Результаты расчета: \( x(A) = 105.76; \ x_{rp} = 0.510A; \)

\[
\mathbf{H} = \begin{bmatrix}
0.5 & 0.5 & 0.5 & 0.5 \\
0.5 & 1.5 & 2.0 & 2.5 \\
0.5 & 2.0 & 4.5 & 7.0 \\
0.5 & 2.5 & 7.0 & 14.5
\end{bmatrix}
\]

Литература

1. Булгаков А. Я. Эффективно вычисляемый параметр качества устойчивости систем линейных дифференциальных уравнений с постоянными коэффициентами.— Сб. мат. журн., 1980, т. 21, № 3, с. 33—41.


6. Карачаров К. А., Пилищук А. Г. Введение в техническую теорию устойчивости динамики.— М.: Физматгиз, 1962.— 244 с.

7. Булгаков А. Я. Вычисление экспоненты от асимптотически устойчивой матрицы.— В кн.: Вычислительные методы линейной алгебры. Новосибирск: Наука, 1985, с. 4—17.

8. Булгаков А. Я., Годунов С. К. Численное определение одного из критериев качества устойчивости систем линейных дифференциальных уравнений с постоянными коэффициентами.— Новосибирск, 1982.— 56 с.— (Препринт/АН СССР, Сиб. отд-ие, ИМ).


УЧЕТ ВЫЧИСЛИТЕЛЬНЫХ ПОГРЕШНОСТЕЙ В ОДНОМ ВАРИАНТЕ МЕТОДА СОПРЯЖЕННЫХ ГРАДΙЄНТОВ

А. Я. БУЛГАКОВ, С. К. ГОДУНОВ

ВВЕДЕНИЕ

В работе рассматривается вариант метода сопряженных градиентов для решения операторного уравнения

\[ \mathbf{H} = C, \]

где \( \mathbf{H} \) — несамосопряженный линейный оператор, действующий в \( N \)-мерном евклидово пространстве. Этот вариант использовался авторами при решении матричного уравнения Ляпунова [1]. Предлагаемый алгоритм — один из возможных вариантов метода сопряженных градиентов. Следует отметить, что разные варианты метода по-разному реагируют на возмущения, возникающие при приближенном выполнении арифметических операций на ЭВМ (см., например, [2—5]). Выбор конкретного варианта сделан с учетом ряда предложенных вычислительных экспериментов.

В первых двух параграфах описывается указанный алгоритм, для которого в предположении точного вычисления всех формул выведены оценки уменьшения нормы невязки за первые \( k \) шагов и конкретно за \( k \)-й шаг. В процессе работы алгоритма строится базис, в котором оператор \( \mathbf{H} \) записывается в виде произведения ортогональной и двуходиагональной.
ной матриц. Это обстоятельство позволяет, вычисляя сингулярные числа двухдиагональной матрицы, получить в процессе решения уравнения оценки ее обусловленности $\mu(\mathbf{A})$. Параграф 3 посвящен численному экспериментальному опробованию алгоритма, где отражены все основные качественные особенности процесса. Расчеты показали, что при стандартном программировании формул (1.2) теряются за счет влияния ошибок округления два его основных свойства: (i) — невозвратие невязки за шаг процесса, (ii) — оценка обусловленности оператора числом обусловленности двухдиагональной матрицы, сформированной в результате работы первых $k$ шагов процесса. В § 4, 5 для сохранения свойства (i) предложены специальные приемы программной реализации алгоритма (дополнительные нормировки векторов и матриц, вычисления с двойной точностью, итерационные уточнения), которые позволяют существенно сократить влияние ошибок округления на окончательные результаты. Здесь же показано, что при не слишком больших $\mu(\mathbf{A})$ алгоритм сходится не хуже метода наискорейшего спуска. Для того чтобы процесс давал возможность оценить $\mu(\mathbf{A})$, вычисляются приближения наибольшего и наименьшего сингулярных чисел $\mathbf{A}$ и определяются соответствующие им приближенные сингулярные векторы. Их определение производится итерационным процессом, каждый этап которого использует двухдиагональные матрицы, возникающие в результате цикла из $k$ шагов метода сопряженных градиентов. В § 6 описан алгоритм решения операторного уравнения. В § 7 приведены результаты расчетов линейных систем § 3.

§ 1. МЕТОД СОПРЯЖЕННЫХ ГРАДИЕНТОВ

Для решения операторного уравнения

$$\mathbf{A} \mathbf{H} = \mathbf{C}$$

в $N$-мерном евклидовом пространстве предлагается итерационный процесс. Метрика подобрана таким образом, что удобно применять оператор $\mathbf{A}^*$, сопряженный к оператору $\mathbf{A}$.

Построим последовательность векторов $H_j, V_j, V_{j+1/2}, W_{j-1/2}, G_{j-1/2}, W_j, G_j, Y_j, \xi_j$ и числовых параметров $c_j, d_j, \eta_j, \rho_j, s_j, \xi_j (j = 1, 2, \ldots)$ по следующим формулам:

$$H_0 = G_0 = W_0 = 0; \quad V_j = \mathbf{A}H_{j-1} - C; \quad c_j = \|V_j\|; \quad V_{j+1/2} = \mathbf{A}^*V_j;$$

если $c_j = 0$, то процесс заканчивается;

$$b_j = \|V_{j+1/2}\|; \quad Y_j = \mathbf{A}V_{j+1/2}; \quad \eta_j = (Y_j, G_{j-1}); \quad s_{j-1} = \eta_j / b_j; \quad W_{j-1/2} = V_{j+1/2} - \eta_i W_{j-1}; \quad G_{j-1/2} = \mathbf{A}W_{j-1/2};$$

$$d_j = \|G_{j-1/2}\|; \quad W_j = d_j^{-1}W_{j-1/2}; \quad G_j = d_j^{-1}G_{j-1/2}; \quad \rho_j = d_j / b_j; \quad \xi_j = (V_j, G_j); \quad H_j = H_{j-1} - \xi_j W_j.$$ 

Из этих правил можно вывести, что при $k \geq 1$ векторы $G_k, V_{k+1/2}, H_k$ представляются следующим образом:

$$H_k = \mathbf{A}^* [c_1^{(k)} I + c_2^{(k)} [\mathbf{A}^*] + \ldots + c_k^{(k)} [\mathbf{A}^*]^{k-1}] C;$$

$$V_{k+1/2} = \mathbf{A}^* [\beta_1^{(k)} I + \beta_2^{(k)} [\mathbf{A}^*] + \ldots + \beta_k^{(k)} [\mathbf{A}^*]^{k-1}] C;$$

$$G_k = [\gamma_1^{(k)} I + \gamma_2^{(k)} [\mathbf{A}^*] + \ldots + \gamma_k^{(k)} [\mathbf{A}^*]^{k-1}] C,$$

вследствие чего

$$-\mathbf{A}H_k + C = [I - c_1^{(k)} [\mathbf{A}^*] - \ldots - c_k^{(k)} [\mathbf{A}^*]^{k}] C.$$
Лемма 1. Если процесс (1.2) при \( i = 1, 2, \ldots, n \) не прерывался, т. е. если \( c_i \neq 0 \) для всех \( i = 1, 2, \ldots, n \), то справедливы соотношения (1.5), (1.6), (1.7):

\[
(V_{i+1}, G_i) = 0 \quad \text{при всех } 1 \leq i \leq j \leq n, \tag{1.5}
\]

\[
(\tilde{V}_i, \tilde{V}_j) = \delta_{ij} \quad \text{при всех } 1 \leq i, j \leq n, \tag{1.6}
\]

\[
(G_i, G_j) = \delta_{ij} \quad \text{при всех } 1 \leq i \leq j \leq n, \tag{1.7}
\]

\( \delta_{ij} \) — символ Кронекера.

Доказательство леммы 1 приведено в конце параграфа.

Покажем, что итерационный процесс (1.2) может оборваться в том и только том случае, когда \( c_j \) — норма невязки процесса — равна нулю, т. е. когда уже получен точное решение уравнения (1.1).

Процесс не может продолжаться в случае равенства нулю нормирующих коэффициентов \( b_j \), \( d_j \). Будет показано, что пока \( c_j \neq 0 \), коэффициенты \( b_j \) и \( d_j \) в нуль не обращаются.

Из (1.2) выводится следующая цепочка неравенств:

\[
b_j = \| V_{j+1/2} \| = \| \xi V_j \| = \| \xi (\xi H_{j-1} - C) \| = \sigma(\xi) \| \xi H_{j-1} - C \| = \sigma(\xi) c_j,
\]

которая вследствие невырожденности оператора \( \xi \) позволяет заключить, что \( b_j = 0 \) лишь при условии \( c_j = 0 \).

Заметим, что если \( d_j = 0 \), то обязательно \( c_j = 0 \). Пусть \( d_j = 0 \), тогда из (1.2) следует, что

\[
0 = G_{j+1/2} = \delta V_{j+1/2} - \eta_j G_{j-1}; \quad (\delta V_{j+1/2}, V_j) = \eta_j (G_{j-1}, V_j).
\]

Осталось показать, что в силу выбора \( \xi_i \) в процессе (1.2) для всех \( j = 1, 2, \ldots, (G_{j-1}, V_j) = 0 \) и, значит, \( c_j = 0 \).

