (36.2), (36.4) находятся лишь один раз. Многократно приходится решать только простые системы (36.3), (36.6) и выполнять преобразования (36.5), (36.7).

Согласно формуле (36.15) точность любого метода полностью определяется точностью разложения матрицы на множители. Но как видно из табл. 34.1, в этом отношении различные разложения отличаются друг от друга не так уж сильно. Поэтому, *если какой-либо из методов не обеспечивает нужной точности решения системы линейных алгебраических уравнений, то нет никаких оснований надеяться на то, что другой метод будет давать для этой же системы существенно лучшие результаты. Скорее всего, такую систему следует рассматривать как неустойчивую и решать ее одним из тех способов, которые обсуждаются в* §§ 39, 41.

УПРАЖНЕНИЯ

1. Привести примеры численных методов решения линейных систем, которые не охватываются оценкой (36.15).

2. Рассмотрим любой численный метод, для которого выполняются оценки (36.8). Доказать, что

$$\|Ax-b\|_E \lesssim (\varphi(n)+\psi(n)+\theta(n))\,p^{-t+1}\,\|A\|_E\,\|x\|_E.$$

3 Провести анализ формулы (36.15) для конкретных ЭВМ и конкретных численных методов. При каких значениях числа обусловленности можно гарантировать относительную погрешность решения не больше, чем 10^{-2}?

§ 37. Системы с матрицами полного ранга

Рассмотрим прямоугольные матрицы размеров $m\times n$, ранг которых совпадает с минимальным из чисел m, n. Такие матрицы называются матрицами *полного ранга*. Отличительная их особенность состоит в том, что они не изменяют свой ранг при любых достаточно малых возмущениях. Системы линейных алгебраических уравнений с матрицами полного ранга имеют много общего с системами с невырожденными матрицами. Такие системы называются *недоопределенными*, если $m<n$, и *переопределенными*, если $m>n$.

Пусть матрица A системы (36.1) является матрицей полного ранга. Переопределенная система может быть несовместной. Недоопределенная система всегда совместна, но имеет не единственное решение. Поэтому будем искать *нормальное псевдорешение* [1] системы (36.1), т. е. такой вектор x_0, который среди всех векторов x, минимизирующих функционал невязки

$$\Phi_0(x)=\|Ax-b\|_E^2,$$

имеет наименьшую евклидову норму.

Инвариантность евклидовой нормы к унитарным преобразованиям позволяет свести задачу отыскания нормального псевдорешения системы общего вида к более простой задаче. Действительно, выполним какое-нибудь преобразование (36.4) с унитарными матрицами L, S. Тогда легко проверить, что в обозначениях (36.5) — (36.7)

$$\Phi_0(x)=\|Ax-b\|_E^2=\|Gu-l\|_E^2=\Phi_0(u),$$

при этом $\|x\|_E=\|u\|_E$. Следовательно, задача определения нормального псевдорешения системы (36.1) эквивалентна решению такой же задачи для системы (36.6). Но преобразование (36.4) всегда можно выбрать так, что матрица G будет достаточно простой, например, треугольной, нормализованной трапецевидной, двухдиагональной и т. п. На основе этого преобразования можно построить достаточно эффективные численные методы.

Однако системы с прямоугольными матрицами можно решать и другим способом. Известно [1], что единственное псевдорешение переопределенной системы (36.1) с матрицей A полного ранга является обычным решением системы

$$A^*Ax=A^*b \tag{37.1}$$

с квадратной невырожденной матрицей A^*A порядка n. Нормальное решение x_0 недоопределенной системы (36.1) получается из решения системы

$$AA^*y=b \tag{37.2}$$

с квадратной невырожденной матрицей AA^* порядка m путем простого преобразования

$$x_0=A^*y. \tag{37.3}$$

Поэтому прямое вычисление матрицы и правой части систем (37.1), (37.2) и последующее решение этих систем любым из исследованных ранее методов также вполне приемлемо для построения численных методов решения исходной задачи.

Формально соотношения (37.1) — (37.3) легко заменить другими. Рассмотрим решение переопределенной системы. Если x_0 — точное псевдорешение, то невязка

$$r_0 = b - Ax_0 \tag{37.4}$$

согласно (37.1) удовлетворяет соотношению

$$A^* r_0 = 0.$$

Поэтому вместо системы (37.1) можно решать систему

$$\begin{bmatrix} E & A \\ A^* & 0 \end{bmatrix} \begin{bmatrix} r_0 \\ x_0 \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}. \tag{37.5}$$

Вместо соотношений (37.2), (37.3) получаем аналогичную систему:

$$\begin{bmatrix} E & A^* \\ A & 0 \end{bmatrix} \begin{bmatrix} x_0 \\ z_0 \end{bmatrix} = \begin{bmatrix} 0 \\ b \end{bmatrix}. \tag{37.6}$$

Здесь — z_0 совпадает с решением y_0 системы (37.2). Матрицы систем (37.5), (37.6) эрмитовы невырожденные порядка $m+n$.

Итак, численные методы решения систем с прямоугольными матрицами полного ранга можно строить тремя различными способами. Первый способ связан с унитарным преобразованием матрицы и минимизацией функционала невязки, второй с решением систем (37.1), (37.2), третий — с решением систем (37.5), (37.6). Какому из способов отдать предпочтение — не очевидно. Тем более, что все они требуют примерно одинаковых затрат как по времени решения, так и по объему используемой памяти ЭВМ. Необходимо их сравнение по точности. Мы начнем исследования с первого способа.

Предположим, что система недоопределенная. Приведем матрицу A с помощью умножения справа на унитарную матрицу S к левой треугольной матрице G. Пусть

$$G = [\tilde{G} \mid 0], \tag{37.7}$$

где $\tilde{G}$ — невырожденная левая треугольная матрица порядка m. Если нормальное решение u_0 системы (36.6) представить в виде

$$u_0 = \begin{bmatrix} u_0' \\ u_0'' \end{bmatrix},$$

где размерность вектора u_0' равна m, то для векторов u_0', u_0'' будем иметь

$$\tilde{G} u_0' = l, \quad u_0'' = 0. \tag{37.8}$$

Решение системы в (37.8) осуществляется по одному из тех алгорифмов, которые были исследованы ранее, например, с помощью обратной подстановки.

Анализ ошибок округления в описанном процессе выполняется по той же схеме, что и для системы с невырожденной матрицей. Единственное отличие заключается в том, что теперь следует учитывать отклонение матриц преобразования от унитарных. Для приведения матрицы A к матрице G наиболее целесообразным является использование преобразований отражения. В этом случае в соответствии с оценкой (16.6) получаем, что реально вычисленный вектор x_0 удовлетворяет соотношению

$$\frac{\| x_0 - x_0 \|_E}{\| x_0 \|_E} \lesssim 9{,}8 m \nu_A^+ \rho^{-t+1}, \tag{37.9}$$

где $\nu_A^+ = \| A \|_{2,E} \| A^+ \|_{2,E}$ или, что то же самое,

$$\nu_A^+ = \| \tilde{G} \|_{2,E} \| \tilde{G}^{-1} \|_{2,E}. \tag{37.10}$$

Предположим, далее, что система переопределенная. Приведем матрицу A с помощью умножения слева на унитарную матрицу L к правой треугольной матрице G. Ясно, что G можно представить в виде

$$G = \begin{bmatrix} \tilde{G} \\ 0 \end{bmatrix}, \tag{37.11}$$

где $\tilde{G}$ — невырожденная правая треугольная матрица. Пусть

$$l = \begin{bmatrix} l' \\ l'' \end{bmatrix},$$

где размерность вектора l' равна n. Тогда единственное псевдорешение u_0 системы (36.6) удовлетворяет уравнению $\tilde{G} u_0 = l'$. Снова в этом процессе наиболее целесообразным является использование преобразований отражения.

Анализ ошибок показывает, что в соответствии с оценкой (16.7) реально вычисленный вектор $\tilde{x}_0$ удовлетворяет соотношению

$$\frac{\|\tilde{x}_0 - x_0\|_E}{\|x_0\|_E} \lesssim 9{,}8n\nu_A p^{-t+1} + 4{,}9n\nu_A^* (\nu_A^* + 1)\frac{\|l''\|_E}{\|l'\|_E} p^{-t+1}. \quad (37.12)$$

Теперь вместо (37.10) имеем

$$\nu_A^* = \|\tilde{G}\|_{2,E}\|\tilde{G}^{-1}\|_{2,E}.$$

Полученная оценка (37.12) показывает, что *при решении переопределенной системы первым способом аналогия в отношении точности с системой, имеющей невырожденную матрицу, имеет место лишь в том случае, когда правая часть хорошо согласована с матрицей*, т. е. отношение $\|l''\|_E/\|l'\|_E$ достаточно мало. Если

$$(\nu_A^* + 1)\|l''\|_E/\|l'\|_E < 2, \quad (37.13)$$

то оценка (37.12), по существу, совпадает с (37.9) при замене, конечно, m на n. Если же

$$(\nu_A^* + 1)\|l''\|_E/\|l'\|_E > 2, \quad (37.14)$$

то оценка (37.12) в целом становится такой:

$$\frac{\|\tilde{x}_0 - x_0\|_E}{\|x_0\|_E} \lesssim 4{,}9n(\nu_A^*)^2\frac{\|l''\|_E}{\|l'\|_E} p^{-t+1}. \quad (37.15)$$

В практических задачах нередко возникают переопределенные системы линейных алгебраических уравнений, в которых выполняется условие (37.14). Для того чтобы ответить на вопрос, с какой же точностью можно решить такие системы, рассмотрим второй способ их решения.

Будем вычислять матрицу и правую часть системы (37.1), используя операцию накопления. Тогда в действительности придется иметь дело с системой

$$(A^*A + \mathrm{E})x = A^*b + e, \quad (37.16)$$

где

$$\|\mathrm{E}\|_E \lesssim \frac{1}{2}p^{-t+1}\|A^*A\|_E,$$
$$\|e\|_E \lesssim \frac{1}{2}p^{-t+1}\|A^*b\|_E. \quad (37.17)$$

Решая систему (37.16) одним из исследованных ранее методов, мы получим некоторый вектор $\tilde{x}_0$, для которого согласно (36.15), (37.17) выполняется соотношение

$$\frac{\|\tilde{x}_0 - x_0\|_E}{\|x_0\|_E} \lesssim (2f(n) + 1)\nu_{A^*A}p^{-t+1}.$$

Для евклидовой нормы $\nu_{A^*A} \leqslant (\nu_A^*)^2$, для спектральной нормы это неравенство обращается в равенство. Поэтому, используя, например, метод квадратного корня для решения системы (37.16), получим, что

$$\frac{\|\tilde{x}_0 - x_0\|_E}{\|x_0\|_E} \lesssim 3(\nu_A)^2 p^{t+1}. \quad (37.18)$$

При выполнении условия (37.14) оценка (37.18) значительно лучше оценки (37.15). Если в условии (37.14) знак «много больше» заменить знаком «больше», то и тогда оценка (37.18) остается более предпочтительной.

Таким образом, сравнивая первые два способа, можно сделать следующий вывод в отношении точности решения переопределенной системы линейных алгебраических уравнений. *Если правая часть системы достаточно хорошо согласована с матрицей, то такую систему целесообразно решать с помощью унитарного преобразования матрицы без перехода к системе* (37.1). *При плохом согласовании более точные результаты дает прямое решение системы* (37.1). Условием хорошего согласования является выполнение неравенства (37.13).

Заметим, что этот вывод относится лишь к точности, определяемой влиянием *ошибок округления*. Что же касается точности, определяемой погрешностью задания *входных данных*, то она, конечно, одинакова при любых способах решения системы и определяется, например, оценкой (16.7).

Оценка (37.9) настолько хороша, что *при решении недоопределенной системы линейных алгебраических уравнений* (36.1) *нет никаких оснований для перехода к системе* (37.2). Если все же находить нормальное решение такой системы согласно (37.2), (37.3) и для решения системы (37.2) использовать метод квадратного корня, то мы получим некоторый вектор $\tilde{x}_0$, для которого снова выполняется оценка (37.18). Но мы не будем останавливаться на ее выводе.

Третий способ решения исходной системы (36.1) *не дает ничего нового* по сравнению с первым. Если матрицу A привести к треугольному виду с помощью унитарного преобразования, то, несмотря на внешнее различие вычислительных схем, результаты будут полностью совпадать, включая ошибки округления. Однако в некоторых случаях этот способ оказывается полезным, например, при уточнении псевдорешения.

УПРАЖНЕНИЯ

1. Доказать, что матрица имеет полный ранг тогда и только тогда, когда никакие возмущения не приводят к увеличению ее ранга?

2. Доказать, что для матрицы A полного ранга и размеров $m \times n$ справедливы соотношения

$$A^+ = A^*(AA^*)^{-1}, \qquad m \leqslant n,$$

$$A^+ = (A^*A)^{-1}A^*, \qquad m \geqslant n.$$

3. Пусть матрица C имеет полный ранг. Доказать, что для того, чтобы матрица A тех же размеров имела полный ранг, достаточно выполнение неравенства

$$\|C^+\| \|A - C\| < 1 \tag{37.19}$$

для какой-нибудь нормы.

4. Доказать, что нормальное псевдорешение системы с матрицей полного ранга является непрерывной функцией элементов матрицы и правой части в достаточно малой окрестности их изменения.

5. Используется ли обратный анализ ошибок в терминах возмущения системы (36.1) при оценке точности ее решения вторым способом?

6. В чем причина различий точности первых двух способов?

7. Пусть матрица A полного ранга имеет размеры $m \times n$. Обозначим

$$B = \begin{cases} \begin{bmatrix} E & A \\ A^* & 0 \end{bmatrix}, & m \geqslant n, \\ \begin{bmatrix} E & A^* \\ A & 0 \end{bmatrix}, & m \leqslant n. \end{cases}$$

Доказать, что выполняются оценки

$$\|B\|_2 \leqslant 1 + \|A\|_2, \quad \|B^{-1}\|_2 \leqslant 1 + \|A^+\|_2 + \|A^+\|_2^2.$$

8. Доказать, что системы (37.5), (37.6) определяют то же нормальное псевдорешение, что и системы с матрицами

$$\begin{bmatrix} \alpha E & A \\ \beta A^* & 0 \end{bmatrix}, \quad \begin{bmatrix} \alpha E & \beta A^* \\ A & 0 \end{bmatrix} \tag{37.20}$$

и правыми частями из (37.5), (37.6). Здесь α, β — любые ненулевые числа.

9. Как выбрать параметры α, β, чтобы матрицы (37.20) имели наименьшее число обусловленности?

10. Как выбрать параметры α, β, чтобы системы с матрицами (37.20) обеспечивали наибольшую точность нормального псевдорешения системы (36.1)?

§ 38. Уточнение решения

Все рассмотренные численные методы решения систем линейных алгебраических уравнений обладают одним общим свойством. Именно, реально вычисленное решение (псевдорешение) является точным для некоторой возмущенной задачи. Выполненные исследования показывают, что эти возмущения весьма малы и нередко сопоставимы с ошибками округления входных данных. Если входные данные получены посредством каких-либо измерений или предварительных расчетов, то обычно они уже содержат значительно большие ошибки. В этом случае *всякая попытка улучшить приближенное решение (псевдорешение) без привлечения дополнительных сведений о точной задаче или ошибках входных данных окажется несостоятельной*, ибо нет никакого критерия предпочтения одного приближенного решения (псевдорешения) другому.

Положение существенно изменяется, если входные данные заданы в ЭВМ точно. Теперь среди всех приближенных решений (псевдорешений), соответствующих определенному уровню эквивалентных возмущений, можно выбрать то, которое наиболее близко к точному. Как правило, это будет правильно округленное точное решение (псевдорешение).

Рассмотрим сначала систему линейных алгебраических уравнений (36.1) с невырожденной матрицей. Пусть x — точное решение, $x^{(k)}$ — некоторое приближение к нему, полученное любым способом. Обозначим

$$x = x^{(k)} + \Delta^{(k)}. \tag{38.1}$$

Подставляя это выражение в (36.1), получим

где

$$A\Delta^{(k)} = r_k, \tag{38.2}$$

$$r_k = b - Ax^{(k)}. \tag{38.3}$$

Будем считать, что способ вычисления невязки (38.3) и численный метод решения системы (38.2) таковы, что реальная поправка $\tilde{\Delta}^{(k)}$ удовлетворяет соотношению

$$\frac{\|\Delta^{(k)}-\tilde{\Delta}^{(k)}\|_E}{\|\Delta^{(k)}\|_E}\leqslant\theta, \tag{38.4}$$

где θ заметно меньше единицы. Основная трудность в удовлетворении этого условия связана с выбором соответствующего способа вычисления невязки. Если $x^{(k)}$ близко к точному решению, то невязка становится малой и прямое ее вычисление с одинарной точностью будет приводить к большим относительным ошибкам. Кроме этого, абсолютная малость невязки может быть причиной значительных ошибок в поправке $\Delta^{(k)}$ из-за нерегулярности поведения ошибок округления вблизи машинного нуля. Поэтому наиболее целесообразно поступать следующим образом.

1. Вычисляем невязку в режиме накопления.
2. Нормируем невязку.
3. Решаем систему (38.2) одним из методов, удовлетворяющих условию (36.15).
4. Умножаем вычисленную поправку на обратную величину нормирующего множителя.

В этом случае поправка будет определена с высокой относительной точностью. Несложные вычисления показывают, что теперь

$$\theta\lessapprox(2f(n)+3/2)\,\nu_A p^{-t+1}. \tag{38.5}$$

Следовательно, для всех матриц, не являющихся патологически плохо обусловленными, число θ в (38.4) действительно можно считать заметно меньшим единицы.

Процесс решения системы (38.2) заканчивается вычислением следующего приближения $x^{(k+1)}$ к точному решению x. Находим

$$x^{(k+1)}=x^{(k)}+\tilde{\Delta}^{(k)}+v_k, \tag{38.6}$$

где v_k есть вектор ошибок, появляющихся от сложения $x^{(k)}$ и $\tilde{\Delta}^{(k)}$. Если число θ заметно меньше единицы, то вектор $x^{(k+1)}$ сонзмерим по величине с точным решением. Поэтому, не ограничивая существенно общности, можно считать, что

$$\|v_k\|_E\approx\frac{1}{2}\|x\|_E p^{-t+1}, \tag{38.7}$$

Далее из (38.1), (38.6) следует

$$\Delta^{(k+1)}=\Delta^{(k)}-\tilde{\Delta}^{(k)}-v_k,$$

откуда, учитывая (38.4), (38.7), получаем, что

$$\|\Delta^{(k+1)}\|_E\leqslant\theta\|\Delta^{(k)}\|_E+\frac{1}{2}\|x\|_E p^{-t+1}. \tag{38.8}$$

Согласно описанному процессу построим последовательность векторов $\{x^{(k)}\}$, исходя из любого вектора x_0, например, нулевого. Из (38.8) заключаем, что

$$\overline{\lim_{k\to\infty}}\|\Delta^{(k)}\|_E=\|x\|_E p^{-t+1}/2(1-\theta).$$

Если θ заметно меньше единицы, то, начиная с некоторого k, все векторы $x^{(k)}$ будут отличаться от точного решения x примерно так же, как отличается от него правильно округленное точное решение. Как правило, последовательность $\{x^{(k)}\}$ будет сходиться к правильно округленному точному решению.

Таким образом, если входные данные системы с невырожденной матрицей заданы точно, то можно построить последовательность векторов $\{x^{(k)}\}$, определяющую исключительно точное приближение к точному решению. Процесс уточнения решения тем эффективнее, чем меньше число θ. Обычно в последовательности $\{x^{(k)}\}$ достаточно построить 2 — 3 вектора, чтобы достичь нужной точности. Но и построение большего числа векторов практически не приводит к заметному увеличению общего времени решения задачи. Многократное решение систем (38.2) может быть осуществлено весьма быстро, если разложение матрицы A на множители, выполненное при решении первой системы, использовать для решения всех последующих систем с другими правыми частями.

С процессом уточнения решения связан один интересный факт. Напомним, что уже первое приближение к решению является точным решением возмущенной системы. При этом возмущения не только малы, но и не зависят практически от обусловленности матрицы. Правильно округленное точное решение также является точным решением некоторой возмущенной системы. И снова возмущения малы и не зависят от обусловленности матрицы. Для наиболее точных методов решения систем эти возмущения

соизмеримы по величине. Поэтому *нет никакого основания ожидать существенного уменьшения норм невязок при последовательном выполнении процесса уточнения решения.* Более того, нормы невязок на некоторых шагах могут даже несколько увеличиться. Несмотря на это, точность последовательных приближений $x^{(k)}$ постоянно повышается. *Описанный процесс уточнения связан не с уменьшением эквивалентных возмущений или величин невязок, а с устранением влияния обусловленности матрицы исходной системы на погрешность в решении.*

Распространение описанного процесса на системы с прямоугольными матрицами полного ранга имеет свои особенности. Они связаны прежде всего с тем, какой из рассмотренных в § 37 способов решения таких систем взят за основу процесса.

Пусть используется первый способ. Предположим, что решается переопределенная несовместная система (36.1). Если $x_0^{(k)}$ есть некоторое приближение к единственному псевдорешению x_0, то обозначим $x_0 = x_0^{(k)} + \Delta_0^{(k)}$. Но

$$\| Ax_0 - b \|_E = \| A\Delta_0^{(k)} - r_k \|_E,$$

где

$$r_k = b - Ax_0^{(k)}. \tag{38.9}$$

Поэтому поправку $\Delta_0^{(k)}$ можно находить из условия минимизации функционала невязки для системы

$$A\Delta_0^{(k)} = r_k. \tag{38.10}$$

Может показаться, что, вычисляя с высокой относительной точностью невязку (38.9), мы найдем с высокой относительной точностью поправку $\Delta_0^{(k)}$ как единственное псевдорешение системы (38.10). Однако в действительности попытка «уточнить» псевдорешение переопределенной системы приводит к следующей ситуации. Чем лучше взято приближение к псевдорешению, тем хуже его невязка будет согласована с матрицей и тем хуже будет относительная точность поправки, полученной по первому способу решения системы (38.10).

Таким образом, *первый способ решения переопределенной системы не может быть положен в основу процесса уточнения. Нельзя его использовать и для уточнения нормального решения недоопределенной системы.* В лучшем

случае здесь можно надеяться на хорошую близость к какому-нибудь решению. При этом отклонение от нормального решения может быть значительным.

Эффективные процессы уточнения для систем с прямоугольными матрицами полного ранга можно построить на основе второго и третьего способов их решения. Напомним, что эти способы связаны с решением систем (37.1), (37.2) или (37.5), (37.6), матрицы которых невырожденные.

Рассмотрим применение второго способа к решению переопределенной системы. Теперь для поправки $\Delta_0^{(k)}$ мы получаем систему

$$A^*A\Delta_0^{(k)} = s_k,$$

где

$$s_k = A^*(b - Ax_0^{(k)}).$$

Чтобы вычислить вектор s_k с высокой относительной точностью уже недостаточно использования операций накопления. При вычислении s_k, по существу, все арифметические операции необходимо выполнять с удвоенной точностью.

Предположим далее, что вторым способом решается недоопределенная система. Пусть $y_0^{(k)}$ — некоторое приближение к решению y_0 системы (37.2). Если

$$y_0 = y_0^{(k)} + \nabla_0^{(k)},$$

то для поправки $\nabla_0^{(k)}$ получаем систему

$$A^*A\nabla_0^{(k)} = t_k,$$

где

$$t_k = b - AA^*y_0^{(k)}.$$

Снова при вычислении вектора t_k все арифметические операции надо выполнять с удвоенной точностью. С удвоенной точностью следует сохранить последнее приближение $y_0^{(k)}$ и выполнить преобразование (37.3).

Необходимость использования вычислений с удвоенной точностью может вызывать определенные трудности при практической реализации методов. От этого недостатка свободны процессы уточнения, основанные на третьем способе решения систем с прямоугольными матрицами.

Процессы уточнения для систем (37.5), (37.6) осуществляются почти одинаково, поэтому мы рассмотрим лишь уточнение решения системы (37.5). Пусть $x_0^{(k)}$, $r_0^{(k)}$ — приближения соответственно к псевдорешению x_0 и его невязке r_0. Обозначим

$$\begin{bmatrix} r_0 \\ x_0 \end{bmatrix} = \begin{bmatrix} r_0^{(k)} \\ x_0^{(k)} \end{bmatrix} + \begin{bmatrix} \sigma_0^{(k)} \\ \delta_0^{(k)} \end{bmatrix}.$$

Тогда для поправок $\sigma_0^{(k)}$, $\delta_0^{(k)}$ получаем систему

$$\begin{bmatrix} E & A \\ A^* & 0 \end{bmatrix} \begin{bmatrix} \sigma_0^{(k)} \\ \delta_0^{(k)} \end{bmatrix} = \begin{bmatrix} g_k \\ h_k \end{bmatrix},$$

где $$g_k = b - Ax_0^{(k)} - r_0^{(k)}, \quad h_k = -A^* r_0^{(k)}.$$

Для вычисления этих векторов с высокой относительной точностью вполне достаточно использования операций накопления.

Многократность решения систем не приводит к заметному увеличению времени счета. Если для систем (37.1), (37.2) применяется метод квадратного корня, то разложения матриц A^*A и AA^*, полученные на первом шаге, используются и на всех остальных шагах. При решении систем (37.5), (37.6) и в процессах уточнения, связанных с ними, целесообразно выполнить предварительно унитарное преобразование матрицы A к матрице G простого вида. Если

$$LAS = G,$$

где L, S — унитарные матрицы, то это определяет и разложения матриц в (37.5), (37.6). Именно,

$$\begin{bmatrix} E & A \\ A^* & 0 \end{bmatrix} = \begin{bmatrix} L^* & 0 \\ 0 & S \end{bmatrix} \begin{bmatrix} E & G \\ G^* & 0 \end{bmatrix} \begin{bmatrix} L & 0 \\ 0 & S^* \end{bmatrix},$$
$$\begin{bmatrix} E & A^* \\ A & 0 \end{bmatrix} = \begin{bmatrix} S & 0 \\ 0 & L^* \end{bmatrix} \begin{bmatrix} E & G^* \\ G & 0 \end{bmatrix} \begin{bmatrix} S^* & 0 \\ 0 & L \end{bmatrix}. \qquad (38.9)$$

Разложения (38.9) используются на всех шагах процесса уточнения. Полезно иметь в виду и такие соотношения:

$$A^*A = S(G^*G)S^*, \quad AA^* = L^*(GG^*)L.$$

Скорость сходимости процессов уточнения определяется соответствующими значениями θ в (38.4). Если использовать рассмотренные выше алгорифмы решения систем линейных алгебраических уравнений с матрицами полного ранга, то для систем (37.1), (37.2)

$$\theta \leqslant 3 \|A\|_2 \|A^+\|_2 p^{-t+1}, \qquad (38.10)$$

для систем (37.5), (37.6)

$$\theta \leqslant 7{,}1q \|A\|_2 (1 + \|A^+\|_2)^2 p^{-t+1}. \qquad (38.11)$$

Здесь $q = \min\{m, n\}$.

Различие оценок (38.10), (38.11) связано прежде всего с принципиальным различием свойств систем (37.1), (37.2) и (37.5), (37.6). Пусть, например, матрица и правая часть исходной системы умножаются на одно и то же число α. Тогда ее псевдорешения* не изменяются, а решение вспомогательной системы (37.2) является однородной функцией α. Поэтому все полученные ранее оценки точности псевдорешений, включая оценку (38.10), инвариантны к такому умножению. Составные же решения систем (37.5), (37.6) не будут однородными функциями α, что и приводит к появлению неоднородной зависимости в правой части (38.11).

Выполненные исследования формально связаны с процессом уточнения, однако в действительности они имеют существенно бо́льшее значение. Заметим, что важным моментом в обосновании процесса является лишь то, что относительные ошибки всех поправок ограничены сверху константой, которая меньше единицы. При этом никак не учитывается способ получения самих последовательных приближений. Поэтому, если мы построим какой-либо процесс, обладающий в отношении поправок аналогичным свойством, то тем самым будет построен некоторый *итерационный метод* решения систем линейных алгебраических уравнений.

Рассмотрим один из распространенных способов построения итерационных методов. Пусть матрица A системы (36.1) представлена в виде

$$A = B + C, \qquad (38.12)$$

где B — невырожденная матрица. Предположим, что известно приближение $x^{(k)}$ к точному решению x. Тогда поправка будет удовлетворять системе (38.2). Заменим систему (38.2) системой

$$B\Delta^{(k)} = r_k. \qquad (38.13)$$

Из (38.12), (38.13) вытекает, что

$$B(\Delta^{(k)} - \bar{\Delta}^{(k)}) = -C\Delta^{(k)}.$$

Следовательно,

$$\frac{\|\Delta^{(k)} - \bar{\Delta}^{(k)}\|}{\|\Delta^{(k)}\|} \leqslant \|B^{-1}C\|.$$

Если выполняется условие

$$\|B^{-1}C\| < 1, \tag{38.14}$$

то последовательность векторов

$$x_{k+1} = x_k + \bar{\Delta}_k \tag{38.15}$$

будет сходиться к точному решению x.

Вычислительная схема таких методов обычно имеет иной вид. Именно, записывается рекуррентное соотношение

$$Bx_{k+1} = b - Cx_k, \tag{38.16}$$

связывающее два последовательных приближения. Это соотношение может быть получено непосредственно из (38.13), (38.15).

При выборе разложения (38.12) необходимо следить не только за выполнением условия (38.14), но и за тем, чтобы системы (38.16) с матрицей B решались значительно проще, чем с матрицей A. Для этого матрицу B выбирают либо достаточно простой, либо такой, чтобы она легко обращалась или легко раскладывалась на простые множители. Конечно, обращение матрицы или ее разложение на множители выполняется только один раз.

УПРАЖНЕНИЯ

1. Пусть матрица A системы (36.1) близка к унитарной. Доказать, что последовательность векторов $x_{k+1} = x_k + A^*(b - Ax_k)$ сходится к точному решению системы (36.1) при любом начальном приближении x_0.

2. Рассмотрим систему линейных алгебраических уравнений вида

$$x = Dx + f, \tag{38.17}$$

где $\|D\| < 1$. Доказать, что последовательность векторов $x_{k+1} = Dx_k + f$ сходится к точному решению системы (38.17) при любом начальном приближении x_0. Этот процесс называется *методом простой итерации*.

3. Пусть диагональные элементы матрицы A являются преобладающими. Возьмем в качестве матрицы B в разложении (38.12) диагональную часть A. В этом случае процесс (38.16) называется *методом Якоби*. Написать достаточные условия сходимости, используя различные матричные нормы и условие (38.14).

4. Предположим, что матрица A системы (37.1) положительно определенная. Пусть матрица B в разложении (38.12) является левой треугольной, а матрица C — строго правой треугольной. В этом случае процесс (38.16) называется *методом Некрасова*. Доказать, что последовательность векторов из (38.16) сходится к точному решению системы (36.1) при любом начальном приближении.

5. Доказать, что для того, чтобы последовательность векторов, определяемая процессом (38.16), сходилась к точному решению системы (36.1) при любом начальном приближении, необходимо и достаточно, чтобы все собственные значения матрицы $B^{-1}C$ были по модулю меньше единицы.

6. Выполнить анализ ошибок, возникающих при реализации процесса (38.16).

§ 39. Особенности решения неустойчивых систем

Выполненные исследования систем линейных алгебраических уравнений показали, что если матрица имеет полный ранг, то при соответствующем ограничении уровня ошибок входных данных и повышении точности вычислений решение (псевдорешение) системы может быть получено с любой заданной точностью. Но эти же исследования показывают и другое. Если фиксированы уровень ошибок входных данных и точность вычислений, то всегда найдутся системы с настолько большими значениями чисел обусловленности, что для них нельзя гарантировать в решении (псевдорешении) никакой точности.

Такие системы называются *неустойчивыми* или *плохо обусловленными*. В целом они характеризуются тем, что незначительное изменение условий счета может привести к недопустимо большим ошибкам в решении. Причина этого явления в основном одна — в пределах изменения ошибок входных данных или эквивалентных возмущений матрица системы либо становится матрицей не полного ранга, либо очень близка к таковой. Все трудности решения неустойчивых систем связаны, по существу, лишь с трудностями решения систем с матрицами неполного ранга в условиях возмущения входных данных и влияния ошибок округления. Таких трудностей немало.

В теоретическом плане решение систем с матрицами не полного ранга не отличается от решения рассмотренных выше систем с прямоугольными матрицами полного ранга. Пусть задана любая система

$$Ax = b \tag{39.1}$$

линейных алгебраических уравнений. Снова будем искать ее нормальное псевдорешение, т. е. такой вектор x_0, который имеет наименьшую евклидову норму среди векторов, минимизирующих функционал невязки:

$$\Phi_0(x) = \|Ax - b\|_E.$$

Известно [1], что в этом случае

$$x_0 = A^+ b,$$

где A^+ — псевдообратная матрица для матрицы A.

Однако *внешнее сходство между системами с матрицами полного и не полного ранга обманчиво.* В действительности между ними существует принципиальное различие. Именно:

Если матрица системы имеет полный ранг, то в некоторой окрестности изменения входных данных нормальное псевдорешение непрерывно. Если же матрица системы не имеет полного ранга, то в любой окрестности изменения входных данных нормальное псевдорешение разрывно.

