V. V. Voevodin, On methods of conjugate direction, *Zh. Vychisl. Mat. Mat. Fiz.*, 1979, Volume 19, Number 5, 1313–1317

ЖУРНАЛ
ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ

Том 19 Сентябрь 1979 Октябрь № 5

УДК 519.61

НАУЧНЫЕ СООБЩЕНИЯ

О МЕТОДАХ СОПРЯЖЕННЫХ НАПРАВЛЕНИЙ

В. В. ВОЕВОДИН

(Москва)

Исследуются обобщения методов сопряженных направлений на системы линейных уравнений с несимметричными матрицами. Указывается простое условие, которому удовлетворяют все наиболее известные варианты этих методов.

Методы сопряженных направлений [1] широко используются для решения систем линейных алгебраических уравнений

(1) $$Ax=f$$

во многих задачах вычислительной математики.

Имеется большое число математически эквивалентных форм реализации методов сопряженных направлений. Их различие оказывается весьма значительным, особенно в отношении устойчивости к ошибкам округления. Понять причину возникновения таких различий из самих расчетных формул трудно.

Классический метод сопряженных градиентов применяется, как правило, к решению систем с вещественной симметричной положительно-определенной матрицей A. Принимая во внимание общий итог исследования устойчивости многих методов решения алгебраических задач [2], можно высказать предположение, что реально вычисленное решение будет точным решением некоторой возмущенной системы, скорее всего с несимметричным возмущением. Поэтому прямое распространение метода на случай систем с комплексными неэрмитовыми матрицами может показать, какие из вариантов методов сопряженных направлений являются основными, а какие – подчиненными. Конечно, подобное расширение значительно увеличит сферу применения методов этой группы.

Пусть R – комплексная матрица порядка n. Упорядоченную систему векторов $s_1,\ldots,s_p$, $p\leqslant n$, назовем *R-псевдоортогональной*, если для $i<j\leqslant p$ выполняются условия

$$(Rs_j,\ s_i)=0,\qquad (Rs_j,\ s_j)\neq 0.$$

Любая R-псевдоортогональная система является линейно-независимой. Легко найти и разложение любого вектора x по базисной R-псевдоортогональной системе $s_1,\ldots,s_n$. Если

$$x=\alpha_1 s_1+\alpha_2 s_2+\ldots+\alpha_n s_n,$$

то для определения коэффициентов $\alpha_1,\ldots,\alpha_n$ необходимо решить систему с левой треугольной матрицей:

$$\begin{aligned}
&\alpha_1(Rs_1,\ s_1) &&=(Rx,\ s_1),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&\alpha_1(Rs_1,\ s_n)+\ldots+\alpha_n(Rs_n,\ s_n) &&=(Rx,\ s_n).
\end{aligned}$$

Из любой линейно-независимой системы векторов $r_0,\ldots,r_{p-1}$ почти всегда можно построить R-псевдоортогональную систему $s_1,\ldots,s_p$ с помощью процесса, ана-

логичного процессу ортогонализации Грамма – Шмидта. Пусть $s_1=r_0$, и предположим, что из векторов $r_0, \ldots, r_{i-1}$ уже получены векторы $s_1, \ldots, s_i$. Вектор s_{i+1} будем искать в виде

$$s_{i+1}=r_i+\sum_{k=1}^{i}\beta_{i+1,k}s_k. \tag{2}$$

Условия R-псевдоортогональности вектора s_{i+1} к векторам $s_1, \ldots, s_i$ снова дают для определения неизвестных коэффициентов $\beta_{i+1,k}$ левую треугольную систему. Если все величины (Rs_i, s_i) окажутся отличными от нуля, полученная система $s_1, \ldots, s_p$ будет R-псевдоортогональной.

Рассмотрим решение системы (1) с комплексной матрицей A порядка n. Следуя [1], возьмем $R=CAB$ для каких-либо матриц C, B, и пусть $s_1, \ldots, s_n$ – некоторая R-псевдоортогональная система. Обозначим через x_0 начальный вектор, и пусть

$$x=x_0+B\sum_{j=1}^{n}a_js_j, \qquad x_i=x_0+B\sum_{j=1}^{i}a_js_j, \qquad r_i=f-Ax_i. \tag{3}$$

Тогда из соотношений

$$x_i=x_{i-1}+a_iBs_i \tag{4}$$

вытекает, что

$$r_i=r_{i-1}-a_iABs_i. \tag{5}$$

Легко показать, что для выбранной CAB-псевдоортогональной системы $s_1, \ldots, s_n$ имеют место равенства

$$(Cr_i, s_k)=0, \qquad 1\leqslant k\leqslant i. \tag{6}$$

Действительно,

$$r_i=f-Ax_i=A(x-x_i)=\sum_{j=i+1}^{n}a_jABs_j,$$

и далее

$$(Cr_i, s_k)=\left(\sum_{j=i+1}^{n}a_jCABs_j, s_k\right)=\sum_{j=i+1}^{n}a_j(CABs_j, s_k)=0$$

для всех $k\leqslant i$.