Итак, установлено, что итерационный процесс (1.2) обрывается лишь на основании получения точного решения (1.1).

Метод сопряженных градиентов, если не учитывать неизбежных вычислительных погрешностей, должен приводить к точному решению за конечное число шагов, не превышающее в нашем случае \( N \).

Опираясь на лемму 1, можно получить следующую оценку скорости сходимости (1.2) к точному решению:

\[
\| \xi (H_k - H) \| = \| \xi H_k - C \| \leq \left( (\mu(\xi) + 1)/(\mu(\xi) - 1) + (\mu(\xi) - 1)/(\mu(\xi) + 1) \right)^{-1} \cdot 2 \| C \|. \tag{1.8}
\]

Приступая к доказательству, заметим, что из формул (1.2)

\[
V_{k+1} = V_1 - \sum_{j=1}^{k} \xi_j G_j.
\]

Так как \( (V_{k+1}, G_i) = 0 \), \( \| G_i \| = 1 \) при всех \( i = 1, 2, \ldots, k \), то \( \xi_i = (V_i, G_i) \).

Легко проверить, что \( \| V_{k+1} \| \) представляет собою минимум квадратичного функционала \( (V, V) \) на множестве всех векторов вида

\[
V = V_1 - \sum_{i=1}^{k} \alpha_i G_i.
\]

В самом деле, форма \( (V, V) \) записывается как

\[
(V, V) = (V_1, V_1) - 2 \sum_{i=1}^{k} \alpha_i (V_1, G_i) + \sum_{i=1}^{k} \alpha_i^2,
\]

и ее экстремум достигается при \( \alpha_i = (V_1, G_i) = \xi_i \). Заметим, что \( V_i = C \), можно заключить, что вектор \( V_{k+1} \), полученный на \( k \)-й итерации процесса (1.2), обеспечивает минимизацию формы \( (V, V) \) на множестве всех векторов:

\[
V = V_1 - \sum_{i=1}^{k} a_i^{(h)} \{ \xi^* \} = \left\{ I - \sum_{i=1}^{k} a_i^{(h)} \{ \xi^* \} \right\} C.
\]

40
где $a^{(k)}_i — произвольные числа. Отсюда следует, что

$$\| \mathbf{H}_1 \mathbf{C} \| \leq \min_{a^{(k)}_i} \max_{\lambda} |a^{(k)}(\lambda)| \cdot |\mathbf{C}|, \quad (1.9)$$

где максимум модуля полинома

$$a^{(k)}(\lambda) = \sum_{i=0}^{k} a_i^{(k)} \lambda^i$$

берется по всем значениям $\lambda$ на отрезке

$$\sigma_1(\mathbf{C}) = \sigma_1(\mathbf{C}^*) \leq \lambda \leq \sigma_N(\mathbf{C}^*) = \sigma_N(\mathbf{C}),$$

'содержащего весь спектр оператора $\mathbf{C}$, и выбирается наименьшее возможное' значение этого максимума среди всех полиномов $a^{(k)}_i(\lambda)$ с равным единице коэффициентом $a^{(k)}_0 = 1$. Неравенства (1.9) позволяют, воспользо- зовавшись известными свойствами полиномов, наименьшее уклошающихся от нуля (см., например, [6, с. 590]), установить неравенства (1.8). Заме- чательной для процесса (1.2) является возможность получения оценок максимальных и минимальных сингулярных чисел оператора $\mathbf{H}$ во вре- мя счета.

Из формул (1.2) следуют равенства

$$d_i \mathbf{G} = \mathbf{G} \mathbf{V}_{i+1} - \eta_i \mathbf{G}_{i-1},$$

эквивалентные (ввиду определения $\hat{\mathbf{V}} = \mathbf{V}_{i+1/2} \mathbf{V}_{i-1/2}^{-1}$)

$$\mathbf{C} \mathbf{G}_i = \rho_i \mathbf{G}_i + s_{i-1} \mathbf{G}_{i-1}. \quad (1.10)$$

Равенства (1.10) означают, как уже замечено в [1], что оператор $\mathbf{C}$ представляетя в виде произведения ортогонального преобразования, пере- водящего ортонормированный базис $\hat{\mathbf{V}}_1, \hat{\mathbf{V}}_2, \ldots, \hat{\mathbf{V}}_h, \ldots$ в ортонорми- рованный базис из векторов $\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_h, \ldots$, на преобразование, задава- емое в этом последнем базисе двухдиагональной матрицей $B_N$:

$$B_N = \begin{bmatrix} \rho_1 & & & 0 \\ & \rho_2 & & \\ & & \ddots & \\ 0 & & & \rho_s \end{bmatrix}.$$

Поэтому $\sigma_1(\mathbf{C}) = \sigma_1(B_N), \quad \sigma_i(\mathbf{C}) = \sigma_i(B_N), \quad \mu(\mathbf{C}) = \mu(B_N)$. Если провести процесс (1.2) до конца, выполнив $N$ шагов, то, вычисляя сингулярные числа двухдиагональной матрицы $B_N$, сумели бы определить $\mu(\mathbf{C})$. На са- мом же деле сингулярные числа даже урезанных матрицы $B_k (k \leq N)$, которые формируются по результатам $k$ шагов процесса (1.2), позволяют получить оценки сингулярных чисел и числа обусловленности $B_N$, так как имеют место следующие, без труда обосновываемые, неравенства:

$$\sigma_i(B_{k-1}) \geq \sigma_i(B_k) \geq \sigma_i(B_N);$$

$$\sigma_N(B_k) \geq \sigma_N(B_k) \geq \sigma_{i-1}(B_{k-1});$$

$$\mu(B_{k-1}) \leq \mu(B_k) \leq \mu(B_N) = \mu(\mathbf{C}).$$

Итак, процесс (1.2) дает возможность получать оценки сингулярных чисел и, следовательно, числа обусловленности оператора $\mathbf{C}$.

Более того, возможно, используя величины процесса (1.2), вычисленные за первые $k$ шагов ($k = 2, 3, \ldots, N$), определить векторы, на которых отношение Радон оператора $\mathbf{H}$ достигает значения $\sigma_i(B_k)$ и $\sigma_i(B_N)$. Кратко описаем алгоритм получения этих векторов. Предположим, что $\sigma — суммарное число и $x = (x_1, x_2, \ldots, x_k)^T$ — соответствующий ему сингулярный вектор матрицы $B_k$ ($\sigma = \|B_k x\| / \|x\|$). Тогда на векторе $X = \begin{bmatrix} \end{bmatrix}$.
отношение Релея оператора $A$ равно $\sigma$. В самом деле, используя (1.10), находим

$$\|X\| = \left(\sum_{i=1}^{k} (x_i^2 + x_i \xi_i) + x_k^2 \rho_k^2\right)^{1/2} = \|B_k x\|;$$

где $\|X\| = \left(\sum_{i=1}^{k} x_i^2\right)^{1/2}$.

Итак, показано, что $\|X\|/\|X\| = \sigma$. Следовательно, использование $\sigma_1(B_k)$ и $\sigma_k(B_k)$ в качестве оценок сингулярных чисел оператора $\mathcal{A}$, а значит, и его числа обусловленности гарантируется указанием векторов, на которых эти оценки достигаются.

Перейдем к доказательству леммы 1. Из формул (1.2) вытекает справедливость цепочки равенств

$$(V_{i+1}, V_{j-1}) = (\mathcal{A}V_{i+1}, W_{j-1} + \eta_i W_{i-1}) = (V_{i+1}, \mathcal{A}W_{j-1} + \eta_i \mathcal{A}W_{i-1}) = (V_{i+1}, d_i G_j + \eta_i G_{j-1}),$$

из которой, в предположении соотношений (1.5), получаем справедливость соотношений (1.6).

Прежде чем переходить к доказательству (1.5), (1.7), следуя работе [2], заметим, что из формул (1.2) вытекает справедливость соотношений

$$\eta_k = (\mathcal{A}V_k, G_{k-1}); \quad d_k G_k = \mathcal{A}V_k - \eta_k G_{k-1};$$

$$\xi_k = (V_k, G_k); \quad V_{k+1} = \xi_k G_k,$$

из которых нетрудно заключить, что выбор $\xi_k$ и $\eta_k$ обеспечивает выполнение равенств

$$(G_k, G_{k-1}) = 0; \quad V_{k+1} = \xi_k G_k.$$

В процессе выполнения (1.2) все $\xi_k \neq 0$ и справедлива следующая цепочка:

$$\xi_k = (V_k, G_k) = d_k^{-1} [(V_k, \mathcal{A}V_k) - \eta_k (V_k, G_k)] =$$

$$= d_k^{-1} (V_k, \mathcal{A}V_k) \geq d_k^{-1} \xi_k \left(\mathcal{A}V_k\right) = d_k^{-1} \xi_k \mathcal{A}V_k > 0.$$
и покажем, что они останутся справедливыми в случае, если $k$ заменить на $k + 1$.