Это различие настолько важно, что заставляет считать исследование зависимости погрешности нормального псевдорешения от возмущения входных данных и ошибок округления неотъемлемой и обязательной частью любого численного метода решения систем с матрицами неполного ранга. Тем не менее такое исследование проводится еще очень редко. По-видимому, немалую роль в этом играет тот гипноз легкости, с которой математика точных вычислений предлагает «эффективные» методы решения систем линейных алгебраических уравнений. Однако эта легкость связана лишь с тем, что не сбращается внимание на сложные проблемы, стоящие совсем рядом.

Согласно многочисленным рецептам можно решать любую систему, например, методом Гаусса с выбором главного элемента по всей матрице. Если матрица имеет полный ранг, то после выполнения всех преобразований будет получена система с треугольной невырожденной матрицей. Если же матрица имеет не полный ранг, то после выполнения меньшего числа преобразований будет получена система с треугольной матрицей, у которой одна или несколько строк нулевые. С точки зрения классической математики это более благоприятный случай, так как никакие дальнейшие преобразования не требуются. В обоих случаях решение системы с треугольной матрицей не вызывает особых трудностей. В процессе преобразований легко устанавливается и факт совместности исходной системы.

Подобные рецепты выглядят весьма привлекательно. Почти не возникает сомнений в том, что требуется лишь незначительная модификация уже известных численных методов и мы получим возможность решать системы общего вида, по крайней мере совместные системы. Кажется ясным и то, как нужно модифицировать методы. В основе этих модификаций лежит следующая идея.

Ошибки округления малы. Как правило, малы и ошибки входных данных. Будем решать систему каким-либо прямым методом, например, методом Гаусса с выбором главного элемента. Если точная матрица имеет не полный ранг, то в процессе реальных преобразований, по-видимому, получится система с треугольной матрицей, у которой все элементы последних строк будут малы. Отбросим эти уравнения и найдем решения полученной системы. Они будут служить достаточно хорошим приближением к решениям точной системы.

На основе этой идеи было опубликовано и продолжает публиковаться огромное число работ. Все отличия их друг от друга связаны лишь с использованием различных преобразований исходной системы и применением различных критериев замены «малых» элементов преобразованной системы нулями. Однако данная идея не сразу привела к эффективному решению систем линейных алгебраических уравнений общего вида. Более того, долгое время было не ясно, можно ли вообще построить устойчивый процесс решения систем с матрицами неполного ранга в условиях возмущения входных данных и влияния ошибок округления. Успех пришел лишь тогда, когда был *детально исследован весь механизм возникновения неустойчивости и найдены гарантированные средства подавления его действия.*

Для определения нормального псевдорешения наиболее целесообразно использовать унитарные преобразования исходной системы. Но, в отличие от систем с матрицами полного ранга, применение этих преобразований теперь не влечет за собой обеспечение общей устойчивости.

Предположим, что в результате унитарных преобразований получена система с двухдиагональной матрицей G порядка n, элементы g_{ij} которой таковы:

$$g_{ij} = \begin{cases} 1, & i = j, \\ \varepsilon^{-1/(n-1)}, & i = j-1, \\ 0 \text{ в остальных случаях.} \end{cases} \tag{39.2}$$

Определитель этой матрицы равен единице, равны единице и все собственные значения. Нельзя считать, что строки этой матрицы близки к линейно зависимым, так как, например, при $\varepsilon = 2^{-(n-1)}$ угол между любой вектором-строкой матрицы G и подпространством, натянутым на все предыдущие векторы-строки, не менее $\pi/8$. Поэтому совсем не ясно, можно ли какие-либо из строк матрицы G заменить нулевыми, не потеряв существенно точность решения.

Возмутим матрицу (39.2), поместив в позицию $(n, 1)$ элемент, равный $(-1)^n\varepsilon$. При $\varepsilon = 2^{-(n-1)}$ это возмущение столь мало, что для $n > 40$ оно становится меньше, чем возмущение от округления одного элемента матрицы до 12 десятичных разрядов после запятой. И все же возмущенная матрица при любом ε оказывается вырожденной. Следовательно, по величине элементов строк и столбцов матрицы (39.2) нельзя сделать правильный вывод о степени ее близости к матрице не полного ранга.

Унитарное преобразование исходной системы к системе с нормализованной трапециевидной матрицей также оказывается не очень эффективным. Рассмотрим левую нормализованную треугольную матрицу G порядка n с элементами

$$g_{ij} = \begin{cases} 1/\sqrt{j}, & i = j, \\ -\sqrt{1/j - 1/(j+1)}, & i > j, \\ 0, & i < j. \end{cases} \tag{39.3}$$

Возмущение на элемент

$$\varepsilon = -\sqrt{\frac{2}{n}}\left(\prod_{j=3}^{n}\left(1 + \frac{1}{\sqrt{j}}\right)\right)^{-1} \tag{39.4}$$

в позиции $(1, n)$ делает эту матрицу вырожденной. Но при больших n

$$\varepsilon \approx -\sqrt{2}e^{-2\sqrt{n}}, \tag{39.5}$$

и снова величина элементов строк и столбцов матрицы (39.3) не отражает степени ее близости к матрице не полного ранга.

Итак, *если матрица точной системы имеет не полный ранг, то малость возмущений входных данных и ошибок округления совсем не обязательно приведет к появлению в процессе преобразования системы каких-либо строк или столбцов, целиком состоящих из таких же малых элементов.* В этом заключается основная, но не единственная трудность построения численных методов решения систем с матрицами не полного ранга, основанных на эквивалентных преобразованиях исходной системы.

Еще одна трудность связана с обоснованием дальнейших преобразований тех систем, матрицы которых имеют строки или столбцы с малыми элементами. Путь ее преодоления, по существу, был уже указан в § 16.

Если входные данные системы с матрицей неполного ранга заданы с ошибками, то никакое повышение точности вычислений и никакие преобразования не могут обеспечить *гарантированной* точности нормального псевдорешения. Как мы уже отмечали, для этого необходимо привлечение *дополнительной* информации о точной задаче. Но предположим все-таки, что после выполнения унитарных преобразований получена система с малыми строками или столбцами. Замена этих строк и столбцов нулевыми эквивалентна малому возмущению матрицы исходной системы. Если мы сможем достаточно точно найти нормальное псевдорешение полученной системы, то согласно результатам § 16 это означает, что достаточно точно будет вычислена проекция нормального псевдорешения точной системы на одно из подпространств, натянутых на правые сингулярные векторы. Нет никаких оснований рассчитывать на получение лучшего результата без привлечения дополнительной информации.

Необходимость использования дополнительной информации при решении неустойчивых систем вызывает определенные трудности при конструировании соответствующих вычислительных алгорифмов. Эта информация весьма *разнообразна* по своей природе. Кажется, что единственная возможность учесть ее для сколько-нибудь широкого класса задач должна состоять в *параметризации* вычислительного алгорифма. В этом случае получение

достоверного приближения к нужному решению исходной задачи будет заключаться в многократном решении параметризованной задачи с целью подбора совокупности параметров согласно дополнительной информации. Два вида такой параметризации рассмотрены во второй главе. В § 16 исследована *дискретная* параметризация, связанная с аппроксимацией матрицы исходной системы близкими матрицами меньшего ранга, в § 17 — *непрерывная* параметризация, связанная с минимизацией регуляризирующего функционала.

Исследования, выполненные в §§ 16, 17, составляют теоретическую основу решения неустойчивых систем линейных алгебраических уравнений. Теперь необходимо построить такие вычислительные алгоритмы, которые сохраняли бы свою устойчивость в условиях влияния ошибок округления и позволяли бы решать параметризованные задачи достаточно быстро.

УПРАЖНЕНИЯ

1. Найти нормальные решения систем

$$x+\alpha y=1,$$
$$\alpha x+2y=\alpha$$

при $\alpha=\sqrt{2}$ и $\alpha\neq\sqrt{2}$. Сравнить эти решения между собой.

2. Найти нормальные псевдорешения систем

$$x+\alpha y=1,$$
$$\alpha x+2y=0$$

при $\alpha=\sqrt{2}$ и $\alpha\neq\sqrt{2}$. Сравнить эти псевдорешения между собой.

3. Рассмотрим любую совместную систему (39.1). Пусть $\tilde{x}$ — некоторый вектор, удовлетворяющий равенству

$$(A+E)\,\tilde{x}=b+\varepsilon, \qquad (39\,6)$$

где E, ε — малые возмущения. Доказать, что ближайшее к $\tilde{x}$ решение x системы (39.1) удовлетворяет асимптотическому соотношению

$$\|\tilde{x}-x\|_E/\|\tilde{x}\|_E \lesssim \nu_A^{\flat}\,(\delta_A+\delta_b),$$

где δ_A, δ_b — относительные величины возмущений в A, b.

4. Если $\tilde{x}$ является нормальным решением системы (39.6), то будет ли нормальным ближайшее к $\tilde{x}$ решение x системы (39.1)?

5. Рассмотрим несовместную систему (39.1) и пусть

$$\min\|Ax-b\|_E^2=\theta^2<\|b\|_E^2.$$

Предположим, что $\tilde{x}$ — некоторый вектор, для которого

$$\|A\tilde{x}-b\|_E^2=\theta^2+\varepsilon^2,$$

где ε — малое неотрицательное число. Доказать, что ближайшее к $\tilde{x}$ псевдорешение x системы (39.1) удовлетворяет асимптотическому соотношению

$$\frac{\|\tilde{x}-x\|_E}{\|\tilde{x}\|_E}\lesssim \nu_A^{\flat}\,\frac{\varepsilon}{\left(\|b\|_E^2-\theta^2\right)^{1/2}}.$$

6. Пусть сингулярные числа $\rho_1, \ldots, \rho_n$ матрицы (39.2) занумерованы в порядке убывания. Доказать, что при $0<\varepsilon<1$ выполняются соотношения

$$\varepsilon^{-1/(n-1)}-1\leqslant\rho_k\leqslant\varepsilon^{-1/(n-1)}+1, \qquad k\neq n, \quad 0<\rho_n\leqslant\varepsilon.$$

Провести анализ этих соотношений для $\varepsilon=2^{-(n-1)}$.

7. Доказать, что для сингулярных чисел матрицы (39.3) выполняются соотношения

$$(\sqrt{2}-1)/2\sqrt{2}\leqslant\rho_k\leqslant\sqrt{n}, \qquad k\neq n, \quad 0<\rho_n\leqslant|\varepsilon|,$$

где ε вычисляется согласно (39.4), (39.5). При этом

$$\sum_{k=1}^{n}\rho_k^2=n.$$

§ 40. Системы с двухдиагональными матрицами

Любая система линейных алгебраических уравнений с помощью унитарных преобразований сводится к системе с квадратной двухдиагональной матрицей. Так как унитарные преобразования реализуются вполне устойчиво, то при конструировании численных методов решения неустойчивых систем можно ограничиться рассмотрением систем с двухдиагональными матрицами.

Предположим для определенности, что G — левая двухдиагональная вещественная матрица порядка n с неотрицательными диагональными элементами. Обозначим G через G_0 и построим последовательность $\{G_k\}$, где матрицы G_k с четными номерами — левые двухдиагональные, с нечетными номерами — правые двухдиагональные. Если k четное, то

$$G_{k+1}=T_{n,n-1}^{(k)}\ \ldots\ T_{32}^{(k)}T_{21}^{(k)}G_k. \qquad (40.1)$$

В случае нечетного k

$$G_{k+1}=G_kT_{12}^{(k)}T_{23}^{(k)}\ \ldots\ T_{n-1,n}^{(k)}. \qquad (40.2)$$

Матрицы вращения $T_{ij}^{(k)}$ находятся из условия исключения элемента G_k в позиции (i, j). Если внедиагональные элементы G_k перенумеровать сверху вниз, то они при

всех k исключаются подряд, начиная с первого. Из процесса исключения сразу же вытекает следующее:

Если все элементы обеих диагоналей матрицы G_0 отличны от нуля, то будут отличны от нуля все элементы обеих диагоналей у каждой из матриц G_k.

Если среди элементов обеих диагоналей матрицы G_0 имеются нулевые, то матрица G_2 будет квазидиагональной, клетки которой диагональные, либо двухдиагональные с ненулевыми элементами на обеих диагоналях. Матрицы G_k имеют аналогичное строение при всех $k \geqslant 2$.

Эти свойства позволяют без уменьшения общности считать, что все элементы на обеих диагоналях матриц G_k являются ненулевыми.

Обозначим диагональные элементы матрицы G_k через $\rho_1^{(k)}, \ldots, \rho_n^{(k)}$, внедиагональные — через $\varepsilon_1^{(k)}, \ldots, \varepsilon_{n-1}^{(k)}$. Преобразование (40.1) унитарное, поэтому у матриц G_k и G_{k+1} из (40.1) совпадают квадраты евклидовых норм векторов-столбцов. Следовательно,

$$\begin{aligned}
&\rho_1^{(k+1)2} = \rho_1^{(k)2} + \varepsilon_1^{(k)2},\\
&\rho_2^{(k+1)2} + \varepsilon_1^{(k+1)2} = \rho_2^{(k)2} + \varepsilon_2^{(k)2},\\
&\ldots\ldots\ldots\ldots\ldots\ldots\\
&\rho_{n-1}^{(k+1)2} + \varepsilon_{n-2}^{(k+1)2} = \rho_{n-1}^{(k)2} + \varepsilon_{n-1}^{(k)2},\\
&\rho_n^{(k+1)2} + \varepsilon_{n-1}^{(k+1)2} = \rho_n^{(k)2}.
\end{aligned} \tag{40.3}$$

У матриц G_k и G_{k+1} из (40.2) совпадают квадраты евклидовых норм векторов-строк, что также приводит к (40.3). Итак, соотношения (40.3) выполняются для всех k. Отсюда получаем:

$$\rho_1^{(k+1)2} = \rho_1^{(0)2} + \sum_{r=0}^{k} \varepsilon_1^{(r)2},$$

$$\rho_2^{(k+1)2} + \sum_{r=1}^{k+1} \varepsilon_1^{(r)2} = \rho_2^{(0)2} + \sum_{r=0}^{k} \varepsilon_2^{(r)2},$$

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

$$\rho_{n-1}^{(k+1)2} + \sum_{r=1}^{k+1} \varepsilon_{n-2}^{(r)2} = \rho_{n-1}^{(0)2} + \sum_{r=0}^{k} \varepsilon_{n-1}^{(r)2},$$

$$\rho_n^{(k+1)2} + \sum_{r=1}^{k+1} \varepsilon_{n-1}^{(r)2} = \rho_n^{(0)2}.$$

Принимая во внимание равенство евклидовых норм матриц G_k, находим далее, что

$$\sum_{r=0}^{k} \varepsilon_q^{(r)2} \leqslant 2\|G\|^2$$

для всех q. Но тогда для всех q

$$\lim_{k\to\infty} \varepsilon_q^{(k)} = 0. \tag{40.4}$$

Предельные соотношения (40.4) означают, что последовательность $\{G_k\}$ сходится к диагональной матрице. Диагональные элементы $\rho_1, \ldots, \rho_n$ этой матрицы совпадают с сингулярными числами исходной матрицы G.

Исследуем скорость сходимости последовательности $\{G_k\}$. В силу унитарности преобразования (40.1) скалярные произведения соседних векторов-столбцов матриц G_k и G_{k+1} совпадают. Для матриц G_k и G_{k+1} из (40.2) совпадают скалярные произведения соседних векторов-строк. Это приводит к соотношениям

$$\rho_q^{(k+1)} \varepsilon_q^{(k+1)} = \rho_{q+1}^{(k)} \varepsilon_q^{(k)}$$

для $1 \leqslant q < n$, из которых вытекает, что

$$\varepsilon_q^{(k+1)} = \frac{\rho_{q+1}^{(k)}}{\rho_q^{(k+1)}} \varepsilon_q^{(k)}.$$

Но для всех q $\lim\limits_{k\to\infty} \rho_q^{(k)} = \rho_q$, поэтому

$$\varepsilon_q^{(k)} = O\left(\left(\frac{\rho_{q+1}}{\rho_q}\right)^k\right). \tag{40.5}$$

Сравнивая (40.4), (40.5), заключаем, что для того, чтобы выполнялись предельные соотношения (40.4), необходимо, чтобы выполнялись неравенства $\rho_{q+1} \leqslant \rho_q$ для всех q. Поэтому последовательность $\{G_k\}$ сходится к такой диагональной матрице, у которой элементы $\rho_1, \ldots, \rho_n$ упорядочены в порядке убывания, т. е.

$$\rho_1 \geqslant \rho_2 \geqslant \ldots \geqslant \rho_n.$$

Если G является матрицей неустойчивой системы, то она имеет одно или несколько малых сингулярных чисел. Предположим, что таких сингулярных чисел $n-g$ и пусть

они «оторваны» от остальных. В этом случае отношение ρ_{g+1}/ρ_g будет достаточно малым и уже после нескольких преобразований (40.1), (40.2) последние $n-g$ строк и столбцов всех матриц G_k будут состоять только из малых элементов. Так, например, матрица (39.2) при $\varepsilon = 2^{-(n-1)}$ имеет лишь одно малое сингулярное число. Но легко проверить, что выполнение только одного преобразования (40.1) приводит к тому, что последний диагональный элемент матрицы становится меньше, чем $2^{-(n-1)}$. В матрице G_2 последняя строка и последний столбец будут состоять только из малых элементов.

В § 16 мы описали применение сингулярного разложения матрицы для решения неустойчивых систем линейных алгебраических уравнений. Однако отметим, что в действительности сингулярное разложение использовалось нами лишь для того, чтобы *аппроксимировать исходную матрицу близкой матрицей меньшего ранга*, для которой легко находится нормальное псевдорешение. Для этих целей вполне пригоден процесс (40.1), (40.2), особенно в том случае, когда матрица системы имеет оторванную группу малых сингулярных чисел. Нужная аппроксимация получается путем замены строк и столбцов матрицы G_k, имеющих малые элементы, нулевыми строками и столбцами.

Процесс преобразований (40.1), (40.2) численно устойчив. Если выполняется s таких преобразований, то при использовании вычислений с одинарной точностью реально полученная матрица $\tilde{G}_s$ будет точно унитарно эквивалентна матрице $G+E_s$, при этом

$$\|E_s\|_E \lesssim \frac{4\sqrt{2}+5}{2}\, sp^{t+1}\, \|G\|_E.$$

Если малые сингулярные числа матрицы G не оторваны от остальных, то сходимость последовательности $\{G_k\}$ к диагональной матрице будет медленной. Процесс (40.1), (40.2) оказывается не очень эффективным, так как приходится запоминать большое число матриц преобразования. В этом случае при решении неустойчивых систем

$$Gu = l$$

целесообразно использовать процессы минимизации регуляризирующего функционала типа (17.1), что приводит к необходимости решать системы

$$(G^*G + \alpha E)\, u_\alpha = G^*l. \tag{40.6}$$

Эти системы решаются настолько быстро, что время подбора для них параметра α почти никогда не становится существенным фактором при решении неустойчивых систем линейных алгебраических уравнений.

УПРАЖНЕНИЯ

1. Доказать, что матрицы G_0, $G_1, \ldots$, построенные согласно (40.1), (40.2), удовлетворяют соотношениям

$$G_0^*G_0 = G_1^*G_1,$$
$$G_1G_1^* = G_2G_2^*,$$
$$\ldots\ldots\ldots\ldots$$

2. Пусть A — трехдиагональная положительно определенная матрица Построим последовательность правых двухдиагональных матриц $\{L_k\}$ по следующему предписанию:

$$A = L_1^*L_1,$$
$$L_1L_1^* = L_2^*L_2,$$
$$\ldots\ldots\ldots\ldots$$
$$L_{k-1}L_{k-1}^* = L_k^*L_k,$$
$$\ldots\ldots\ldots\ldots$$

Доказать, что при всех k матрицы $L_kL_k^*$ подобны матрице A и последовательность $\{L_kL_k^*\}$ сходится к диагональной матрице.

3 Выполнить анализ ошибок при решении систем (40.6).

§ 41. Тактика решения систем общего вида

Опираясь на выполненные исследования, мы можем теперь разработать некоторую тактику действий по решению систем линейных алгебраических уравнений общего вида. Применение этой тактики целесообразно в тех случаях, когда имеющихся сведений о системе недостаточно для того, чтобы сделать выбор численного метода и гарантировать его устойчивость.

Мы не будем накладывать какие-либо ограничения на исходную систему. Она может быть как совместной, так и несовместной, как хорошо, так и плохо обусловленной. Ранг матрицы системы может быть произвольным. *Вычислительный процесс устроен таким образом, что чем «лучше»*

исходная система, тем раньше он прекратится, давая приближенное решение. Оценка точности будет зависеть от свойств системы, обнаруживаемых по ходу процесса, и от некоторой априорной информации. Описываемая совокупность действий легко реализуется на ЭВМ и, по-видимому, является оптимальной, как по объему вычислительной работы, так и по использованию памяти ЭВМ.

Итак, пусть решается система из m линейных алгебраических уравнений с n неизвестными. Будем считать, что вместо точной системы

$$Ax = b \tag{41.1}$$

задана возмущенная система

$$\tilde{A}x = \tilde{b} \tag{41.2}$$

и при этом известны оценки вида

$$\|A - \tilde{A}\|_E \leqslant \varepsilon_A \|\tilde{A}\|_E, \quad \|b - \tilde{b}\|_E \leqslant \varepsilon_b \|\tilde{b}\|_E \tag{41.3}$$

евклидовых норм возмущений. Задача состоит в отыскании по системе (41.2) некоторого приближения $\tilde{x}_0$ к нормальному псевдорешению x_0 системы (41.1) и получении максимально возможной достоверной информации о степени близости $\tilde{x}_0$ к x_0.

Первый этап предлагаемой тактики действий всегда заключается в унитарном преобразовании матрицы $\tilde{A}$ к двухдиагональному виду. Это означает, что мы находим такие матрицы $\tilde{L}$, $\tilde{S}$ и двухдиагональную матрицу $\tilde{G}$, для которых

$$\tilde{L}(\tilde{A} + E)\tilde{S} = \tilde{G}.$$

Матрицы $\tilde{L}$, $\tilde{S}$ представлены в виде произведения вычисленных матриц отражения, а эквивалентное возмущение E удовлетворяет соотношению

$$\|E\|_E \lesssim 2{,}9\,(m+n)\,p^{-t+1}\|\tilde{A}\|_E. \tag{41.4}$$

Следующий этап связан с проверкой матриц A, $\tilde{A}$ на полноту их ранга. Согласно (37.19), (41.3), (41.4) они имеют полный ранг, если имеет полный ранг матрица $\tilde{G}$ и выполняется хотя бы одно из условий:

$$\|\tilde{G}^+\|_2 \|\tilde{G}\|_E (\varepsilon_A + 2{,}9\,(m+n)\,p^{-t+1}) \gtrsim 1,$$
$$\|\tilde{G}^+\|_E \|\tilde{G}\|_E (\varepsilon_A + 2{,}9\,(m+n)\,p^{-t+1}) \gtrsim 1.$$

Эти соотношения носят асимптотический характер. Поэтому для того, чтобы гарантировать правильность выводов, мы будем проверять выполнение более сильных условий

$$\|\tilde{G}^+\|_2 \|\tilde{G}\|_E (\varepsilon_A + 2{,}9\,(m+n)\,p^{-t+1}) \leqslant 0{,}1,$$
$$\|\tilde{G}^+\|_E \|\tilde{G}\|_E (\varepsilon_A + 2{,}9\,(m+n)\,p^{-t+1}) \leqslant 0{,}1. \tag{41.5}$$

Основная трудность в проверке (41.5) связана с вычислением нормы матрицы $\tilde{G}^+$. Если $m > n$, то

$$\tilde{G} = \begin{bmatrix} \check{G} \\ 0 \end{bmatrix},$$

где $\check{G}$ — квадратная двухдиагональная матрица порядка n. Но тогда

$$\tilde{G}^+ = [\check{G}^+ \,|\, 0].$$

Если же $m < n$, то

$$\tilde{G} = [\hat{G} \,|\, 0],$$

где $\hat{G}$ — квадратная двухдиагональная матрица поряда m, и в этом случае

$$\tilde{G}^+ = \begin{bmatrix} \hat{G}^+ \\ 0 \end{bmatrix}.$$

Следовательно, при вычислении $\tilde{G}^+$ или ее нормы можно ограничиться рассмотрением квадратной двухдиагональной матрицы.

Прямоугольная матрица $\tilde{G}$ имеет полный ранг тогда и только тогда, когда соответствующая матрица $\check{G}$ или $\hat{G}$ невырожденная. Поэтому первое, что мы должны сделать, это проверить их невырожденность. Такую проверку можно осуществить, формально вычисляя норму обратной матрицы и контролируя ее величину в процессе вычислений.

Евклидову норму матрицы, обратной к двухдиагональной, удобно находить, используя соотношения типа (26.2). Пусть для определенности $m \geqslant n$. Обозначим через $\rho_1^{(0)}, \ldots, \rho_n^{(0)}$ диагональные элементы матрицы $\tilde{G}$, а через $\varepsilon_1^{(0)}, \ldots, \varepsilon_{n-1}^{(0)}$ — внедиагональные. В соответствии с (26.2) имеем

$$\alpha_1^1 = 1/\rho_1^{(0)2},$$
$$\alpha_q^1 = \left(\alpha_{q-1}^1 \varepsilon_{q-1}^{(0)2} + 1\right)/\rho_q^{(0)2}, \qquad q = 2, 3, \ldots, n.$$

Тогда

$$\|\tilde{G}\|_E^2 = \sum_{q=1}^{n} \alpha_q^2.$$

Второе из соотношений (41.5) несколько слабее первого, но проверяется легче. Если оно не выполняется, то может выполняться другое. Обычно левая часть первого соотношения в (41.5) меньше, чем во втором, но не более, чем в $\sqrt{n}$ раз. Для вычисления спектральной нормы матрицы $\tilde{G}^+$ и проверки на полноту ранга необходимо вычислить минимальное сингулярное число матрицы $\tilde{G}$. Его можно определить с помощью процесса (40.1), (40.2). Реализация процесса не вызывает особых трудностей, так как в данном случае не нужно запоминать матрицы преобразования. Как правило, при этом невелико и время счета по крайней мере по сравнению со временем преобразования матрицы $\tilde{A}$ к двухдиагональному виду.

Предположим теперь, что выполняется одно из условий (41.5) и установлена полнота ранга матриц A, $\tilde{A}$. Отсюда вытекает, что эти матрицы не изменяют свой ранг в пределах возмущения Δ, где

$$\|\Delta\|_E \lessapprox (\varepsilon_A + 2{,}9\,(m+n)\,p^{-t+1})\,\|\tilde{A}\|_E.$$

Определяем вектор

$$\tilde{l} = \mathrm{fl}_2(\tilde{L}\tilde{b}), \tag{41.6}$$

находим нормальное псевдорешение $\tilde{u}_0$ системы

$$\tilde{G}u = \tilde{l} \tag{41.7}$$

и вычисляем вектор

$$\tilde{x}_0 = \mathrm{fl}_2(\tilde{S}\tilde{u}_0). \tag{41.8}$$

Он будет служить приближением к нормальному псевдорешению x_0 точной системы (41.1). Все вычислительные операции были подробно рассмотрены ранее. Решение системы (41.7) сводится к решению системы с двухдиагональной невырожденной матрицей, преобразования (41.6), (41.8) осуществляются по алгорифмам, исследованным в §§ 20, 21.

Выполнение любого из условий (41.5) позволяет дать асимптотически правильные оценки точности. Если $m = n$, то в соответствии с (10.10), (36.15),

$$\frac{\|\tilde{x}_0 - x_0\|_E}{\|x_0\|_E} \lessapprox \nu_{\tilde{G}}\,(\varepsilon_A + \varepsilon_b + 5{,}8\,(m+n)\,p^{-t+1}). \tag{41.9}$$

Если $m < n$, то согласно (16.6), (36.15) и аналогично (37.9)

$$\frac{\|\tilde{x}_0 - x_0\|_E}{\|x_0\|_E} \lessapprox \nu_{\tilde{G}}\,(\varepsilon_A + \varepsilon_b + 9{,}8\,(m+n)\,p^{-t+1}). \tag{41.10}$$

В случае $m > n$, в соответствии с (16.7), (37.12),

$$\frac{\|\tilde{x}_0 - x_0\|_E}{\|x_0\|_E} \lessapprox \nu_{\tilde{G}}\,(\varepsilon_A + \varepsilon_b + 9{,}8\,(m+n)\,p^{-t+1}) + \nu_{\tilde{G}}\,(\nu_{\tilde{G}}\,(\varepsilon_A + 4{,}9\,(m+n)\,p^{-t+1}) + \varepsilon_b + 4{,}9\,(m+n)\,p^{-t+1})\,\frac{\|\tilde{l}''\|_E}{\|\tilde{l}'\|_E}. \tag{41.11}$$

В последней оценке $\tilde{l}'$ есть вектор, содержащий первые n координат вектора $\tilde{l}$ из (41.6), $\tilde{l}''$ есть вектор, содержащий последние $m-n$ координат $\tilde{l}$.

Процесс получения оценок (41.9) — (41.11) почти не отличается от процесса получения других аналогичных оценок, поэтому мы не будем останавливаться на нем подробно. Во всех оценках $\nu_{\tilde{G}}$ есть наименьшая из величин $\|\tilde{G}^+\|_E\|\tilde{G}\|_E$, $\|\tilde{G}^+\|_3\|\tilde{G}\|_E$, вычисленных при проверке выполнения условий (41.5).

Если входные данные системы (41.2) заданы без ошибок или эти ошибки значительно меньше, чем эквивалентные возмущения, возникающие при переходе к системе (41.7), то вычисленное псевдорешение $\tilde{x}_0$ можно уточнить. В этом состоит очередной этап тактики действий. Уточнение осуществляется по тем алгорифмам, которые были рассмотрены в § 38, причем эффективно используется уже выполненное преобразование матрицы $\tilde{A}$ к двухдиагональному виду. Если входные данные системы (41.2) заданы с большими ошибками, то никакой процесс не может гарантировать уточнение без привлечения дополнительной информации.

Тогда

$$\|\tilde{G}^{-}\|_E^2 = \sum_{q=1}^{n} \alpha_q^2.$$

Второе из соотношений (41.5) несколько слабее первого, но проверяется легче. Если оно не выполняется, то может выполняться другое. Обычно левая часть первого соотношения в (41.5) меньше, чем во втором, но не более, чем в $\sqrt{n}$ раз. Для вычисления спектральной нормы матрицы $\tilde{G}^{-}$ и проверки на полноту ранга необходимо вычислить минимальное сингулярное число матрицы $\tilde{G}$. Его можно определить с помощью процесса (40.1), (40.2). Реализация процесса не вызывает особых трудностей, так как в данном случае не нужно запоминать матрицы преобразования. Как правило, при этом невелико и время счета по крайней мере по сравнению со временем преобразования матрицы $\tilde{A}$ к двухдиагональному виду.

Предположим теперь, что выполняется одно из условий (41.5) и установлена полнота ранга матриц A, $\tilde{A}$. Отсюда вытекает, что эти матрицы не изменяют свой ранг в пределах возмущения Δ, где

$$\|\Delta\|_E \lesssim (\varepsilon_A + 2{,}9\,(m+n)\,p^{-t+1})\,\|\tilde{A}\|_E.$$

Определяем вектор

$$\tilde{l} = \mathrm{fl}_2(\tilde{L}\tilde{b}), \tag{41.6}$$

находим нормальное псевдорешение $\tilde{u}_0$ системы

$$\tilde{G}u = \tilde{l} \tag{41.7}$$

и вычисляем вектор

$$\tilde{x}_0 = \mathrm{fl}_2(\tilde{S}\tilde{u}_0). \tag{41.8}$$

Он будет служить приближением к нормальному псевдорешению x_0 точной системы (41.1). Все вычислительные операции были подробно рассмотрены ранее. Решение системы (41.7) сводится к решению системы с двухдиагональной невырожденной матрицей, преобразования (41.6), (41.8) осуществляются по алгорифмам, исследованным в §§ 20, 21.

Выполнение любого из условий (41.5) позволяет дать асимптотически правильные оценки точности. Если $m = n$, то в соответствии с (10.10), (36.15),

$$\frac{\|\tilde{x}_0 - x_0\|_E}{\|x_0\|_E} \lesssim \nu_{\tilde{G}}\,(\varepsilon_A + \varepsilon_b + 5{,}8\,(m+n)\,p^{-t+1}). \tag{41.9}$$

Если $m < n$, то согласно (16.6), (36.15) и аналогично (37.9)

$$\frac{\|\tilde{x}_0 - x_0\|_E}{\|x_0\|_E} \lesssim \nu_{\tilde{G}}\,(\varepsilon_A + \varepsilon_b + 9{,}8\,(m+n)\,p^{-t+1}). \tag{41.10}$$

В случае $m > n$, в соответствии с (16.7), (37.12),

$$\frac{\|\tilde{x}_0 - x_0\|_E}{\|x_0\|_E} \lesssim \nu_{\tilde{G}}\,(\varepsilon_A + \varepsilon_b + 9{,}8\,(m+n)\,p^{-t+1}) + \nu_{\tilde{G}}^{2}\,(\nu_{\tilde{G}}\,(\varepsilon_A + 4{,}9\,(m+n)\,p^{-t+1}) + \varepsilon_b + 4{,}9\,(m+n)\,p^{-t+1})\,\frac{\|\tilde{l}''\|_E}{\|\tilde{l}'\|_E}. \tag{41.11}$$

В последней оценке $\tilde{l}'$ есть вектор, содержащий первые n координат вектора $\tilde{l}$ из (41.6), $\tilde{l}''$ есть вектор, содержащий последние $m-n$ координат $\tilde{l}$.