Из (2) заключаем, что если последовательность векторов s_i получена путем CAB-псевдоортогонализации векторов r_i, то r_k есть линейная комбинация векторов $s_1, \ldots, s_{k+1}$. Следовательно, скалярное произведение (Cr_i, r_k) – линейная комбинация чисел $(Cr_i, s_1), \ldots, (Cr_i, s_{k+1})$, и оно равно нулю в соответствии с (6) для $k<i$, т. е.

$$(Cr_i, r_k)=0, \qquad k<i. \tag{7}$$

Это означает, что если $(Cr_i, r_i)\neq 0$ для всех i, то $r_0, \ldots, r_{n-1}$ есть C-псевдоортогональная система.

В общем случае векторы $s_1, \ldots, s_n$ находятся по рекуррентному соотношению (2) и для определения коэффициентов $\beta_{i+1,k}$ приходится решать систему с треугольной матрицей. Коэффициенты a_i из (4) всегда определяются очень просто. Согласно (5) – (7), имеем

$$a_i=\frac{(Cr_{i-1}, r_{i-1})}{(CABs_i, r_{i-1})}=\frac{(Cr_{i-1}, s_i)}{(CABs_i, s_i)}. \tag{8}$$

В классической схеме метода сопряженных градиентов для вещественной симметричной матрицы A среди коэффициентов $\beta_{i+1,k}$ из (2) только $\beta_{i+1,i}$ отличен от нуля. Аналогичным свойством обладает семейство матриц, удовлетворяющих условию

$$(CABC^{-1})^*=\alpha E+\beta AB \tag{9}$$

для некоторых чисел α, β. Пусть

$$s_{i+1}=r_i+f_i s_i, \tag{10}$$

причем

$$f_i=-\frac{(CABr_i, s_i)}{(CABs_{i+1}, s_i)}=-\bar{\beta}\frac{(Cr_i, ABs_i)}{(CABs_i, s_i)}. \tag{11}$$

(Здесь и ниже чертой сверху обозначаются комплексно-сопряженные величины.) Условие $(CABs_{i+1}, s_i)=0$ выполняется из-за выбора коэффициента f_i. Принимая во внимание (9), получаем для $k<i$ из (5) – (7), что

$$(CABs_{i+1}, s_k)=((CABC^{-1})Cr_i, s_k)=(Cr_i, (CABC^{-1})^* s_k)=$$

$$=(Cr_i, (\alpha E+\beta AB)s_k)=\bar{\alpha}(Cr_i, s_k)+\frac{\bar{\beta}}{a_k}\{(Cr_i, r_{k-1})-(Cr_i, r_k)\}=0.$$

Если матрица AB из (9) имеет хотя бы два различных собственных значения λ_i, λ_j, то $|\beta|=1$. Действительно, приведем матрицы левой и правой частей соотношения (9) одновременно к каноническому виду Жордана и приравняем между собой диагональные элементы. Тогда либо при некоторой нумерации собственных значений будут выполняться равенства

$$\bar{\lambda}_1=\alpha+\beta\lambda_1, \qquad \bar{\lambda}_2=\alpha+\beta\lambda_2, \tag{12}$$

где $\lambda_1\neq\lambda_2$, либо при какой-то другой нумерации собственных значений будем иметь

$$\bar{\lambda}_1=\alpha+\beta\lambda_2, \quad \bar{\lambda}_2=\alpha+\beta\lambda_3, \ldots, \quad \bar{\lambda}_k=\alpha+\beta\lambda_1, \tag{13}$$

где снова $\lambda_1\neq\lambda_2$. Исключая α в каждой из групп соотношений, заключаем, что либо $|\bar{\lambda}_1-\bar{\lambda}_2|=|\beta|\,|\lambda_1-\lambda_2|$, либо $|\bar{\lambda}_1-\bar{\lambda}_2|=|\beta|^k|\lambda_1-\lambda_2|$. Так как $\lambda_1\neq\lambda_2$, то $|\beta|=1$ в обоих случаях.