Прежде всего отметим, что $(G_{h+1}, G_h) = 0$ в силу выбора числа $\eta_{h+1}$ на $(k+1)$-й итерации. Докажем далее, что $(G_{k+1}, G_j) = 0$ для всех $1 \leq j \leq k - 1$. В самом деле,

$$ (G_{k+1}, G_j) = d_k^{-1} (\mathcal{G}*V_{h+1} - \eta_{k+1} G_h, G_j) = d_k^{-1} (\mathcal{G}*V_{h+1}, G_j), $$

так как $(G_h, G_j) = 0$ по предположению индукции. Подставляя сюда $\mathcal{G}*G_j$ из формул (1.13):

$$ \mathcal{G}*G_j = \alpha_j G_{j+1} + \beta_j G_j + \gamma_j G_{j-1} $$

и учитывая, что по предположению индукции при любом $1 \leq j \leq k - 1$

$$(V_{k+1}, G_{h+1}) = (V_{k+1}, G_j) = (V_h, G_{j-1}) = 0,$$

получаем требуемое равенство $(G_{k+1}, G_{j+1}) = 0$.

Приведенное доказательство «прямо не проходит» при $j = 1$, однако соответствующие изменения рассуждений очевидны и на них не будем останавливаться. Остается доказать, что $(V_{k+2}, G_j) = 0$ при всех $1 \leq j \leq k + 1$. Напомним, что $V_{k+2} = V_{k+1} - \xi_{k+1} G_{k+1}$. Поэтому при $1 \leq j \leq k$ имеем

$$(V_{k+2}, G_j) = (V_{k+1}, G_j) - \xi_{k+1} (G_{k+1}, G_j) = 0,$$

так как $(V_{k+1}, G_j) = 0$ по предположению индукции, а $(G_{k+1}, G_j) = 0$, как только что доказано. Выполнение равенства $(V_{k+2}, G_{k+1}) = 0$ обеспечивается выбором $\xi_{k+1}$ на $(k+1)$-й итерации. Проведение шага индукции полностью завершено, а с ним и доказательство леммы 1.

§ 2. ШАГ МЕТОДА СОПРЯЖЕННЫХ ГРАДИЕНТОВ

Продолжим рассмотрение метода сопряженных градиентов. Будут выведены несколько вспомогательных неравенств и оценка уменьшения нормы невязки $\|V_h\|$ за $k$-й шаг.

Из формул (1.2) получаем соотношения

$$ \eta_k = (\mathcal{G}*V_h, G_{h+1}); $$
$$ G_{k-1/2} = \mathcal{G}*V_h - \eta_k G_{h-1}; $$
$$ d_k = \|G_{k-1/2}\|; $$
$$ G_k = d_k^{-1} G_{k-1/2}; $$
$$ V_{k+1} = (V_h, G_h); $$
$$ V_{k+1} = V_h - \xi_k G_h. $$

Мы вправе считать, что $(V_h, G_{h-1}) = 0$. Это обеспечивается выбором $\xi_k$. Из (2.6) и (2.5) следует, что

$$ \|V_{k+1}\| \leq \|V_k\| $$

и, значит, процесс (2.1) — (2.6) является процессом невозрастания невязки. Более того, опираясь на условие $(V_h, G_{h-1}) = 0$, для процесса (2.1) — (2.6) можно вывести оценку

$$ \|V_{k+1}\| \leq \left[ \frac{1}{\mu} \left( \frac{1}{\mu^2(\epsilon)} + 1 \right) \right] \|V_h\| - \eta_k \frac{(V_h, \mathcal{G} V_h)^2}{\|\mathcal{G} V_h\|^2}. $$

Рассмотрим пучок векторов

$$ z = V_h + \alpha \mathcal{G}*V_h + \beta G_{h-1}. $$
Тогда для любых $\alpha$, $\beta$ выполнено неравенство $\|z\| \geq \|V_{h+1}\|$, что следует из равенств

$$
\frac{\partial \|z\|^2}{\partial \alpha} \bigg|_{z=V_{h+1}} = 2 (V_{h+1}, \xi \xi*V_k) = 0;
$$
$$
\frac{\partial \|z\|^2}{\partial \beta} \bigg|_{z=V_{h+1}} = 2 (V_{h+1}, \xi \xi*V_k) = 0.
$$

Кроме того, имеет место цепочка равенств

$$
\|z\|^2 = 2 \beta (z, G_{h-1}) + \beta \|G_{h-1}\|^2 = \|V_k + \alpha \xi \xi*V_k\|^2;
$$

$$
(z, G_{h-1}) = (V_h, G_{h-1}) + \alpha (\xi \xi*V_h, G_{h-1}) + \beta = \alpha (\xi \xi*V_h, G_{h-1}) + \beta,
$$

которая позволяет заключить, что $(z, G_{h-1}) = 0$ при условии $\beta = -\alpha (\xi \xi*V_h, G_{h-1}) = -\alpha \eta_h$. Тогда при $\alpha_0 = (V_h, \xi \xi*V_h)/\|\xi \xi*V_h\|^2$; $\beta_0 = -\alpha_0 \eta_h$ из (2.9) получаем

$$
\|V_{h+1}\|^2 \leq \|z_0\|^2 = \|V_h + \alpha_0 \xi \xi*V_h\|^2 - \beta_0^2 \leq
$$

$$
\leq \left[ \left( \mu^2 (\xi^2) - 1 \right)/\left( \mu^2 (\xi^2) + 1 \right) \right]^2 \|V_h\|^2 - \eta_h (V_h, \xi \xi*V_h)/\|\xi \xi*V_h\|^2.
$$

Доказанная оценка сходимости метода (2.1) — (2.6), а следовательно, и метода (1.2) сильнее оценки уменьшения нормы невязки метода наискорейшему спуску.

В заключение отметим два несложных утверждения

$$
\|\xi \xi*V_h\|^2 = \|G_{h-1/2}\|^2 + \eta_h,
$$

$$
(\xi \xi*V_h, V_h)/\|V_h\|^2 \leq \|G_{h-1/2}\| \leq \|\xi \xi*V_h\|.
$$

(2.10) (2.11)

Равенство (2.10) непосредственно следует из формул (2.1) — (2.2). Перейдем к выводу (2.11).

Выбор $\xi_{k-1}$ обеспечивает выполнение равенства $(V_k, G_{k-1}) = 0$. Умножив (2.2) на $V_k$ и пользуясь равенством $(V_k, G_{k-1}) = 0$, находим

$(G_{h-1/2}, V_h) = (\xi \xi*V_h, V_h)$.

Отсюда вместе с (2.10) получаем справедливость неравенства (2.11).

§ 3. ВЛИЯНИЕ ОШИБОК ОКРУГЛЕНИЯ

В МАШИННОЙ РЕАЛИЗАЦИИ МЕТОДА СОПРЯЖЕННЫХ ГРАДИЕНТОВ

Приведены результаты численного опробования метода сопряженных градиентов (см. § 1, 2) при решении на ЭВМ систем линейных уравнений $Lu = f$. Для этого применялось стандартное программирование формул (1.2). Чтобы итерации не прерывались на $k$-м шаге даже в случае получения точного решения уравнения $Lu = f$, при $d_h = 0$ задавалось $d_h$ равным машинной константе $\epsilon_1$. (1 + $\epsilon_1$ — наименьшее машинное число, близкое к 1).

Пример 1. В качестве исходных данных выбраны матрица $L$ и вектор $f$:

$$
f = (-1.1364325, -0.8070552, -1.2363081, -0.7271438),
$$
$$
L = \begin{bmatrix}
0.2317832 & -0.2017895 & -0.4012386 & -0.0405040 \\
-0.1206849 & 0.1389269 & 0.0378899 & -0.0242303 \\
-0.9645644 & -0.0906907 & -0.0957022 & 0.0186665 \\
0.6364325 & -0.6570810 & 0.1444403 & -0.0408918
\end{bmatrix}
$$

Сингулярные числа матрицы $L$ известны: $\sigma_1 (L) \approx 1$; $\sigma_3 (L) \approx 0.7071067$; $\sigma_2 (L) \approx 0.4386913$; $\sigma_1 (L) \approx 0.25$. 44
В процессе расчета получены значения невязки $c_k = \|L u_{k-1} - f\|$ ($k = 1, 2, 3, 4, 5, 6$) и вычислены сингулярные числа двухдиагональных матриц $B_m$ ($m = 3, 4, 5, 6$), формируемых на $m$-м шаге процесса.

Получены результаты:

- $c_1 = 2; c_2 = 1.2705; c_3 = 0.8603; c_4 = 0.4862; c_5 = 10^{-11}; c_6 = 2 \times 10^{-7}$;
- $\sigma_1(B_3) = 0.3618; \sigma_2(B_4) = 0.6867; \sigma_3(B_5) = \sigma_4(L) \cdot (1 - 3 \cdot 10^{-4})$;
- $\sigma_4(B_4) = \sigma_5(L); \sigma_5(B_5) = \sigma_6(L); \sigma_6(B_6) = \sigma_7(L)$;
- $\sigma_1(B_6) = 6 \times 10^{-2}; \sigma_2(B_7) = \sigma_3(L) (1 + 0.1 \cdot 2); \sigma_4(B_8) = \sigma_2(L) (1 + 0.1 \cdot 4)$;
- $\sigma_5(B_8) = \sigma_1(L) (1 + 0.1 \cdot 8); \sigma_6(B_8) = \sigma_5(L) (1 + 0.1 \cdot 6)$;
- $\sigma_7(B_8) = \sigma_6(L) (1 + 0.1 \cdot 8); \sigma_8(B_8) = 4.2699.91$.

