V. V. Voevodin, The accuracy of solving systems of equations by direct methods, *Zh. Vychisl. Mat. Mat. Fiz.*, 1968, Volume 8, Number 5, 1094–1096

ЖУРНАЛ
ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ и МАТЕМАТИЧЕСКОЙ ФИЗИКИ

Том 8 Сентябрь 1968 Октябрь № 5

НАУЧНЫЕ СООБЩЕНИЯ

УДК 518 : 512.25

О ТОЧНОСТИ РЕШЕНИЯ СИСТЕМ УРАВНЕНИЙ ПРЯМЫМИ МЕТОДАМИ

В. В. ВОЕВОДИН

(Москва)

Ниже рассматриваются некоторые общие вопросы оценки эквивалентных возмущений [1, 2] при решении систем линейных алгебраических уравнений. Описывается класс методов, для которого получены эффективные априорные (при вычислении с фиксированной запятой) и апостериорные (при вычислении с плавающей запятой) оценки. В качестве частного результата получены априорные оценки для метода ортогонализации. Эти оценки лучше, чем все известные в настоящее время, полученные для других методов.

Пусть решается система линейных алгебраических уравнений каким-либо из прямых методов. Не ограничивая существенно общности, можно предположить, что теоретически метод сводится к построению последовательности систем $A_p x = b_p$, эквивалентных исходной системе $A_0 x = b_0$ и связанных между собой рекуррентными соотношениями

$$A_{p+1} = L_p A_p, \qquad b_{p+1} = L_p b_p, \qquad p = 0, 1, \ldots, m-1. \tag{1}$$

Здесь L_p — невырожденные матрицы.

Обычно реальный вычислительный алгорифм состоит из следующих этапов.

1. По элементам матрицы A_p (или A_p и b_p) вычисляются элементы L_p. Эти элементы содержат ошибки, зависящие от способа их вычислений.

2. С реально вычисленными A_p, b_p и L_p совершается преобразование (1). Матрица A_{p+1} и вектор b_{p+1} будут содержать некоторые ошибки, появившиеся из-за округления результатов промежуточных вычислений.

3. Принудительно поддерживается форма матрицы A_m.

4. Решается система с матрицей A_m.

Из всех этапов некоторого пояснения требует лишь третий этап. Элементы матрицы L_p вычисляются, как правило, из соотношений, накладываемых на элементы матрицы A_{p+1}. Так как элементы L_p вычисляются с ошибкой, то в действительности эти соотношения выполняться не будут. Реализация третьего этапа заключается в том, что формально мы считаем их выполненными.

Рассмотрим теперь влияние каждого этапа на величину эквивалентного возмущения. Будем считать для простоты, что вычисления ведутся с фиксированной запятой [2], и ограничимся сравнением метода Гаусса и метода ортогонализации без нормировки строк [3].

Каковы бы ни были ошибки в вычислении элементов матрицы L_p, они практически не влияют на ошибку второго этапа, которая в методе Гаусса есть величина порядка $n^2 2^{-t}$, а в методе ортогонализации — порядка $n 2^{-t}$ [2]. При этом из [2] следует, что наиболее вероятная ошибка в методе Гаусса будет все же величиной порядка $n^{3/2} 2^{-t}$, т. е. больше, чем в методе ортогонализации.

Ошибка в вычислении элементов матрицы L_p сказывается в методе Гаусса лишь на третьем этапе и приводит к эквивалентному возмущению порядка $n2^{-t}$ [1], если применялся процесс с выбором главного элемента. В противном случае ошибка может быть существенно больше. В методе ортогонализации ошибки первого этапа не приводят к дополнительным ошибкам на третьем этапе, но, как показано в [4], § 13, могут в значительной степени нарушить ортогональность строк матрицы A_m.

Принудительное сохранение формы матрицы A_m имеет определенные достоинства, так как позволяет легко решать систему на четвертом этапе. В методе Гаусса матрица A_m треугольная, и решение системы с такой матрицей может быть осуществлено весьма легко и точно с помощью обратной подстановки [1]. Что же касается метода ортогонализации, то формальное предположение об ортогональности строк матрицы может привести к громадным ошибкам в решении. Тем не менее точное решение системы $A_m x = b_m$ в этом методе более близко (в смысле величины эквивалентного возмущения) к точному решению исходной системы $A_0 x = b_0$, чем в методе Гаусса.

Таким образом, основная потеря точности в методе ортогонализации происходит на четвертом этапе. Для того чтобы устранить этот недостаток и сделать матрицу A_m ортогональной, можно воспользоваться процессом, описанным в [4], § 13. Там матрица A_{p+1} и вектор b_{p+1} вычисляются с помощью следующего итерационного процесса:

$$A_{p+1} = \lim_{t\to\infty} A_p^{(t)}, \qquad A_p^{(t)} = L_p^{(t)} A_p^{(t-1)},$$

$$b_{p+1} = \lim_{t\to\infty} b_p^{(t)}, \qquad b_p^{(t)} = L_p^{(t)} b_p^{(t-1)}, \tag{2}$$

$$A_p^{(0)} = A_p, \qquad b_p^{(0)} = b_p.$$

На каждом шаге элементы $L_p^{(t)}$ вычисляются по тем же формулам, что и на первом шаге, но с заменой матрицы A_p матрицей $A_p^{(t-1)}$. Сходимость процесса доказана в [4], § 13. Обычно для достижения нужной точности требуется выполнить не более двух — трех итераций.