Не ограничивая существенно общности, можно считать, что $\beta=1$. Выполнения этого условия легко достичь умножением матрицы A на множитель $\beta^{1/2}$. Поэтому в дальнейшем можно было бы рассматривать класс матриц, удовлетворяющих такому условию:

$$(CABC^{-1})^*=\alpha E+AB. \tag{14}$$

Из соотношений (12), (13) следует, что число α в (14) чисто мнимое. Кроме этого, нетрудно установить некоторые свойства спектра матрицы AB из (14). Пусть λ является собственным значением AB из (14). Если $\operatorname{Im}\lambda=i\alpha/2$, то при существовании собственного значения $\bar{\lambda}$ его кратность обязательно меньше кратности λ. Если $\operatorname{Im}\lambda\neq i\alpha/2$, то наряду с λ существует собственное значение вида $\bar{\lambda}-i\alpha/2$ такой же кратности. Других ограничений на распределение собственных значений матрицы AB и на число α соотношение (14) не накладывает.

Отметим некоторые из важнейших классов матриц, для которых условие (9) выполняется: A, B, C – эрмитовы, $C=B$; CAB, C – эрмитовы; матрица C перестановочна с AB, AB – нормальная и ее спектр лежит на прямой линии; $A=M+N$, где $M=M^*$, $N=-N^*$, при этом $C=B=M^{-1}$ или $C=B=N^{-1}$.

При выполнении соотношения (9) общий процесс метода сопряженных направлений может осуществляться по следующему предписанию:

$$s_1=r_0, \qquad r_i=r_{i-1}-a_i ABs_i, \qquad s_{i+1}=r_i+f_i s_i, \qquad x_i=x_{i-1}+a_i Bs_i.$$

Здесь x_0 – произвольный начальный вектор, коэффициенты a_i, f_i вычисляются согласно (8), (11). Если обозначить $u_i=Bs_i$, то процесс будет таким:

$$u_1=Br_0, \quad r_i=r_{i-1}-a_iAu_i, \quad u_{i+1}=Br_i+f_iu_i, \quad x_i=x_{i-1}+a_iu_i,$$

при этом

$$a_i=\frac{(Cr_{i-1}, r_{i-1})}{(CAu_i, r_{i-1})}=\frac{(B^{-1*}Cr_{i-1}, u_i)}{(B^{-1*}CAu_i, u_i)},$$

$$f_i=-\frac{(B^{-1*}CABr_i, u_i)}{(B^{-1*}CAu_i, u_i)}=-\beta\frac{(Cr_i, Au_i)}{(B^{-1*}CAu_i, u_i)}.$$

В частных случаях данные формулы могут оказаться более простыми. Например, при $C=E$, $B=A^*$ имеем

$$a_i=\frac{(r_{i-1}, r_{i-1})}{(Au_i, r_{i-1})}=\frac{(r_{i-1}, r_{i-1})}{(u_i, u_i)}>0,$$

$$f_i=-\frac{(r_i, Au_i)}{(u_i, u_i)}=\frac{(r_i, r_i)}{(r_{i-1}, r_{i-1})}>0.$$

Если же $C=AA^*$, $B=A^*$, то

$$a_i=\frac{(A^*r_{i-1}, A^*r_{i-1})}{(Au_i, Au_i)}>0, \qquad f_i=\frac{(A^*r_i, A^*r_i)}{(A^*r_{i-1}, A^*r_{i-1})}>0.$$

Как уже отмечалось, последовательность векторов r_i является C-псевдоортогональной. Следовательно, невязка r_n заведомо будет нулевой, а вектор x_n – точным решением системы (1). Единственное, что может помешать реализации метода,– это обращение в нуль одного из скалярных произведений $(CABs_i, s_i)$. Однако данной ситуации всегда можно избежать путем выбора нового начального приближения x_0.

Векторы r_i и s_{i+1} – линейные комбинации векторов $r_0, ABr_0, \ldots, (AB)^ir_0$. Поэтому если в разложении вектора по каноническому базису Жордана матрицы AB присутствуют не все составляющие, то обращение невязки в нуль произойдет раньше. Процесс оканчивается особенно быстро, если матрица AB к тому же имеет простую структуру и большое число совпадающих собственных значений. Именно, если в разложении вектора r_0 по собственным векторам матрицы AB ненулевые составляющие соответствуют m попарно различным собственным значениям, то $r_m=0$.