Пример 2. В качестве $L$ взята матрица

$$L = \begin{bmatrix}
0.0331119 & -0.0003639 & -0.3657_{10} & -0.107_{10} & -6 \\
-0.0172407 & 0.0002506 & 0.3450_{10} & -0.616_{10} & -6 \\
-0.1377948 & -0.0001635 & -0.8720_{10} & -0.470_{10} & -8 \\
0.0052046 & -0.0011852 & 0.1317_{10} & -0.403_{10} & -6
\end{bmatrix}$$

Правая часть задается вектором $f$:

$$f = (-1.1364325, -0.8070552, -1.2360811, -0.7271348).$$

Известны сингулярные числа матрицы $L$:

- $\sigma_1(L) \approx 0.142; \sigma_2(L) \approx 0.00127; \sigma_3(L) \approx 0.3999_{10} - 5; \sigma_4(L) \approx 0.635_{10} - 6$.

В результате расчета получено нарушение одного из основных свойств процесса (1.2) — невозрастание нормы невязки. Приведем значения норм невязок за первые 6 шагов ($c_k = \|L u_{k-1} - f\|$; $c_1 = 2; c_2 = 1.732; c_3 = 1.414; c_4 = 316.144; c_5 = 132.202; c_6 = 0.974$). Начиная с 6-й итерации норма практически стабилизировалась (постоянно убывая с 6-й итерации, на 80-й итерации она достигла значения $c_{80} = 0.973$).

Итак, приведенные примеры показывают, что если для реализации алгорита метода сопряженных градиентов использовать -стандартное программирование формулу, то влияние ошибок округления приводит к тому, что теряются два основных свойства алгорита:

1) невозрастание невязки за шаг процесса (1.2);
2) оценка $\mu(2)$ числом обусловленности двухдиагональной матрицы $B_m$, сформированной по результатам первых $k$ шагов процесса.

§ 4. ПОГРЕШНОСТИ МАШИННЫХ ВЫЧИСЛЕНИЯ

Прежде чем перейти к анализу влияния ошибок округления на процесс (1.2), изучим, опираясь на работы [7—9], ошибки округления, возникающие при выполнении операций процесса (1.2) с использованием «арифметики вычисленных порядков» (см. § 2 из [9]). Здесь, как и в [9], для формального описания арифметики используются операторы $\mathfrak{R}$ и $\mathcal{P}$, заданные на пространстве векторов длины $N$, и операторы $\mathfrak{w}$, $\mathfrak{p}$, заданные на поле вещественных чисел.

В § 2 [9] показано, что использование «арифметики вычисленных порядков» позволяет вычислить нормировку вектора $x$ и разность векторов $x$ и $\alpha y$ ($\alpha$ — скаляр), с точностью:

$$\|\alpha(\|x\| \cdot x)_{y, \mathfrak{w}} - \|x\| \cdot x\| \|x\| \cdot x\| + 1,01e_1\|x\| - 1,01e_1\|x\|$$

(4.1)
если \((2N^{1/2} + 2N^{1/2} + 2\sqrt{N}/\gamma) \varepsilon_1 / \varepsilon_2 < 0.01\), где \(\varepsilon_1\) — машинная постоянная (1 + \(\varepsilon_1\) — наименьшее машинное число, превосходящее 1). Пусть \(\gamma\) — основание системы счисления используемой ЭВМ и \(\varepsilon_2\) — машинная постоянная такая, что \(1/\gamma^2 (1 - \varepsilon_1 / \gamma) \varepsilon_2\) — наименьшее по модулю отличное от нуля машинное число.


\[
| (x, y) - [(x, y)]_{\text{выч}} | \leq 1.01 \varepsilon_1 \|x\| \cdot \|y\|. \tag{4.3}
\]

Оно вытекает из более общего неравенства (см. § 21 [7])

\[
| (x, y) - [(x, y)]_{\text{выч}} | \leq \varepsilon_1 |(x, y)| + N\varepsilon_2^2 / \gamma \|x\| \cdot \|y\| + N\varepsilon_2 / \gamma^2, \tag{4.4}
\]

полученного для случая накопления скалярного произведения векторов с двойной точностью.

Итак, осталось оценить погрешности выполнения двух операций: вычисления образа оператора \(H\) и вычисления невязки.

Предположим, что погрешности этих операций оцениваются неравенствами

\[
\| (2H)_{\text{выч}} - 2H \| \leq 2\varepsilon_1 \|2H\| \tag{4.5}
\]

и

\[
\| (2H - C)_{\text{выч}} - (2H - C) \| \leq 2\varepsilon_1 \|2H - C\|. \tag{4.6}
\]

В [1] выведено, что при достаточно слабых ограничениях на размерность матричного оператора Ляпунова и его числа обусловленности можно гарантировать выполнение неравенств (4.5) и (4.6). Здесь мы выведем подобные оценки и в случае, если оператор \(H\) задается матрицей \(L\) размерности \(N \times N\).

Пусть \(L_i\) — \(i\)-я строка матрицы \(L\). Тогда, используя неравенство (4.4), можно записать, что

\[
| [(L_i, H)]_{\text{выч}} - (L_i, H) | \leq \varepsilon_1 |(L_i, H)| + N\varepsilon_2^2 / \gamma \|L_i\| \cdot \|H\| + N\varepsilon_2 / \gamma^2 \varepsilon_2.
\]

Следовательно,

\[
\| (LH)_{\text{выч}} - LH \| \leq \varepsilon_1 \|LH\| + N^{3/2} \varepsilon_2^2 / \gamma \|L\| \cdot \|H\| + N^{3/2} / \gamma^2 \varepsilon_2.
\]

Если максимальный по модулю элемент вектора \(H\) лежит в интервале от \(1/\gamma\) до 1 (что справедливо для матрицы с вычисленным порядком максимального элемента), то, принимая во внимание неравенство \(\sigma_1(L) \geq \|H\| / \|LH\|\), последнюю оценку можно упростить:

\[
\| (LH)_{\text{выч}} - LH \| \leq \varepsilon_1 \|LH\| \left[ 1 + N^{3/2} \varepsilon_2^2 / \gamma \cdot \mu(L) + N^{3/2} / (\gamma \sigma_1(L) \varepsilon_2) \right].
\]

Итак, если выполнены условия

\[
2N^{3/2} \varepsilon_2^2 / \gamma \mu(L) < 1; 2N^{3/2} \varepsilon_2^2 / (\gamma \sigma_1(L) \varepsilon_2) < 1, \tag{4.7}
\]

tо выполнено неравенство

\[
\| (LH)_{\text{выч}} - LH \| \leq 2\varepsilon_1 \|LH\|. \tag{4.8}
\]

Для определения невязки \(LH - C\) можно использовать технологию вычисления с двойной точностью, подробно описанную в § 16 [8]. Смысл ее заключается в накоплении с двойной точностью следующих сумм \((j = 1, 2, \ldots, N)\):

\[
V_j = (LH)_j + C_j = (L_j, H) + 1 \cdot C_j.
\]

Это позволяет утверждать выполнение неравенств \((j = 1, 2, \ldots, N)\):

\[
| (V_j)_{\text{выч}} - V_j | \leq \varepsilon_1 |V_j| + (N + 1) \varepsilon_1^2 / \gamma \sqrt{|L_j|^2 + C_j^2} \sqrt{|H|^2 + \bar{1} + (N + 1) / \gamma^2 \varepsilon_2}.
\]
и, значит,
$$\|V\| \leq e_1 \|V\| + (N + 1) \varepsilon \sqrt{N} \|L\|^2 + \|C\|^2 \sqrt{H}^2 + 1 + (N + 1)^{3/2}/\gamma^2 e_2.$$ 

Заметим, что поскольку \(1/\gamma < \|C\| < \sqrt{N}\) и \((N + 1)^{3/2}/\gamma^2 e_2 < (N + 1)^{3/2}/\gamma e_2 \|C\|\), то, предположив, что \(\|C\|/\|V\| < \delta_1\), (4.9)

где \(\delta\), выбирается удовлетворяющим неравенству \(2(N + 1)^{3/2}/\gamma^2 e_2 \|\gamma e_2 \delta\| < 1\), можно гарантировать выполнение оценки \((N + 1)^{3/2}/\gamma^2 e_2 < 1/2e_1\|V\|\).

Аналогично, воспользовавшись очевидной цепочкой неравенств
$$\|V\| + \sqrt{N} \|L\|^2 + \|C\| \leq \sqrt{N} \|H\| + \|L\| + \|C\| \leq \|V\| + \sqrt{N} \|L\|^2 + \|C\| \leq \|V\|,$$

можем, взяв в (4.9) \(\delta\), удовлетворяющим неравенству
$$2(N + 1) \varepsilon \sqrt{N} \|V\|/(\|C\| + \sqrt{N} \|L\|)(1 + \delta_1) + (1 + \sqrt{N} \|L\|)/(\gamma \delta_1) < 1,$$

записать, что \(\|LH - C\| < 2\varepsilon \|LH - C\|\).