Процесс получения оценок (41.9) — (41.11) почти не отличается от процесса получения других аналогичных оценок, поэтому мы не будем останавливаться на нем подробно. Во всех оценках $\nu_{\tilde{G}}$ есть наименьшая из величин $\|\tilde{G}^{+}\|_E\|\tilde{G}\|_E$, $\|\tilde{G}^{+}\|_2\|\tilde{G}\|_E$, вычисленных при проверке выполнения условий (41.5).

Если входные данные системы (41.2) заданы без ошибок или эти ошибки значительно меньше, чем эквивалентные возмущения, возникающие при переходе к системе (41.7), то вычисленное псевдорешение $\tilde{x}_0$ можно уточнить. В этом состоит очередной этап тактики действий. Уточнение осуществляется по тем алгорифмам, которые были рассмотрены в § 38, причем эффективно используется уже выполненное преобразование матрицы $\tilde{A}$ к двухдиагональному виду. Если входные данные системы (41.2) заданы с большими ошибками, то никакой процесс не может гарантировать уточнение без привлечения дополнительной информации.

Тогда

$$\|\tilde{G}^+\|_E^2 = \sum_{q=1}^{n} \alpha_q^2.$$

Второе из соотношений (41.5) несколько слабее первого, но проверяется легче. Если оно не выполняется, то может выполняться другое. Обычно левая часть первого соотношения в (41.5) меньше, чем во втором, но не более, чем в $\sqrt{n}$ раз. Для вычисления спектральной нормы матрицы $\tilde{G}^+$ и проверки на полноту ранга необходимо вычислить минимальное сингулярное число матрицы $\tilde{G}$. Его можно определить с помощью процесса (40.1), (40.2). Реализация процесса не вызывает особых трудностей, так как в данном случае не нужно запоминать матрицы преобразования. Как правило, при этом невелико и время счета по крайней мере по сравнению со временем преобразования матрицы $\tilde{A}$ к двухдиагональному виду.

Предположим теперь, что выполняется одно из условий (41.5) и установлена полнота ранга матриц $\tilde{A}$, $\hat{A}$. Отсюда вытекает, что эти матрицы не изменяют свой ранг в пределах возмущения Δ, где

$$\|\Delta\|_E \lessapprox (\varepsilon_A + 2{,}9\,(m+n)\,p^{-t+1})\,\|\hat{A}\|_E.$$

Определяем вектор

$$\tilde{l} = \mathrm{fl}_2(\tilde{L}\tilde{b}), \tag{41.6}$$

находим нормальное псевдорешение $\tilde{u}_0$ системы

$$\tilde{G}u = \tilde{l} \tag{41.7}$$

и вычисляем вектор

$$\tilde{x}_0 = \mathrm{fl}_2(\tilde{S}\tilde{u}_0). \tag{41.8}$$

Он будет служить приближением к нормальному псевдорешению x_0 точной системы (41.1). Все вычислительные операции были подробно рассмотрены ранее. Решение системы (41.7) сводится к решению системы с двухдиагональной невырожденной матрицей, преобразования (41.6), (41.8) осуществляются по алгорифмам, исследованным в §§ 20, 21.

Выполнение любого из условий (41.5) позволяет дать асимптотически правильные оценки точности. Если $m = n$, то в соответствии с (10.10), (36.15),

$$\frac{\|\tilde{x}_0 - x_0\|_E}{\|x_0\|_E} \lessapprox \nu_{\tilde{G}}^{+}\,(\varepsilon_A + \varepsilon_b + 5{,}8\,(m+n)\,p^{-t+1}). \tag{41.9}$$

Если $m < n$, то согласно (16.6), (36.15) и аналогично (37.9)

$$\frac{\|\tilde{x}_0 - x_0\|_E}{\|x_0\|_E} \lessapprox \nu_{\tilde{G}}^{+}\,(\varepsilon_A + \varepsilon_b + 9{,}8\,(m+n)\,p^{-t+1}). \tag{41.10}$$

В случае $m > n$, в соответствии с (16.7), (37.12),

$$\begin{aligned}\frac{\|\tilde{x}_0 - x_0\|_E}{\|x_0\|_E} \lessapprox\ & \nu_{\tilde{G}}^{+}\,(\varepsilon_A + \varepsilon_b + 9{,}8\,(m+n)\,p^{-t+1}) + \\ & + \nu_{\tilde{G}}\,(\nu_{\tilde{G}}^{+}\,(\varepsilon_A + 4{,}9\,(m+n)\,p^{-t+1}) + \\ & + \varepsilon_b + 4{,}9\,(m+n)\,p^{-t+1})\,\frac{\|\tilde{l}''\|_E}{\|\tilde{l}'\|_E}.\end{aligned} \tag{41.11}$$

В последней оценке $\tilde{l}'$ есть вектор, содержащий первые n координат вектора $\tilde{l}$ из (41.6), $\tilde{l}''$ есть вектор, содержащий последние $m-n$ координат $\tilde{l}$.

Процесс получения оценок (41.9)—(41.11) почти не отличается от процесса получения других аналогичных оценок, поэтому мы не будем останавливаться на нем подробно. Во всех оценках $\nu_{\tilde{G}}$ есть наименьшая из величин $\|\tilde{G}^+\|_E\,\|\tilde{G}\|_E$, $\|\tilde{G}^+\|_2\,\|\tilde{G}\|_E$, вычисленных при проверке выполнения условий (41.5).

Если входные данные системы (41.2) заданы без ошибок или эти ошибки значительно меньше, чем эквивалентные возмущения, возникающие при переходе к системе (41.7), то вычисленное псевдорешение $\tilde{x}_0$ можно уточнить. В этом состоит очередной этап тактики действий. Уточнение осуществляется по тем алгорифмам, которые были рассмотрены в § 38, причем эффективно используется уже выполненное преобразование матрицы $\hat{A}$ к двухдиагональному виду. Если входные данные системы (41.2) заданы с большими ошибками, то никакой процесс не может гарантировать уточнение без привлечения дополнительной информации.

Предположим теперь, что не выполняется ни одно из условий (41.5) или точность, определяемая оценками (41.9)—(41.11), недостаточна. Любой из этих случаев означает, что свойства точной системы (41.1) таковы, что для того, чтобы по возмущенной системе (41.2) получить достоверную информацию о точном псевдорешении $\bar{x}_0$, необходимо привлечение дополнительной информации либо о точной системе, либо о точном псевдорешении.

Плохая определимость нормального псевдорешения в условиях возмущения входных данных в конечном счете связана с плохой определимостью его проекций на правые сингулярные векторы матрицы системы, соответствующие малым сингулярным числам. Смысл привлечения дополнительной информации состоит в том, чтобы тем или иным способом исключить влияние этих проекций, не потеряв существенно точность нормального псевдорешения.

Независимо от того, имеется ли дополнительная информация или нет, целесообразно осуществить еще один этап тактики действий. В общем случае он направлен на получение более полной информации о внутренней структуре системы. При благоприятном стечении обстоятельств на этом этапе может быть получено не только решение системы, но и оценки точности.

Выполним несколько шагов процесса (40.1), (40.2), взяв в качестве начальной матрицу $\bar{G}$. Это означает, что будут получены матрицы M_k, N_k и двухдиагональная матрица $\bar{G}_k$, для которых

$$M_k(\bar{G}+\Pi)N_k=\bar{G}_k. \tag{41.12}$$

Матрицы M_k, N_k представлены как произведения матриц вращения, а эквивалентное возмущение Π удовлетворяет соотношению

$$\|\Pi\|_E \le \frac{\sqrt{q}(\sqrt{2}+0.5)}{2}\,kp^{t+1}\|\bar{A}\|_E.$$

Если матрица точной системы имеет оторванную группу из $n-g$ малых сингулярных чисел, то после выполнения небольшого числа шагов последние $n-g$ строк и столбцов матрицы $\bar{G}_k$ обычно становятся малыми. Заменив их нулевыми строками и столбцами, мы получим близкую к $\bar{G}_k$ двухдиагональную матрицу P_k.

В соответствии с преобразованием (41.12) запишем систему (41.7) в следующем виде:

$$(M_k^*\bar{G}_kN_k^*)\,u=\bar{l},$$

и рассмотрим близкую к ней систему

$$(M_k^*P_kN_k^*)\,u=\bar{l}. \tag{41.13}$$

Для матрицы P_k легко находится псевдообратная матрица P_k^+, причем $\|P_k^+\|$ невелика. Поэтому нормальное псевдорешение системы (41.13) можно вычислить с высокой точностью. Определяем вектор

$$\bar{u}_0=\Pi(N_kP_k^+M_k\tilde{l})$$

и затем вектор $\tilde{x}_0$ согласно (41.8). Этот вектор берем в качестве приближения к нормальному псевдорешению x_0 точной системы (41.1).

Как показано в § 16, вектор $\tilde{x}_0$ будет близок к проекции x_0 на подпространство, натянутое на правые сингулярные векторы матрицы системы, соответствующие большим сингулярным числам. Кроме того, для $\tilde{x}_0$ гарантируется малость невязки и устойчивость к ошибкам входных данных и ошибкам округления. Различные количественные характеристики таких параметров точности могут быть вычислены в случае необходимости в процессе счета.

Если правая часть системы достаточно хорошо согласована с матрицей, то вектор $\tilde{x}_0$ будет близок также к x_0. Однако количественные оценки этой близости нельзя получить из результатов счета без привлечения дополнительной информации.

При практической реализации описанного этапа тактики действий процесс (40.1), (40.2) следует проводить до тех пор, пока матрицу $\bar{G}_k$ нельзя расщепить на сумму

$$\bar{G}_k=P_k+F_k, \tag{41.14}$$

где норма F_k мала, а норма P_k^+ невелика. Но если матрица исходной системы имеет оторванную группу малых сингулярных чисел, то такое расщепление, как правило, наступает очень быстро. Поэтому при проведении процесса

(40.1), (40.2) можно ограничиться выполнением небольшого числа шагов, тем более, что все оценки точности, связанные с вектором $\tilde{x}_0$, наиболее эффективны именно при оторванности малых сингулярных чисел. Если расщепление (41.14) не наступило после 8 — 10 шагов, то, скорее всего, это говорит о том, что сингулярные числа матрицы не имеют достаточного разделения.

В этом случае переходим к последнему этапу. Будем решать систему (41.2) с помощью процесса минимизации регуляризирующего функционала

$$\Phi_\alpha(x) = \alpha \|x\|_E^2 + \|\tilde{A}x - \tilde{b}\|_E.$$

Преобразование, выполненное на первом этапе, сводит эту задачу к минимизации более простого функционала

$$\Phi_\alpha(u) = \alpha \|u\|_E + \|\tilde{G}u - \tilde{l}\|_E, \tag{41.15}$$

причем сохраняются соотношения (41.6) — (41.8). Но минимизация функционала (41.15) приводит к решению систем вида

$$(\tilde{G}^*\tilde{G} + \alpha E)\, u_\alpha = \tilde{G}^*\tilde{l} \tag{41.16}$$

с трехдиагональными положительно определенными матрицами.

Если нет никакой дополнительной информации, то можно указать лишь устойчивый способ вычисления такого вектора $\tilde{x}_0$, который асимптотически будет близок к искомому вектору x_0.

Пусть нормы суммарных возмущений матрицы и правой части системы не превосходят положительного числа ε. Возьмем $\alpha = \varepsilon^{1/2}$ и определим из системы (41.16) соответствующий вектор u_α. Положим $\tilde{u}_0 = u_\alpha$ и вычислим $\tilde{x}_0$ согласно (41.8). Как вытекает из формулы (17.4), вектор $\tilde{x}_0$ будет близок к x_0 с точностью порядка $\varepsilon^{1/2}$, независимо от того, была ли исходная система совместной или несовместной. В случае совместности исходной системы можно взять $\alpha = \varepsilon^{1/3}$. При этом будет обеспечена асимптотическая близость вектора $\tilde{x}_0$ к x_0 с точностью порядка $\varepsilon^{1/3}$.

Количественные оценки этой близости снова невозможны без привлечения *дополнительной* информации о точной задаче.

УПРАЖНЕНИЯ

1. Пусть x_0 есть псевдорешение системы (41.1), обеспечивающее минимум функционала (Bx, x) с положительно определенной матрицей B. Доказать, что $x_0 = C^{-1}y_0$, если y_0 есть нормальное псевдорешение системы

$$AC^{-1}y = b,$$

а матрица C связана с B равенством $B = C^*C$.

2. Рассмотрим в условиях упражнения 1 регуляризирующий функционал

$$\Phi_\alpha^*(x) = \alpha(Bx, x) + \|Ax - b\|_E^2 \tag{41.17}$$

и пусть x_α есть вектор, обеспечивающий его минимум. Доказать, что

$$\lim_{\alpha \to +0} x_\alpha = x_0.$$

3. Доказать, что минимизация функционала (41.17) сводится к решению системы

$$(A^*A + \alpha B)\, x_\alpha = A^*b.$$

4. Пусть известно, что $\|A^{-1}\|_2 \leq \alpha$. Предположим, что в процессе реализации тактики действий выяснилось, что матрица $\tilde{G}$ имеет оторванную группу сингулярных чисел значительно меньших, чем α^{-1}. Получить гарантированные оценки близости вычисленного вектора $\tilde{x}_0$ к нормальному псевдорешению x_0 системы (41.1).

5. Пусть известны оценки сверху для евклидовых норм нормальных решений систем

$$(A^*A)^p x = A^*b$$

для каких-либо двух различных значений $p \geq 1$. Получить в процессе реализации тактики действий гарантированные оценки близости $\tilde{x}_0$ к x_0.

6. Пусть известен ранг матрицы системы (41.1). В каком случае, пользуясь этой информацией, можно получить гарантированные оценки близости $\tilde{x}_0$ к x_0 в процессе реализации тактики действий?

7. Пусть известна евклидова норма нормального псевдорешения системы (41.1). Предположим, что в процессе реализации тактики действий выяснилось, что матрица $\tilde{G}$ имеет оторванную группу малых сингулярных чисел. В каком случае по этой информации можно получить гарантированные оценки близости $\tilde{x}_0$ к x_0?

8. Что меняется в упражнениях 4 — 7, если дополнительно известен факт совместности или несовместности системы (41.1)?

§ 42. Некоторые замечания

Мы рассмотрели различные методы решения систем линейных алгебраических уравнений, основанные на разложении матрицы на множители. С решением систем и разложением матрицы тесно связаны многие другие задачи линейной алгебры. На некоторых из этих задач мы сейчас остановимся.

Вычисление определителя. Выполнение преобразований матрицы в процессе решения системы уравнений позволяет без больших дополнительных затрат получить значение определителя. Пусть имеет место (36.2) или (36.4). Для разложения (36.2)

$$\det A = \det B \cdot \det C.$$

Если матрицы B, C треугольные, то их определители равны произведению диагональных элементов. Как правило, среди матриц B, C одна имеет единичные диагональные элементы. Для разложения (36.4)

$$\det A = \frac{\det G}{\det L \cdot \det S}.$$

Определители матриц L, S не требуют каких-либо вычислений. Для преобразований вида (24.3), (24.9) и преобразований вращения они равны единице, для преобразований отражения они равны $(-1)^r$, где r — число преобразований. Матрица G чаще всего принадлежит к одному из типов, описанных в § 26, и ее определитель находится без особого труда.

Как бы ни вычислялся определитель, мы можем утверждать лишь то, что полученное его значение будет совпадать с точным значением определителя некоторой возмущенной матрицы. Рассмотрим матрицы A, $A+E$ и пусть $\rho_1, \ldots, \rho_n$ и $\hat{\rho}_1, \ldots, \hat{\rho}_n$ — их сингулярные числа. Легко проверить, что

$$\frac{\det (A+E)}{\det A} = \prod_{i=1}^{n} \frac{\hat{\rho}_i}{\rho_i} = 1 + \sum_{i=1}^{n} \frac{\hat{\rho}-\rho_i}{\rho_i}.$$

Если обозначить

$$\frac{\det (A+E)}{\det A} = 1 + \theta,$$

то, воспользовавшись неравенством Коши — Буняковского и соотношением (15.11), будем иметь

$$|\theta| \leqslant \|E\|_E \|A^{-1}\|_{E'}. \tag{42.1}$$

Для возмущений, связанных с разложениями из табл. 34.1, оценку (42.1) можно записать в таком виде:

$$|\theta| \leqslant f(n)\, \nu_A p^{t+1}.$$

Здесь ν_A есть евклидово число обусловленности матрицы A.

Обращение матрицы. Разложения (36.2), (36.4) можно использовать для вычисления обратной матрицы. Из (36.2) вытекает, что

$$A^{-1} = C^{-1} B^{-1}, \tag{42.2}$$

а из (36.4) имеем

$$A^{-1} = S G^{-1} L. \tag{42.3}$$

Поэтому, если выполнено преобразование (36.2) или (36.4), то для получения матрицы A^{-1} остается лишь обратить одну или две матрицы простого вида и осуществить перемножение матриц согласно (42.2) или (42.3).

С формальной точки зрения для обращения матрицы можно использовать любое из разложений, описанных в табл. 34.1. Однако в практическом отношении не все они равноценны. Основное различие между ними связано с объемом требуемой памяти ЭВМ. Чтобы вычислить обратную матрицу по формуле (42.2) или (42.3), необходимо после выполнения преобразования матрицы A запомнить все матрицы из (36.2) или (36.4), кроме самой A. Как видно из табл. 34.1, уже на этом этапе некоторые из разложений требуют значительной дополнительной памяти. Разложения, в которых матрица G двухдиагональная, трехдиагональная или почти треугольная, требуют большой дополнительной памяти и на этапе вычисления G^{-1}, так как матрица G^{-1} будет полной. Поэтому в действительности из всех разложений в табл. 34.1 для обращения матрицы наиболее удобны лишь те, которые связаны с ее разложением на треугольные множители или ее приведением к треугольному виду с помощью преобразований отражения.

К задаче вычисления обратной матрицы можно подойти несколько иначе. Матрица A^{-1} является единственным решением матричного уравнения $AX = E$.

Обозначим через $x_1, \ldots, x_n$ векторы-столбцы матрицы A^{-1}. Тогда x_i является решением системы линейных алгебраических уравнений

$$Ax_i = e_i, \tag{42.4}$$

где e_i — координатный вектор с единицей на i-м месте. Снова для решения системы (42.4) оказываются полезными разложения (36.2), (36.4).

С практической точки зрения почти безразлично, вычислять ли обратную матрицу по формулам (42.2), (42.3) или с помощью решения систем (42.4). Мы отдадим

предпочтение второму способу только потому, что все вопросы, связанные с решением систем, уже исследованы. При реализации этого способа может потребоваться некоторое изменение вычислительной схемы методов, вызванное необходимостью одновременного решения систем (42.4) с многими правыми частями.

Рассмотрим два наиболее интересных примера. Пусть получено разложение (36.2), где матрица B левая треугольная, а C — правая треугольная. При последовательном решении первых систем из (36.3) с правыми частями $e_1, \ldots, e_n$ информация о самих решениях может быть размещена в ЭВМ на том же месте, где находится матрица B. Дополнительно требуется не более n слов памяти. Вторые системы из (36.3) будем решать параллельно, определяя одну за другой координаты с одинаковыми номерами для всех систем. В этом случае элементы матрицы A^{-1} могут быть получены в ЭВМ на месте матрицы A примерно за $2n^3$ арифметических операций.

Предположим, что выполнено преобразование (36.4), причем матрица G — правая треугольная, а матрица S представлена в виде произведения $U_1 \ldots U_{n-2} U_{n-1}$ матриц отражения. Последовательное решение систем (36.6) с правыми частями $e_n, \ldots, e_1$ снова позволяет разместить всю информацию о решениях на месте соответствующих столбцов матрицы G. Преобразования (36.5) будем осуществлять параллельно, умножая сначала все векторы на U_{n-1}, затем на U_{n-2} и наконец на U_1. При этом следует учитывать как специальный вид преобразуемых векторов, так и специальный вид самих преобразований. Элементы матрицы A^{-1} опять могут быть получены на месте матрицы A примерно за $3n^3$ арифметических операций.

Если соблюдается режим вычислений, при котором была получена оценка (36.15), то нахождение обратной матрицы с помощью любого из рассмотренных в § 36 численных методов решения систем (42.4) гарантирует выполнение оценки (36.15) для каждого столбца реально вычисленной матрицы $\bar{A}^{-1}$. Поэтому аналогичная оценка справедлива и для самой матрицы $\bar{A}^{-1}$. Именно,

$$\frac{\|\bar{A}^{-1} - A^{-1}\|_E}{\|A^{-1}\|_E} \lesssim 2\nu_A f(n)\, p^{-t+1}. \qquad (42.5)$$

Здесь ν_A есть евклидово число обусловленности матрицы A.

Применение систем (42.4) для вычисления матрицы A^{-1} позволяет в случае необходимости уточнить отдельные или все ее столбцы. Техника выполнения этого процесса детально описана в § 38.

Вычисление псевдообратной матрицы. Нормальное псевдорешение x_0 системы (41.1) связано с матрицей и правой частью соотношением

$$x_0 = A^+ b.$$

Отсюда сразу вытекает, что если мы будем находить x_i из системы (42.4) как ее нормальное псевдорешение, то получим ни что иное, как i-й вектор-столбец псевдообратной матрицы A^+. Поэтому вычисление псевдообратной матрицы лишь деталями отличается от рассмотренной выше задачи обращения матрицы.

Определение псевдообратной матрицы всегда сводится к случаю $m \leqslant n$, ибо при $m \geqslant n$ можно воспользоваться формулой

$$(A')^+ = (A^+)'.$$

Если матрица A не очень близка к матрице неполного ранга, то для решения систем (42.4) наиболее целесообразно применить первый из трех способов, описанных в § 37. В этом случае реально вычисленная матрица $\bar{A}^+$ будет удовлетворять согласно (37.9) соотношению

$$\frac{\|\bar{A}^+ - A^+\|_E}{\|A^+\|_E} \lesssim 9{,}8 \min\{m, n\}\, \nu_A p^{-t+1}. \qquad (42.6)$$

Близость матрицы A к матрице неполного ранга значительно усложняет задачу вычисления A^+. Снова необходима дополнительная информация, а для решения систем (42.4) приходится применять методы, предназначенные для неустойчивых систем линейных алгебраических уравнений.

УПРАЖНЕНИЯ

1. Как влияют на значение определителя перестановки, выполненные в процессе преобразования матрицы?
2. Рассмотреть преобразования матрицы, позволяющие вычислить все ее главные миноры.
3. Удобно ли использовать для вычисления определителя разложение, в котором есть полная унитарная матрица?

4. Влияет ли очередность решения систем (42.4) при выполненном разложении (36.2) или (36.4) на объем дополнительной памяти?

5. Доказать, что для любой невырожденной матрицы A и любой матрицы X того же порядка выполняется соотношение

$$\frac{\|X-A^{-1}\|}{\|A^{-1}\|} \leqslant \min\{\|AX-E\|, \|XA-E\|\}.$$

6. Рассмотрим невырожденную матрицу A и любую матрицу X, для которой выполняется условие $\|AX-E\|<1$. Обозначим $X_0=X$ и построим последовательность матриц $\{X_k\}$ согласно предписанию

$$X_k = X_{k-1}(2E-AX_{k-1}).$$

Доказать, что последовательность $\{X_k\}$ сходится к матрице A^{-1}; при этом

$$\frac{\|X_k-A^{-1}\|}{\|A^{-1}\|} \leqslant \|AX-E\|^{2^k}.$$

7. Рассмотреть аналог упражнения 6 в случае выполнения условия $\|XA-E\|<1$.

8. Сравнить формулы (36.15), (37.12) и (42.5), (42.6). Почему в первом случае имеется большое различие, а во втором — большое сходство?

9. Исследовать различные численные методы вычисления псевдообратной матрицы, основанные на разложениях матрицы на множители.

ГЛАВА VI

РЕШЕНИЕ ПРОБЛЕМЫ СОБСТВЕННЫХ ЗНАЧЕНИЙ

Вычисление собственных значений и собственных векторов матрицы является одной из самых сложных задач линейной алгебры. Численные методы для решения проблемы собственных значений должны быть итерационными, так как в конечном счете они связаны с определением корней алгебраического многочлена.

В этих методах собственные значения вычисляются как пределы некоторых числовых последовательностей без предварительного определения коэффициентов характеристического многочлена. Как правило, одновременно находятся и собственные векторы или другие векторы, связанные с ними простыми соотношениями.

Мы рассмотрим некоторые из численных методов решения проблемы собственных значений. Все они эффективны, но достаточно трудоемки. Их развитие и применение в практике вычислений стало возможно лишь после создания быстродействующих вычислительных машин.

§ 43. Метод вращений

Рассмотрим вещественную симметричную матрицу A порядка n. Определение ее собственных значений и собственных векторов равносильно отысканию такой ортогональной матрицы T, для которой

$$\Lambda = T'AT \tag{43.1}$$

есть диагональная матрица. В этом случае столбцы T будут собственными векторами матрицы A, а диагональные элементы Λ — ее собственными значениями.

Среди всех ортогональных преобразований подобия преобразование (43.1) минимизирует сумму квадратов внедиагональных элементов. Поэтому попытаемся находить

4. Влияет ли очередность решения систем (42.4) при выполненном разложении (36.2) или (36.4) на объем дополнительной памяти?

5. Доказать, что для любой невырожденной матрицы A и любой матрицы X того же порядка выполняется соотношение

$$\frac{\|X-A^{-1}\|}{\|A^{-1}\|} \leqslant \min\{\|AX-E\|, \|XA-E\|\}.$$

6. Рассмотрим невырожденную матрицу A и любую матрицу X, для которой выполняется условие $\|AX-E\|<1$. Обозначим $X_0=X$ и построим последовательность матриц $\{X_k\}$ согласно предписанию

$$X_k = X_{k-1}(2E-AX_{k-1}).$$

Доказать, что последовательность $\{X_k\}$ сходится к матрице A^{-1}; при этом

$$\frac{\|X_k-A^{-1}\|}{\|A^{-1}\|} \leqslant \|AX-E\|^{2^k}.$$

7. Рассмотреть аналог упражнения 6 в случае выполнения условия $\|XA-E\|<1$.

8. Сравнить формулы (36.15), (37.12) и (42.5), (42.6). Почему в первом случае имеется большое различие, а во втором — большое сходство?

9. Исследовать различные численные методы вычисления псевдообратной матрицы, основанные на разложениях матрицы на множители.

ГЛАВА VI

РЕШЕНИЕ ПРОБЛЕМЫ СОБСТВЕННЫХ ЗНАЧЕНИЙ

Вычисление собственных значений и собственных векторов матрицы является одной из самых сложных задач линейной алгебры. Численные методы для решения проблемы собственных значений должны быть итерационными, так как в конечном счете они связаны с определением корней алгебраического многочлена.

В этих методах собственные значения вычисляются как пределы некоторых числовых последовательностей без предварительного определения коэффициентов характеристического многочлена. Как правило, одновременно находятся и собственные векторы или другие векторы, связанные с ними простыми соотношениями.

Мы рассмотрим некоторые из численных методов решения проблемы собственных значений. Все они эффективны, но достаточно трудоемки. Их развитие и применение в практике вычислений стало возможно лишь после создания быстродействующих вычислительных машин.

§ 43. Метод вращений

Рассмотрим вещественную симметричную матрицу A порядка n. Определение ее собственных значений и собственных векторов равносильно отысканию такой ортогональной матрицы T, для которой

$$\Lambda = T'AT \tag{43.1}$$

есть диагональная матрица. В этом случае столбцы T будут собственными векторами матрицы A, а диагональные элементы Λ — ее собственными значениями.

Среди всех ортогональных преобразований подобия преобразование (43.1) минимизирует сумму квадратов внедиагональных элементов. Поэтому попытаемся находить

матрицу T на основе какого-либо процесса минимизации данной суммы. Будем строить последовательность матриц

$$A_0 = A,\ A_1,\ \ldots,\ A_\nu,\ \ldots, \tag{43.2}$$

каждая из которых получается из предыдущей с помощью выполнения преобразования подобия, содержащего лишь одну матрицу вращения.

Для упрощения записи опустим индекс ν и исследуем результат преобразования

$$\hat{A} = T_{ij}' A T_{ij}. \tag{43.3}$$

Обозначим через a_{kl}, $\hat{a}_{kl}$ элементы матриц A, $\hat{A}$ и пусть, в соответствии с (18.22), угол поворота матрицы вращения T_{ij} есть α. Матрица $\hat{A}$ отличается от A двумя строками и двумя столбцами с номерами i, j. Если к тому же принять во внимание инвариантность евклидовой нормы к унитарным преобразованиям, то из (43.3) легко получить соотношение

$$\sum_{k\neq l} |\hat{a}_{kl}|^2 = \sum_{k\neq l} a_{kl}^2 - 2a_{ij}^2 + \frac{1}{2}\{(a_{jj}-a_{ii})\sin 2\alpha + 2a_{ij}\cos 2\alpha\}^2, \tag{43.4}$$

связывающее суммы квадратов внедиагональных элементов матриц $\hat{A}$ и A.

Соотношение (43.4) означает, что для максимального уменьшения суммы квадратов внедиагональных элементов необходимо матрицу вращения T_{ij} выбрать так, чтобы выполнялись два условия:

$$|a_{ij}| = \max_{k\neq l} |a_{kl}|$$

и

$$(a_{jj}-a_{ii})\sin 2\alpha + 2a_{ij}\cos 2\alpha = 0.$$

Второе из условий дает

$$\operatorname{tg} 2\alpha = \frac{2a_{ij}}{a_{ii}-a_{jj}}, \quad |\alpha| \leqslant \frac{\pi}{4}. \tag{43.5}$$

Ясно, что после выполнения преобразования (43.3) внедиагональные элементы матрицы $\hat{A}$, находящиеся в позициях (i, j) и (j, i), будут равны нулю.

Пусть t_ν^2 есть сумма квадратов внедиагональных элементов $a_{kl}^{(\nu)}$ матрицы A_ν из последовательности (43.2). В силу формулы (43.4) и выбора угла поворота

$$t_{\nu+1}^2 = t_\nu^2 - 2\left(a_{i_\nu j_\nu}^{(\nu)}\right)^2.$$

Если при каждом преобразовании вращения исключать максимальный по модулю внедиагональный элемент, то

$$\left(a_{i_\nu, j_\nu}^{(\nu)}\right)^2 \geqslant \frac{t_\nu^2}{n(n-1)},$$

и далее будем иметь

$$t_{\nu+1}^2 \leqslant t_\nu^2\left(1-\frac{2}{n(n-1)}\right) \leqslant \ldots \leqslant t_0^2\left(1-\frac{2}{n(n-1)}\right)^{\nu+1}. \tag{43.6}$$

Следовательно,

$$\lim_{\nu\to\infty} t_\nu^2 = 0. \tag{43.7}$$

Согласно (43.3) любая матрица A_ν из последовательности (43.2) связана с исходной матрицей A ортогональным подобным преобразованием

$$A_\nu = (T_{i_0 j_0} T_{i_1 j_1} \ldots T_{i_{\nu-1}, j_{\nu-1}})' A (T_{i_0 j_0} T_{i_1 j_1} \ldots T_{i_{\nu-1}, j_{\nu-1}}).$$

Поэтому диагональные элементы A_ν и соответствующие столбцы матрицы

$$T_\nu = T_{i_0 j_0} T_{i_1 j_1} \ldots T_{i_{\nu-1}, j_{\nu-1}} \tag{43.8}$$

являются точными собственными значениями и точными собственными векторами некоторой симметричной матрицы $A + E_\nu$, где

$$\|E_\nu\|_E = t_\nu$$

для всех ν. Теперь предельное равенство (43.7) и результаты теории возмущений из § 13 позволяют сделать важный вывод. Именно, для всех достаточно больших ν диагональные элементы матриц A_ν будут близки с любой заданной точностью к собственным значениям матрицы A, а столбцы матриц T_ν — к ее собственным векторам. Этот способ решения полной проблемы вещественной симметричной матрицы и называется *методом вращений*.

Может показаться, что неравенства (43.6) свидетельствуют о весьма слабых свойствах построенного метода

в отношении скорости его сходимости. Однако в действительности они не совсем правильно отражают существо процесса. Докажем, что *независимо от наличия кратных собственных значений данный метод асимптотически обладает квадратичной сходимостью.*

Предположим, что процесс проведен настолько далеко, что все внедиагональные элементы матрицы A_v стали величинами порядка ε и малыми по сравнению с ненулевыми разностями собственных значений матрицы A. Представим матрицу A_v в виде

$$A_v = \Lambda + (A_v - \Lambda).$$

Обозначим $A_v - \Lambda$ через Ω и воспользуемся результатами теории возмущений из § 13. Матрица A_v имеет собственные значения, совпадающие с диагональными элементами матрицы Λ. Поэтому в обозначениях § 13 из формулы (13.8) вытекает, что все элементы матриц Ω_{kk} являются величинами порядка ε^2. В терминах построенного метода это означает следующее. Во-первых, диагональные элементы матрицы A_v приближают собственные значения матрицы A с точностью порядка ε^2 независимо от их кратности. Кроме этого, если диагональные элементы $a_{ii}^{(v)}$ и $a_{jj}^{(v)}$ близки к одному и тому же собственному значению, то внедиагональные элементы $a_{ij}^{(v)}$ и $a_{ji}^{(v)}$ на самом деле являются величинами порядка ε^2.