Пусть F_0 и g_0 — матрица и вектор эквивалентных возмущений при вычислении A_m и b_m в процессе (2). Не меняя рассуждений, изложенных в [2], можно показать, что элементы $f_{ij}^{(0)}$ и $g_i^{(0)}$ удовлетворяют соотношениям

$$|f_{ij}^{(0)}| \leqslant \frac{k_i}{2} 2^{-t}, \qquad |g_i^{(0)}| \leqslant \frac{k_i}{2} 2^{-t} \tag{3}$$

при вычислении с фиксированной запятой и

$$|f_{ij}^{(0)}| \leqslant k_i |a_{ij}^{(m)}| 2^{-t}, \qquad |g_i^{(0)}| \leqslant k_i |b_i^{(m)}| 2^{-t} \tag{4}$$

при вычислении с плавающей запятой. Через k_i обозначено количество итераций в процессе (2) при вычислении A_i, $a_{ij}^{(m)}$ и $b_i^{(m)}$ — элементы реально вычисленных A_m, b_m.

Как уже отмечалось, обычно $k_i \leqslant 3$, поэтому

$$\|F_0\|_E \leqslant \frac{3n}{2} 2^{-t}, \qquad \|g_0\|_E \leqslant \frac{3\sqrt{n}}{2} 2^{-t} \tag{5}$$

и

$$\|F_0\|_E \leqslant 3\|A_m\|_E 2^{-t}, \qquad \|g_0\|_E \leqslant 3\|b_m\|_E 2^{-t}, \tag{6}$$

соответственно, при вычислении с фиксированной и плавающей запятой. В соотношениях (4), (6) отброшены члены порядка 2^{-2t}. Через $\|\ \|_E$ обозначена евклидова норма.

Интересно отметить следующее обстоятельство. Как вытекает из [4], стр. 86, в процессе ортогонализации без нормировки никакая строка матрицы не может увеличить своей евклидовой нормы. Поэтому метод ортогонализации выгодно применять при вычислении с фиксированной запятой. Кроме этого, $\|A_m\|_E \leqslant \|A_0\|_E$. Следовательно,

$$\|F_0\|_E \leqslant 3\,\|A_0\|_E \cdot 2^{-t} \tag{7}$$

Оценки (5), (7) весьма интересны, так как они лучше, чем оценки, полученные для других методов [1, 5, 6], и лишь множителем 3 отличаются от оценок эквивалентных возмущений, полученных при округлении коэффициентов матрицы A_0 до t знаков. Из соотношений (4), в частности, следует, что в режиме плавающей запятой метод ортогонализации в самые слабые места матрицы A_0 вкладывает самые малые ошибки. Действительно, если i-я строка A_0 близка к линейно зависимой с первыми $i-1$ строками, то все коэффициенты $a_{ij}^{(m)}$, $j = 1, 2, \ldots, n$, будут малы.

Процесс (2) может быть применен и для реализации других методов. Он позволяет уменьшать ошибки на третьем этапе и сохранять вид матрицы A_m. Особый интерес представляет применение этого процесса для тех методов, в которых при вычислении A_{p+1} к $(p+1)$-й строке прибавляется линейная комбинация первых p строк. Для таких методов справедливы оценки (3)—(6). Кроме метода ортогонализации, к ним относится, например, метод исключения по типу компактной схемы [3].

Поступила в редакцию 3.07.1967

Цитированная литература

1. J. H. Wilkinson. The algebraic eigenvalue problem. Oxford, Clarendon Press, 1965.
2. В. В. Воеводин. Об асимптотическом распределении ошибок округления при линейных преобразованиях. Ж. вычисл. матем. и матем. физ., 1967, 7, № 5, 965—976.
3. Д. К. Фаддеев, В. Н. Фаддеева. Вычислительные методы линейной алгебры. М., Физматгиз, 1963.
4. В. В. Воеводин. Численные методы алгебры. М., «Наука», 1966.
5. J. H. Wilkinson. A priori error analysis of algebraic processes. Тезисы докладов по приглашению. Междунар. конгресс математиков, Москва, 1966.
6. В. В. Воеводин. Об одном порядке исключения неизвестных. Ж. вычисл. матем. и матем. физ., 1966, 6, № 4, 758—760.

УДК 518:512.25

О МИНИМИЗАЦИИ ВЫЧИСЛИТЕЛЬНЫХ АЛГОРИТМОВ ПРИ РЕШЕНИИ ЗАДАЧИ ИСКЛЮЧЕНИЯ

Н. И. КОКОВКИН-ЩЕРБАК

(Пятигорск)

Задачу исключения (см. [1], гл. 2, § 22) можно сформулировать в следующей форме: определить значение $y = d$, удовлетворяющее системе

$$\begin{aligned} &a_{i1}x_1 + \ldots + a_{in}x_n + 0y = a_{i,n+2}, \qquad i = 1, 2, \ldots, n, \\ &-c_1x_1 - \ldots - c_nx_n + y = 0, \end{aligned} \tag{1}$$

имеющей единственное решение.