§ 5. УЧЕТ ВЫЧИСЛИТЕЛЬНЫХ ПОГРЕШНОСТЕЙ ПРИ РАСЧЕТЕ ШАГА МЕТОДА СОПРЯЖЕННЫХ ГРАДΙЕНТОВ

Рассмотрим влияние ошибок округления на уменьшение нормы невязки за шаг метода сопряженных градиентов (1.2). При вычислениях на машине по формулам (1.2) неизбежны погрешности. Для их учета достаточно предположить, что векторы и скаляры в машинной реализации процесса (1.2) связаны между собой соотношениями

$$V_k = \varphi H_{k-1} - C + \varphi_1; \quad \alpha_k = \|V_k\| + \alpha_1; \quad V_{k+1/2} = \varphi \star V_k + \varphi_2;$$
$$b_k = \|V_{k+1/2}\| + \alpha_2; \quad V_k = \varphi \star V_{k+1/2} + \varphi_3; \quad \eta_k = (Y_k, G_{k-1}) + \alpha_3;$$
$$s_k = \eta_k/b_k + \alpha_4; \quad W_{k+1/2} = V_{k+1/2} - \eta_k W_{k+1/2} + \varphi_4;$$

(5.1)

$$G_{k-1/2} = \varphi W_{1/2} - \varphi_5; \quad d_k = \|G_{k-1/2}\| + \alpha_5; \quad W_k = d_k^{-1}W_{k-1/2} + \varphi_6;$$
$$G_k = d_k^{-1}G_{k-1/2} + \varphi_7; \quad \rho_k = d_k/b_k + \alpha_6; \quad \xi_k = (V_k, G_k) + \alpha_7;$$

$$H_k = H_{k-1} - \xi_k W_k + \varphi_8;$$

в которых \(\varphi_i; \eta_j; (i = 1, 2, \ldots, 8; \quad j = 1, 2, \ldots, 7)\) обозначают погрешности, допущенные при вычислениях по формулам (1.2). Оценки \(\|\varphi\|\) и \(\|\alpha\|\) зависят от конкретно выбранного способа расчета формул процесса (1.2). Будем рассматривать ошибки округления, возникающие при вычислениях формул (1.2) в "арифметике вынесенных порядков" и при накоплении скалярных производных с двойной точностью. Расчет формул (1.2) подробно представлен в § 4. Считая выполненными условия (4.5) — (4.6), приведем оценки из § 4 для некоторых \(\|\varphi\|, \|\alpha\|\) процесса (5.1):

$$\|\varphi_1\| \leq 2\varepsilon, \|V_k\|; \quad \|\varphi_2\| \leq 2\varepsilon, \|\varphi \star V_k\|;$$
$$\|\varphi_3\| \leq 2\varepsilon, \|V_{k+1/2}\|; \quad \|\alpha_1\| \leq 2\varepsilon, \|V_k\| = 2\varepsilon, \|V_{k+1/2}\|;$$
$$\|\varphi_4\| \leq 2\varepsilon, \|V_{k+1/2}\| + 1,01\varepsilon, \|V_{k+1/2} - \eta_k W_{k+1/2}\|;$$

(5.2)

$$\|\varphi_5\| \leq 2\varepsilon, \|\varphi \star W_{k-1/2}\|; \quad \|\alpha_3\| \leq 2\varepsilon, \|G_{k-1/2}\|;$$
$$\|\varphi_6\| \leq 2\varepsilon, \|d_k^{-1}W_{k-1/2}\|; \quad \|\varphi_7\| \leq 2\varepsilon;$$
$$\|\alpha_5\| \leq 2\varepsilon, \|V_k\|; \quad \|\varphi_8\| \leq 1,01\varepsilon, \|\xi_k W_k\| + 1,01\varepsilon, \|H_{k-1} - \xi_k W_k\|.$$
тогда справедлива оценка
\[ \|V_{k+1}\| \leq \|V_k\| + 506_1 + 7\sigma_2 \| \mu(8) e_1 \| V_k \|, \]  
(5.4)
выводу которой посвящен этот параграф.

Оценка (5.4) показывает, что при выполнении (5.3) на шаге метода сопряженных градиентов норма невязки убывает не хуже, чем в методе найкращего спуска. Если же введено одно из условий (5.3) (что легко проверяется в процессе расчета), то предлагается останавливать процесс, предварительно заполнив приближенное решение \( H_{k-1} \), а затем переходять к его уточнению, а именно для решения уравнения \( \delta H = C \), где \( C = \delta H_{k-1} - C \), использовать рассмотренный метод сопряженных градиентов. В результате для синтеза сопряженных градиентов и итерационного уточнения справедлива оценка (5.4) уменьшения нормы невязки за шаг.

Перейдем к выводу оценки (5.4) при условии (5.3). Из формул (5.1) вытекает справедливость следующих соотношений:

\[ \eta_k = (\delta \delta \* V_k, G_{k-1}) + \beta_1; \quad G_{k-1} = \delta \delta \* V_k - \eta_k G_{k-1} + \psi_1; \]  
(5.5)

\[ d_k = \| G_{k-1/2} \| + \beta_2; \quad G_k = d_k^{-1} G_{k-1/2} + \psi_2; \]

\[ \xi_k = (V_k, G_k) + \beta_3; \quad V_{k+1} = V_k - \xi_k G_k + \psi_3 \]

с явными выражениями для \( \beta_i, \psi_i \):

\[ \beta_1 = (\delta \delta \* G_{k-1}) + (\psi_k, G_{k-1}) + \alpha_5; \]

\[ \beta_2 = \alpha_5; \quad \beta_3 = \alpha_5; \quad \psi_2 = \psi_1; \]

\[ \psi_1 = \delta \delta \* \eta_k (G_{k-1} - \delta \delta W_{k-1}) + \delta \delta \* \psi_4 + \psi_5; \]

\[ \psi_3 = \psi_1 - \xi_k (\delta \delta \* G_k) + \psi_4. \]  
(5.6)

Здесь \( \tilde{\psi}_i \) — погрешность вычисления невязки, полученной в результате работы \((k - 1)\)-го шага процесса (5.1).

Отметим, что, производя оценки, будем пренебрегать членами малости порядка \( e_1^2 \) по сравнению с членами порядка \( e_1 \).

Оценки для \( \psi_1, \beta_i \) \((i = 1, 2, 3)\) нетрудно получить из (5.6), (5.2), зная оценки \( \| G_k - \delta \delta W_k \|, |\eta_k|, \| G_{k-1} - \delta \delta W_{k-1} \|, |\xi_k| \), к выводу которых сейчас и перейдем.

Из (5.5) следует:

\[ |\eta_k| \leq \| \delta \delta \* V_k \| + |\beta_1|; \]  
(5.7)

\[ |\xi_k| \leq \| V_k \| + |\beta_2|. \]  
(5.8)

Наконец, показаем, что

\[ \| \delta \delta W_k - G_k \| \leq 3\mu(\delta) e_1. \]  
(5.9)

В самом деле, из (5.1) имеем ценочку равенств

\[ G_k = d_k^{-1} G_{k-1/2} + \psi_2 = d_k^{-1} (\delta \delta W_{k-1/2} + \psi_2) + \psi_2 = \delta \delta W_k - \delta \delta \* \psi_4 + d_k^{-1} \psi_2 + \psi_2, \]

позволяющую заключить, следуя (5.2), что

\[ \| \delta \delta W_k - G_k \| \leq \| d_k^{-1} \psi_2 \| + \| \psi_2 \| + \| \psi_4 \| \leq 2e_1/(1 - e_1) + 2e_1 + 2\mu(\delta) e_1. \]

Полученное неравенство гарантирует справедливость (5.9) при условии \( \mu(\delta) > 4 \). Если \( \mu(\delta) < 4 \), то влияние ошибок округления незначительно и процесс близок по поведению к точному методу сопряженных градиентов.

Итак, из (5.2) и (5.6)—(5.9) следует справедливость оценок

\[ |\psi_1| \leq 4e_1 \| \delta \delta \* V_k \| + 2\mu(\delta) e_1 |\eta_k| + 2\mu(\delta) e_1 \| G_{k-1/2} \|; \]

\[ |\beta_1| \leq 2e_1 \| \delta \delta \* V_k \| + 4e_1 \| \delta \delta \* V_k \|; \]

\[ |\beta_3| \leq 2e_1 \| G_{k-1/2} \|; \quad |\beta_3| \leq 2e_1 \| V_k \|; \]

\[ |\psi_2| \leq 2e_1; \quad |\psi_3| \leq 5\mu(\delta) e_1 \| V_k \| + 7\mu(\delta) e_1 \| C \|. \]  
(5.10)
Таким образом, вопрос уменьшения нормы невязки процесса (5.1), (5.2) сводится к вопросу уменьшения нормы невязки процесса (5.5) при условии (5.10). Векторы и скаляры, определяемые на k-м шаге метода сопряженных градиентов, можно считать близкими (степень их близости будет вычислена ниже) соответствующим величинам, вычисляемом по известным
\[ V_k = V_{k+1} \cdot G_{k+1} = G_{k+1} - (G_{k+1} \cdot V_{k+1})/\|V_{k+1}\|^2 \cdot V_{k+1} \]  
(5.11)
полученным формулам
\[ \eta_k = (\mathcal{E} \mathcal{V} \cdot G_{k+1}) / \|V_{k+1}\|^2 \cdot V_{k+1} \]
(5.12)
отметим, что \((V_{k+1}, G_{k+1}) = 0\) в силу (5.11). Следовательно, из сказанного в § 2 имеем оценку
\[ \|V_{k+1}\| = \left[ \mu^2(\mathcal{E}) - 1 \right] / \left[ \mu^2(\mathcal{E}) + 1 \right] \cdot \|V_{k+1}\|, \]
и для доказательства справедливости (5.4) при условии (5.3) осталось показать, что если
\[ \|C\| \cdot \|V_k\| \leq \delta_1; \quad |\eta_k| / d_{k+1} \leq \delta_2, \]
то
\[ \|V_{k+1} - V_k\| \leq \mu(\mathcal{E}) e_1 (50\delta_1 + 7\delta_2) \|V_k\|. \]
(5.13)
Для выяснения степени близости векторов \(V_{k+1}, V_k\) (оценка \(\|V_{k+1} - V_k\|\)) требуется оценить близость векторов и скаляров в (5.5) и (5.12).