Для максимального по модулю внедиагонального элемента матрицы A_v соответствующие диагональные элементы не могут быть близкими. Следовательно, угол поворота, вычисляемый согласно (43.5), будет порядка ε. Но тогда

$$\cos\alpha = 1 + O(\varepsilon^2), \quad \sin\alpha = O(\varepsilon).$$

Теперь легко показать, что при исключении максимального внедиагонального элемента все остальные меняющиеся внедиагональные элементы изменятся на величины порядка ε^2. В частности, элемент, исключенный на предыдущем шаге, будет порядка ε^2. Поэтому не более чем через $\frac{n(n-1)}{2}$ шагов все внедиагональные элементы станут величинами порядка ε^2, что и доказывает квадратичную сходимость.

Реализация описанного варианта метода вращений требует выбора максимального по модулю внедиагонального элемента матрицы на каждом шаге. При выполнении этой операции на ЭВМ требуется значительная затрата машинного времени. Поэтому необходимость указанного выбора является существенным недостатком метода с точки зрения удобства его реализации на ЭВМ.

Более удобными оказываются *циклические* процессы и, в частности, циклические процессы с *барьерами*. При циклическом процессе выбирается определенная нумерация внедиагональных элементов матрицы и их исключение происходит по циклам. В течение каждого цикла исключаются по очереди все внедиагональные элементы в порядке их нумерации. Чаще всего элементы нумеруются подряд по строкам слева направо и сверху вниз или по столбцам сверху вниз и слева направо. При этом, конечно, нумеруются только наддиагональные или поддиагональные элементы.

Недостатком такого процесса является то, что приходится исключать малые внедиагональные элементы, в то время когда в матрице еще присутствуют большие. Это обстоятельство значительно уменьшает скорость работы.

Отмеченный недостаток частично устраняется введением барьеров. Вводится монотонно убывающая к нулю последовательность положительных чисел $\alpha_1, \alpha_2, \ldots$, называемых барьерами, и при циклическом просмотре исключаются лишь те из внедиагональных элементов, которые по модулю не меньше α_1. После того как все внедиагональные элементы станут по модулю меньше α_1, барьер α_1 заменяется на α_2 и процесс продолжается.

Этот процесс позволяет решать полную проблему собственных значений значительно быстрее, чем процесс с выбором максимального элемента. Однако практическое его использование встречает ряд трудностей, связанных с оптимальным выбором барьеров. Если барьер выбрать очень большим, то будет затрачено много времени на просмотр малых элементов. Если же его выбрать очень малым, то будет затрачено много времени на исключение малых элементов, которые, по существу, не влияют на скорость сходимости.

Особого внимания заслуживает следующий способ выбора элемента, подлежащего исключению. Если в матрице A_v исключается элемент, стоящий в позиции (i_v, j_v), то суммы квадратов внедиагональных элементов в каждой

строке матрицы $A_{\nu+1}$ будут такие же, как у матрицы A_ν, кроме строк с номерами i_ν, j_ν. Поэтому если в начале процесса вычислить суммы квадратов внедиагональных элементов строк матрицы A, то в дальнейшем в полученной последовательности $\{\beta_p\}$ из n чисел при каждом элементарном шаге будут меняться лишь два числа.

Этот факт позволяет находить *оптимальный* для исключения элемент путем просмотра всего лишь $2n-1$ чисел. Определяется он следующим образом. Сначала в последовательности $\{\beta_p\}$ находим максимальный элемент β_{i_ν}, а затем в i_ν-й строке отыскиваем максимальный по модулю элемент a_{i_ν, j_ν}. Очевидно, что он будет близок к наибольшему по модулю и во всяком случае не меньше, чем среднее квадратичное всех внедиагональных элементов. Для того чтобы подготовить последовательность $\{\beta_p\}$ к следующему шагу, необходимо пересчитать числа β_{i_ν} и β_{j_ν}. Вся теория метода вращений с выбором максимального элемента полностью переносится на процесс с выбором оптимального элемента.

Мы уже неоднократно рассматривали различные преобразования, основанные на использовании матриц вращения, и всегда устанавливали их устойчивость к ошибкам округления. Поэтому можно с уверенностью утверждать, что описанные процессы также должны обладать соответствующей устойчивостью.

Обратим внимание на некоторые детали, связанные с вычислением самих матриц вращения. Обозначим

$$x = 2a_{ij}, \quad y = a_{ii} - a_{jj}.$$

Если $y = 0$, то берем

$$\cos\alpha = \sin\alpha = \sqrt{2}/2.$$

В случае $y \neq 0$ из (43.5) получаем, что

$$\cos\alpha = \left(\frac{1}{2}\left(1 + \frac{|y|}{(x^2+y^2)^{1/2}}\right)\right)^{1/2},$$
$$\sin\alpha = \operatorname{sign}(xy)\left(\frac{1}{2}\left(1 - \frac{|y|}{(x^2+y^2)^{1/2}}\right)\right)^{1/2}. \tag{43.9}$$

Первая из этих формул вполне пригодна для вычислений, вторая же будет давать большие относительные ошибки,

когда $|x|$ мал по сравнению с $|y|$. Но заметим, что

$$\sin\alpha \cdot \cos\alpha = \frac{1}{2}\operatorname{sign}(xy)\frac{|x|}{(x^2+y^2)^{1/2}}.$$

Следовательно,

$$\sin\alpha = \frac{\operatorname{sign}(xy)\,|x|}{2\cos\alpha\,(x^2+y^2)^{1/2}}. \tag{43.10}$$

Теперь первая из формул (43.9) совместно с (43.10) позволяет вычислить все элементы матрицы вращения с высокой относительной точностью. Если выражения

$$\frac{|x|}{(x^2+y^2)^{1/2}}, \quad \frac{|y|}{(x^2+y^2)^{1/2}}$$

находить в соответствии с алгорифмом, описанным в § 18, то опорные элементы реально полученной матрицы вращения будут иметь вид (18.4). При этом

$$(\tilde{c}^2+\tilde{s}^2)^{1/2} = 1+\nu,$$

где

$$|\nu| \leqslant \frac{15}{8}\,p^{t+1}, \tag{43.11}$$

если умножение и деление на 2 осуществляются неточно, и

$$|\nu| \leqslant \frac{3}{2}\,p^{t+1}$$

в противном случае. Проверку правильности этих соотношений мы предлагаем провести читателю в качестве упражнений.

Влияние ошибок округления приведет к тому, что вместо матриц A_ν, T_ν из (43.2), (43.8) реально будут вычислены некоторые другие матрицы $\tilde{A}_\nu$, $\tilde{T}_\nu$, которые связаны с матрицей A соотношением $\tilde{A}_\nu = \tilde{T}_\nu^{\mathrm{T}}(A+\Delta_\nu)\tilde{T}_\nu$. Величина эквивалентного возмущения Δ_ν в общем случае будет зависеть от порядка исключения внедиагональных элементов.

Предположим, что элементы исключаются в циклическом порядке с использованием барьеров или без них. Будем считать также, что вычисляется лишь половина всех внедиагональных элементов A_ν, а остальным элементам приписываются принудительные значения из сооб-

ражений симметрии. Если матрица A_ν получена на r-м цикле, то, используя оценку (43.11) и результаты исследований в §§ 19, 23, 32, заключаем, что

$$\|\Delta_\nu\|_E \lesssim 16rnp^{-t+1}\|A\|_E. \qquad (43.12)$$

Опыт практического применения метода вращений показывает, что независимо от порядка матрицы обычно требуется выполнить не более 5—6 полных циклов для максимального уменьшения суммы квадратов внедиагональных элементов. Однако ошибки округления оказывают основное влияние лишь на первых 2—3 циклах.

Для процессов с выбором максимального или оптимального элемента не удается получить оценку лучше, чем (43.12), или хотя бы сравнимую с ней. Как правило, вместо n появляется множитель n^2. Но, по-видимому, это связано лишь с трудностью получения хорошей оценки, а не с тем, что такие процессы менее точны, чем циклические.

В заключение приведем одно полезное следствие из (43.12). Обозначим через $\lambda_1, \ldots, \lambda_n$ точные собственные значения, через $\bar\lambda_1, \ldots, \bar\lambda_n$ — реально вычисленные. Если в формуле (43.12) взять $r=3$, то из (13.9) вытекает, что

$$\left(\frac{\sum_{i=1}^{n}(\lambda_i-\bar\lambda_i)^2}{\sum_{i=1}^{n}\lambda_i^2}\right)^{1/2} \lesssim 48np^{-t+1}.$$

УПРАЖНЕНИЯ

1. Доказать, что последовательность (43.2) сходится к фиксированной матрице.
2. Доказать, что при циклическом исключении внедиагональных элементов с использованием барьеров метод вращений сходится.
3. При любом ли выборе барьеров метод вращений будет сходиться асимптотически с квадратичной скоростью?
4. Пусть матрица A имеет кратные собственные значения. Доказать, что в матрицах последовательности (43.2) существуют фиксированные места, на которых, начиная с некоторого ν, не будут находиться ни максимальный, ни оптимальный элементы.
5. Доказать, что при выборе максимального или оптимального элемента последовательность матриц (43.8) сходится к фиксированной матрице.

6. Будет ли последовательность (43.8) сходиться к фиксированной матрице, если используется циклический вариант метода вращений без барьеров?
7. Исследовать асимптотическую скорость сходимости метода вращений с выбором максимального или оптимального элемента в зависимости от числа кратных собственных значений и их кратности.
8. Доказать, что для метода вращений сохраняется не только сходимость, но и ее асимптотическая квадратичность, если углы поворота α брать в пределах $|\alpha| \leqslant \pi/4$ в соответствии с формулами

$$\operatorname{tg}\alpha = \begin{cases} 1, & |a_{ij}| > |a_{ii}-a_{jj}|, \\ a_{ij}/(a_{ii}-a_{jj}), & |a_{ij}| \leqslant |a_{ii}-a_{jj}|, \end{cases}$$

или

$$\operatorname{tg}\frac{\alpha}{2} = \begin{cases} \sqrt{2}-1, & |a_{ij}| > (2\sqrt{2}-2)|a_{ii}-a_{jj}|, \\ a_{ij}/2(a_{ii}-a_{jj}), & |a_{ij}| \leqslant (2\sqrt{2}-2)|a_{ii}-a_{jj}|. \end{cases}$$

В чем достоинство этих формул?
9. Показать, что метод вращений распространяется на комплексные эрмитовы матрицы.
10. Рассмотрим комплексную нормальную матрицу A порядка n с элементами

$$a_{kl} = \begin{cases} -\frac{n-2}{2}e^{i(\gamma_k+\gamma_l)}, & l=k, \\ e^{i(\gamma_k+\gamma_l)}, & l\neq k, \end{cases}$$

где $\gamma_p = (p-1)\pi/n$ для $p=1, 2, \ldots, n$. Доказать, что при $n \geqslant 6$ никакое подобное преобразование этой матрицы с помощью комплексного аналога матрицы вращения не позволяет уменьшить сумму квадратов модулей внедиагональных элементов.

§ 44. Метод бисекций

Пусть A — вещественная симметричная матрица. Предположим, что для невырожденной матрицы T матрица

$$\Lambda = T'AT \qquad (44.1)$$

является диагональной. Тогда, в соответствии с законом инерции квадратичных форм [1], можно утверждать, что число нулевых, положительных и отрицательных диагональных элементов Λ не зависит от способа приведения матрицы A из (44.1) к диагональному виду, т. е. не зависит от матрицы T.

Возьмем в качестве T ортогональную матрицу собственных векторов из (43.1). В этом случае матрица Λ в (44.1) будет матрицей собственных значений. Следова-

тельно, если мы для какой-нибудь другой матрицы T сможем подсчитать число нулевых, положительных и отрицательных элементов соответствующей матрицы Λ, то тем самым будет определено число нулевых, положительных и отрицательных собственных значений матрицы A. Эту задачу можно решить весьма эффективно, даже не вычисляя явно матрицы T и Λ.

Предположим пока, что матрица A имеет ненулевые главные миноры. Тогда существует невырожденная правая треугольная матрица S и диагональная матрица D с элементами ± 1, для которых справедливо равенство

$$A = S'DS.$$

Согласно сказанному выше число положительных и отрицательных элементов D равно соответственно числу положительных и отрицательных собственных значений матрицы A. Но элементы матрицы D легко определить. Действительно, используя формулу Бине — Коши [1], находим, что для всех r

$$A\begin{bmatrix}1 & 2 & \dots & r\\ 1 & 2 & \dots & r\end{bmatrix} = \prod_{i=1}^{r} s_{ii}^2 d_{ii},$$

поэтому

$$d_{11} = \operatorname{sign} A\begin{bmatrix}1\\1\end{bmatrix}, \quad d_{rr} = \operatorname{sign} \frac{A\begin{bmatrix}1 & 2 & \dots & r\\ 1 & 2 & \dots & r\end{bmatrix}}{A\begin{bmatrix}1 & 2 & \dots & r-1\\ 1 & 2 & \dots & r-1\end{bmatrix}}.$$

Итак, знаки главных миноров симметричной матрицы позволяют установить число ее положительных и отрицательных собственных значений.

Знаки главных миноров матрицы $A - \lambda E$ при любом вещественном λ определяют число собственных значений матрицы A соответственно больших λ и меньших λ. Беря различные значения λ, можно найти число собственных значений, лежащих на произвольном отрезке вещественной оси и, следовательно, в нужной мере локализовать любое собственное значение матрицы A. На этой идее основан рассматриваемый ниже численный метод нахождения собственных значений симметричной матрицы, называемый *методом бисекций*.

Будем считать, что симметричная матрица A имеет трехдиагональную форму с ненулевыми внедиагональными

элементами. Такие матрицы называются *матрицами Якоби* и имеют вид

$$A = \begin{bmatrix} \alpha_1 & \beta_1 & & & & 0\\ \beta_1 & \alpha_2 & \beta_2 & & & \\ & \beta_2 & \alpha_3 & \beta_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \beta_{n-2} & \alpha_{n-1} & \beta_{n-1}\\ 0 & & & & \beta_{n-1} & \alpha_n \end{bmatrix}. \tag{44.2}$$

В рассмотрении матриц Якоби нет особого ограничения. В самом деле, мы показали в § 32, что симметричную матрицу можно привести к трехдиагональной форме с помощью ортогонального подобного преобразования. Если окажется, что некоторые из внедиагональных элементов равны нулю, то трехдиагональная матрица распадется в прямую сумму диагональных матриц и трехдиагональных матриц с ненулевыми внедиагональными элементами. Для диагональных матриц решение проблемы собственных значений очевидно, поэтому остается решить эту проблему для матриц вида (44.2).

Обозначим через $\sigma_1, \dots, \sigma_n$ главные миноры трехдиагональной матрицы A. Аналогично формуле (26.6) находим

$$\begin{aligned} &\sigma_0 = 1, \quad \sigma_1 = \alpha_1,\\ &\sigma_r = \alpha_r\sigma_{r-1} - \beta_{r-1}^2\sigma_{r-2}, \qquad 2 \leqslant r \leqslant n. \end{aligned} \tag{44.3}$$

Из этих соотношений и вида матрицы A вытекает ряд полезных следствий. Например,

никакие два соседних главных минора матрицы (44.2) *не могут одновременно равняться нулю;*

если минор σ_r, $1 < r < n$, *равен нулю, то соседние миноры* σ_{r-1}, σ_{r+1}, *отличны от нуля и имеют противоположные знаки;*

все собственные значения матрицы (44.2) *простые.*

Некоторые затруднения может вызвать лишь доказательство последнего утверждения. Предположим, что собственное значение λ является кратным. Тогда ранг матрицы $A - \lambda E$ должен быть не больше $n-2$. Но он заведомо не меньше $n-1$, так как внедиагональные элементы отличны от нуля и, следовательно, отличен от нуля минор, расположенный в первых $n-1$ столбцах и последних $n-1$ строках. Полученное противоречие означает, что все собственные значения матрицы Якоби простые.

возмущение E', соответствующее выполненной замене, очень мало́. Именно,

$$\|E'\|_1 \leqslant 4p\omega. \tag{44.6}$$

Обозначим через Ω максимальное положительное число, представимое на ЭВМ. Обычно Ω лишь незначительно меньше числа $(p\omega)^{-1}$. Выберем какое-либо число M, близкое к Ω, например,

$$M = (4p\omega)^{-1}.$$

Это число потребуется для дальнейших вычислений.

Рассмотрим теперь типичный шаг вычислительного процесса. Он заключается в нахождении величин

$$u = \gamma(\alpha x - \beta^2 y), \quad v = \gamma x.$$

Здесь α, β удовлетворяют условиям (44.5), числа x, y не равны нулю одновременно, параметр γ выбирается по ходу вычислений.

Предположим сначала, что $x = 0$. В этом случае не возникают никакие ошибки, если взять $\gamma = (\beta^2 |y|)^{-1}$ и положить

$$u = -\operatorname{sing} y, \quad v = 0.$$

Если же $x \neq 0$, то процесс вычислений будем осуществлять в такой последовательности:

$$\begin{aligned}
&s = \mathrm{fl}(\beta y) \equiv \beta y (1+\varepsilon_1),\\
&z = \mathrm{fl}(\beta s) \equiv \beta^2 y (1+\varepsilon_1)(1+\varepsilon_2),\\
&\theta = \max\{|x|, |z|\},\\
&v = \mathrm{fl}\left(\frac{Mx}{\theta}\right) \equiv \frac{Mx}{\theta}(1+\varepsilon_3),\\
&q = \mathrm{fl}(\alpha v) \equiv \alpha v (1+\varepsilon_4),\\
&r = \mathrm{fl}(M\beta) \equiv M\beta(1+\varepsilon_5),\\
&m = \mathrm{fl}(r\beta) \equiv M\beta^2(1+\varepsilon_5)(1+\varepsilon_6),\\
&l = \mathrm{fl}\left(\frac{mv}{\theta}\right) \equiv \frac{mv}{\theta}(1+\varepsilon_7),\\
&\tilde{u} = \mathrm{fl}(q-l) \equiv (q-l)(1+\varepsilon_8).
\end{aligned} \tag{44.7}$$

Ни одна из величин слева не превосходит по модулю Ω. Чтобы на промежуточных этапах не возникало переполнение или неоправданное появление машинного нуля, требуется лишь аккуратно вычислить θ и l. Это можно сделать, например, таким способом:

$$\theta = \begin{cases} \mathrm{fl}\left(\left(\frac{M}{\theta}\right)x\right), & \theta \geqslant 1,\\ \mathrm{fl}\left(\left(\frac{x}{\theta}\right)M\right), & \theta < 1, \end{cases}$$

$$l = \begin{cases} \mathrm{fl}\left(\left(\frac{m}{\theta}\right)y\right), & m<1,\ \theta<1 \text{ или } m\geqslant 1,\ \theta\geqslant 1,\\ \mathrm{fl}\left(\left(\frac{y}{\theta}\right)m\right), & m<1,\ \theta\geqslant 1 \text{ или } m\geqslant 1,\ \theta<1. \end{cases}$$

Исследуем более детально возникающие ошибки. Заметим, что ε_1 и ε_2 влияют незначительно лишь на выбор θ и не влияют на остальные вычисления в (44.7). Поэтому будем рассматривать ошибки ε_i только для $i \geqslant 3$.

Пусть $\varepsilon_3 = -1$. Тогда, в соответствии с этим предположением, должно выполняться асимптотическое неравенство

$$\left|\frac{Mx}{\theta}\right| \lesssim \omega. \tag{44.8}$$

Но (44.8) не имеет места, когда $\theta = |x|$. Если же $\theta = |z|$, то будут справедливы такие соотношения:

$$\left|\frac{Mx}{\theta}\right| = \left|\frac{Mx}{\beta^2 y}\right| \geqslant \frac{(4p\omega)^{-1}\omega}{(1+4)^2(p\omega)^{-1}} = 4\omega.$$

И снова (44.8) не имеет места. Следовательно, $\varepsilon_3 \neq -1$.

Ошибки ε_5, ε_6 заведомо не равны -1 в силу того, что $M\beta^2 \geqslant p\omega$. Не равна -1 и ошибка ε_8. Действительно, если $\theta = |x|$, то $q = 0$ при $\alpha = 0$ и $q \geqslant (4p)^{-1}$ при $\alpha \neq 0$. В обоих случаях $\varepsilon_8 \neq -1$ независимо от величины l. Если же $\theta = |z|$, то $|l| \cong M$ и $\varepsilon_8 \neq -1$ при любых q.

Если все $\varepsilon_i \neq -1$, то, принимая во внимание описанный процесс вычислений и величины ошибок округления от выполнения отдельных арифметических операций, заключаем, что

$$|\varepsilon_i| \lesssim \begin{cases} \frac{1}{2}p^{t+1}, & i \neq 3,7,\\ p^{t+1}, & i = 3,7. \end{cases} \tag{44.9}$$

Однако подчеркнем, что ε_6 и ε_7 могут оказаться равными -1.

возмущение E', соответствующее выполненной замене, очень мало́. Именно,

$$\|E'\|_1 \leqslant 4p\omega. \tag{44.6}$$

Обозначим через Ω максимальное положительное число, представимое на ЭВМ. Обычно Ω лишь незначительно меньше числа $(p\omega)^{-1}$. Выберем какое-либо число M, близкое к Ω, например,

$$M = (4p\omega)^{-1}.$$

Это число потребуется для дальнейших вычислений.

Рассмотрим теперь типичный шаг вычислительного процесса. Он заключается в нахождении величин

$$u = \gamma(\alpha x - \beta^2 y), \quad v = \gamma x.$$

Здесь α, β удовлетворяют условиям (44.5), числа x, y не равны нулю одновременно, параметр γ выбирается по ходу вычислений.

Предположим сначала, что $x = 0$. В этом случае не возникают никакие ошибки, если взять $\gamma = (\beta^2 |y|)^{-1}$ и положить

$$u = -\operatorname{sing} y, \quad v = 0.$$

Если же $x \neq 0$, то процесс вычислений будем осуществлять в такой последовательности:

$$\begin{aligned}
&s = \mathrm{fl}(\beta y) \equiv \beta y (1+\varepsilon_1),\\
&z = \mathrm{fl}(\beta s) \equiv \beta^2 y (1+\varepsilon_1)(1+\varepsilon_2),\\
&\theta = \max\{|x|, |z|\},\\
&v = \mathrm{fl}\left(\frac{Mx}{\theta}\right) \equiv \frac{Mx}{\theta}(1+\varepsilon_3),\\
&q = \mathrm{fl}(\alpha v) \equiv \alpha v (1+\varepsilon_4),\\
&r = \mathrm{fl}(M\beta) \equiv M\beta (1+\varepsilon_5),\\
&m = \mathrm{fl}(r\beta) \equiv M\beta^2 (1+\varepsilon_5)(1+\varepsilon_6),\\
&l = \mathrm{fl}\left(\frac{my}{\theta}\right) \equiv \frac{my}{\theta}(1+\varepsilon_7),\\
&u = \mathrm{fl}(q-l) \equiv (q-l)(1+\varepsilon_8).
\end{aligned} \tag{44.7}$$

Ни одна из величин слева не превосходит по модулю Ω. Чтобы на промежуточных этапах не возникало переполнение или неоправданное появление машинного нуля, требуется лишь аккуратно вычислить v и l. Это можно сделать, например, таким способом:

$$v = \begin{cases} \mathrm{fl}\left(\left(\frac{M}{\theta}\right)x\right), & \theta \geqslant 1,\\ \mathrm{fl}\left(\left(\frac{x}{\theta}\right)M\right), & \theta < 1, \end{cases}$$

$$l = \begin{cases} \mathrm{fl}\left(\left(\frac{m}{\theta}\right)y\right), & m < 1,\ \theta < 1 \text{ или } m \geqslant 1,\ \theta \geqslant 1,\\ \mathrm{fl}\left(\left(\frac{y}{\theta}\right)m\right), & m < 1,\ \theta \geqslant 1 \text{ или } m \geqslant 1,\ \theta < 1. \end{cases}$$

Исследуем более детально возникающие ошибки. Заметим, что ε_1 и ε_2 влияют незначительно лишь на выбор θ и не влияют на остальные вычисления в (44.7). Поэтому будем рассматривать ошибки ε_i только для $i \geqslant 3$.

Пусть $\varepsilon_3 = -1$. Тогда, в соответствии с этим предположением, должно выполняться асимптотическое неравенство

$$\left|\frac{Mx}{\theta}\right| \gtrsim \omega. \tag{44.8}$$

Но (44.8) не имеет места, когда $\theta = |x|$. Если же $\theta = |z|$, то будут справедливы такие соотношения:

$$\left|\frac{Mx}{\theta}\right| = \left|\frac{Mx}{\beta^2 y}\right| \geqslant \frac{(4p\omega)^{-1}\omega}{(1.4)^2 (p\omega)^{-1}} = 4\omega.$$

И снова (44.8) не имеет места. Следовательно, $\varepsilon_3 \neq -1$.

Ошибки ε_5, ε_6 заведомо не равны -1 в силу того, что $M\beta^2 \geqslant p\omega$. Не равна -1 и ошибка ε_8. Действительно, если $\theta = |x|$, то $q = 0$ при $\alpha = 0$ и $q \geqslant (4p)^{-1}$ при $\alpha \neq 0$. В обоих случаях $\varepsilon_8 \neq -1$ независимо от величины l. Если же $\theta = |z|$, то $|l| \gtrsim M$ и $\varepsilon_8 \neq -1$ при любых q.

Если все $\varepsilon_i \neq -1$, то, принимая во внимание описанный процесс вычислений и величины ошибок округления от выполнения отдельных арифметических операций, заключаем, что

$$|\varepsilon_i| \lesssim \begin{cases} \frac{1}{2}\, p^{t+1}, & i \neq 3,7,\\ p^{t+1}, & i = 3,7. \end{cases} \tag{44.9}$$

Однако подчеркнем, что ε_4 и ε_7 могут оказаться равными -1.

Согласно общей идее обратного анализа ошибок покажем теперь, что реально вычисленные величины $\tilde{u}, \tilde{v}$ удовлетворяют точным равенствам

$$\tilde{u} = \gamma((\alpha+\varepsilon)x - (\beta+\eta)^2 y), \quad \tilde{v} = \gamma x$$

и дадим оценки для эквивалентных возмущений ε, η.

Объединяя последовательно результаты вычислений в (44.7), мы получим независимо от величины ошибок ε_i следующие выражения для $\tilde{u}$, $\tilde{v}$:

$$\tilde{u} = \frac{M}{\theta}(1+\varepsilon_9)\left(\alpha x(1+\varepsilon_4)(1+\varepsilon_8) - \beta^2 y \times \right.$$
$$\left. \times \frac{(1+\varepsilon_5)(1+\varepsilon_6)(1+\varepsilon_7)(1+\varepsilon_8)}{1+\varepsilon_9}\right), \quad (44.10)$$
$$\tilde{v} = \left(\frac{M}{\theta}(1+\varepsilon_9)\right)x.$$

Это означает, что

$$\gamma = \frac{M}{\theta}(1+\varepsilon_9).$$

Так как $\varepsilon_9 \neq -1$, то $\gamma > 0$.

При оценке эквивалентных возмущений ε, η нам придется рассмотреть несколько случаев. Предположим сначала, что все ошибки $\varepsilon_i \neq -1$. Тогда из (44.9), (44.10) вытекает, что

$$|\varepsilon| \lesssim |\alpha| p^{-t+1}, \quad |\eta| \lesssim \frac{7}{4}|\beta| p^{-t+1}. \quad (44.11)$$

Если $\theta = |x|$, то при всех α ошибка $\varepsilon_4 \neq -1$. Допустим, что $\varepsilon_7 = -1$, т. е. выполняется асимптотическое неравенство

$$|M\beta^2 y/\theta| \lesssim \omega\omega. \quad (44.12)$$

Представив $\tilde{u}$ в следующем виде:

$$\tilde{u} = \frac{M}{\theta}(1+\varepsilon_9)\left(\left(\alpha(1+\varepsilon_4)(1+\varepsilon_8) + \frac{\beta^2 y}{x}\right)x - \beta^2 y\right)$$

и воспользовавшись неравенством (44.12), находим, что

$$|\beta^2 y/x| \lesssim \omega^2/4M p.$$

Поэтому для эквивалентных возмущений получаем такие оценки:

$$|\varepsilon| \lesssim |\alpha| p^{-t+1}, \quad |\eta| = 0.$$

Пусть $\theta = |z|$. В этом случае $|l| \cong M$ и $\varepsilon_7 \neq -1$. Если ошибка $\varepsilon_4 = -1$, то должно выполняться асимптотическое неравенство

$$|\omega M x/\theta| \lesssim \omega,$$

откуда следует, что

$$|\alpha x| \lesssim |\beta^2 y|\omega^2/4p. \quad (44.13)$$

Записав $\tilde{u}$ в виде

$$\tilde{u} = \frac{M}{\theta}(1+\varepsilon_9)\left(\alpha x - \left(\beta^2 y \frac{(1+\varepsilon_5)(1+\varepsilon_6)(1+\varepsilon_7)(1+\varepsilon_8)}{1+\varepsilon_9} - \alpha x\right)\right),$$

и приняв во внимание соотношение (44.13), заключаем, что теперь

$$|\varepsilon| = 0, \quad |\eta| \lesssim \frac{7}{4}|\beta| p^{-t+1}.$$

Окончательное сравнение полученных результатов показывает, что для эквивалентных возмущений ε, η всегда гарантируется выполнение оценок (44.11).

Таким образом, вычисляя по описанному алгоритму знаки главных миноров якобиевой матрицы A, нормированной согласно (44.5), мы в действительности получим следующее. Реально вычисленная последовательность знаков будет точно совпадать с последовательностью знаков главных миноров некоторой возмущенной матрицы $A + E''$. Если A имеет вид (44.2), то возмущение E'' будет симметричной трехдиагональной матрицей, имеющей аналогичный вид:

$$E'' = \begin{bmatrix} \varepsilon_1'' & \eta_1'' & & & \\ \eta_1'' & \varepsilon_2'' & \eta_2'' & & 0 \\ & \eta_2'' & \varepsilon_3'' & \eta_3'' & \\ & & \ddots & \ddots & \ddots \\ & 0 & & \eta_{n-2}'' & \varepsilon_{n-1}'' & \eta_{n-1}'' \\ & & & & \eta_{n-1}'' & \varepsilon_n'' \end{bmatrix}.$$

При этом элементы ε_i'', η_i'' удовлетворяют неравенствам

$$|\varepsilon_i''| \lesssim |\alpha_i| p^{-t+1}, \quad |\eta_i''| \lesssim \frac{7}{4}|\beta_i| p^{-t+1}. \quad (44.14)$$

Отсюда, в частности, вытекает, что возмущенная матрица будет также якобиевой и, следовательно, по знакам ее главных миноров можно правильно определить число нулевых, положительных и отрицательных собственных

значений матрицы $A+E''$. Собственные значения матрицы A асимптотически отличаются от собственных значений матрицы $A+E''$ не более, чем на $(15/8)\,p^{-t+1}$, так как в соответствии с (44.5), (44.14) имеем

$$\|E''\|_1 \leqslant \frac{15}{8}\,p^{-t+1}. \qquad (44.15)$$

Теперь мы можем перейти непосредственно к описанию численного метода отыскания собственных значений матрицы Якоби. Будем считать, что матрица A нормирована и ее коэффициенты для всех i удовлетворяют неравенствам

$$|\alpha_i| \leqslant 1/4, \quad 2p\omega < |\beta_i| \leqslant 1/4. \qquad (44.16)$$

Как уже отмечалось, основной операцией метода является вычисление знаков главных миноров матриц $A-\lambda E$ для различных λ. Но все собственные значения матрицы A с элементами (44.16) не превосходят по модулю $3/4$. Поэтому достаточно брать λ из сегмента $[-3/4, +3/4]$. В этом случае коэффициенты матриц $A-\lambda E$ будут удовлетворять неравенствам (44.5).

Вычисляя знаки главных миноров матрицы $A-\lambda E$, мы в действительности правильно определим лишь знаки главных миноров некоторой матрицы $A-\lambda E+E_\lambda$. Возмущение E_λ складывается из возмущений (44.6), (44.15) и возмущения E''', возникающего при вычислении матрицы $A-\lambda E$. Очевидно, что матрица E''' является диагональной, причем

$$\|E'''\|_1 \leqslant (1/2)\,p^{-t+1}.$$

Следовательно, для суммарного возмущения E_λ получаем оценку

$$\|E_\lambda\|_1 \leqslant (19/8)p^{-t+1} \qquad (44.17)$$

для всех значений λ.

Предположим, что собственные значения λ матрицы A занумерованы в порядке алгебраического убывания, т. е.

$$\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n.$$

Покажем, как определить k-е по номеру собственное значение λ_k независимо от остальных.