Классический метод сопряженных градиентов является не только конечным, но и итерационным, так как на каждом шаге минимизирует функционал ошибок. Описанная группа методов в общем случае таким свойством не обладает. Но если матрица CAB эрмитова и положительно-определенная, то на каждом шаге будет минимизироваться обобщенный функционал ошибок с матрицей $B^{-1*}CA$. Доказательство этого факта ничем не отличается от соответствующего доказательства в [1].

Если матрица A вырожденная, то метод всегда позволяет получить одно из решений, а в случае положительной определенности матрицы CAB – решение с минимальным обобщенным функционалом ошибок.

Как и в методе сопряженных градиентов, в данной группе методов имеют место трехчленные соотношения для векторов s_i, r_i. Из (5), (10) вытекает, что

$$s_{i+1}=(1+f_i)s_i-a_iABs_i-f_{i-1}s_{i-1},$$

$$r_{i+1}=\left(1+\frac{f_ia_{i+1}}{a_i}\right)r_i-a_{i+1}ABr_i-\frac{f_ia_{i+1}}{a_i}r_{i-1}.$$

Отсюда можно получить и другие соотношения, например такое:

$$x_{i+1}=x_{i-1}+\omega_{i+1}(\alpha_i Br_i+x_i-x_{i-1}), \tag{15}$$

где ω_{i+1}, α_i – подходящим образом выбранные числа.

Все наиболее известные варианты методов сопряженных направлений с точностью до обозначений и порядка выполнения операций укладываются в описанную выше схему. Рассмотрим некоторые из примеров.

1. A – положительно-определенная эрмитова матрица, $C=B=E$. Классический метод сопряженных градиентов, его вещественный вариант описан в [1]. На каждом шаге минимизируется функционал ошибок с матрицей A.

2. A – произвольная матрица, $C=E$, $B=A^*$. Математически метод совпадает с методом AA^*-минимальных итераций при специальном выборе начального вектора. Вещественный вариант описан в [1]. На каждом шаге минимизируется функционал ошибок с матрицей E, т. е. евклидова норма самой ошибки.

3. A – произвольная матрица, $C=AA^*$, $B=A^*$. Математически метод совпадает с методом A^*A-минимальных итераций при специальном выборе начального вектора. Вещественный вариант описан в [1]. На каждом шаге минимизируется функционал ошибок с матрицей A^*A, т. е. евклидова норма вектора невязки.

4. A, B – эрмитовы матрицы, $C=B$. Вещественный вариант метода в форме (15) для случая положительно-определенных матриц A, B описан в [3]. На каждом шаге минимизируется функционал ошибок с матрицей A, если $A\geqslant 0$.

5. A – произвольная матрица. Если $A=M+N$, где $M=M^*$, $N=-N^*$, то $B=C=M^{-1}$ или $B=C=N^{-1}$. Вещественный вариант метода в форме (15) для случая положительно-определенной матрицы M описан в [4].

6. A – нормальная матрица, спектр которой лежит на прямой, $C=A^*$, $B=E$. Вещественный вариант метода для симметричной матрицы A в несколько иной форме описан в [5]. На каждом шаге минимизируется функционал ошибок с матрицей A^*A, т. е. евклидова норма вектора невязки.

Если матрица A прямоугольная или квадратная неэрмитова, то можно переходить к системам с неотрицательно-определенными эрмитовыми матрицами с помощью первой или второй трансформации Гаусса. Однако заметим, что в случае отказа от явного вычисления матриц A^*A и AA^* при реализации метода сопряженных градиентов расчетные формулы оказываются совпадающими с формулами для методов A^*A- и AA^*-минимальных итераций.

Поступила в редакцию 20.04.1979

Цитированная литература

1. *Д. К. Фаддеев, Н. В. Фаддеева.* Вычислительные методы линейной алгебры. М., «Наука», 1963.
2. *В. В. Воеводин.* Вычислительные основы линейной алгебры. М., «Наука», 1977.
3. *P. Concus, G. Golub, D. O'Leary.* A generalized conjugate method for the numerical solution of elliptic partial differential equations. In «Sparse Matrix Computations». New York, Acad. Press, 1976, 309–332.
4. *P. Concus, G. Golub.* A generalized conjugate method for nonsymmetric systems of linear equations. Berkeley, Univ. of California, Techn. Rept, 1975.
5. *O. Axelsson.* Solution of linear systems of equations: iterative methods. In «Sparse Matrix Techniques». New York, Springer, 1977, 1–51.