Предварительно заметим, что из (5.5) и (5.10) следует
\[ |(V_{k+1}, G_{k+1})| \leq 5.5 \mu(\mathcal{E}) e_1 \|V_{k+1}\| + \mu(\mathcal{E}) e_1 \|C\|, \]
откуда
\[ |(V_{k+1}, G_{k+1})| \leq 5.5 \mu(\mathcal{E}) e_1 \|V_{k+1}\| + \mu(\mathcal{E}) e_1 \|C\|; \]
(5.14)
Полученные неравенства позволяют, используя (5.11), заключить, что
\[ \|G_{k+1} - G_k\| \leq \|V_{k+1}\| \cdot \left\{ \|5.5 \mu(\mathcal{E}) e_1 \|V_{k+1}\| + \mu(\mathcal{E}) e_1 \|C\| \right\}. \]
(5.15)
Далее, опираясь на (5.10), (5.11), (5.15) и равенство
\[ \eta_k - \eta_k = \beta_1 + (\mathcal{E} \mathcal{E} \mathcal{E} \mathcal{V} \cdot \bar{G}_{k+1} - G_{k+1}), \]
нетрудно вывести оценку
\[ |\eta_k - \eta_k| \leq 2e_1 \|\mathcal{E}\| \cdot \|\mathcal{E} \mathcal{E} \mathcal{E} \mathcal{V} \cdot \| + 4e_1 \|\mathcal{E} \mathcal{E} \mathcal{E} \mathcal{V} \cdot \| + |(G_{k+1} - G_k) - (G_{k+1} - G_k)\| / \|V_{k+1}\|^2 \leq \]
\[ \leq 2e_1 \|\mathcal{E}\| \cdot \|\mathcal{E} \mathcal{E} \mathcal{E} \mathcal{V} \cdot \| + 4e_1 \|\mathcal{E} \mathcal{E} \mathcal{E} \mathcal{V} \cdot \| + \|\mathcal{E} \mathcal{E} \mathcal{E} \mathcal{V} \cdot \|^2 / \|V_{k+1}\|^2 \|5.5 \mu(\mathcal{E}) e_1 \|V_{k+1}\| + \mu(\mathcal{E}) e_1 \|C\| + \]
\[ + 6.5 \mu(\mathcal{E}) e_1 \|C\| \cdot \|\mathcal{E} \mathcal{E} \mathcal{E} \mathcal{V} \cdot \| / \|V_{k+1}\|^2. \]
(5.16)
Аналогично, используя равенство
\[ G_{k+1} = G_{k+1} - \bar{G}_{k+1} \cdot \eta_k G_{k+1} - \eta_k G_{k+1} - \psi_k = \]

\[ = (\eta_k - \bar{\eta}_k) G_{k+1} - \eta_k (G_{k+1} - G_{k+1}) - \psi_k. \]
вытекающее из (5.5) и (5.12), и опираясь на неравенства (5.6), (5.15) — (5.16), легко найти оценку
\[ \|G_{k+1} - G_{k+1}\| \leq 2e_1 \|\mathcal{E}\| \cdot \|\mathcal{E} \mathcal{E} \mathcal{E} \mathcal{V} \cdot \| + 4e_1 \|\mathcal{E} \mathcal{E} \mathcal{E} \mathcal{V} \cdot \| + \]
\[ + 6.5 \mu(\mathcal{E}) e_1 \|\mathcal{E} \mathcal{E} \mathcal{E} \mathcal{V} \cdot \|^2 / \|V_{k+1}\|^2 + |\eta_k| / \|V_{k+1}\|, \]

4 Заказ № 547
49
из которой непосредственно получаем

$$
\| \mathcal{G}_{k-1/2} - \mathcal{G}_{k-1/2} \| / \max \{ \| \mathcal{G}_{k-1/2} \|, \| \mathcal{G}_{k-1/2} \| \} \leq
$$

$$
\leq 6,5 \mu (\varepsilon) e \| \mathcal{C} \| / \| \mathcal{V}_k \| + (2e, \| \mathcal{V}_k \| \cdot \| \mathcal{V}_k \| + 4e, \| \mathcal{V}_k \| + \| \eta \| / \| \mathcal{V}_k \| \cdot 6,5 \mu (\varepsilon) e \| \mathcal{C} \| / \max \{ \| \mathcal{G}_{k-1/2} \|, \| \mathcal{G}_{k-1/2} \| \}).
$$

(5.17)

При выводе (5.17) использовалось неравенство

$$
2 \| \mathcal{V}_k \| / \| \mathcal{V}_k \| \leq \| \mathcal{G}_{k-1/2} \|,
$$

которое было выведено в § 2.

Перейдем к оценке близости векторов $\mathcal{G}_k$ и $\mathcal{G}_k'$. Заметим, что в силу (5.5) и (5.12) справедливы равенства

$$
\mathcal{G}_k - \mathcal{G}_k = \mathcal{G}_k - \mathcal{G}_k - \mathcal{G}_k - \mathcal{G}_k + \mathcal{V}_k =
$$

$$
= \| \mathcal{G}_{k-1/2} \|^{-1} \mathcal{G}_{k-1/2} - \mathcal{G}_{k-1/2} - \mathcal{G}_{k-1/2} - \mathcal{G}_{k-1/2} + \mathcal{V}_k,
$$

откуда следует, что

$$
\| \mathcal{G}_k - \mathcal{G}_k \| \leq 2 \| \mathcal{G}_{k-1/2} - \mathcal{G}_{k-1/2} \| / \max \{ \| \mathcal{G}_{k-1/2} \|, \| \mathcal{G}_{k-1/2} \| \} +
$$

$$
+ \| \alpha_1 \| / \| \mathcal{G}_{k-1/2} \| - \| \alpha_1 \| + \| \mathcal{V}_k \| \leq 2e, / (1 + e) + 2e, + \| \mathcal{V}_k \| / \max \{ \| \mathcal{G}_{k-1/2} \|, \| \mathcal{G}_{k-1/2} \| \}.
$$

(5.18)

Тем самым получена оценка близости векторов $\mathcal{G}_k$ и $\mathcal{G}_k'$.

Наконец, перейдем к оценке близости векторов $\mathcal{V}_{k+1}$ и $\mathcal{V}_{k+1}$. В силу (5.5) и (5.12) справедлива цепочка равенств

$$
\mathcal{V}_{k+1} - \mathcal{V}_{k+1} = \mathcal{V}_{k+1} - \mathcal{V}_{k+1} - \mathcal{V}_{k+1} = [\mathcal{V}_k, \mathcal{G}_k - \mathcal{G}_k] + \mathcal{V}_k - \mathcal{G}_k - \mathcal{G}_k - \mathcal{G}_k - \mathcal{G}_k,
$$

откуда

$$
\| \mathcal{V}_{k+1} - \mathcal{V}_{k+1} \| \leq \| \mathcal{V}_k \| + \| \mathcal{V}_k \| + \| \mathcal{V}_k \| / \max \{ \| \mathcal{G}_{k-1/2} \|, \| \mathcal{G}_{k-1/2} \| \}.
$$

Следовательно, ввиду неравенств (5.6), (5.17) — (5.18) находим, что

$$
\| \mathcal{V}_{k+1} - \mathcal{V}_{k+1} \| \leq 6,2 \varepsilon, / \| \mathcal{V}_k \| + 21 \mu (\varepsilon) \varepsilon, / \| \mathcal{V}_k \| + 27 \mu (\varepsilon) \varepsilon, / \| \mathcal{C} \| +
$$

$$
+ 6,5 \mu (\varepsilon) \varepsilon, / \| \eta \| / \max \{ \| \mathcal{G}_{k-1/2} \|, \| \mathcal{G}_{k-1/2} \| \}.
$$

Полученное неравенство доказывает справедливость неравенства (5.13) при условии (5.3), а тем самым и справедливость неравенства (5.4).