Обозначим через $n_+(\lambda)$ число собственных значений матрицы A строго больших, чем λ. Если известны такие числа a_0, b_0, что

$$b_0 > a_0, \quad n_+(a_0) \geqslant k, \quad n_+(b_0) < k,$$

то λ_k заведомо принадлежит полуинтервалу $(a_0, b_0]$. Заметим, что в качестве a_0 можно взять любое число, меньшее $-3/4$, в качестве b_0 — любое число, большее $+3/4$. Положим теперь

$$c_0 = \frac{1}{2}(a_0 + b_0)$$

и определим $n_+(c_0)$. Если $n_+(c_0) \geqslant k$, то λ_k принадлежит полуинтервалу $(c_0, b_0]$, если же $n_+(c_0) < k$, то λ_k принадлежит полуинтервалу $(a_0, c_0]$. Поэтому всегда можно указать полуинтервал длины $(1/2)(b_0-a_0)$, содержащий λ_k. Продолжая этот процесс, мы получим систему вложенных полуинтервалов $(a_s, b_s]$, содержащих λ_k, причем

$$(b_s - a_s) = 2^{-s}(b_0 - a_0).$$

Это позволяет локализовать собственное значение λ_k с любой нужной точностью.

Сделанный вывод справедлив лишь в случае точных вычислений. Ошибки округления, конечно, вносят свои коррективы. Рассмотрим реально построенный полуинтервал $(a_s, b_s]$. Строго говоря, можно утверждать лишь то, что k-е собственное значение $\tilde{\lambda}_k$ некоторой матрицы $A+E_{a_s}$ больше a_s, а k-е собственное значение $\tilde{\lambda}_k$ другой матрицы $A+E_{b_s}$ меньше или равно b_s, причем для возмущений E_{a_s}, E_{b_s} выполняются неравенства (44.17). Но в соответствии с (44.17) заключаем, что

$$|\tilde{\lambda}_k - \lambda_k| \leqslant \frac{19}{8}\,p^{-t+1}, \quad |\tilde{\lambda}_k - \lambda_k| \leqslant \frac{19}{8}\,p^{-t+1}.$$

Следовательно, должны выполняться соотношения

$$a_s - \frac{19}{8}\,p^{-t+1} < \lambda_k \leqslant b_s + \frac{19}{8}\,p^{-t+1}. \qquad (44.18)$$

Если в качестве приближения к λ_k всегда брать точку c_s, являющуюся серединой полуинтервала $(a_s, b_s]$, то из (44.18) вытекает, что

$$|c_s - \lambda_k| \leqslant \frac{19}{8}\,p^{-t+1} + \frac{b_s - a_s}{2}.$$

В частности, при указанном выше выборе начальных значений a_0, b_0

$$|c_s - \lambda_k| \lesssim \frac{19}{8} p^{-t+1} + \frac{3}{4} 2^{-s}.$$

Нет никакого смысла в выполнении очень большого числа шагов деления полуинтервалов пополам. Обычно достаточно взять $s = [t \log_2 p]$, где $[\cdot]$ означает целую часть числа. Принимая во внимание, что $p \geqslant 2$, получаем такую оценку:

$$|c_s - \lambda_k| \lesssim \frac{25}{8} p^{-t+1}. \qquad (44.19)$$

Рассмотренный метод определения собственных значений матрицы Якоби обладает исключительной универсальностью. Его можно использовать не только для нахождения заданного по номеру собственного значения, но и для вычисления всех или части собственных значений, принадлежащих любой заданной области, для исследования общего распределения собственных значений и т. п. *На его реализацию не оказывает никакого влияния наличие близких и кратных собственных значений и даже очень большое их скопление.* При этом достижимая точность (44.19) не зависит от размеров матрицы.

Все эти свойства кажутся особенно удивительными, если вспомнить, что, в конечном счете, метод связан с распознаванием нулевых и ненулевых чисел, причем распознавание осуществляется в условиях влияния ошибок округления.

УПРАЖНЕНИЯ

1. Пусть одно из собственных значений трехдиагональной симметричной матрицы имеет кратность p. Доказать, что по крайней мере $p-1$ поддиагональных элементов этой матрицы равны нулю.

2. Предположим, что трехдиагональная симметричная матрица имеет несколько нулевых внеднагональных элементов. Означает ли это, что матрица имеет кратные собственные значения?

3. Пусть A — матрица (44.2). Обозначим через A_r матрицу главного минора порядка r. Доказать, что для всех $r > 1$ матрицы A_r и A_{r-1} не имеют общих собственных значений.

4. В условиях упражнения 3 доказать, что для всех $r > 1$ между каждыми соседними собственными значениями A_r находится одно собственное значение A_{r+1} и одно собственное значение A_{r-1}.

5. Доказать, что при неограниченном увеличении (уменьшении) коэффициента α_n матрицы (44.2) все собственные значения остаются ограниченными, кроме максимального (минимального).

6. Доказать, что при неограниченном увеличении модуля коэффициента β_{n-1} матрицы (44.2) все собственные значения остаются ограниченными, кроме максимального и минимального.

7. Исследовать процессы вычисления знаков главных миноров матрицы Якоби, основанные на исключении элементов с помощью матриц вращения или элементарных неунитарных матриц.

8. Будут ли матрицы эквивалентных возмущений, возникающие при выполнении упражнения 7, симметричными и трехдиагональными?

9. Установите связь между близостью собственных значений и малостью минимального по модулю внедиагонального элемента матрицы Якоби.

10. Доказать, что при больших n максимальное собственное значение матрицы Якоби $\mathfrak{W}_{2n+1}$ порядка $2n+1$ с элементами

$$\alpha_i = n - i + 1, \quad \beta_i = 1 \qquad (44.20)$$

для всех i отличается от ближайшего собственного значения на величину порядка $(n!)^{-2}$.

11. Показать, что метод бисекций распространяется на комплексные эрмитовы матрицы.

§ 45. *QR*-алгорифм

Пусть A — произвольная вещественная матрица порядка n. Построим последовательность ортогональных матриц Q_k и правых треугольных матриц R_k по следующим рекуррентным формулам:

$$\begin{aligned} A &= Q_1 R_1, & A_1 &= R_1 Q_1, \\ A_1 &= Q_2 R_2, & A_2 &= R_2 Q_2, \\ &\cdots\cdots\cdots\cdots & & \\ A_{k-1} &= Q_k R_k, & A_k &= R_k Q_k, \\ &\cdots\cdots\cdots\cdots & & \end{aligned} \qquad (45.1)$$

Легко показать, что для всех k матрицы A_k из (45.1) подобны исходной матрице A. Действительно,

$$A_k = Q_k^{-1} (Q_k R_k) Q_k =$$
$$= Q_k^{-1} (R_{k-1} Q_{k-1}) Q_k = \ldots = (Q_1 \ldots Q_k)^{-1} A (Q_1 \ldots Q_k).$$

Обозначив $Q_1 \ldots Q_k = P_k$, заключаем, что

$$A_k = P_k^{-1} A P_k. \qquad (45.2)$$

Так как матрицы Q_k ортогональные, то будут ортогональными и матрицы P_k. Поэтому A_k ортогонально подобны A.

Соотношения (45.1), (45.2) позволяют получить еще одно следствие. Рассмотрим произведение правых треугольных матриц

$$R_k \dots R_1 = U_k;$$

имеем

$$P_k U_k = P_{k-1}(Q_k R_k) U_{k-1} = P_{k-1}(R_{k-1} Q_{k-1}) U_k =$$
$$= P_{k-1}(P_{k-1}^{-1} A P_{k-1}) U_{k-1} = A(P_{k-1} U_{k-1}) = \dots = A^k.$$

Следовательно,

$$P_k U_k = A^k,$$

т. е. для всех k матрицы P_k, U_k являются ортогональными и правыми треугольными сомножителями в соответствующих разложениях степеней матрицы A.

Исследуем теперь строение матриц A^k при больших k. Будем считать, что матрица A невырожденная. Представим ее в виде

$$A = Q \Lambda Q^{-1}, \qquad (45.3)$$

где Λ — правая каноническая матрица Жордана, и пусть $D = \{\lambda_1, \dots, \lambda_n\}$ — диагональная матрица собственных значений. Предположим, что существует разложение $Q^{-1} = LU$, где L — левая треугольная матрица с единичными диагональными элементами, U — правая треугольная матрица. Ясно, что $A^k = Q \Lambda^k Q^{-1}$, поэтому из равенств

$$A^k = P_k U_k = Q \Lambda^k Q^{-1} = (Q \Lambda^k L D^{-k}) D^k U$$

заключаем, что матрица

$$\Delta_k = P_k^{-1} Q \Lambda^k L D^{-k} = U_k U^{-1} D^{-k}$$

является правой треугольной. Далее находим

$$P_k^{-1} = \Delta_k D^k L^{-1} \Lambda^{-k} Q^{-1}, \quad P_k = Q \Lambda^k L D^{-k} \Delta_k^{-1},$$

откуда в соответствии с (45.2), (45.3) вытекает, что

$$A_k = \Delta_k D^k L^{-1} \Lambda^{-k} Q^{-1} Q \Lambda Q^{-1} Q \Lambda^k L D^{-k} \Delta_k^{-1} =$$
$$= \Delta_k \{D^k (L^{-1} \Lambda L) D^{-k}\} \Delta_k^{-1}. \qquad (45.4)$$

Не ограничивая общности, можно считать, что собственные значения матрицы A расположены на диагонали матрицы Λ в порядке убывания модулей, т. е.

$$|\lambda_1| = \dots = |\lambda_{r_1}| > |\lambda_{r_1+1}| = \dots = |\lambda_{r_2}| > \dots$$
$$\dots > |\lambda_{r_{s-1}+1}| = \dots = |\lambda_{r_s}|. \qquad (45.5)$$

Рассмотрим матрицу $B = L^{-1} \Lambda L$. Принимая во внимание вид L и Λ, нетрудно установить, что B является левой почти треугольной матрицей и имеет над главной диагональю такие же элементы, как Λ. Если b_{ij} — элементы B, то матрицы $C_k = D^k (L^{-1} \Lambda L) D^{-k}$ из (45.4) имеют элементы

$$c_{ij}^{(k)} = b_{ij} (\lambda_i / \lambda_j)^k. \qquad (45.6)$$

Разобьем матрицы C_k на клетки таким образом, чтобы диагональные клетки были квадратными и имели те же размеры, что и группы равных по модулю собственных значений в (45.5). При таком разбиении матрицы C_k будут клеточными левыми треугольными. Поэтому, если

$$C_k = \begin{bmatrix} C_{11}^{(k)} & & & \\ C_{21}^{(k)} & C_{22}^{(k)} & & \\ \dots & \dots & \dots & \\ C_{m1}^{(k)} & C_{m2}^{(k)} & \dots & C_{mm}^{(k)} \end{bmatrix},$$

то из (45.5), (45.6) следует, что при неограниченном увеличении числа k элементы диагональных клеток не меняют свои модули, а элементы поддиагональных клеток сходятся к нулю. Обозначив через τ_s величину модулей собственных значений из s-й группы в (45.5), заключаем, что элементы поддиагональных клеток $C_{ij}^{(k)}$ убывают как величины

$$\gamma_{ij}^{(k)} = O((\tau_i / \tau_j)^k), \quad i > j. \qquad (45.7)$$

Таким образом, при всех k элементы C_k остаются ограниченными, а сами матрицы C_k с ростом k приближаются по форме к клеточно-диагональной матрице. Скорость приближения определяется соотношениями (45.7).

Напомним, что матрицы Δ_k в (45.4) являются правыми треугольными. Если бы элементы Δ_k и Δ_k^{-1} оставались ограниченными при всех k, то из формулы (45.4) и вида матриц C_k сразу вытекало бы, что при неограниченном увеличении k матрицы A_k со скоростью (45.7) будут приближаться по форме к клеточной правой треугольной матрице с диагональными блоками таких же размеров, как у матриц C_k.

Матрицы Δ_k, Δ_k^{-1} заведомо ограничены, если A имеет простую структуру. В самом деле, в этом случае $\Lambda = D$ и тогда

$$\Delta_k = P_k^{-1} Q (D^k L D^{-k}), \quad \Delta_k^{-1} = (D^k L^{-1} D^{-k}) Q^{-1} P_k.$$

Матрицы P_k, P_k^{-1} являются ограниченными, так как они ортогональные, матрицы Q, Q^{-1} — постоянные. Элементы же матриц D^kLD^{-k} и $D^kL^{-1}D^{-k}$ не превосходят по модулю соответствующих элементов матриц L и L^{-1} в силу соотношений типа (45.6), связывающих эти элементы.

В общем случае исследование матриц Δ_k, Δ_k^{-1} осуществляется несколько сложнее. Запишем Δ_k, Δ_k^{-1} в следующем виде:

$$\Delta_k = P_k^{-1}Q(\Lambda^kD^{-k})(D^kLD^{-k}),$$
$$\Delta_k^{-1} = (D^kL^{-1}D^{-k})(D^k\Lambda^{-k})Q^{-1}P_k.$$

Ясно, что если элементы матриц Δ_k, Δ_k^{-1} растут, то по порядку они растут не быстрее, чем элементы матриц Λ^kD^{-k} и $D^k\Lambda^{-k}$.

Лемма 45.1. *Пусть матрица Λ невырожденная и максимальный порядок канонического ящика Жордана для нее равен s. Тогда при больших k элементы матриц Λ^kD^{-k} и $D^k\Lambda^{-k}$ суть величины $O(k^{s-1})$.*

Доказательство. Матрица Λ клеточно-диагональная и ее клетками являются канонические ящики Жордана. Поэтому утверждение леммы достаточно установить лишь для того случая, когда Λ представляет собой один ящик Жордана.

Итак, предположим, что Λ есть ящик Жордана порядка s с ненулевым собственным значением λ. Очевидно, что этому ящику будет соответствовать матрица $D = \lambda E$. Воспользуемся [2] асимптотическим представлением матрицы Λ^k

$$\Lambda^k = \lambda^k G_k^{-1}(R + e_k)G_k.$$

Здесь G_k — диагональная матрица с элементами

$$g_{ii}^{(k)} = (k/\lambda)^i, \quad i = 1, 2, \ldots, s, \tag{45.8}$$

R — правая треугольная матрица с элементами

$$r_{ij} = 1/(j-i)!, \quad j \geqslant i;$$

все элементы матрицы e_k стремятся к нулю с ростом k. Теперь получаем, что

$$\Lambda^kD^{-k} = G_k^{-1}(R + e_k)G_k,$$
$$D^k\Lambda^{-k} = G_k^{-1}(R + e_k)^{-1}G_k.$$

Согласно (45.8) максимальный элемент этих матриц при больших k находится в позиции $(1.s)$ и имеет величину $O(k^{s-1})$.

Если вернуться к оценке возможной скорости роста элементов матриц Δ_k, Δ_k^{-1}, то из доказанной леммы вытекает, что

Элементы матриц Δ_k, Δ_k^{-1} по порядку роста не превосходят k^{s-1}, если порядок жордановых ящиков матрицы Λ не превосходит s.

Построенные в соответствии с (45.1) матрицы A_k унитарно подобны марице A, поэтому элементы этих матриц ограничены равномерно по k. Представим матрицы A_k в клеточном виде

$$A_k = \begin{bmatrix} A_{11}^{(k)} & A_{12}^{(k)} & \ldots & A_{1m}^{(k)} \\ A_{21}^{(k)} & A_{22}^{(k)} & \ldots & A_{2m}^{(k)} \\ \ldots & \ldots & \ldots & \ldots \\ A_{m1}^{(k)} & A_{m2}^{(k)} & \ldots & A_{mm}^{(k)} \end{bmatrix}, \tag{45.9}$$

где диагональные клетки $A_{ii}^{(k)}$ квадратные и имеют те же размеры, что и группы равных по модулю собственных значений в (45.4). Согласно формуле (45.4), установленным свойствам матриц C_k и оценкам роста элементов матриц Δ_k, Δ_k^{-1} заключаем, что элементы поддиагональных клеток $A_{ij}^{(k)}$ сходятся к нулю как величины

$$v_{ij}^{(k)} = O(k^{2(s-1)}(\tau_i/\tau_j)^k), \tag{45.10}$$

если порядок жордановых ящиков матрицы A не превосходит s. При этом собственные значения диагональных клеток $A_{ii}^{(k)}$ сходятся к собственным значениям i-й группы в (45.5).

Для того чтобы матрицы A_k приближались к клеточной правой треугольной матрице подобным образом, достаточно, чтобы при упорядочении диагональных элементов матрицы Λ в соответствии с (45.5) были отличны от нуля главные миноры матрицы Q^{-1} в разложении (45.3).

Проведем процесс (45.1) настолько далеко, чтобы все элементы поддиагональных клеток $A_{ij}^{(k)}$ матрицы A_k стали малыми. Заменив эти элементы нулями, мы получим клеточную правую треугольную матрицу $\hat{A}_k$. Для нее реше-

ние проблемы собственных значений осуществляется значительно проще, чем для матрицы A, так как обычно группы равных по модулю собственных значений не бывают большими. В частности, если все собственные значения матрицы A различны по модулю, что матрица $\hat{A}_k$ оказывается треугольной.

Используя результаты теории возмущений, можно утверждать, что собственные значения и корневые векторы матрицы $\hat{A}_k$ могут служить хорошими приближениями для собственных значений и корневых векторов матрицы A_k. Поэтому, решив проблему собственных значений для матрицы $\hat{A}_k$, получаем в соответствии с подобным преобразованием (45.2) приближенное решение этой же проблемы для исходной матрицы A.

Численный метод решения проблемы собственных значений, основанный на построении последовательности матриц A_k согласно (45.1), называется *QR-алгорифмом*. Однако обычно под QR-алгорифмом понимают нечто большее, включая в него всю совокупность приемов ускорения

УПРАЖНЕНИЯ

1. Рассмотрим процесс (45.1), в котором матрицы R_k правые треугольные, а матрицы Q_k таковы, что нормы соответствующих матриц P_k, P_k^{-1} равномерно по k ограничены. Исследовать асимптотическое поведение матриц A_k.

2. Значительно ли сокращается объем вычислений при получении произведений R_kQ_k в (45.1) за счет треугольного вида матриц R_k?

3. В чем меняется исследование процесса (45.1) в случае вырожденности матрицы A?

4. Обозначим через $L_l(\lambda)$ многочлен, корнями которого являются все собственные значения из l-й группы в (45.5), через $L_l^{(k)}(\lambda)$ — характеристический многочлен клетки $A_{ll}^{(k)}$ матрицы A_k, представленной согласно (45.9). Используя лемму 11.2, доказать, что

$$L_1^{(k)}(\lambda)=L_1(\lambda)+O\left(k^{2(s-1)}(\tau_2/\tau_1)^k\right),$$

$$L_l^{(k)}(\lambda)=L_l(\lambda)+O\left(k^{2(s-1)}(\tau_l/\tau_{l-1})^k\right)+O\left(k^{2(s-1)}(\tau_{l+1}/\tau_l)^k\right),\quad l\neq 1,\ m,$$

$$L_m^{(k)}(\lambda)=L_m(\lambda)+O\left(k^{2(s-1)}(\tau_m/\tau_{m-1})^k\right).$$

5. Пусть для некоторого l все корни многочлена $L_l(\lambda)$ равны между собой. Доказать, что диагональные клетки $A_{ll}^{(k)}$ матриц (45.9) приближаются к треугольным. Какова скорость этого приближения?

6. Пусть матрица A нормальная и QR-алгорифм выполняется по предписанию (45.1). Доказать, что последовательность матриц A_k приближается к клеточно-диагональной матрице.

7. В обозначениях упражнения 4 и в условиях упражнения 6 доказать, что

$$L_1^{(k)}(\lambda)=L_1(\lambda)+O\left((\tau_2/\tau_1)^{2k}\right),$$

$$L_l^{(k)}(\lambda)=L_l(\lambda)+O\left((\tau_l/\tau_{l-1})^{2k}\right)+O\left((\tau_{l+1}/\tau_l)^{2k}\right),\quad l\neq 1,\ m,$$

$$L_m^{(k)}(\lambda)=L_m(\lambda)+O\left((\tau_m/\tau_{m-1})^{2k}\right).$$

8. Показать, что QR-алгорифм распространяется на комплексные матрицы.

§ 46. Ускорение QR-алгорифма

Одним из важнейших факторов ускорения QR-алгорифма является *инвариантность процесса* (45.1) *к правой почти треугольной форме матрицы*. В самом деле, пусть A правая почти треугольная. Тогда можно считать, что матрица Q_1 есть произведение $T_{12}T_{23}\dots T_{n-1,n}$ матриц вращения. Но в этом случае матрица A_1 будет снова правой почти треугольной. Конечно, такой же вид будут иметь все матрицы A_k.

Если матрица A полная и не обладает какой-либо спецификой, то реализация одного шага процесса (45.1) требует выполнения около $(10/3)n^3$ арифметических операций. Если же A правая почти треугольная, то на одном шаге нужно выполнить только $6n^2$ операций. Поэтому каждый шаг процесса (45.1) будет осуществляться примерно в $n/2$ раз быстрее, если предварительно матрицу A привести подобным преобразованием к правой почти треугольной форме. Такое преобразование было описано и исследовано в § 32.

Предположим, что элемент $a_{j+1,j}$ правой почти треугольной матрицы A равен нулю. Разобьем A на такие четыре клетки, чтобы диагональные клетки были квадратными и клетка в левом верхнем углу имела порядок j. Но тогда клетка в нижнем левом углу будет нулевой, т. е.

$$A=\begin{bmatrix}\alpha_{11} & \alpha_{12}\\ 0 & \alpha_{22}\end{bmatrix}. \tag{46.1}$$

Процесс (45.1) инвариантен в этой форме, причем на всех его шагах клетка в нижнем правом углу преобра-

вуется независимо от остальных клеток. Следовательно, нет особого смысла в применении QR-алгорифма непосредственно к таким матрицам.

Если мы интересуемся только собственными значениями матрицы A, то достаточно определить собственные значения клеток α_{11}, α_{22}. Даже при определении собственных векторов матрицы (46.1) целесообразно сначала определить собственные значения клеток α_{11}, α_{22}. Поэтому каждый раз, когда какие-нибудь из поддиагональных элементов правой почти треугольной матрицы по тем или иным причинам обратятся в нули, мы будем продолжать применение QR-алгорифма только к соответствующим матрицам меньших размеров.

Это, конечно, затрудняет использование QR-алгорифма для определения собственных векторов. Однако мы покажем в дальнейшем, что, вычислив собственные значения, можно весьма эффективно получить собственные векторы.

Всюду в дальнейшем мы будем считать, что QR-алгорифм применяется к правым почти треугольным матрицам с ненулевыми поддиагональными элементами.

Если какое-нибудь собственное значение такой матрицы является кратным, то оно должно входить только в один канонический ящик Жордана. В самом деле, пусть собственному значению λ соответствует более одного ящика Жордана матрицы A. Тогда ранг матрицы $A-\lambda E$ должен быть не больше $n-2$. Но он заведомо не меньше $n-1$, так как поддиагональные элементы отличны от нуля и, следовательно, отличен от нуля минор, расположенный в первых $n-1$ столбцах и последних $n-1$ строках. Полученное противоречие подтверждает справедливость высказанного свойства.

Из этого свойства, в частности, вытекает, что *если правая почти треугольная матрица с ненулевыми поддиагональными элементами имеет простую структуру, то все ее собственные значения различны.* Как показывает пример матрицы (44.20), мы не можем надеяться на то, что наличие очень близких собственных значений обязательно приведет к появлению очень малых поддиагональных элементов.

Рассмотрим теперь любую последовательность чисел $\nu_1, \nu_2, \ldots$. Снова построим ортогональные матрицы Q_k и

правые треугольные матрицы R_k по рекуррентным формулам

$$\begin{aligned} A-\nu_1 E &= Q_1 R_1, & A_1 &= R_1 Q_1 + \nu_1 E, \\ A_1-\nu_2 E &= Q_2 R_2, & A_2 &= R_2 Q_2 + \nu_2 E, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots & & \\ A_{k-1}-\nu_k E &= Q_k R_k, & A_k &= R_k Q_k + \nu_k E, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots & & \end{aligned} \tag{46.2}$$

Если A правая почти треугольная матрица, то аналогичный вид будут иметь все матрицы A_k. Как и раньше, легко показать, что матрицы A_k ортогонально подобны исходной матрице A; при этом

$$\begin{gathered} A_k = (Q_1 \ldots Q_k)' A (Q_1 \ldots Q_k), \\ (Q_1 \ldots Q_k)(R_k \ldots R_1) = (A-\nu_1 E)(A-\nu_2 E)\ldots(A-\nu_k E). \end{gathered} \tag{46.3}$$

Предположив, что при некоторой, вообще говоря, зависящей от k нумерации собственных значений $\lambda_1^{(k)}, \ldots, \lambda_n^{(k)}$ выполняются неравенства

$$\prod_{i=1}^{k} |\lambda_1^{(k)} - \nu_i| \geqslant \prod_{i=1}^{k} |\lambda_2^{(k)} - \nu_i| \geqslant \ldots \geqslant \prod_{i=1}^{k} |\lambda_n^{(k)} - \nu_i|, \tag{46.4}$$

получим, что поддиагональные элементы $a_{i+1,i}^{(k)}$ матрицы A_k суть величины такого порядка:

$$a_{i+1,i}^{(k)} = O\left(\left|\frac{\lambda_{i+1}^{(k)} - \nu_i}{\lambda_1^{(k)} - \nu_i}\right| \prod_{l=1}^{k} |\lambda_{i+1}^{(k)} - \nu_l| / |\lambda_1^{(k)} - \nu_l|\right).$$

Числа $\nu_1, \nu_2, \ldots$ в (46.2) называются *сдвигами*. Выбирая их подходящим образом, можно в значительной мере ускорить сходимость QR-алгорифма.

По существу, все известные стратегии выбора сдвигов состоят из двух различных этапов. На первом этапе каким-либо способом обеспечивается заметное уменьшение одного из поддиагональных элементов матрицы A_k. Теоретическое обоснование этого этапа для матриц общей структуры, как правило, отсутствует, но проводится экспериментальное подтверждение его эффективности. Малость поддиагонального элемента позволяет начать на втором этапе целенаправленный выбор сдвигов. При этом обеспечивается дальнейшее уменьшение поддиагонального элемента с существенно большей скоростью.

Пусть элемент $a^{(m)}_{j+1,\,j}$ не превосходит по модулю малого числа ε. Покажем, как выбрать теперь сдвиги, чтобы скорость убывания элементов $a^{(k)}_{j+1,\,j}$ стала не менее, чем квадратичной.

Разобьем матрицу A_m на клетки таких же размеров, как в матрице (46.1). Если обозначить это разбиение через

$$A_m = \begin{bmatrix} \alpha^{(m)}_{11} & \alpha^{(m)}_{12} \\ \alpha^{(m)}_{21} & \alpha^{(m)}_{22} \end{bmatrix},$$

то диагональные клетки $\alpha^{(m)}_{11}$, $\alpha^{(m)}_{22}$ являются правыми почти треугольными, а клетка $\alpha^{(m)}_{21}$ имеет лишь один ненулевой элемент порядка ε в верхнем правом углу.

Вычислим каким-либо способом, например, по формуле (26.5), характеристический многочлен $f_m(\lambda)$ матрицы $\alpha^{(m)}_{22}$. Принимая во внимание неравенства (46.4) при $k=m$, заключаем, что согласно лемме 11.2 величины $f_m(\lambda^{(m)}_q)$ будут иметь порядок по крайней мере ε при $q>j$ и не будут малыми при $q\leqslant j$. Обозначим через $\nu_{m+1}, \ldots, \nu_{m+n-j}$ корни многочлена $f_m(\lambda)$ и выполним дополнительно $n-j$ шагов процесса (46.2), взяв $\nu_{m+1}, \ldots, \nu_{m+n-j}$ в качестве сдвигов. Рассмотрим соответствующие неравенства (46.4) при $k=m+n-j$. Для малых ε совокупности собственных значений $\lambda^{(m)}_q$, $\lambda_q^{(m+n-j)}$ при $q>j$ будут одинаковыми.

Но в этом случае

$$a^{(m+n-j)}_{j+1,\,j} = O\left((m+n-j)^{2(s-1)} \prod_{l=1}^{m+n-j} \frac{\lambda^{(m+n-j)}_{j+1}-\nu_l}{\lambda^{(m+n-j)}_{j}-\nu_l}\right) =$$

$$= O\left(m^{2(s-1)} \prod_{l=1}^{m} \frac{\lambda^{(m+n-j)}_{j+1}-\nu_k}{\lambda^{(m+n-j)}_{j}-\nu_l} \times\right.$$

$$\left.\times\, O\left((m+n-j)/m)^{2(s-1)} \prod_{l=m+1}^{m+n-j} \frac{\lambda^{(m+n-j)}_{j+1}-\nu_l}{\lambda^{(m+n-j)}_{j}-\nu_l}\right)\right) \leqslant$$

$$\leqslant O\left(m^{2(s-1)} \prod_{l=1}^{m} \frac{\lambda^{(m)}_{j+1}-\nu_l}{\lambda^{(m)}_{j}-\nu_l}\right) \times$$

$$\times\, O\left(((m+n-j)/m)^{2(s-1)} \frac{f_m\left(\lambda^{(m+n-j)}_{j+1}\right)}{f_m\left(\lambda^{(m+n-j)}_{j}\right)}\right) =$$

$$= a^{(m)}_{j+1,\,j} O(\varepsilon) = O(\varepsilon^2).$$

Циклически повторяя описанный выбор сдвигов через каждые $n-j$ шагов процесса (46.2), мы обеспечим не менее, чем квадратичную скорость убывания поддиагональных элементов $a^{(k)}_{j+1,\,j}$. Как только элемент в позиции $(j+1,\ j)$ станет достаточно малым, его можно заменить нулем, не потеряв при этом существенно в точности, и продолжать применение QR-алгорифма к матрицам меньших размеров.

Практическая реализация процесса ускорения в соответствии с изложенной схемой не всегда оказывается эффективной. Предположим, что вещественная матрица A имеет комплексные собственные значения. В этом случае среди корней вещественного многочлена $f_m(\lambda)$ могут оказаться комплексно сопряженные пары. Но тогда некоторые из промежуточных матриц A_k будут комплексными, несмотря на то, что матрица A_{m+n-j} является вещественной. Появление комплексных матриц весьма нежелательно как с точки зрения времени счета, так и с точки зрения использования памяти ЭВМ. Поэтому покажем сейчас, как можно вычислять матрицу A_{m+n-j}, минуя получение промежуточных матриц A_k. Для этого нам потребуется

Лемма 46.1. *Пусть для матрицы A выполнены унитарно подобные преобразования*

$$C_1 = T_1^* A T_1, \quad C_2 = T_2^* A T_2,$$

причем матрицы C_1, C_2 — правые почти треугольные с ненулевыми поддиагональными элементами. Если первые столбцы матриц T_1, T_2 совпадают, то существует такая диагональная матрица S с элементами, равными по модулю единице, что

$$T_2 = T_1 S, \quad C_2 = S^* C_1 S. \tag{46.5}$$

Доказательство. По существу, нам достаточно показать, что если унитарная матрица T и правая почти треугольная матрица C с ненулевыми поддиагональными элементами связаны соотношением

$$AT = TC, \tag{46.6}$$

то при заданном первом столбце T матрицы T, C определяются в основном однозначно.

Обозначим через $t_1, \ldots, t_n$ столбцы матрицы T, через c_{ij} элементы матрицы C. Приравнивая первые столбцы левой и правой частей в (46.6), имеем

$$At_1 = c_{11}t_1 + c_{21}t_2.$$

Так как T унитарна, то

$$\begin{aligned} c_{11} &= (At_1,\ t_1),\\ c_{21} &= \alpha_{21}\,\|At_1 - c_{11}t_1\|_E,\\ t_2 &= (\bar{\alpha}_{21}/c_{21})\,(At_1 - c_{11}t_1), \end{aligned}$$

где α_{21} — произвольное комплексное число, по модулю равное единице. Приравнивая вторые столбцы в (46.6), получим

$$At_2 = c_{12}t_1 + c_{22}t_2 + c_{32}t_3,$$

откуда вытекает, что

$$\begin{aligned} c_{12} &= (At_2,\ t_1),\\ c_{22} &= (At_2,\ t_2),\\ c_{32} &= \alpha_{31}\,\|At_2 - c_{12}t_1 - c_{22}t_2\|_E,\\ t_3 &= (\bar{\alpha}_{32}/c_{32})\,(At_2 - c_{12}t_1 - c_{22}t_2). \end{aligned}$$

Здесь снова α_{32} — произвольное комплексное число, по модулю равное единице.

Продолжая процесс, устанавливаем, что все столбцы матрицы T, начиная со второго, определяются однозначно с точностью до умножения на комплексные числа, по модулю равные единице. Это и подтверждает справедливость соотношений (46.5).

Теперь перейдем к обоснованию процесса вычисления матрицы A_{m+n-j}. Ясно, что

$$A_{m+n-j} = T'_{n-j}A_mT_{n-j}, \quad T_{n\,j}L_{n-j} = f_m(A_m)$$

для некоторой ортогональной матрицы $T_{n\,j}$ и правой треугольной матрицы L_{n-j}.

Предположим, что мы каким-либо способом нашли такую ортогональную матрицу T, у которой первый столбец совпадает с первым столбцом $T_{n\,j}$ и при этом матрица $C = T'A_mT$ является правой почти треугольной. Возможны две ситуации. Если все поддиагональные элементы матрицы C отличны от нуля, то согласно лемме 46.1

$$T = T_{n-j}S, \quad C = S'A_{m+n-j}S,$$

где S — диагональная матрица с элементами, равными по модулю единице. Модули соответствующих элементов матриц $A_{m-n\,j}$ и C одинаковы, поэтому безразлично, с какой из них продолжать QR-алгорифм. Мы будем продолжать его с матрицей C. Если же какие-то поддиагональные элементы матрицы C равны нулю, то это более благоприятный случай, так как можно продолжать QR-алгорифм с матрицами меньших размеров.