§ 6. ОПИСАНИЕ ОБЩЕГО АЛГОРИТМА РЕШЕНИЯ УРАВНЕНИЯ $\$ = C$

Опишем общую схему решения уравнения

$$
\$ = C,
$$

(6.1)

все этапы которой рассматривались в предыдущих параграфах. При построении алгоритма особое внимание будем уделять арифметике процесса. В § 4 отмечено, что для реализации алгоритма (4.2) предполагается использовать «арифметику вычисленных порядков». Предположим, что в нашем распоряжении имеются два оператора: $\mathcal{M}$ и $\mathcal{P}$, позволяющие каждому действительному числу $\alpha$ поставить в соответствие целое число $p(\alpha)$ — порядок числа $\alpha$ и $\mathcal{W}(\alpha)$ — мантиссу числа $\alpha$ ($1/\gamma \leq |\mathcal{W}(\alpha)| < 1$ и $\alpha = \mathcal{W}(\alpha) \gamma^{(\alpha)}$, где $\gamma$ — основание системы счисления используемой ЭВМ. Назовем канонической парой представления числа $\alpha$ пару $[\alpha^0, \alpha^1]$. Если $\alpha^0$ — целое число, $1/\gamma \leq |\alpha^0| < 1$ и $\alpha = \alpha^0 \gamma^{(\alpha)}$, (\alpha, \text{равное нулю, представляется парой } [0, 0]). Используя операторы $\mathcal{M}$ и $\mathcal{P}$, каждому действительному числу $\alpha$ поставим в соответствие каноническую пару $[\mathcal{M}(\alpha), p(\alpha)]$. Аналогично для вектора введены операторы $\mathcal{R}$ и $\mathcal{P}$, позволяющие вычислить его каноническую пару. Подробно эти операторы рассмотрены в работе [9].
Операторы $\mathcal{M}$, $\mathcal{P}$, $\mathbb{m}$, $p$ дают возможность формально описать «арифметику вынесенных порядков». Ниже перечислены основные этапы алгоритма, выполняемые последовательно один за другим. Скалярное произведение $(x, y)$ в пространстве векторов считается подобным так, чтобы возможно более простой вид имел оператор $\mathbb{E}^* — сопряженный к оператору $\mathbb{E}$.

1. Входные данные. В машину вводятся следующие величины: $N$ — целое число; $C = \{C_i\}$ — вектор правой части размерности $N$; $H = \{H_i\}$ — вектор начального приближения размерности $N$; $M$ — целое число, ограничитель итераций; $\delta_1, \delta_2$ — параметры, задаваемые для уменьшения влияния ошибок округления (см. § 5); $\rho$ — требуемая точность.

2. Задаем новую правую часть. С двойной точностью накапливаются суммы $(i = 1, 2, \ldots, N)$:

\[
(F_1)_i = (\mathcal{E} H_0)_i + C_i; \quad F_0^1 = \mathcal{M}(F_1); \quad F_1^1 = \mathcal{P}(F_1).
\]

2.1. Предположим, что для некоторого $i > 0$ вычислен вектор $H_{i-1}$ такой, что каноническая пара $[F_0^i, F_1^i]$ задает вектор-невязку $\delta H_{i-1} - \gamma F_{i-1}^1$. Далее, начиная с пункта 3, описывается итерационный процесс получения приближенного решения операторного уравнения (6.1) с вектором $F_0^i$ в правой части.

3. Задаем следующие канонические пары, необходимые для первого шага итерационного процесса (1.2):

\[
[U_0^0, U_0^0] = [0, 0]; \quad [W_0^0, W_1^0] = [0, 0]; \quad [G_0^0, G_1^0] = [0, 0].
\]

Напомним, что означают эти формальные равенства:

\[
(U_0^0)_i = (G_0^0)_i = (W_0^0)_i = 0; \quad (i = 1, 2, \ldots, N),
\]

\[
U_0^i = G_0^1 = W_0^1 = 0.
\]

4. Определяем следующие канонические пары (как и шаг, $k \geq 1$) итерационного алгоритма (1.2). Пусть имеются полученные в предыдущих вычислениях:

\[
[U_{k-1}^0, U_{k-1}^1] = (k - 1)\text{-е приближение;}
\]

\[
[G_{k-1}^0, G_{k-1}^1]; \quad [W_{k-1}^0, W_{k-1}^1] — вспомогательные векторы, задаваемые в виде канонических пар.
\]

Стандартный шаг состоит в переходе к каноническим параметрам $[U_0^0, U_1^1], [G_0^0, G_1^1], [W_0^0, W_1^1]$. 4.1. Вычисляется $\bar{V}_k$ — невязка $(k - 1)$-го шага по схеме § 4. С двойной точностью накапливаются суммы $(i = 1, 2, \ldots, N)$:

\[
(\bar{V}_k)_i = (\mathcal{E} U_{k-1}^0)_i - \gamma U_{k-1}^1 (F_0^i)_i; \quad V_k^1 = \mathcal{P}(\bar{V}_k) + U_{k-1}^1; \quad V_k^0 = \mathcal{M}(\bar{V}_k).
\]

4.2. Определяется $c_k$ — норма невязки на $k$-м шаге:

\[
\bar{c}_k = \left( \sum_{i=1}^{N} (V_0^1)_i^2 \right)^{1/2}; \quad c_k = \bar{c}_k \gamma^{1/2}.
\]

4.3. Проверяется справедливость неравенства $\delta_1 c_k < \gamma^{-1}$. Если это выполнено, то-управление передается п. 5, в противном случае начинается работа п. 4.4. В случае выполнения проверяемого неравенства тем более верна оценка $\delta_1 || V || < || F_0^i ||$, и, значит, для уменьшения влияния ошибок округления (см. § 5) итерации по формулам (1.2) прекращаются.

4.4. Вычисляется вектор $V_{k+1/2}$:

\[
\bar{V}_k^{1/2} = \mathbb{E}^* V_k; \quad V_k^{1/2} = \mathcal{M}(\bar{V}_k^{1/2}); \quad V_k^{1/2} = \mathcal{P}(\bar{V}_k^{1/2}) + V_k^1.
\]

4.5.
4.5. Находимся $b_k$ (норма $V_{h+1/2}$):

$$
\overline{b}_k = \left( \sum_{j=1}^{N} (V_{h+1/2,j}^2)^{1/2} \right)^{1/2}; \quad b_k^1 = p (\overline{b}_k) + V_{h+1/2}; \quad b_k^0 = m (\overline{b}_k).
$$

$$
Y_k = \mathcal{P} (\overline{Y}_k) + V_{h+1/2}; \quad Y_k^0 = \mathcal{M} (\overline{Y}_k);
$$

$$
\eta_k = (Y_k^0, G_{h-1}); \quad \eta_k^0 = m (\eta_k); \quad \eta_k^1 = p (\eta_k) + Y_k^1 + G_{h-1}.
$$

4.6. Определяется $\eta$:

4.7. Вычисляем $s_{h-1} = (k - 1)$-й элемент побочной диагонали матрицы $B_h$ (см. § 1). Если $k \geq 2$, то подаем $s_{h-1} = \eta_k/b_k$, затем задается $s_{h-1} = \eta_k^1/b_k^1$.

4.8. Определяется вектор $W_{h-1/2}$ (ищется линейная комбинация канонических пар):

$$
q = \max \{ V_{h+1/2}, \eta_k^1 + W_{h-1} \} \quad (j = 1, 2, \ldots, N);
$$

$$
(W_{h-1/2})_j = \gamma V_{h+1/2,-q} (V_{h+1/2})_j - \gamma \eta_k^1 + W_{h-1} \eta_k^1 (W_{h-1})_j;
$$

$$
W_{h-1/2}^0 = \mathcal{M} (W_{h-1/2}); \quad W_{h-1/2}^1 = q + \mathcal{P} (W_{h-1/2}).
$$

4.9. Вычисляются $G_{h-1}$ и $d_h = \|G_{h-1}\|$:

$$
\overline{G}_{h-1} = \mathcal{P} W_{h-1/2}; \quad G_{h-1} = \mathcal{M} (\overline{G}_{h-1}); \quad G_{h-1}^1 = \mathcal{P} (\overline{G}_{h-1}) + W_{h-1/2};
$$

$$
\overline{d}_h = \left( \sum_{j=1}^{N} (G_{h-1,j})^2 \right)^{1/2}; \quad d_h^1 = p (\overline{d}_h) + G_{h-1/2}; \quad d_h^0 = m (\overline{d}_h).
$$

4.10. Проверяем неравенство $\eta_k^1 - d_h < p (d_h) - 1$. Если оно выполнено, то управление передается п. 5, иначе начинают работу 4.11. Заметим, что проверяемое неравенство в случае выполнения влечет за собой справедливость неравенства $\eta_k/d_h \leq \delta_h$. Значит, для уменьшения влияния ошибок округления (см. § 5) необходимо прекратить итерации.