Теоретически матрицу $T_{n\,j}$ можно получить как произведение матриц вращения при исключении поддиагональных элементов $f_m(A_m)$, например, по столбцам сверху вниз. Будем считать, что

$$T_{n\,1,\,n}\ldots T_{12}f_m(A_m) = L_{n-j}.$$

В силу строения матриц вращения первая строка произведения $T_{1n}\ldots T_{12}$ совпадает с первой строкой произведения $T_{n\,1,\,n}\ldots T_{12}$ и, следовательно, с первой строкой матрицы T'_{n-j}. Но матрицы $T_{1n}, \ldots, T_{12}$ определяются лишь первым столбцом $f_m(A_m)$, у которого могут быть отличны от нуля только первые $n-j+1$ элементов. Поэтому в действительности первый столбец матрицы T_{n-j} совпадает с первым столбцом произведения $T'_{12}\ldots T'_{1,\,n-j+1}$.

Вычислим первый столбец матрицы $f_m(A_m)$ и определим по нему соответствующие матрицы вращения $T_{12}, \ldots, T_{1,\,n\,j+1}$. Получим, далее, матрицу

$$B = T_{1,\,n\,j+1}\ldots T_{12}A_mT'_{12}\ldots T'_{1,\,n-j+1} \qquad (46.7)$$

и приведем ее с помощью алгорифма, описанного в § 32, к ортогонально подобной правой почти треугольной матрице. Это и будет нужная нам матрица C.

Действительно, новая почти треугольная матрица получена как произведение вида $P'BP$. Согласно построению ортогональная матрица P имеет первые столбец и строку, совпадающие с первыми столбцом и строкой единичной матрицы. Следовательно, первый столбец матрицы $T_{n\,j}$ будет совпадать с первым столбцом матрицы $T'_{12}\ldots T'_{1,\,n\,j+1}P$. Последняя матрица не только ортогональная, но и является матрицей подобного преобразования, приводящего A_m к правой почти треугольной, т. е. обладает свойствами матрицы T.

Матрица B в (46.7) имеет много нулевых элементов и отличается от правой почти треугольной тем, что почти

все элементы главного минора порядка $n-j+2$ могут быть отличны от нуля. Специальный вид матрицы B легко учесть при ее приведении к почти треугольной форме. На всех этапах приведения промежуточные матрицы будут отличаться от правой почти треугольной тем, что почти все элементы некоторого минора порядка $n-j+2$, опирающегося на главную диагональ, могут быть отличны от нуля. По мере выполнения процесса приведения этот минор будет перемещаться по диагонали вниз.

В целом прямое получение матрицы A_{m+n-j} из матрицы A_m требует выполнения примерно такого же объема вычислений, как и последовательное ее получение за $n-j$ шагов процесса (46.2) с вещественными сдвигами. При этом не нужно находить корни многочленов. Однако объем дополнительной памяти ЭВМ, необходимой для хранения результатов промежуточных вычислений, быстро растет с уменьшением j.

Описанное ускорение сходимости QR-алгорифма особенно эффективно, когда $j=n-1$. В этом случае очередной сдвиг v_{m+1} совпадает с последним диагональным элементом матрицы A_m и заведомо будет вещественным. Поэтому прямое получение матрицы A_{m+1} не является обязательным, а квадратичное убывание поддиагонального элемента будет происходить на каждом шаге процесса (46.2). Если же вещественная матрица A имеет комплексные собственные значения, прямое получение матрицы A_{m+n-1} может оказаться необходимым при условии, конечно, что мы хотим выполнять лишь вещественные вычисления. В этом случае прямое получение матрицы A_{m+n-j} наиболее эффективно при $j=n-2$. Обязательное появление малых поддиагональных элементов с меньшим значением j связано в основном с наличием у матрицы A канонических ящиков Жордана больших размеров.

Подводя итоги, перечислим еще раз рассмотренные приемы ускорения времени счета при реализации QR-алгорифма.

1. Предварительное приведение матрицы A к правой почти треугольной форме. Без этого приведения QR-алгорифм обычно не применяется.

2. Использование сдвигов для повышения скорости убывания поддиагональных элементов. При наличии комплексных собственных значений наиболее эффективно прямое получение матриц A_k из (46.2).

3. Замена малых поддиагональных элементов нулями. Это позволяет продолжать применение QR-алгорифма к матрицам меньших размеров.

В приемах ускорения остается не ясным только один вопрос. Именно, как начинать выбор сдвигов, чтобы достаточно быстро получить малый поддиагональный элемент в позиции $(j+1, j)$ с наибольшим по возможности значением j? За исключением некоторых специальных классов матриц, пока на него не получено убедительного ответа. Стратегии выбора сдвигов, гарантирующие убывание поддиагональных элементов, обычно оказываются слишком медленными.

Мы опишем сейчас одну из практических процедур выбора сдвигов. Она эффективна, хотя ее применение связано с некоторым риском.

Процесс начинается с вычисления матрицы A_1 при $v_1=0$. Предположим, что уже получены матрицы A_{m-1} и A_m. Проверяем выполнение неравенства

$$|a_{nn}^{(m)}-a_{nn}^{(m-1)}|\leqslant\frac{1}{3}|a_{nn}^{(m-1)}|.$$

Если оно справедливо, то находим матрицу A_{m+1}, беря $v_{m+1}=a_{nn}^{(m)}$. Если же это неравенство не имеет места, то проверяем выполнение другого неравенства:

$$\|\alpha^{(m)}-\alpha^{(m-1)}\|\leqslant\frac{1}{3}\|\alpha^{(m-1)}\|,$$

где $\alpha^{(m-1)}$, $\alpha^{(m)}$ — клетки второго порядка матриц A_{m-1}, A_m, находящиеся в нижнем правом углу. Если это неравенство справедливо, то вычисляем характеристический многочлен матрицы $\alpha^{(m)}$ и находим матрицу A_{m+2}, используя прямой способ ее получения. Если же и последнее неравенство не имеет места, то находим матрицу A_{m+1}, беря $v_{m+1}=0$. Конечно, при первой же возможности заменяем малые поддиагональные элементы нулями и продолжаем применение QR-алгорифма к матрицам меньших размеров.

Применение этой процедуры показало, что среднее число итераций на каждое собственное значение, как правило, не превосходит 5. Процесс исключительно устойчив. Анализ ошибок округления в нем полностью охватывается тем анализом, который был выполнен ранее. Если счи-

тать, что на каждое собственное значение матрицы A требуется не более пяти итераций, то вычисленные собственные значения будут точными для некоторой возмущенной матрицы $A+E$, причем

$$\|E\|_E \lesssim \frac{20\sqrt{2}+25}{2} n^2 p^{-t+1} \|A\|_E.$$

УПРАЖНЕНИЯ

1. Пусть A — эрмитова ленточная матрица. Доказать, что все матрицы A_k из (46.2), полученные при вещественных сдвигах, будут эрмитовыми и ленточными такой же ширины.

2. Уточнить лемму 46.1, если дополнительно известно, что поддиагональные элементы матриц C_1, C_2 положительны.

3. Рассмотреть применение преобразований отражения при прямом получении матрицы A_{m+n-j} из матрицы A_m. Выгодно ли их применение, если $n-j \geqslant 2$?

4. Пусть матрица A нормальная и выполняется QR-алгорифм со сдвигами. Доказать, что при описанном выше способе выбора сдвигов обеспечивается кубическая скорость убывания поддиагональных элементов.

5. Пусть задана некоторая совокупность из k сдвигов, не являющихся собственными значениями матрицы A. Предположим, что в процессе (46.2) для двух различных порядков выбора сдвигов из этой совокупности получены матрицы A'_k и A''_k. Доказать, что при точных вычислениях $A'_k = S_k^* A''_k S_k$, где S_k — диагональная матрица с элементами, равными по модулю единице.

6. Пусть вместо матриц A'_k, A''_k из упражнения 5 реально получены матрицы $\bar{A}'_k$, $\bar{A}''_k$. Доказать, что в общем случае не существует такой функции $f(k, n)$, не зависящей от A, при которой для всех A выполняется неравенство

$$\min_{\tilde{S}_k} \|\bar{A}'_k - \tilde{S}_k^* \bar{A}''_k \tilde{S}_k\| \lesssim f(k, n)\, p^{-t+1} \|A\|.$$

Здесь $\tilde{S}_k$ — диагональные матрицы с элементами, равными по модулю единице.

7. В каком смысле матрицы $\bar{A}'_k$, $\bar{A}''_k$ из упражнения 6 можно назвать близкими?

§ 47. Определение собственных векторов

Рассмотренные численные методы решения проблемы собственных значений лучше приспособлены для определения собственных значений, чем собственных векторов. Так, например, применение метода вращений или QR-алгорифма для нахождения собственных векторов заставляет дополнительно вычислять и запоминать матрицу результирующего преобразования подобия. Эта операция оказывается очень невыгодной, если нужно определить лишь несколько векторов. К тому же вычисление матрицы преобразования усложняет численный метод. Особенно усложняется QR-алгорифм, так как теперь в нем гораздо труднее осуществлять переход к матрицам меньших размеров при появлении нулевых поддиагональных элементов. Применение метода бисекций вообще не дает никакой явной информации относительно собственных векторов.

Все эти причины побуждают нас более внимательно рассмотреть задачу определения собственных векторов по предварительно вычисленным собственным значениям.

Предположим, что находятся собственные векторы матрицы A. Возьмем произвольный вектор u_0 и построим последовательность векторов

$$u_k = \alpha_k A u_{k-1}, \tag{47.1}$$

для $k \geqslant 1$ при некоторых ненулевых числах α_k. Ясно, что

$$u_k = \beta_k A^k u_0, \qquad \beta_k = \prod_{l=1}^{k} \alpha_l.$$

Разложим матрицу A в произведение (45.3) и будем считать, что собственные значения на диагонали матрицы Λ упорядочены согласно (45.5). Тогда имеем

$$u_k = \beta_k Q (\Lambda^k D^{-k}) D^k Q^{-1} u_0. \tag{47.2}$$

Здесь D — диагональная матрица собственных значений.

Пусть в соответствии с (45.5) L есть подпространство, натянутое на первые r_1 векторов-столбцов матрицы Q, $L^\perp$ — его ортогональное дополнение. Представим векторы u_k в виде

$$u_k = u_k^{(L)} + u_k^{(L^\perp)},$$

где $u_k^{(L)} \in L$, $u_k^{(L^\perp)} \in L^\perp$. Если $u_0^{(L)} \neq 0$, то из (47.2) следует, что $u_k^{(L)} \neq 0$ для всех k. Воспользовавшись результатами и обозначениями § 45, заключаем далее из (47.2), что

$$\|u_k^{(L^\perp)}\|_E / \|u_k^{(L)}\|_E = O(k^{s-1} (\tau_2/\tau_1)^k). \tag{47.3}$$

Рассмотрим наиболее важные случаи распределения максимальных по модулю собственных значений матрицы A.

1. Все максимальные по модулю собственные значения совпадают и ни один из них не входит в канонический ящик Жордана выше первого порядка.

2. Все максимальные по модулю собственные значения совпадают и некоторые из них входят в канонические ящики Жордана порядка не выше s.

3. Максимальные по модулю собственные значения вещественной матрицы образуют простую комплексно сопряженную пару.

Из формулы (47.2) получаем, что в первом случае все проекции $u_k^{(L)}$ коллинеарны. Следовательно, с точностью до нормирующего множителя последовательность векторов u_k сходится со скоростью (47.3) к одному из собственных векторов, соответствующих максимальному по модулю собственному значению. Этот собственный вектор определяется только проекцией $u_0^{(L)}$. Меняя вектор u_0, можно находить различные собственные векторы.

Во втором случае проекции $u_k^{(L)}$ не будут оставаться коллинеарными, и их положение в L изменяется в соответствии с изменением максимальных элементов матрицы $\Lambda^k D^k$. Однако при каждом k вектор u_k с точностью (47.3) будет близок к некоторому вектору, принадлежащему корневому подпространству, соответствующему максимальному по модулю собственному значению. Вообще говоря, будет иметь место и сходимость к одному из собственных векторов. Однако большого практического значения она не имеет, так как ее скорость всего лишь порядка k^{-1}.

Существенно по другому ведет себя последовательность векторов u_k в третьем случае. Ее характерные особенности можно видеть на следующем примере. Пусть в качестве A взята матрица вращения (18.2). Нетрудно проверить, что

$$A^k = \begin{bmatrix} \cos k\alpha & -\sin k\alpha \\ \sin k\alpha & \cos k\alpha \end{bmatrix},$$

и тогда

$$u_k = \beta_k \| u_0 \|_E \begin{bmatrix} \cos(k\alpha + \theta) \\ \sin(k\alpha + \theta) \end{bmatrix}$$

для некоторого числа θ. Это означает, что при больших k координаты векторов u_k будут осциллировать. Тем не

менее вектор u_k с точностью (47.3) будет близок к некоторому вектору, являющемуся линейной комбинацией комплексно сопряженных собственных векторов, соответствующих максимальным по модулю комплексно сопряженным собственным значениям.

Процесс (47.1) принято называть *прямыми итерациями*. Он применяется в основном для определения корневого базиса, соответствующего максимальным по модулю собственным значениям. Используя сдвиги, можно несколько увеличить скорость сходимости. Значительного же ускорения нельзя получить, так как невозможно с помощью сдвигов сделать достаточно малым отношение τ_2/τ_1.

Прямые итерации можно использовать и для определения корневых векторов, соответствующих минимальным по модулю собственным значениям, если матрицу A в (47.1) заменить матрицей A^{-1}.

Обозначим через L подпространство, натянутое на последние столбцы матрицы Q, соответствующие согласно (45.5) минимальным по модулю собственным значениям. Возьмем произвольный вектор u_0 и построим последовательность векторов

$$u_k = \alpha_k A^{-1} u_{k-1}. \tag{47.4}$$

Собственные значения матрицы A^{-1} являются обратными величинами по отношению к собственным значениям матрицы A. Поэтому в соответствии с изложенным выше близость векторов u_k к корневому подпространству L в процессе (47.4) определяется соотношением

$$\frac{\| u_k^{(L-1)} \|_E}{\| u_k^{(L)} \|_E} = O\left(k^{s-1}\left(\frac{\tau_m}{\tau_{m-1}}\right)^k\right).$$

С точки зрения скорости сходимости положение изменилось принципиально, так как теперь с помощью сдвигов можно сделать отношение τ_m/τ_{m-1} достаточно малым.

При построении векторов u_k из (47.4), как правило, не вычисляют матрицу A^{-1}, а находят эти векторы путем решения систем линейных алгебраических уравнений

$$A u_k = \alpha_k u_{k-1}.$$

Такой процесс принято называть *обратными итерациями*. Именно обратные итерации являются одним из самых эффективных численных методов определения корневых

векторов матрицы по предварительно вычисленным ее собственным значениям.

Предположим, что для собственного значения λ матрицы A известно достаточно точное приближение $\tilde{\lambda}$. Построим последовательность векторов

$$(A-\tilde{\lambda}E)\,u_k=\alpha_k u_{k-1}, \tag{47.5}$$

исходя из некоторого вектора u_0. Пусть L означает корневое подпространство матрицы A, соответствующее λ. Если

$$\varepsilon=|\tilde{\lambda}-\lambda|, \quad a=\min_{\lambda_i\neq\lambda}|\tilde{\lambda}-\lambda_i|,$$

то, очевидно, что при $a>\varepsilon$

$$\|u_k^{(L^\perp)}\|_E/\|u_k^{(L)}\|_E=O\left(k^{n-1}(\varepsilon/a)^k\right).$$

На первый взгляд кажется, что влияние ошибок округления в реальных вычислениях должно существенно изменить свойства обратных итераций в процессе (47.5). В самом деле, при реализации этого процесса приходится решать системы уравнений. Конечно, мы можем надеяться на то, что реально вычисленные векторы $\tilde{u}_k$ будут удовлетворять соотношениям вида

$$(A-\tilde{\lambda}E+E_k)\tilde{u}_k=\tilde{\alpha}_k(\tilde{u}_{k-1}+\eta_{k-1}), \tag{47.6}$$

где евклидовы нормы E_k, η_k ограничены малыми величинами E, η. Если $\tilde{\lambda}$ близко к λ, матрицы систем (47.6) будут плохо обусловленными. Поэтому вектор $\tilde{u}_k$ из (47.6) будет значительно отличаться от вектора $\tilde{u}_k$ из (47.5). Но тогда вроде бы нельзя ожидать получения из последовательности векторов $\tilde{u}_k$ надежной информации о корневом подпространстве, соответствующем λ.

В целом правильные аргументы привёли нас к неправильному выводу. Если решение системы (47.5) содержит большую ошибку, то вектор ошибок будет в основном принадлежать именно тому подпространству, которое мы пытаемся определить. Чем больше ошибка в вычисленном векторе, тем с большей точностью этот вектор принадлежит нужному подпространству.

Ошибки округления не могут существенно изменить общую скорость процесса (47.5) и влияют лишь на величину достижимой точности. Теперь

$$\overline{\lim_{k\to\infty}}\left(\|\tilde{u}_k^{(L^\perp)}\|_E/\|\tilde{u}_k^{(L)}\|_E\right)=q, \tag{47.7}$$

где q, вообще говоря, отлично от нуля. Покажем, как оценить главный член числа q.

Согласно теореме Шура [1] существует унитарная матрица R, для которой матрица $C=RAR^*$ является правой треугольной, причем первыми на диагонали C стоят собственные значения, равные λ. Сделав замену

$$z_k=R\tilde{u}_k, \tag{47.8}$$

получим вместо (47.6) соотношение

$$(C-\tilde{\lambda}E+T_k)\,z_k=\tilde{\alpha}_k(z_{k-1}+\xi_{k-1}), \tag{47.9}$$

где

$$T_k=RE_kR^*, \quad \xi_{k-1}=R\eta_{k-1}.$$

Предположим, что кратность собственного значения λ равна r. В соответствии с результатами теории возмущений существует малая матрица вида

$$H_k=\begin{bmatrix}0 & H_{12}^{(k)}\\ -H_{12}^{(k)*} & 0\end{bmatrix}$$

такая, что

$$B_k=(E+H_k)(C-\tilde{\lambda}E+T_k)(E+H_k)^{-1}$$

является клеточной правой треугольной. Если

$$B_k=\begin{bmatrix}B_{11}^{(k)} & B_{12}^{(k)}\\ 0 & B_{22}^{(k)}\end{bmatrix},$$

то клетки матриц H_k, B_k в верхнем левом углу имеют порядок r. Собственные значения $B_{11}^{(k)}$ не превосходят по модулю малого числа $\tilde{e}$, зависящего от степени близости $\tilde{\lambda}$ к λ, величины E и, конечно, от структуры матрицы A. Модули собственных значений $B_{22}^{(k)}$ ограничены снизу некоторым числом, близким к a. Сделав очередную замену

$$v_k=(E+H_k)^{-1}z_k, \tag{47.10}$$

получим вместо (47.9) новое соотношение

$$B_k v_k = \tilde{\alpha}_k (v_{k-1} + \theta_{k-1}), \tag{47.11}$$

где

$$\theta_{k-1} = (E + H_k)\,\xi_{k-1} + (H_k - H_{k-1})\,z_{k-1}.$$

Пусть евклидовы нормы θ_k, $H_{ik}^{(k)}$ ограничены малыми величинами θ, H. Для любого вектора ω через ω', ω'' обозначим векторы, составленные из первых r и последних $n-r$ координат ω. Теперь из (47.11) вытекает, что в общем случае

$$\overline{\lim_{k\to\infty}} (\| v_k'' \|_E / \| v_k' \|_E) = O(\varepsilon\theta).$$

Но тогда из (47.10) следует

$$\overline{\lim_{k\to\infty}} (\| z_k'' \|_E / \| z_k' \|_E) \lessapprox H. \tag{47.12}$$

Согласно (47.8) справедливы равенства

$$\| \tilde{u}_k^{(L-1)} \|_E = \| z_k' \|_E, \quad \| \tilde{u}_k^{(L)} \|_E = \| z_k' \|_E,$$

поэтому из (47.7), (47.12) находим, что $q \leqslant H$.

Величина H зависит от внутренней структуры матрицы A. Если A имеет базис из собственных векторов, то в соответствии с (13.10) имеем

$$H \lessapprox (\nu_Q / 2a)\,E,$$

где ν_Q есть спектральное число обусловленности матрицы собственных векторов Q из (45.3). Для нормальных матриц можно считать $\nu_Q = 1$. Следовательно, в этом случае

$$q \lessapprox E/2a. \tag{47.13}$$

УПРАЖНЕНИЯ

1. Построить алгорифмы определения максимальных по модулю собственных значений матрицы A на основе процесса (47.1).

2. Построить алгорифмы уточнения собственных значений матрицы A на основе процесса (47.5).

3. Доказать, что для матрицы $A - \tilde{\lambda}E$ из (45.7) выполняется неравенство

$$\| (A - \tilde{\lambda}E)^{-1} \| \geqslant \varepsilon^{-1}.$$

4. Пусть для матрицы A максимальный размер жорданова ящика с собственным значением λ равен s. Доказать, что

$$\| (A - \tilde{\lambda}E)^{-1} \| = O(\varepsilon^{-s}).$$

5. Пусть в процессе (47.5) параметр α_k выбирается из условия $\| u_k \| = 1$. Доказать, что в общем случае

$$\overline{\lim_{k\to\infty}} |\alpha_k| = O(\varepsilon^{-1}),$$

$$\sup_k |\alpha_k| = O(\varepsilon^{-s}).$$

6. Доказать, что в условиях упражнений 4, 5 при $s > 1$ вектор u_k будет близок к собственному вектору, соответствующему λ, с точностью порядка εk^{-1}.

7. Исследовать поведение норм невязок $(A - \tilde{\lambda}E)\,u_k$ в процессе (47.5).

§ 48. Особенности вычислений

Перейдем к рассмотрению вычислительного процесса решения систем (47.5). С теоретической точки зрения свойства обратных итераций не зависят от чисел α_k. Однако это не означает, что можно брать $\alpha_k = 1$ и применять какой-либо из численных методов, описанных в гл. V, даже в том случае, когда матрица $A - \tilde{\lambda}E$ невырожденная. Если собственные значения вычислены с большой точностью, матрица $A - \tilde{\lambda}E$ будет очень плохо обусловленной. Поэтому при $\alpha_k = 1$ возможен значительный рост элементов промежуточных вычислений и возникает реальная опасность переполнения. Эта опасность устраняется соответствующим выбором чисел α_k.

Исследуем сначала решение системы типа (47.5) с треугольной матрицей. Пусть заданы правая треугольная матрица C и вектор l порядка n. Определим такой вектор u, для которого выполняется равенство

$$Cu = \alpha l, \tag{48.1}$$

при некотором числе α, где $|\alpha| \leqslant 1$.

Обозначим через c_{ij}, l_i элементы матрицы C и вектора l. Будем находить координаты вектора u с помощью процесса, напоминающего обратную подстановку с одновременной нормировкой правой части. В случае $|c_{nn}| > |l_n|$ положим

$$u_n^{(n)} = l_n / c_{nn}, \quad \alpha^{(n)} = 1,$$

иначе

$$u_n^{(n)} = 1, \quad \alpha^{(n)} = c_{nn} / l_n.$$

Если c_{nn}, l_n равны нулю одновременно, то

$$u_n^{(n)} = 1, \quad \alpha^{(n)} = 1.$$

Предположим, что уже вычислены числа $u_{i+1}^{(i+1)}, \ldots, u_n^{(i+1)}, \alpha^{(i+1)}$. Находим

$$\gamma_i = \alpha^{(i+1)} l_i - \sum_{s=i+1}^{n} c_{is} u_s^{(i+1)}.$$

В случае $|c_{ii}| > |\gamma_i|$ положим

$$u_s^{(i)} = \begin{cases} \gamma_i / c_{ii}, & s = i, \\ u_s^{(i+1)} & s > i, \end{cases} \tag{48.2}$$

иначе

$$u_s^{(i)} = \begin{cases} 1, & s = i, \\ (c_{ii}/\gamma_i)\, u_s^{(i+1)}, & s > i, \end{cases} \tag{48.3}$$

$$\alpha^{(i)} = (c_{ii}/\gamma_i)\, \alpha^{(i+1)}.$$

Если c_{ii}, γ_i равны нулю одновременно, считаем $c_{ii}/\gamma_i = 1$. В качестве u берем вектор с координатами $u_1^{(1)}, \ldots, u_n^{(1)}$.

Оценим влияние ошибок округления. Пусть на всех этапах вычислений машинные нули не появляются. Предположим, что реально полученные числа $\tilde{u}_{i+1}^{(i+1)}, \ldots, \tilde{u}_n^{(i+1)}, \tilde{\alpha}^{(i+1)}$ можно трактовать как полученные при точных вычислениях, исходя из возмущенных данных

$$\begin{aligned} c_{gh}^{(i+1)} &= c_{gh}\left(1 + \varepsilon_{gh}^{(i+1)}\right), \quad g \geqslant i+1, \ h \geqslant g, \\ l_g^{(i+1)} &= l_g\left(1 + \eta_g^{(i+1)}\right), \quad g \geqslant i+1, \end{aligned} \tag{48.4}$$

где все величины $\varepsilon_{gh}^{(i+1)}$, $\eta_g^{(i+1)}$ имеют порядок p^{-t+1}. Это предположение заведомо выполняется на первом шаге, причем

$$|\varepsilon_{nn}^{(n)}| \lesssim (1/2)\, p^{-t+1}, \quad \eta_n^{(n)} = 0.$$

Определяем

$$\tilde{\gamma}_i = \mathrm{fl}_2\left(\tilde{\alpha}^{(i+1)} l_i - \sum_{s=i+1}^{n} c_{is} \tilde{u}_s^{(i+1)}\right) =$$

$$= \left(\tilde{\alpha}^{(i+1)} l_i - \sum_{s=i+1}^{n} c_{is} \tilde{u}_s^{(i+1)}\right)(1 + \sigma_i),$$

Если имеет место (48.2), то находим

$$\tilde{u}_s^{(i)} = \begin{cases} \mathrm{fl}\,(\tilde{\gamma}_i / c_{ii}) = (\tilde{\gamma}_i / c_{ii})(1 + \varkappa_i), & s = i, \\ \tilde{u}_s^{(i+1)}, & s > i. \end{cases}$$

Поэтому новые величины $\tilde{u}_i^{(i)}, \ldots, \tilde{u}_n^{(i)}, \tilde{\alpha}^{(i)}$ можно считать точно полученными из возмущенных данных:

$$c_{gh}^{(i)} \cong \begin{cases} c_{gh}(1 - \varkappa_i), & g = i, \ h = i, \\ c_{gh}(1 + \sigma_i), & g = i, \ h > i, \\ c_{gh}^{(i+1)}, & g \geqslant i+1, \ h \geqslant g, \end{cases}$$

$$l_g^{(i)} = \begin{cases} l_g(1 + \sigma_i), & g = i, \\ l_g^{(i+1)}, & g \geqslant i+1. \end{cases}$$

В случае, когда имеет место (48.3), находим

$$\beta_i = \begin{cases} 1, & \tilde{\gamma}_i = 0, \\ \mathrm{fl}\,(c_{ii}/\tilde{\gamma}_i) = (c_{ii}/\tilde{\gamma}_i)(1 - \varkappa_i), & \tilde{\gamma}_i \neq 0, \end{cases}$$

$$\tilde{u}_s^{(i)} = \begin{cases} 1, & s = i, \\ \mathrm{fl}\,(\beta_i \tilde{u}_s^{(i+1)}) = (c_{ii}/\tilde{\gamma}_i)\, \tilde{u}_s^{(i+1)}(1 - \varkappa_i)(1 + \nu_s), & s > i, \end{cases}$$

$$\tilde{\alpha}^{(i)} = \mathrm{fl}\,(\beta_i \alpha^{(i+1)}) = (c_{ii}/\tilde{\gamma}_i)\, \alpha^{(i+1)}(1 - \varkappa_i)(1 + \nu_i).$$

Теперь новые величины можно считать точно полученными из данных:

$$c_{gh}^{(i)} = \begin{cases} c_{gh}(1 - \varkappa_i)(1 + \nu_i), & g = i, \ h = i, \\ c_{gh}(1 + \sigma_i), & g = i, \ h > i, \\ c_{gh}^{(i+1)}(1 + \nu_i)/(1 + \nu_h), & g \geqslant i+1, \ h \geqslant g, \end{cases}$$

$$l_g^{(i)} = \begin{cases} l_g(1 + \sigma_i), & g = i, \\ l_g^{(i+1)}, & g \geqslant i+1. \end{cases}$$

Принимая во внимание ограниченность ошибок величиной $(1/2)\, p^{-t+1}$, заключаем из всех этих соотношений, что реально вычисленный вектор $\tilde{u}$ удовлетворяет равенству

$$(C + \Delta)\, \tilde{u} = \tilde{\alpha}\, l\, (1 + \delta). \tag{48.5}$$

Здесь $\tilde{\alpha}$ совпадает с $\tilde{\alpha}^{(1)}$, а для возмущений Δ, δ справедливы оценки

$$\begin{aligned} \|\Delta\|_E &\lesssim np^{-t+1}\|C\|_E, \\ \|\delta\|_E &\lesssim p^{-t+1}\|l\|_E. \end{aligned} \tag{48.6}$$

Если на промежуточных этапах вычислений машинные нули появляются, оценки (48.6) остаются такими же, так как их правые части изменяются лишь на величины порядка ω. Вместо (48.4) возмущенные данные на всех шагах процесса будут иметь вид

$$\begin{aligned} c_{gh}^{(i+1)} &= c_{gh}\left(1+e_{gh}^{(i+1)}\right)+\tau_{gh}^{(i+1)}, \quad g \geqslant i+1,\ h \geqslant g, \\ l_g^{(i+1)} &= l_g\left(1+\eta_g^{(i+1)}\right)+\xi_g^{(i+1)}, \quad g \geqslant i+1, \end{aligned}$$

где величины $e_{gh}^{(i+1)}$, $\eta_g^{(i+1)}$ имеют порядок p^{-t+1}, а величины $\tau_{gh}^{(i+1)}$, $\xi_g^{(i+1)}$ — порядок ω. Снова выполняется равенство (48.5), но теперь $\tilde{\alpha}$ совпадает с $\tilde{\alpha}^{(1)}$ не всегда. Если $\tilde{\alpha}^{(1)}=0$, но все диагональные элементы матрицы C отличны от нуля, то это означает, что равенство (48.5) выполняется при таком $\tilde{\alpha}$, которое по модулю меньше ω.

Не ограничивая общности, можно считать, что матрица A системы (47.5) правая почти треугольная. Аналогичный вид будет иметь матрица $A-\tilde{\lambda}E$. Поэтому систему (47.5) целесообразно решать следующим образом. Сначала приводим ее с помощью умножения слева на подходящим образом выбранную последовательность матриц вращения к системе с правой треугольной матрицей. Затем решение полученной системы находим с помощью описанного выше процесса. Реально вычисленный вектор $\bar{u}_k$ удовлетворяет (47.6). Принимая во внимание неравенство $\|A-\tilde{\lambda}E\| \lesssim 2\|A\|$, а также оценки (35.13), (48.6), получаем

$$\begin{aligned} \|E_k\|_E &\lesssim (\sqrt{2}+1)np^{-t+1}\|A\|_E, \\ \|\eta_{k-1}\| &\lesssim \sqrt{2}np^{-t+1}\|\bar{u}_k\|_E. \end{aligned} \tag{48.7}$$

Напомним, что $\|\bar{u}_k\|_1=1$ для всех $k>1$.

Обратные итерации особенно эффективны, когда матрица A симметричная. В этом случае без ограничения общности можно считать, что матрица A трехдиагональная и тогда, в соответствии с (35.14),

$$\begin{aligned} \|E_k\|_E &\lesssim (3\sqrt{2}+4)p^{-t+1}\|A\|_E, \\ \|\eta_{k-1}\|_E &\lesssim \sqrt{2}np^{-t+1}\|\bar{u}_{k-1}\|_E. \end{aligned}$$

При реализации обратных итераций приходится многократно решать системы (47.5). Однако их решение находится достаточно быстро, так как разложение матрицы $A-\tilde{\lambda}E$ на множители осуществляется только один раз.

Во многих случаях удается вычислить корневые векторы исключительно точно. Предположим, что с помощью обратных итераций уже получен достаточно точный вектор $\bar{u}$. Рассмотрим систему

$$(A-\tilde{\lambda}E)\omega = \bar{u}. \tag{48.8}$$

Используя имеющееся разложение матрицы $A-\tilde{\lambda}E$ на множители, мы можем попытаться применить процесс уточнения, описанный в § 38, к системе (48.8). Если этот процесс удастся реализовать, то реально найденный вектор $\tilde{\omega}$, по существу, будет совпадать с правильно округленным корневым вектором.

С точки зрения точных вычислений такое уточнение совпадает с выполнением еще одного шага обратных итераций без нормировок. Однако его *практическая реализация осуществляется иначе и позволяет исключить влияние эквивалентного возмущения в матрице A на достижимую точность.* Единственное препятствие, которое может возникнуть при решении системы (48.8), связано с большим ростом элементов промежуточных вычислений. Но порядок роста, как правило, не превосходит e^{-1}, и, если $\tilde{\lambda}$ не слишком близко к λ, процесс уточнения можно реализовать.

Обратим внимание на следующее обстоятельство. Если вещественная матрица A имеет комплексно сопряженные собственные значения λ, $\bar{\lambda}$, то диагональные элементы матриц $A-\lambda E$, $A-\bar{\lambda}E$ будут комплексными. Конечно, можно осуществлять обратные итерации целиком в комплексной арифметике. Однако это невыгодно в отношении как скорости счета, так и величины используемой памяти ЭВМ.

Более целесообразно выполнять обратные итерации несколько иначе. Будем определять векторы из прямой суммы L корневых подпространств матрицы A, соответствующих λ, $\bar{\lambda}$. Рассмотрим вещественную матрицу

$$B(\tilde{\lambda}) = (A - \tilde{\lambda}E)(A - \bar{\tilde{\lambda}}E) = A^2 - 2\operatorname{Re}\tilde{\lambda}A + |\tilde{\lambda}|^2 E.$$

Она имеет малые собственные значения $(\lambda - \tilde{\lambda})(\lambda - \bar{\tilde{\lambda}})$ и $(\bar{\lambda} - \tilde{\lambda})(\bar{\lambda} - \bar{\tilde{\lambda}})$. Прямая сумма ее корневых подпространств, соответствующих этим собственным значениям, совпадает с L. Поэтому обратные итерации с матрицей $B(\tilde{\lambda})$ позволяют эффективно находить векторы из L.

Для многих задач вполне достаточно определения векторов из L. Если все же необходимо найти корневые векторы матрицы A, то можно поступить следующим образом.

Подпространство L инвариантно относительно A и собственные значения оператора $A|L$, индуцированного на L оператором A, равны лишь λ и $\bar{\lambda}$. Предположим, что вычислен ортонормированный базис $v_1, \ldots, v_l$ подпространства L. Обозначим через V матрицу размера $n \times l$, столбцы которой совпадают с $v_1, \ldots, v_l$. В этом базисе оператор $A|L$ имеет матрицу

$$S = V^* A V \tag{48.9}$$

порядка l. Если координаты корневого вектора матрицы S, соответствующего λ, равны $s_1, \ldots, s_l$, то корневой вектор x матрицы A, соответствующий λ, будет таким:

$$x = \sum_{i=1}^{l} s_i v_i.$$

Заметим, что в случае простого собственного значения λ матрица (48.9) имеет второй порядок.

УПРАЖНЕНИЯ

1. Пусть система (47.5) решается описанным выше способом. Что означает появление нулевого α_k?

2. Исследовать вычислительный процесс выполнения обратных итераций с матрицей $B(\tilde{\lambda})$.

3. Предположим, что на этапе решения системы (48.8) мы изменили $\tilde{\lambda}$. Как влияет это изменение на величину достижимой точности и рост элементов при решении системы?

4. Предположим, что мы меняем $\tilde{\lambda}$ на каждом шаге выполнения обратных итераций. Как это отражается на времени счета?

5. Сравнить вычислительный процесс в упражнении 4 и QR-алгоритм со сдвигами.

6. Пусть матрица A вещественная и $\tilde{\lambda}$ есть приближение к комплексному собственному значению λ. Рассмотреть различные способы сведения комплексной системы (47.5) порядка n к вещественной системе порядка $2n$.

7. В условиях упражнения 6 исследовать вычислительный процесс выполнения обратных итераций.

8. Сравнить результаты выполнения упражнений 2, 7 между собой

§ 49. Апостериорные оценки точности

В общем случае нельзя найти эффективные априорные оценки точности решения проблемы собственных значений, особенно в отношении собственных векторов. Однако по приближенным собственным значениям и собственным векторам иногда удается получить достаточно хорошие апостериорные оценки.

Пусть A — нормальная матрица. Обозначим через $\lambda_1, \ldots, \lambda_n$ ее собственные значения, через $x_1, \ldots, x_n$ — соответствующие ортонормированные собственные векторы. Предположим, что приближенно вычислено собственное значение $\tilde{\lambda}$ и собственный вектор $\tilde{x}$. Рассмотрим невязку

$$r = (A - \tilde{\lambda}E)\tilde{x}.$$

Если

$$\tilde{x} = \sum_{i=1}^{n} \alpha_i x_i, \tag{49.1}$$

то

$$r = \sum_{i=1}^{n} (\lambda_i - \tilde{\lambda})\alpha_i x_i$$

и в силу ортонормированности векторов x_i

$$|r|^2 = \sum_{i=1}^{n} |\lambda_i - \tilde{\lambda}|^2 |\alpha_i|^2. \tag{49.2}$$

Обозначим через S целочисленное множество, для которого выполняется неравенство

$$\max_{i \in S} |\lambda_i - \tilde{\lambda}| \geqslant a_s > 0,$$

через L — подпространство, натянутое на собственные векторы x_i, для которых $i \notin S$. Ясно, что проекция $\mathrm{pr}_L \tilde{x}$ вектора $\tilde{x}$ на L задается формулой

$$\mathrm{pr}_L \tilde{x} = \sum_{i \notin S} \alpha_i x_i.$$

Если $\|r\|_E = \varepsilon$, то из (49.2) вытекает

$$\varepsilon^2 = \sum_{i=1}^{n} |\lambda_i - \tilde{\lambda}_i|^2 |\alpha_i|^2 \geqslant \sum_{i \notin S} |\lambda_i - \tilde{\lambda}_i|^2 |\alpha_i|^2 \geqslant a_s^2 \sum_{i \notin S} |\alpha_i|^2. \tag{49.3}$$

Следовательно,

$$\|\mathrm{pr}_L \tilde{x}\|_E = \left(\sum_{i \notin S} |\alpha_i|^2 \right)^{1/2} \leqslant \varepsilon / a_s. \tag{49.4}$$

Для нормальной матрицы A величина a_s легко оценивается по приближенно вычисленным собственным значениям $\tilde{\lambda}_i$ и априорно известной оценке эквивалентного возмущения матрицы A. Поэтому неравенство (49.4) представляет собой эффективную апостериорную оценку точности собственных векторов нормальной матрицы.

Оценка (49.4) зависит от величины нормы невязки. Если задан вектор $\tilde{x}$, то можно выбрать число μ_R так, чтобы невязка $(A - \mu_R E)\tilde{x}$ имела наименьшую норму. Имеем

$$\|(A - \mu_R E)\tilde{x}\|_E^2 = ((A - \mu_R E)\tilde{x},\ (A - \mu_R E)\tilde{x}) =$$
$$= (A\tilde{x},\ A\tilde{x}) - (A\tilde{x},\ \mu_R\tilde{x}) - (\mu_R\tilde{x},\ A\tilde{x}) + (\mu_R\tilde{x},\ \mu_R\tilde{x}) =$$
$$= (A\tilde{x},\ A\tilde{x}) - \frac{|(A\tilde{x},\ \tilde{x})|^2}{(\tilde{x},\ \tilde{x})} + \left(\mu_R - \frac{(A\tilde{x},\ \tilde{x})}{(\tilde{x},\ \tilde{x})}\right)\overline{\left(\mu_R - \frac{(A\tilde{x},\ \tilde{x})}{(\tilde{x},\ \tilde{x})}\right)}(\tilde{x},\ \tilde{x}).$$

Очевидно, что минимальное значение нормы невязки достигается при

$$\mu_R = \frac{(A\tilde{x},\ \tilde{x})}{(\tilde{x},\ \tilde{x})}. \tag{49.5}$$

Правая часть этого равенства называется *отношением Рэлея*, соответствующим вектору $\tilde{x}$. Если

$$\|(A - \tilde{\lambda}E)\tilde{x}\|_E = \varepsilon,$$

то всегда

$$\|(A - \mu_R E)\tilde{x}\|_E = \varepsilon' \leqslant \varepsilon.$$

Отношение Рэлея определено для любой матрицы, но особое значение оно имеет для нормальной. Предположим, что при некотором номере s соответствующее отношение Рэлея μ_R удовлетворяет условию

$$\max_{i \neq s} |\lambda_i - \mu_R| \geqslant a_s' > 0.$$

Согласно (49.1), (49.5) находим

$$\mu_R = \frac{\sum_{i=1}^{n} \lambda_i |\alpha_i|^2}{\sum_{i=1}^{n} |\alpha_i|^2}.$$

Следовательно,

$$\mu_R \sum_{i=1}^{n} |\alpha_i|^2 = \sum_{i=1}^{n} \lambda_i |\alpha_i|^2,$$

откуда

$$(\mu_R - \lambda_s)|\alpha_s|^2 = \sum_{i \neq s} (\lambda_i - \mu_R)|\alpha_i|^2 = \sum_{i \neq s} \frac{|\lambda_i - \mu_R|^2 |\alpha_i|^2}{(\lambda_i - \mu_R)}.$$

Аналогично (49.3) получаем

$$\sum_{i \neq s} |\lambda_i - \mu_R|^2 |\alpha_i|^2 \leqslant \varepsilon'^2,$$
$$|\alpha_s|^2 \geqslant \|\tilde{x}\|_E^2 - \varepsilon'^2/a_s'^2,$$

поэтому окончательно будем иметь

$$|\mu_R - \lambda_s| \leqslant \left(\frac{\varepsilon'^2}{a_s'}\right) \Big/ \left(\|\tilde{x}\|_E^2 - \frac{\varepsilon'^2}{a_s'^2}\right). \tag{49.6}$$

Итак, при $a_s' \geqslant \varepsilon'$ отношение Рэлея приближает изолированное собственное значение нормальной матрицы с точностью порядка ε'^2.

Вычисление отношений Рэлея дает возможность не только более точно находить отдельные собственные значения, но и оценить их погрешность. Мы не можем так же просто уточнить отдельные собственные векторы нормальной матрицы. Отношение Рэлея позволяет лишь несколько улучшить оценку (49.4), заменив ее более точной оценкой

$$\|\mathrm{pr}_L \tilde{x}\|_E \leqslant \varepsilon'/a_s'.$$

Для ненормальной матрицы трудно получить даже апостериорные оценки точности. Чтобы лучше понять

причину возникновения трудностей, определим главные члены поправок к приближенному решению проблемы собственных значений, выразив их через известные величины.

Пусть собственные значения $\lambda_1, \ldots, \lambda_n$ матрицы A попарно различны. Обозначим через $x_1, \ldots, x_n$ и $y_1, \ldots, y_n$ нормированные собственные векторы матриц A и A^*. Имеем

$$Ax_i = \lambda_i x_i, \qquad A^* y_i = \bar{\lambda}_i y_i. \tag{49.7}$$

Если известны приближенные величины $\tilde{\lambda}_i$, $\tilde{x}_i$, $\tilde{y}_i$, то

$$\lambda_i = \tilde{\lambda} + \Delta\lambda_i, \qquad x_i = \tilde{x}_i + \Delta x_i, \qquad y_i = \tilde{y}_i + \Delta y_i, \tag{49.8}$$

где все поправки, вообще говоря, малы.

В силу предположения о попарном различии собственных значений векторы $\tilde{x}_1, \ldots, \tilde{x}_n$ и $\tilde{y}_1, \ldots, \tilde{y}_n$ будут линейно независимыми. Поэтому x_i и y_i можно представить как суммы

$$x_i = \sum_{j=1}^{n} h_{ij}\tilde{x}_j, \quad y_i = \sum_{j=1}^{n} k_{ij}\tilde{y}_j.$$

Здесь h_{ii}, k_{ii} близки к единице, а остальные коэффициенты малы.

Собственные векторы определяются с точностью до скалярного множителя. Следовательно, можно считать, что $h_{ii} = k_{ii} = 1$. Теперь

$$\Delta x_i = \sum_{j \neq i} h_{ij}\tilde{x}_j, \quad \Delta y_i = \sum_{j \neq i} k_{ij}\tilde{y}_j.$$

Рассмотрим невязки

$$r_i = (A - \tilde{\lambda}E)\tilde{x}_i, \quad q_i = (A^* - \bar{\tilde{\lambda}}E)\tilde{y}_i.$$

Подставив в первое уравнение (49.7) значения λ_i, x_i из (49.8) и отбросив члены второго порядка малости, получим

$$r_i - \Delta\lambda_i\tilde{x}_i \cong -A\Delta x_i + \tilde{\lambda}_i\Delta x_i. \tag{49.9}$$

Далее находим

$$(A\Delta x_i, \tilde{y}_j) = (\Delta x_i, A^*\tilde{y}_j) = (\Delta x_i, \bar{\tilde{\lambda}}_j\tilde{y}_j + q_j) = \bar{\tilde{\lambda}}_j(\Delta x_i, \tilde{y}_j) + (\Delta x_i, q_j) = \bar{\tilde{\lambda}}_j \sum_{k \neq i} h_{ik}(\tilde{x}_k, \tilde{y}_j) + (\Delta x_i, q_j), \tag{49.10}$$

$$(\tilde{\lambda}_i\Delta x_i, \tilde{y}_j) = \tilde{\lambda}_i(\Delta x_i, \tilde{y}_j) = \tilde{\lambda}_i \sum_{k \neq i} h_{ik}(\tilde{x}_k, \tilde{y}_j).$$

Точные векторы x_k, y_j ортогональны [1] при $k \neq j$, поэтому $(\tilde{x}_k, \tilde{y}_j) = 0$ с точностью до членов первого порядка малости. Умножая (49.9) скалярно на $\tilde{y}_j$ и учитывая (49.10), будем иметь

$$(r_i, \tilde{y}_i) \cong \Delta\lambda_i(\tilde{x}_i, \tilde{y}_i),$$

$$(r_i, y_j) \cong h_{ij}(\tilde{\lambda}_i - \tilde{\lambda}_j)(\tilde{x}_j, \tilde{y}_j), \qquad i \neq j.$$

Отсюда следует, что

$$\Delta\lambda_i \cong \frac{(r_i, \tilde{y}_i)}{(\tilde{x}_i, \tilde{y}_i)}, \qquad h_{ij} \cong \frac{(r_i, \tilde{y}_j)}{(\tilde{\lambda}_i - \tilde{\lambda}_j)(\tilde{x}_j, \tilde{y}_j)}. \tag{49.11}$$

Аналогичная формула справедлива и для коэффициентов k_{ij}, определяющих поправку Δy_i:

$$k_{ij} \cong \frac{(q_i, \tilde{x}_j)}{(\tilde{\lambda}_i - \tilde{\lambda}_j)(\tilde{y}_i, \tilde{x}_j)} \tag{49.12}$$

Из первой формулы (49.11) вытекает, что с точностью до членов второго порядка малости

$$\lambda_i = \frac{(A\tilde{x}_i, \tilde{y}_i)}{(\tilde{x}_i, \tilde{y}_i)}.$$

Правая часть этого равенства называется *обобщенным отношением Рэлея*. Оно снова дает возможность более точно находить отдельные собственные значения. Однако из-за неортогональности системы собственных векторов матрицы A теперь нельзя получить гарантированные оценки точности типа (49.6).

Если матрица A нормальная, то можно считать, что $\tilde{x}_i = \tilde{y}_i$, а система векторов $\tilde{x}_1, \ldots, \tilde{x}_n$ близка к ортонормированной. В этом случае

$$\|\Delta x_i\|_E^2 \cong \sum_{j \neq i} |h_{ij}|^2, \quad \sum_{j \neq i} |(r_i, \tilde{y}_j)|^2 \leq \|r_i\|_E^2.$$

Пусть

$$\max_{j \neq i} |\tilde{\lambda}_i - \tilde{\lambda}_j| \geq a_i > 0.$$

Из вторых формул (49.11) легко получаем неравенство

$$\|\Delta x_i\|_E \lesssim \|r_i\|_E / a_i,$$

в основном совпадающее с (49.4).

Если матрица A не является нормальной, то все поправки в (49.11), (49.12) для ненормированных векторов x_i, y_i, по существу, пропорциональны величинам

$$c_i = \frac{\|x_i\|_E \|y_i\|_E}{|(x_i, y_i)|}.$$

Число c_i принято называть *коэффициентом перекоса* матрицы A, соответствующим собственному значению λ_i. Для вещественных векторов x_i, y_i $c_i = 1/|\cos\varphi_i|$, где φ_i есть угол между x_i и y_i. Ясно, что всегда $c_i \geqslant 1$.

Трудности получения гарантированных апостериорных оценок точности для ненормальной матрицы связаны с тем, что ее коэффициенты перекоса могут быть как угодно большими. Поэтому формулами (49.11), (49.12) можно пользоваться лишь тогда, когда априори известно, что все коэффициенты перекоса не очень велики. В этом случае формулы (49.11), (49.12) позволяют не только более точно найти собственные значения и собственные векторы, но и оценить главные члены ошибок.

УПРАЖНЕНИЯ

1. Пусть $r = (A - \mu E)v$. Доказать, что μ является собственным значением, а v — собственным вектором матрицы $A - (v, v)^{-1} rv^*$.

2. Пусть μ_R есть отношение Рэлея для вектора v и $r = (A - \mu_R E)v$. Доказать, что $(v, r) = 0$.

3. В условиях упражнения 2 доказать, что μ_R является собственным значением, а v — собственным вектором матрицы $A - (v, v)^{-1}(rv^* + vr^*)$.

4. Что можно сказать о возмущениях матрицы A в упражнениях 1, 3?

5. Рассмотрим эрмитову матрицу A и пусть λ_1, λ_n суть ее наибольшее и наименьшее собственные значения. Доказать, что

$$\lambda_1 = \max_{v \neq 0} \frac{(Av, v)}{(v, v)}, \quad \lambda_n = \min_{v \neq 0} \frac{(Av, v)}{(v, v)}.$$

6. В условиях упражнения 5 получить формулы, выражающие остальные собственные значения через отношение Рэлея.

7. Доказать, что вместо соотношения (49.6) в действительности выполняется более сильное неравенство

$$|\mu_R - \lambda_s| \leqslant \frac{r^*r}{a_s |x^s|}.$$

8. Оценить влияние ошибок округления при вычислении отношений Рэлея. Что следует предпринять для предотвращения потери точности?

9. Пусть X есть матрица собственных векторов для A. Доказать, что $Y = X^{*-1}$ есть матрица собственных векторов для A^*.

10. Доказать, что матрица является нормальной тогда и только тогда, когда все ее коэффициенты перекоса равны единице.

§ 50. Некоторые замечания

Выполненные исследования позволяют высказать некоторые рекомендации по применению рассмотренных алгорифмов для решения проблемы собственных значений.

Пусть матрица A не эрмитова и не имеет какой-либо особой специфики в своих элементах. В этом случае последовательность действий определена, по существу, однозначно.

Сначала с помощью подобного унитарного преобразования приводим матрицу A к правой почти треугольной. Если мы будем определять собственные векторы матрицы A, то должны запомнить преобразование подобия, в противном случае его можно не запоминать.

Следующий этап связан с применением QR-алгорифма к правой почти треугольной матрице. На этом этапе вычисляются только собственные значения. Преобразования подобия не запоминаются.

Используя вычисленные собственные значения, находим далее собственные векторы с помощью обратных итераций. Решение проблемы собственных значений заканчивается восстановлением собственных векторов исходной матрицы по собственным векторам почти треугольной матрицы.

Рассмотрим теперь эрмитову трехдиагональную матрицу A. Собственные значения такой матрицы можно определять либо с помощью QR-алгорифма, либо методом бисекций. На наш взгляд, предпочтение следует отдать методу бисекций, особенно в том случае, когда матрица имеет большой порядок и нужно находить не все собственные значения. Собственные векторы трехдиагональной матрицы в любом случае определяются с помощью обратных итераций.

Если эрмитова матрица A полная, то проблему собственных значений для нее можно решать двумя способами. Первый из них связан с приведением матрицы A к унитарно подобной трехдиагональной эрмитовой матрице, решением проблемы собственных значений для этой мат-

рицы и восстановлением собственных векторов матрицы A по собственным векторам трехдиагональной матрицы. Второй способ основан на применении метода вращений.

В том случае, когда нужно вычислить только собственные значения, первый способ по всем параметрам превосходит второй, за исключением единообразия вычислительной схемы. Если же нужно найти и собственные векторы, то у метода вращений появляется некоторое преимущество. С помощью обратных итераций очень трудно получить ортогональные собственные векторы, особенно при наличии большого скопления близких собственных значений. Собственные векторы, определяемые методом вращений, всегда почти ортогональны.

Рассмотренные алгорифмы, кроме метода вращений, особенно эффективны при решении последовательности спектральных задач, зависящих от некоторого параметра. Собственные значения, полученные для предыдущего значения параметра, могут служить хорошими приближениями для собственных значений, определяемых при последующем значении параметра. При этом значительно сокращается общее время счета.

Таблица 50.1

Сравнительная характеристика алгорифмов

Тип матрицы, алгорифм	Режим вычисл.	Число операций	Точность	Дополн. память
Произвольная				
приведение к почти треуг.	fl_2	$(10/3)\,n^3$	$5{,}0n$	$2n$
восстановл. собств. векторов	fl_2	$2n^3$	$5n$	$2n$
Почти треугольная				
QR-алгорифм	fl	$10n^3$	$26{,}5n^2$	n
обратные итерации	fl_3	$3n^3$	$2{,}4n$	n^3
Эрмитова полная				
приведение к трехдиагон.	fl_2	$(4/3)n^3$	$18{,}5n$	$2n$
метод вращ. (соб. зн.)	fl	$18n^3$	$48n$	0
метод вращ. (соб. зн., соб. векг.)	fl	$36n^3$	$81n$	n^3
Эрмитова трехдиагональная				
метод бисекций	fl	$10tn^2$	2,5	0
обратные итерации	fl	$52n^2$	8,3	n^2
QR-алгорифм	fl	$73n^2$	$85n$	$3n$

Сравнительные характеристики алгорифмов для решения проблемы собственных значений приведены в табл. 50.1. Она составлена в полном соответствии с табл. 34.1 и не нуждается в особых комментариях. Заметим лишь, что точность указана для отдельных собственных значений и собственных векторов, а число операций и дополнительная память—для полной проблемы. Все характеристики получены при следующих предположениях:

На каждое собственное значение в QR-алгорифме требуется пять итераций, на каждый собственный вектор в обратных итерациях требуется три итерации, для реализации метода вращений требуется шесть циклов.

УПРАЖНЕНИЯ

1. Исследовать алгорифмы решения проблемы собственных значений косоэрмитовой матрицы.

2. Исследовать алгорифмы решения проблемы собственных значений унитарной матрицы.

3. Построить алгорифм унитарно подобного преобразования $(2m+1)$-диагональной эрмитовой матрицы к трехдиагональной, без существенного использования дополнительной памяти для хранения результатов промежуточных вычислений.

4. Для алгорифма из упражнения 3 оценить влияние ошибок округления.

5. На основе исследованных алгорифмов для эрмитовых матриц построить численный метод решения проблемы собственных значений произведения двух эрмитовых матриц, из которых одна положительно определенная.

6. Для алгорифма из упражнения 5 оценить влияние ошибок округления.

7. Пусть A, B — произвольные квадратные матрицы. Доказать, что существуют унитарные матрицы Q, Z, для которых матрицы QAZ, QBZ правые треугольные.

8. Построить алгорифм нахождения унитарных матриц R, S, для которых RAS — правая почти треугольная, а RBS — правая треугольная.

9. Для алгорифма из упражнения 8 оценить влияние ошибок округления.

рицы и восстановлением собственных векторов матрицы A по собственным векторам трехдиагональной матрицы. Второй способ основан на применении метода вращений.

В том случае, когда нужно вычислить только собственные значения, первый способ по всем параметрам превосходит второй, за исключением единообразия вычислительной схемы. Если же нужно найти и собственные векторы, то у метода вращений появляется некоторое преимущество. С помощью обратных итераций очень трудно получить ортогональные собственные векторы, особенно при наличии большого скопления близких собственных значений. Собственные векторы, определяемые методом вращений, всегда почти ортогональны.

Рассмотренные алгорифмы, кроме метода вращений, особенно эффективны при решении последовательности спектральных задач, зависящих от некоторого параметра. Собственные значения, полученные для предыдущего значения параметра, могут служить хорошими приближениями для собственных значений, определяемых при последующем значении параметра. При этом значительно сокращается общее время счета.

Таблица 50.1

Сравнительная характеристика алгорифмов

Тип матрицы, алгорифм	Режим вычисл.	Число операций	Точность	Дополн. память
Произвольная				
приведение к почти треуг.	fl_2	$(10/3)n^3$	$5{,}9n$	$2n$
восстановл. собств. векторов	fl_2	$2n^3$	$5n$	$2n$
Почти треугольная				
QR-алгорифм	fl	$10n^3$	$26{,}5n^2$	n
обратные итерации	fl_2	$8n^3$	$2{,}4n$	n^2
Эрмитова полная				
приведение к трехдиагон.	fl_2	$(4/3)n^3$	$18{,}5n$	$2n$
метод вращ. (соб. зн.)	fl	$18n^3$	$48n$	0
метод вращ. (соб. зн., соб. вект.)	fl	$36n^3$	$84n$	n^2
Эрмитова трехдиагональная				
метод бисекций	fl	$10ln^2$	2,5	0
обратные итерации	fl	$52n^2$	8,3	n^2
QR-алгорифм	fl	$73n^2$	$85n$	$3n$

Сравнительные характеристики алгорифмов для решения проблемы собственных значений приведены в табл. 50.1. Она составлена в полном соответствии с табл. 34.1 и не нуждается в особых комментариях. Заметим лишь, что точность указана для отдельных собственных значений и собственных векторов, а число операций и дополнительная память—для полной проблемы. Все характеристики получены при следующих предположениях:

На каждое собственное значение в QR-алгорифме требуется пять итераций, на каждый собственный вектор в обратных итерациях требуется три итерации, для реализации метода вращений требуется шесть циклов.

УПРАЖНЕНИЯ

1. Исследовать алгорифмы решения проблемы собственных значений косоэрмитовой матрицы.

2. Исследовать алгорифмы решения проблемы собственных значений унитарной матрицы.

3. Построить алгорифм унитарно подобного преобразования $(2m+1)$-диагональной эрмитовой матрицы к трехдиагональной, без существенного использования дополнительной памяти для хранения результатов промежуточных вычислений.

4. Для алгорифма из упражнения 3 оценить влияние ошибок округления.

5. На основе исследованных алгорифмов для эрмитовых матриц построить численный метод решения проблемы собственных значений произведения двух эрмитовых матриц, из которых одна положительно определенная.

6. Для алгорифма из упражнения 5 оценить влияние ошибок округления.

7. Пусть A, B — произвольные квадратные матрицы. Доказать, что существуют унитарные матрицы Q, Z, для которых матрицы QAZ, QBZ правые треугольные.

8. Построить алгорифм нахождения унитарных матриц R, S, для которых RAS — правая почти треугольная, а RBS — правая треугольная.

9. Для алгорифма из упражнения 8 оценить влияние ошибок округления.

ПРИЛОЖЕНИЕ I

О РАСПРЕДЕЛЕНИИ ОШИБОК ОКРУГЛЕНИЯ

Мы сравнивали точность численных методов линейной алгебры по мажорантным оценкам норм эквивалентных возмущений. Но мажорантные оценки достигаются не так уж часто. Поэтому в целях создания более полной картины распределения ошибок округления весьма заманчиво считать отдельные ошибки случайными независимыми величинами. Заманчиво потому, что подобная гипотеза производит к вероятностным оценкам, лучшим по сравнению с мажорантными. Однако не менее заманчиво считать отдельные ошибки зависимыми случайными величинами, так как можно предположить, что знание характера зависимости также приведет к лучшим оценкам. Но тогда какими же их считать и каковы они в действительности?

В общем случае ответ на этот вопрос связан со сложными теоретико-числовыми исследованиями, выполнение которых не входит сейчас в нашу задачу. Мы ограничимся здесь лишь изложением нескольких наиболее простых фактов. Тем не менее даже эти факты позволят показать интересные свойства ошибок округления и дадут веские основания для выбора правдоподобной гипотезы совместного распределения всей совокупности ошибок округления вычислительного процесса. Для знакомства с технической стороной исследований мы отсылаем читателей к монографии [3].

Изучение вероятностных свойств ошибок округления невозможно без внесения в их поведение некоторого элемента случайности. Эту случайность нередко связывают с многократным решением одной и той же задачи на различных ЭВМ, с решением задачи при случайном числе верных знаков в промежуточных вычислениях и даже при случайном округлении результатов выполнения арифметических операций. Однако в условиях реальных вычислений внесение случайности можно осуществить, вообще говоря, единственным способом.

На всех современных ЭВМ операция округления является детерминированной. Следовательно, ошибка округления при выполнении любой арифметической операции однозначно определяется значениями аргументов самой операции. Поэтому при фиксированных входных данных задачи и фиксированном алгорифме ее решения вся совокупность ошибок округления определена однозначно и никакой случайности в поведении самих ошибок не возникает. Если вычислительный алгорифм не связан с не зависящими от него случайными процессами, то единственным источником случайности в ошибках округления может служить лишь случайность входных данных задачи.

Мы будем изучать распределение ошибок округления, рассматривая их как функции случайно заданных входных данных и предполагая

выполнение всех вычислений в режиме плавающей запятой с правильным округлением. Так как сейчас нас интересует в основном качественная картина распределения, то мы ограничимся *асимптотическими* исследованиями при $t \to \infty$.

Исследование зависимости ошибок округления от входных данных связано с преодолением многих трудностей, причину возникновения которых можно заметить уже на следующем примере. Представим величину x в виде ap^b, где a — ее мантисса, b — порядок. Легко проверить, что справедливо равенство

$$fl(x) = x + p^{-t} F(x)\, \varepsilon(x, t).$$

Здесь

$$F(x) = p^b, \quad |\varepsilon(x, t)| \leqslant 1/2.$$

Функция $F(x)$ не зависит от t. Она кусочно-постоянная и имеет разрывы в точках $x = p^s$ при всех целых s. Функция же $\varepsilon(x, t)$ является кусочно-линейной с периодом p^{-t} и, следовательно, с таким же периодом имеет разрывы. Поэтому нет никаких оснований ожидать, что в общем случае ошибки округления будут какими-либо гладкими функциями входных данных. Более того, почти очевидно, что при $t \to \infty$ множество точек разрыва ошибок округления будет всюду плотно на множестве входных данных. Это обстоятельство и определяет сложность исследований.

Ошибки округления не очень удобны для исследования. Вместо них мы будем изучать величины типа $\varepsilon(x, t)$ и называть их *нормализованными* ошибками округления.

Если входные данные являются случайными величинами, то нормализованная ошибка выполнения любой арифметической операции будет случайной величиной, распределенной каким-то образом на полусегменте $(-1/2, +1/2]$. Одна из важнейших задач заключается в том, чтобы понять, какими свойствами будет обладать асимптотическое распределение нормализованных ошибок округления.

Любой вычислительный процесс начинается с ввода входных данных в ЭВМ. Возникающие при этом ошибки округления описывает

Теорема 1. *Пусть в ЭВМ вводится случайная величина, имеющая непрерывную плотность распределения. Тогда, независимо от того, какова была плотность распределения, нормализованная ошибка округления при вводе асимптотически распределена равномерно на полусегменте* $(-1/2, +1/2]$.

Дальнейшие шаги процесса связаны с вычислением различных функций от одной или нескольких случайных величин, введенных в ЭВМ. Имеет место

Теорема 2. *Пусть в ЭВМ вводятся случайные величины, имеющие непрерывную плотность совместного распределения. Предположим, что введенные величины являются аргументами некоторой гладкой функции, у которой почти всюду хотя бы одна из координат градиента принимает иррациональное значение. Тогда, независимо от того, какова была плотность распределения входных данных, нормализованная ошибка округления при вычислении этой функции асимптотически распределена равномерно на полусегменте* $(-1/2, +1/2]$.

Условиям этой теоремы удовлетворяют функции, у которых в любой конечной области градиент совпадает не более, чем в конечном числе точек. К ним, очевидно, относятся обратная величина,

произведение и деление, степенная и логарифмическая функции, все тригонометрические функции и многие другие. Для этих функций нормализованная ошибка округления при их вычислении распределена согласно теореме 2. Важно подчеркнуть, что асимптотический характер распределения ошибки не зависит от распределения входных данных, от основания системы счисления, и в известной мере даже от вычисляемой функции. Это свойство ошибок является одним из самых замечательных.

Перечисленные факторы не влияют на вид асимптотического распределения, но, конечно, влияют на характер сходимости реального распределения к асимптотическому. Можно понять некоторые особенности, если рассмотреть ошибки вычисления функций, не удовлетворяющих условиям теоремы 2.

Наиболее интересным примером таких функций является сложение. Асимптотическое распределение его ошибок округления описывает

Теорема 3. *Пусть в ЭВМ вводятся две случайные величины, имеющие непрерывную плотность совместного распределения. Предположим, что во всей области определения разность порядков этих величин постоянна и равна $r \neq 0$. Тогда, независимо от того, какова была плотность распределения входных данных, нормализованная ошибка округления при вычислении суммы введенных величин дискретно распределена на полусегменте $(-1/2, +1/2]$ и принимает на нем асимптотически равновероятно значения вида ip^{-t} для целых i, удовлетворяющих неравенствам $-p^t/2 < i \leqslant +p^t/2$.*

Асимптотическое распределение нормализованной ошибки округления будет дискретным и при вычислении линейной комбинации любого числа введенных в ЭВМ случайных величин, если только все коэффициенты линейной комбинации являются рациональными числами. Однако в общем случае уже сложнее описать множество допустимых значений ошибки и вероятности их появления.

Из последней теоремы вытекает одно интересное следствие, показывающее неравноправие различных оснований систем счисления с точки зрения свойств ошибок округления. Именно, *математическое ожидание нормализованной ошибки округления при сложении двух чисел равно нулю при любом нечетном основании и равно $p^{-t}/2$ при любом четном основании.*

Данный факт требует некоторого пояснения. Строго говоря, утверждение теоремы 3 справедливо лишь для тех реализаций правильного округления, при которых ошибка округления однозначно определяется «хвостом» мантиссы, выходящим за t разрядов. И дело не в том, что рассматривается именно правильное округление. Можно показать, что математическое ожидание ошибки при сложении двух чисел асимптотически отлично от нуля и при любых других способах округления, если только ошибка округления однозначно определяется «хвостом».

Эти свойства систем счисления с четным основанием объясняются в конце концов тем, что невозможно построить округление, основанное на анализе лишь «хвоста» мантиссы таким образом, чтобы ошибки асимптотически компенсировали друг друга. Одним из лучших способов округления для этих систем счисления является классический, однако и здесь ошибка округления мантиссы, имеющей «хвост» величиной в $(1/2)\,p^{-t}$, не может быть скомпенсирована.

Мы уже отмечали в главе I, что правильное округление в системе с четным основанием может быть реализовано неоднозначно. Это связано с неоднозначностью округления чисел, мантиссы которых имеют «хвост», равный $(1/2)\,p^{-t}$. Чтобы исключить появление систематического смещения, необходимо мантиссу с «хвостом» величиной $(1/2)\,p^{-t}$ округлять равновероятно как с избытком, так и с недостатком. При этом датчик случайного округления должен быть связан с каким-либо из последних разрядов мантиссы, чтобы иметь возможность получить два одинаковых результата при повторном счете. Например, мантиссу с «хвостом» $(1/2)\,p^{-t}$ можно округлять с избытком, если ее t-й разряд принимает четное значение, и с недостатком, если нечетное. Такая модификация операции округления особенно удобна на ЭВМ с двоичной системой счисления, так как не требует выполнения дополнительного сложения.

Большинство современных ЭВМ работает в двоичной системе счисления. Правильная реализация округления на таких машинах вызывает определенные трудности и известно не так уж много ЭВМ, где эти трудности преодолены. Экспериментальная проверка показала наличие систематического смещения в ошибках округления почти на всех ЭВМ с двоичной системой. На некоторых из них величина смещения в несколько раз превышает максимальное значение ошибки правильного округления. В этом отношении выгодно выделяются ЭВМ, использующие представление чисел в сокращенной троичной системе. На таких машинах ошибки округления не имеют смещения.

Если операция округления реализована неправильно, то асимптотические распределения нормализованных ошибок при вычислении похожих функций $x+y$ и $x+1{,}2y$ в действительности оказываются совершенно различными. В первом случае распределение дискретное и имеет заметное смещение, во втором — равномерное и без смещения. На ЭВМ с неправильным округлением особенно полезно использование операций накопления. Это не только снижает общий уровень ошибок, но и во многих случаях позволяет устранить систематическое смещение.

Мы рассмотрели только первые шаги вычислительного процесса, связанные с вводом данных в ЭВМ и вычислением различных функций от введенных данных. При этом изучение ошибок осуществлялось лишь на основе вычислительного алгоритма без привлечения каких-либо дополнительных предположений о поведении самих ошибок. Аналогичные исследования можно продолжить и дальше. Однако приходится констатировать, что технические трудности проведения вычислительств начинают расти неизмеримо более быстрыми темпами, чем новые результаты.

Не касаясь описания этих исследований, заметим, что уже сейчас можно высказать некоторые соображения о том, какими должны быть результаты изучения дальнейших шагов вычислительного процесса.

Одним из важнейших свойств, обнаруженных у нормализованных ошибок округления, является независимость вида асимптотического распределения от входных данных. Пусть входные данные имеют непрерывную плотность совместного распределения. Как правило, при точных вычислениях любая совокупность промежуточных результатов будет также иметь непрерывную плотность совместного распре-

деления. Следовательно, кажется правдоподобным предположение, что почти все результаты приближенных вычислений на любых этапах можно трактовать как результаты ввода в ЭВМ некоторых величин, имеющих непрерывную плотность совместного распределения. В этом случае нормализованные ошибки округления при дальнейших вычислениях будут снова вести себя согласно теоремам 1, 2. Кажется правдоподобным и то, что независимость вида асимптотического распределения ошибок от входных данных должна влечь за собой независимость ошибок округления как случайных величин. При этом трудно лишь надеяться на практическую независимость ошибок в совокупности.

Таким образом, выполненные исследования и приведенные выше аргументы показывают, что при оценке суммарного влияния ошибок округления в массовых вычислениях, по-видимому, может быть использована следующая

Гипотеза. *Все нормализованные ошибки округления вычислительного процесса в режиме плавающей запятой являются случайными попарно независимыми величинами, распределение которых не зависит от входных данных и результатов промежуточных вычислений. Они распределены на полусегменте* $(-1/2, +1/2]$ *дискретно для операций сложения и вычитания и равномерно для большинства других операций. За исключением некоторых случаев можно считать, что математическое ожидание нормализованных ошибок округления равно нулю, а дисперсия не превосходит* $1/12$.

При практическом применении этой гипотезы следует проявлять определенную осторожность в отношении предполагаемых значений математического ожидания и дисперсии, что связано лишь с особенностями распределения нормализованных ошибок в операциях типа сложения. Гипотеза подтверждена рядом теоретических исследований, связанных с линейными преобразованиями векторов, разложением матриц на множители, итерационными процессами в линейной алгебре, вычислением определенных интегралов. Информация об этих исследованиях имеется в монографии [3] и библиографическом указателе [8].

Рассмотрим теперь примеры использования высказанной гипотезы для вывода некоторых вероятностных оценок, связанных с ошибками округления.

Важнейшим моментом в изучении устойчивости численных методов линейной алгебры было получение мажорантных оценок евклидовых норм эквивалентных возмущений M при разложении матрицы A на множители. Согласно формуле (34.1) эти оценки таковы:

$$\|M\|_E \lesssim f(n)\,2^{-t+1}\|A\|_E.$$

Для всех вероятностных исследований аналогичное значение имеют оценки вида

$$(M\|M\|_E^2)^{1/2} \lesssim \varphi(n)\,2^{-t+1}\|A\|_E,$$

где $M\|M\|_E^2$ есть математическое ожидание $\|M\|_E^2$. Как следует из табл. 34.1, функции $f(n)$ по порядку зависимости от n принимают значения для различных методов между n^0 и n^1. Однако $\varphi(n)$ при тех же режимах вычислений принимают свои значения уже только между n^0 и $n^{1/2}$.

Функции $\varphi(n)$ находятся по тем же самым схемам, что и функции $f(n)$, за исключением однотипных изменений, связанных с вероятностным анализом. Поэтому мы ограничимся здесь лишь кратким рассмотрением некоторых односторонних преобразований.

Пусть в соответствии с обозначениями § 19 над вектором z выполняется последовательность преобразований с реально заданными матрицами вращения $\bar{T}_{i_1 j_1}, \ldots, \bar{T}_{i_N j_N}$. Согласно гипотезе относительно распределения ошибок имеем

$$(M\|E\|_E^2)^{1/2} \lesssim \sqrt{\frac{2}{3}}\,2^{-t+1}\left(\sum_{k=0}^{N-1}\left(z_{i_k}^2 + z_{j_k}^2\right)\right)^{1/2}.$$

Ясно, что для циклических индексов в последовательности матриц вращения будет выполняться неравенство

$$(M\|E\|_E^2)^{1/2} \leqslant \sqrt{\frac{2}{3}}\,n^{1/2}2^{-t+1}\|z\|_E.$$

Это означает, что для всех рассмотренных ранее разложений матрицы на множители с помощью преобразований вращения функции $\varphi(n)$ не превосходят по порядку $n^{1/2}$. Пример из § 22 показывает неулучшаемость оценок для $\varphi(n)$.

Предположим далее, что в соответствии с обозначениями § 21 над вектором z выполняется последовательность преобразований с реально заданными матрицами отражения $\bar{U}_1, \ldots, \bar{U}_{n-1}$. По аналогии с формулой (21.3) будем иметь

$$(M\|T\|_E^2)^{1/2} \lesssim \left(\sum_{k=0}^{n-2} M\|\tau_k\|_E^2\right)^{1/2},$$

где τ_k — эквивалентное возмущение, возникающее при реализации $k+1$-го шага. Вероятностные оценки ошибок одного шага приводят теперь к неравенству

$$(M\|T\|_E^2)^{1/2} \lesssim \frac{5}{\sqrt{12}}\,n^{1/2}2^{-t+1}\|z\|_E.$$

Снова $\varphi(n)$ не превосходит по порядку $n^{1/2}$ и эта оценка для нее не может быть улучшена.

Если при реализации преобразований отражения не использовать накопление скалярных произведений, то соответствующий анализ ошибок показывает, что функция $f(n)$ принимает значения порядка n^2. Однако вероятностный анализ приводит к оценкам порядка n или $(n\log_2 n)^{1/2}$ для функции $\varphi(n)$ в зависимости от того, какой из способов суммирования, описанных в § 6, используется при вычислении скалярных произведений. Отсюда, пожалуй, можно прийти к заключению, что в *реальных вычислениях применение операций накопления скалярных произведений не должно давать столь же значительный эффект в отношении точности, как при получении гарантированных оценок*. Этот вывод подтверждается анализом самых различных алгорифмов, содержащих вычисление скалярных произведений.

Вероятностный анализ ошибок округления в преобразованиях Гаусса основан на формуле (24.7), из которой сразу же вытекает, что

$$(M\|B\|_E^2)^{1/2} = \left(\sum_{s=1}^{r} M\|B_{s-1}\|_E^2\right)^{1/2}.$$

Опять для функции $\varphi(n)$ получаем неулучшаемую оценку порядка $n^{1/2}$.

Для остальных типов разложений функции $\varphi(n)$ и $f(n)$ имеют один и тот же порядок n^0. Это объясняется тем, что в этих разложениях каждый элемент матриц эквивалентных возмущений определяется лишь одной элементарной ошибкой округления. Поэтому вероятностные оценки норм эквивалентных возмущений не могут быть существенно лучше, чем мажорантные.

Вероятностные оценки дают возможность сделать более точные выводы о соотношении между собой различных методов разложения матрицы на множители. Если относительно большой разброс значений функций $f(n)$ позволял еще отдать предпочтение одному из методов, то малый разброс значений $\varphi(n)$ говорит, что

С точки зрения устойчивости к ошибкам округления между прямыми методами разложения матрицы на множители нет принципиального различия.

Конечно, этот вывод в полной мере относится и к прямым методам решения систем линейных алгебраических уравнений. Согласно (36.15) мажорантная оценка погрешности в решении системы определяется формулой

$$\frac{\|\bar{x}-x^*\|_E}{\|x^*\|_E} \lesssim 2\nu_A f(n)\, p^{-t+1}.$$

Вероятностный анализ приводит к соотношению

$$\left(M\frac{\|\bar{x}-x^*\|_E}{\|x^*\|_E}\right)^{1/2} \lesssim 2\nu_A \varphi(n)\, p^{-t+1}.$$

Аналогический вывод можно сделать и относительно любых других численных методов, основанных на прямых разложениях матрицы на множители.

На этом заканчивается наше краткое знакомство с вероятностным анализом. Мы надеемся, что приведенное описание основ распределения ошибок округления поможет правильно ориентироваться при чтении соответствующей литературы. К тому же оно открывает перед читателем широкие возможности для более точного изучения вычислительных процессов.

ПРИЛОЖЕНИЕ II

РЕШЕНИЕ БОЛЬШИХ ЗАДАЧ ЛИНЕЙНОЙ АЛГЕБРЫ

Говоря о решении задач линейной алгебры на ЭВМ, мы всюду молчаливо предполагали, что входные данные и результаты промежуточных вычислений расположены в оперативной памяти (ОП). Однако в силу ряда причин одновременное хранение всей информации в ОП часто бывает невозможно или нецелесообразно. Такое положение может возникать при решении задач линейной алгебры как на малых вычислительных машинах, так и на больших машинах с многопрограммным управлением.

В этом случае приходится прибегать к использованию внешней памяти (ВП) и решать ряд новых вопросов, касающихся организации обмена информацией между ОП и ВП, математической модификации методов, уменьшения объема промежуточных результатов и т. д. На некоторых из этих вопросов мы сейчас и остановимся.

Всюду под большими задачами мы будем подразумевать такие задачи линейной алгебры, при решении которых общими методами появляется необходимость запоминать гораздо больше информации, чем может быть размещено в предоставленной пользователю части ОП. В зависимости от имеющейся специфики повышение эффективности процесса решения больших задач может осуществляться различными способами. Мы проиллюстрируем их на примере решения прямыми методами систем линейных алгебраических уравнений, матрицы которых принадлежат к одному из следующих видов: полная без специфики, клеточно-тепличева, разреженная с произвольным расположением нулевых элементов.

Пусть матрица A системы уравнений полная и не имеет какой-либо четко выраженной специфики в своих элементах. По существу, эту систему приходится решать одним из рассмотренных ранее методов, модифицировав их таким образом, чтобы обмен информацией между ОП и ВП осуществлялся наиболее эффективно.

Предположим, что вычислительная машина обладает ВП типа магнитного барабана или магнитного диска. Обозначим через τ и ν среднее время ожидания и время выборки одного кода. Тогда время T, затрачиваемое на обмен N кодами между ОП и ВП, будет определяться формулой $T = \tau + N\nu$. Будем считать, что для организации обменов отведена часть ОП величиной m слов, не считая памяти для хранения программы.

Важной характеристикой вычислительной схемы метода является общее время, затрачиваемое на обмен информацией. Сокращение этого времени имеет большое значение для ЭВМ как с однопрограммным, так и с многопрограммным управлением. Ясно, что чем полнее исполь-

ается часть ОП, предназначенная для обмена, тем время обмена будет меньше. Но и при максимальном использовании ОП мы будем получать совершенно различные времена в зависимости от выбранной вычислительной схемы.

Возьмем, например, за основу процесса решения системы метод Жордана [2]. Пусть порядок системы удовлетворяет неравенству $n \leqslant m/2$. Вызовем с ВП первые два столбца расширенной матрицы и коэффициенты второго из них преобразуем по формулам

$$a^*_{l,\ l+k} = a_{ll}^{-1} a_{l,\ l+k}, \qquad a^*_{j,\ l+k} = a_{j,\ l+k} - a_{jl} a_{ll}^{-1} a_{l,\ l+k}$$

для всех $j \neq l$ при $l=1$, $k=1$. Запишем преобразованный второй столбец на прежнее место в ВП, а в ОП вместо него спишем третий. Этот столбец также преобразуем согласно приведенным формулам при $l=1$, $k=2$. После того как будут преобразованы все столбцы ($k=1, \ldots, n$), первый столбец больше не потребуется. В дальнейших преобразованиях ($l=2, \ldots, n$) роль первого столбца последовательно будут выполнять второй, третий и т. д. Через n шагов матрица системы будет приведена к единичной, а решение получится на месте столбца свободных членов.

Сразу видны недостатки этой схемы. При $n < m/2$ ОП используется не полностью, а при $n > m/2$ решать системы таким способом невозможно. Но при $n = m/2$ схема кажется безупречной, так как число арифметических операций не увеличивается и не делается никаких лишних обменов между ОП и ВП. Нетрудно подсчитать, что общее время обмена в данном случае с точностью до главного члена будет определяться формулой

$$T^{(1)} = n^2\tau + n^2\nu.$$

Допустим далее, что матрица системы разбита на клетки A_{ij} порядка p, где $n = pr$. Будем считать, что $p = (m/3)^{1/2}$, т. е. одновременно в ОП можно разместить три клетки. За основу новой вычислительной схемы возьмем первую схему, заменив в формулах элементы a_{ij} на клетки A_{ij}. Хотя одновременно в ОП нельзя целиком расположить два клеточных столбца, ничто не мешает осуществлять операции над ними последовательно.

Как уже отмечалось, первая схема наиболее эффективна при $n = m/2$. Поэтому возьмем во второй схеме $m = 2n$. Простые вычисления показывают, что теперь общее время обмена для второй схемы будет таким: $T^{(2)} = 2{,}8n^{3/2}\tau + 1{,}8n^{5/2}\nu$.

Сравнивая $T^{(1)}$ и $T^{(2)}$ заключаем, что для второй схемы время обмена информацией между ОП и ВП примерно в $n^{1/2}/2$ раз меньше. Столь ощутимый выигрыш появился лишь вследствие более удачного решения организационных вопросов. Рассмотренный пример наглядно показывает, что непродуманная реализация обменов между ОП и ВП может привести к большим непроизводительным затратам машинного времени.

Очевидно, что для решения любой системы уравнений клеточным методом с использованием ВП потребуется больше времени, чем для решения той же системы обычным методом с записью всей матрицы

в ОП. Увеличение времени будет происходить как за счет новой схемы вычислений, так и за счет работы с ВП, причем тем больше, чем меньше часть ОП, предназначенная для обмена информацией. Для оценки проигрыша во времени введем *коэффициент потерь*, который будем вычислять по формуле

$$P = (T_k + T)\, T_0^{-1}.$$

Здесь T_k — время, затрачиваемое на выполнение арифметических операций в клеточном методе, T — время, затрачиваемое на обмен информацией между ОП и ВП, T_0 — время, затрачиваемое на выполнение арифметических операций в обычном методе.

Исследование коэффициента потерь обнаруживает удивительный факт. Для средних машин типа БЭСМ-4 он близок к 2 уже при $m \geqslant 300$ и $n \geqslant 200$. Следовательно, используя ВП для хранения матрицы и всего лишь 300 ячеек ОП для организации обмена, можно на таких ЭВМ весьма успешно решать большие задачи. При этом мы затратим только в 2 раза больше времени по сравнению с тем случаем, если бы смогли записать в ОП всю матрицу.

Клеточные аналоги построены почти для всех численных методов, рассмотренных в настоящей книге. Они с успехом применяются для решения тех задач линейной алгебры, в которых матрицы достаточно велики и не имеют особой специфики. Конечно, для таких задач могут оказаться эффективными и другие принципы организации обменов информацией между ОП и ВП, а также другие модификации вычислительных схем. Важно лишь обеспечить устойчивость численного метода и относительную близость коэффициента потерь к единице.

Не каждую специфику задачи удается учесть с помощью подходящего выбора численного метода. Однако в некоторых случаях оказывается возможным добиться огромного эффекта, особенно для задач с матрицами больших размеров.

Рассмотрим сначала в качестве примера решение систем линейных алгебраических уравнений с так называемыми клеточно-теплицевыми матрицами. Пусть матрица A разбита на клетки A_{ij} порядка p и снова $n = pr$. Матрица A называется *клеточно-теплицевой*, если $A_{ij} = A_{sl}$ при $i - j = s - l$. Системы с такими матрицами возникают в различных задачах акустики, статистики, электродинамики и т. п. В частности, они появляются при решении интегральных уравнений Фредгольма с ядрами, зависящими от расстояния между точками, путем сведения к алгебраическим системам. Важно подчеркнуть, что такие системы могут состоять из нескольких сотен и даже тысяч уравнений.

Клеточно-теплицева матрица обладает ярко выраженной спецификой и может быть задана массивом чисел величиной $2n^2/r$, а не n^2, как в случае полной матрицы. Но если систему с такой матрицей решать, например, с помощью какого-нибудь из вариантов метода Жордана, то уже после первых преобразований специфика матрицы будет нарушена.

Эффективные по скорости и использованию памяти ЭВМ численные методы решения систем с клеточно-теплицевыми матрицами появились сравнительно недавно.

Пусть заданы квадратные матрицы $a_{-r+1}, \ldots, a_0, \ldots, a_{r-1}$ порядка p. Рассмотрим клеточно-теплицевы матрицы

$$A_k = \begin{bmatrix} a_0 & a_1 & a_2 & \ldots & a_k \\ a_{-1} & a_0 & a_1 & \ldots & a_{k-1} \\ a_{-2} & a_{-1} & a_0 & \ldots & a_{k-2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{-k} & a_{-k+1} & a_{-k+2} & \ldots & a_0 \end{bmatrix}$$

и обозначим через α_k и β_k соответственно первый и последний клеточные столбцы матрицы A_k^{-1}.

Пусть

$$\alpha_k = \begin{bmatrix} \alpha_{0,\,k} \\ \alpha_{1,\,k} \\ \vdots \\ \alpha_{k,\,k} \end{bmatrix}, \quad \beta_k = \begin{bmatrix} \beta_{0,\,k} \\ \beta_{1,\,k} \\ \vdots \\ \beta_{k,\,k} \end{bmatrix}.$$

Ясно, что при $k=0$ эти столбцы совпадают и содержат лишь одну клетку a_0^{-1}. Предположим, что известны α_{k-1} и β_{k-1}. Будем искать α_k и β_k в таком виде:

$$\alpha_k = \begin{bmatrix} \alpha_{0,\,k-1} \\ \alpha_{1,\,k-1} \\ \vdots \\ \alpha_{k-1,\,k-1} \\ 0 \end{bmatrix} U_k + \begin{bmatrix} 0 \\ \beta_{0,\,k-1} \\ \vdots \\ \beta_{k-2,\,k-1} \\ \beta_{k-1,\,k-1} \end{bmatrix} V_k,$$

$$\beta_k = \begin{bmatrix} \alpha_{0,\,k-1} \\ \alpha_{1,\,k-1} \\ \vdots \\ \alpha_{k-1,\,k-1} \\ 0 \end{bmatrix} R_k + \begin{bmatrix} 0 \\ \beta_{0,\,k-1} \\ \vdots \\ \beta_{k-2,\,k-1} \\ \beta_{k-1,\,k-1} \end{bmatrix} S_k,$$

где U_k, V_k, R_k, S_k — некоторые матрицы порядка p. Так как α_k и β_k являются клеточными столбцами матрицы A_k^{-1}, то

$$A_k\alpha_k = \begin{bmatrix} E \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad A_k\beta_k = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ E \end{bmatrix}.$$

Здесь E — единичная матрица порядка p. Обозначив

$$F_{1k} = \sum_{t=-k}^{-1} a_t\alpha_{t+k,\,k-1}, \qquad F_{2k} = \sum_{t=-k}^{-1} a_t\beta_{t-1,\,k-1},$$

получаем из последних соотношений, что

$$\left.\begin{aligned} U_k + F_{2k}V_k &= E, \\ F_{1k}U_k + V_k &= 0, \end{aligned}\right\} \qquad \left.\begin{aligned} R_k + F_{2k}S_k &= 0, \\ F_{1k}R_k + S_k &= E, \end{aligned}\right\}$$

и далее находим

$$U_k = (E - F_{2k}F_{1k})^{-1}, \qquad V_k = -F_{1k}(E - F_{2k}F_{1k})^{-1},$$
$$R_k = -F_{2k}(E - F_{1k}F_{2k})^{-1}, \qquad S_k = (E - F_{1k}F_{2k})^{-1}.$$

Таким образом, для матриц $A_0^{-1}, \ldots, A_{r-1}^{-1}$ можно последовательно определить первые и последние клеточные столбцы, не вычисляя элементы остальных столбцов.

Предположим теперь, что решается система линейных алгебраических уравнений $Ax = b$ с клеточно-теплицевой матрицей $A = A_{r-1}$. Представим векторы x, b в виде

$$x = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{r-1} \end{bmatrix}, \quad b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{r-1} \end{bmatrix}$$

и рассмотрим усеченные системы

$$A_k y_k = d_k,$$

где

$$y_k = \begin{bmatrix} y_{0,\,k} \\ y_{1,\,k} \\ \vdots \\ y_{k,\,k} \end{bmatrix}, \quad d_k = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix}.$$

Все векторы, выписанные здесь в квадратных скобках, имеют один и тот же порядок p. Пусть

$$\begin{bmatrix} y_{0,\,k} \\ \vdots \\ y_{k-1,\,k} \\ y_{k,\,k} \end{bmatrix} = \begin{bmatrix} y_{0,\,k-1} \\ \vdots \\ y_{k-1,\,k-1} \\ 0 \end{bmatrix} + \begin{bmatrix} z_{0,\,k} \\ \vdots \\ z_{k-1,\,k} \\ z_{k,\,k} \end{bmatrix}.$$

Подставив это выражение в уравнение

$$A_k y_k = d_k$$

и учитывая, что вектор y_{k-1} и удовлетворяет уравнению

$$A_{k-1}y_{k-1} = d_{k-1},$$

заключаем, что вектор поправки z_k с элементами $z_{0,\,k}, \ldots, z_{k,\,k}$ является решением системы $A_k z_k = f_k$, где

$$f_k = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ f_{kk} \end{bmatrix}, \qquad f_{kk} = b_k - \sum_{t=-k}^{-1} a_t y_{t+k,\,k-1}.$$

Вектор z_k есть линейная комбинация последних столбцов матрицы A_k^{-1}, координаты вектора f_{kk} являются коэффициентами данной линейной комбинации. Следовательно, для рекуррентного вычисления векторов y_k достаточно рекуррентно вычислять последний клеточный столбец матриц A_k^{-1}. Заметим, что вектор x совпадает с y_{r-1}, а вектор $u_0 = a_0^{-1} b_0$.

Этот метод решения систем с клеточно-теплицевыми матрицами весьма эффективен по сравнению с другими прямыми методами. Для его реализации при больших r необходимо выполнить лишь порядка $n^3 r$ арифметических операций, занимая при этом $4n^2 r$ слов памяти ЭВМ. Если клетки матрицы связаны между собой соотношениями $a_{-i} = la_i u$ для некоторых симметричных матриц перестановок l, u, то объем вычислений и необходимую память можно уменьшить примерно в два раза.

Рассмотрим один частный случай клеточно-теплицевой матрицы. Матрица называется *клеточно-циркулянтной*, если ее клетки связаны соотношениями $a_{-i} = a_{r-i}$ при всех i. Для решения систем уравнений с такими матрицами существует более эффективный метод. В основе его лежит тот факт, что любая клеточно-циркулянтная матрица унитарно подобна клеточно-диагональной матрице. При этом матрица подобия не зависит от элементов исходной матрицы, а клеточно-диагональная матрица легко вычисляется. Построенный на этом разложении численный метод требует выполнения порядка $n^2 (p + r) r$ арифметических операций, занимая около $2n^2 r$ слов памяти ЭВМ.

Некоторые системы со сложными матрицами можно успешно решать путем сведения к системам, имеющим более простое строение. Предположим, что матрица разбита на квадратные клетки. Будем называть это разбиение первым уровнем. Пусть далее каждая из клеток первого уровня разбита, в свою очередь, одинаковым образом на квадратные клетки. Это разбиение назовем вторым уровнем и т. д.

Введем в рассмотрение классы γ_r, σ_r, η_r соответственно клеточно-диагональных, клеточно-циркулянтных и клеточно-теплицевых матриц клеточного порядка r. Под π_r будем понимать класс любых других клеточных матриц такого же порядка. Теперь рассмотрим системы, матрицы которых задаются последовательностью $\theta_{n_1}, \theta_{n_2}, \ldots, \theta_{n_r}$, где в каждой позиции вместо θ стоит одна из букв γ, σ, η, π. Это означает, что разбиение первого уровня определяется символом θ_{n_1}, второго — символом θ_{n_2}, последнего — символом θ_{n_r}.

Допустим, что s' букв в данной последовательности совпадают либо с γ, либо с σ и соответствующие порядки разбиений равны $n_{q_1}, \ldots, n_{q_s}$. Оказывается, что решение исходной системы с помощью унитарного преобразования, не зависящего от значений элементов матрицы, сводится к решению $n_{q_1} \ldots n_{q_s}$ систем. Структура этих систем одинакова и определяется последовательностью символов, полученной из заданной путем формального вычеркивания всех букв γ и σ.

Рассмотренные матрицы нередко встречаются в приложениях, особенно при решении многомерных интегральных уравнений. Обычно они имеют столь большой порядок, что сведение к системам меньшего порядка и тщательный учет специфики оказывается единственно возможным способом их решения.

Значительное число теоретических и прикладных задач приводит к решению больших алгебраических систем с *разреженными* матрицами. Эти матрицы имеют много нулевых элементов. Если ненулевые элементы расположены согласно рис. 27.3, то для решения таких систем исключительно эффективным оказывается применение метода Гаусса. При этом эффективность метода будет тем выше, чем меньше на рис. 27.3 площадь заштрихованной части и число нулевых элементов в ней.

Последнее замечание лежит в основе многих способов предварительного преобразования разреженной матрицы. Как правило, эти преобразования состоят лишь в перестановке строк и столбцов и направлены либо на сокращение общего времени счета задачи, либо на уменьшение объема необходимой памяти ЭВМ. Эффект от таких преобразований может быть очень большим. Рассмотрим, например, матрицы вида

$$\begin{bmatrix} * & * & * & * & * \\ * & * & & & 0 \\ * & & * & & \\ * & & & * & \\ * & & 0 & & * \end{bmatrix}, \quad \begin{bmatrix} * & & & & 0 & * \\ & * & & & & * \\ & & * & & * & * \\ & 0 & & & * & * \\ * & * & * & * & * & * \end{bmatrix}$$

порядка n, каждая из которых получается из другой с помощью перестановки строк и столбцов. Если первую матрицу разложить на треугольные множители, то потребуется выполнить $(2/3) n^3$ арифметических операций, заняв при этом n^2 слов памяти, так как оба треугольных множителя будут полными. Для разложения второй матрицы нужно выполнить $2n$ операций, имея всего $3n$ слов памяти.

Если ненулевые элементы разреженной матрицы расположены без какого-нибудь явного порядка, то найти соответствующие матрицы перестановок очень трудно. Для решения этой задачи нередко приходится привлекать различные методы комбинаторики, теории графов, целочисленного программирования и т. п. Однако затраты на предварительное преобразование разреженных матриц вполне окупаются, если сами матрицы имеют большие размеры и системы с ними решаются многократно. Такие ситуации возникают при оперативном управлении энергетическими сетями, транспортными потоками, технологическими процессами. Конечно, аналогичные преобразования можно выполнять и с клеточными разреженными матрицами.

Большие задачи на собственные значения возникают значительно реже, чем большие системы линейных алгебраических уравнений. Особенно редко приходится решать полную проблему для больших матриц. Однако в случае необходимости эти задачи также успешно решаются.

Не требует каких-либо модификаций для больших задач метод бисекций. Известны случаи изучения с его помощью распределения собственных значений матриц Якоби, порядок которых превышал десятки тысяч. Полную матрицу большого порядка целесообразно

привести к почти треугольной с помощью одного из клеточных вариантов преобразований вращения или отражения. Затем к ней можно применить один из численных методов, например, *QR*-алгорифм или обратные итерации. Снова для уменьшения коэффициента потерь целесообразно использовать клеточные варианты методов.

Обсуждая проблемы решения больших задач линейной алгебры, мы не ставили перед собой задачу дать подробный анализ существующих методов. Однако нам хотелось бы обратить внимание на то, что такие задачи могут решаться достаточно эффективно. Для более полного знакомства с этой тематикой мы отсылаем читателей к обзору [7] и библиографическому указателю [8].

ЛИТЕРАТУРА

1. Воеводин В. В. Линейная алгебра, М., «Наука», 1974.
2. Воеводин В. В. Численные методы алгебры (теория и алгорифмы), М., «Наука», 1966.
3. Воеводин В. В. Ошибки округления и устойчивость в прямых методах линейной алгебры, М., Изд-во МГУ, 1969.
4. Икрамов Х. Д. Задачник по линейной алгебре, М., «Наука», 1975.
5. Уилкинсон Дж. Х. Алгебраическая проблема собственных значений, М., «Наука», 1970.
6. Фаддеев Д. К., Фаддеева В. Н. Вычислительные методы линейной алгебры, М.—Л., Физматгиз, 1963.
7. Фаддеев Д. К., Фаддеева В. Н. Вычислительные методы линейной алгебры, Зап. науч. семинаров Ленингр. отд. Матем. ин-та АН СССР, 1975, 54
8. Фаддеева В. Н., Кузнецов Ю. А. и др. Вычислительные методы линейной алгебры, Библиографический указатель, 1828—1974 гг., Новосибирск, 1976.
9. Форсайт Дж., Молер К. Численное решение систем линейных алгебраических уравнений, М., «Мир», 1969.

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

Валентин Васильевич Воеводин

ВЫЧИСЛИТЕЛЬНЫЕ ОСНОВЫ ЛИНЕЙНОЙ АЛГЕБРЫ

М., 1977 г., 304 стр. с илл.

Редактор *Г. Д. Ким*
Техн. редактор *Е. В. Морозова*
Корректор *Т. С. Вайсберг*

Сдано в набор 9/III 1977 г. Подписано к печати 12/VIII 1977 г. Формат 84×108¹/₃₂. Физ. печ. л. 9,5. Условн. печ. л. 15,96. Уч.-изд. л. 14,5. Тираж 40 000 экз. Цена книги 75 коп. Заказ № 1140.

Издательство «Наука»
Главная редакция физико-математической литературы
117071, Москва, В-71, Ленинский проспект, 15

Отпечатано с матриц, изготовленных Ордена Трудового Красного Знамени Ленинградским производственно-техническим объединением «Печатный Двор» имени А. М. Горького Союзполиграфпрома при Государственном комитете Совета Министров СССР по делам издательств, полиграфии, книжной торговли. 197136, Ленинград, П-136, Гатчинская ул., 26, в 4-й типографии изд-ва «Наука». 630077, Новосибирск, 77, Станиславского, 25. Заказ № 685.