4.11. Вычисляются векторы $G_h, W_h (j = 1, 2, \ldots, N)$:

$$
(W_h)_j = (W_{h-1/2})_j/d_h^0; \quad W_h = \mathcal{M} (W_h);
$$

$$
(G_h)_j = (G_{h-1/2})_j/d_h^0; \quad G_h = \mathcal{M} (G_h);
$$

$$
W_h^1 = \mathcal{P} (W_h) + W_{h-1/2} - d_h^1; \quad G_h^1 = \mathcal{P} (G_h) + G_{h-1} - d_h^1.
$$

4.12. Определяем $\rho_h$ (k-й элемент главной диагонали матрицы $B_h$) (см. § 1):

$$
\widetilde{\rho}_h = d_h^0/b_h; \quad \rho_h = \gamma b_h^1 \rho_h.
$$

4.13. Находим $\xi_k$:

$$

\xi_k = (V_{h}^2, G_h^2); \quad \xi_h^1 = p (\xi_h) + V_h^1 + G_h^1; \quad m (\xi_h) = \xi_h^2.
$$

4.14. Вычисляем $k$-е приближение $U_h = U_{h-1} - \xi_h W_h$ (аналогично п. 4.8). Поляем

$$
q = \max \{ U_{h-1}, \xi_h + W_h \};
$$

$$
(U_h)_j = \gamma V_{h,-q} (U_{h-1})_j - \gamma (\xi_h + W_{h-1}) (W_h)_j,
$$

$$
(j = 1, 2, \ldots, N); \quad U_h = \mathcal{M} (U_h); \quad U_h^1 = q + \mathcal{P} (U_h).
$$

На этом основании k-й шаг итерации процесса завершен. Заметим, что $k$ меняется от 1 до $M$, где $M \leq M_1$, так как процесс может обраться в п. 4.3, 4.10 еще до того, как $k$ станет равным $M$. После завершения описанного процесса определена следующая выходная информация: $M_1, \rho_1, \rho_2, \ldots, \rho_{M_1}, s_1, s_2, \ldots, s_{M_1-1}, U_{M_1}^0, U_{M_1}^1, F_i, F_i^1$.
5°. Вычисляется $H_i$ — приближенное решение уравнения (6.1), полученное после завершения четырех этапов процесса $(j = 1, 2, \ldots, N)$:

$$
(H_{i})_j = (H_{i-1})_j + \gamma \frac{U_{M_1}}{F_1} \left( \frac{U_{\rho}}{M_{ij}} \right);
$$

$$
\varepsilon_5 = \varepsilon_5 \gamma F_1
$$

— норма погрешности найденного приближения.

5°.1. Проверяется неравенство $\varepsilon_5 \leq \rho$. Если оно выполнено, то полагается $H = H_i$ и уравнение передается п. 9°, в противном случае начинает работу п. 6°.

6°. Входная информация пункта совпадает с выходной информацией п. 4°. Здесь вычисляются $\sigma_1(B_{M_1})$ — минимальное и $\sigma_{M_1}(B_{M_1})$ — максимальное сингулярные числа, а также соответствующие им сингулярные векторы матрицы $B_{M_1}$ (см. § 1). При этом используются стандартные процедуры линейной алгебры. Полученные сингулярные векторы обозначаем $x$ и $\bar{x}$ ($\bar{x}$ соответствует $\sigma_1(B_{M_1})$, $x - \sigma_{M_1}(B_{M_1})$). Выходная информация: $x_1, x_2, \ldots, x_{M_1}$, $x_+ x_2, \ldots, x_{M_1}$, $\sigma_1(B_{M_1})$, $\sigma_{M_1}(B_{M_1})$, $\rho_1, \rho_2, \ldots, \rho_{M_1}$, $s_1, \varepsilon_2, \ldots, s_{M_1-1}$.

7°. Входная информация совпадает с выходной информацией 6°. Здесь будут получены векторы $X$ и $\bar{X}$, числа $\sigma(X)$ и $\sigma(\bar{X})$ ($\sigma(X) = \|X\|/\|\bar{X}\|$) такие, что если бы не было ошибок округления, то выполнялись бы равенства $\sigma(X) = \sigma_1(B_{M_1})$ и $\sigma(\bar{X}) = \sigma_{M_1}(B_{M_1})$. Но поскольку ошибки округления неизбежны, указанные равенства не всегда выполняются даже приближенно. Векторы $X$ строятся по формуле $X = x_1 \bar{V}_1 + x_2 \bar{V}_2 + \ldots + x_{M_1} \bar{V}_{M_1}$ (см. § 1). Следовательно, необходимо знать векторы $\bar{V}_j$ ($j = 1, 2, \ldots, M_1$), что связано с большими затратами машинной памяти. Для экономии памяти предлагается, повторив $M_1$ итераций метода (1.2), вычислить последовательность векторов $X_t = x_t \bar{V}_1 + x_t \bar{V}_{M_1}$, таких, что $X_0 = 0$, $X_{M_1} = X$.

7°.1. Совпадает с п. 2°.
7°.2. Задаются $X_0$ и $\bar{X}_0$: $X_0 = 0; \bar{X}_0 = 0$.

7°.3. Начинается описание $k$-го шага ($1 \leq k \leq M_1$) итерационного повторного расчета формул (1.2). Считаются вычисленными векторы $X_{k-1}$, $X_{k-1}$ и канонические пары $[U_{k-1}, U_{k-1}]$, $[G_{k-1}, G_{k-1}]$, $[W_{k-1}, W_{k-1}]$. Последовательно работают формулами, совпадающими с формулами п. 4°.4, 4°.4, 4°.5.

7°.4. Вычисляем векторы $X_k$ и $X_k$ ($j = 1, 2, \ldots, N$):

$$(X_k)_j = x_k b_k (V^0)_{k+1/2} \gamma^{b_k - V^1_{k+1/2}} + (X_{k-1})_j;$$

$$(X_k)_j = x_k b_k (V^0)_{k+1/2} \gamma^{b_k - V^1_{k+1/2}} + (X_{k-1})_j;$$

если $k = M_1$, то начинает работу п. 8°.

7°.5. Работают формулы, совпадающие с формулами в последовательности 4°.6, 4°.8 — 4°.9, 4°.11, 4°.13, 4°.14. Этим $k$-й шаг итерационного повторного расчета формул (1.2) завершен. Заметим, что $k$ меняется от 1 до $M_1$. Таким образом, имеем следующую выходную информацию: векторы $X$ и $\bar{X}$, каноническую пару $[V_{M_1}, V_{M_1+1}]$.

8°. Входная информация совпадает с выходной информацией п. 7°. Вычисляются $\sigma(X)$ и $\sigma(\bar{X})$ — гарантированные (векторами $X, \bar{X}$) оценки максимального и минимального сингулярных чисел оператора $\bar{E}$. Здесь же рассчитываются данные для возможности итерационного уточнения, полученного в п. 5° приближенного решения.

53
8.1. Вычисляем \( \sigma(\overline{X}), \sigma(X) : \)
\[ z = \overline{z} \overline{X}; \quad \sigma(\overline{X}) = \|z\|/\|X\|; \]
\[ \overline{z} = \overline{z} \overline{X}; \quad \sigma(X) = \|\overline{z}\|/\|\overline{X}\|. \]

8.2. Проверяется неравенство
\[ \frac{\sigma^2(X) - \sigma^2(\overline{X})}{\sigma^2(\overline{X}) + \sigma^2(X)} \sigma(\overline{X}) + (50\delta_1 + 7\delta_2) \sigma(X) \varepsilon_1 < \sigma(\overline{X}), \]
где \( \varepsilon_1 \) — машинная постоянная (см. § 4). Если это выполнено, то процесс заканчивается без указания вычисленного приближенного решения и его результатом является утверждение, что оператор \( \Phi \) плохо обусловлен. Оценка обусловленности гарантируется указанием векторов \( \overline{X}, X \) и чисел \( \sigma(\overline{X}), \sigma(X) \).

8.3. Вычисляется новая правая часть для итерационного уточнения полученного приближенного решения:
\[ F_{i+1}^0 = \hat{V}_{M+1}^0; \quad F_{i+1}^1 = \hat{V}_{M+1}^1 + F_1. \]
Затем в качестве \( i \) берется \( i + 1 \), и управление переключается к п. 2.1.

9. Выходная информация. В результате работы алгоритма выдаются следующие данные:
1) \( \sigma \) — оценка снизу максимального сингулярного числа оператора \( \Phi \)
и \( \overline{\sigma} \) — оценка сверху минимального сингулярного числа оператора \( \Phi \);
2) векторы \( \overline{X}, X \), на которых отношение Рэлея оператора \( \Phi \) достигает соответственно значений \( \sigma \) и \( \overline{\sigma} \);
3) \[ \overline{\Phi} = (\overline{\Phi})_{j=1}^N \] — полученное приближенное решение (6.1);
4) \( \varepsilon_3 \) — норма невязки полученного приближения (\( \varepsilon_3 \leq \rho \)).

§ 7. РЕЗУЛЬТАТЫ ЧИСЛЕННЫХ ЭКСПЕРИМЕНТОВ

Приведены результаты решения модельных задач \( Lu = f \), рассмотренных в § 3, методом сопряженных градиентов. При решении использовалась "арифметика вынесенных порядков" и накопление скалярных произведений векторов с двойной точностью.

Пример 1: Для примера 1 § 3 получены результаты:
\[ c_1 = 2; \quad c_2 = 1.2705; \quad c_3 = 0.86029; \quad c_4 = 0.48627; \]
\[ c_5 = 0.216 - 9; \quad c_6 = 0.216 - 10; \]
\[ \sigma_1(B_2) = \sigma_2(L) (1 - 0.15); \quad \sigma_3(B_3) = \sigma_2(L) (1 - 0.014); \]
\[ \sigma_4(B_4) = \sigma_4(L) (1 - 0.0002); \quad \sigma_1(B_5) = \sigma_3(L); \]
\[ \sigma_2(B_4) = \sigma_3(L); \quad \sigma_2(B_5) = \sigma_4(L); \quad \sigma_4(B_4) = \sigma_4(L); \]
\[ \sigma_1(B_4) = \sigma_1(L) (1 - 0.11 - 7); \quad \sigma_3(B_3) = \sigma_3(L) (1 - 0.1 - 8); \]
\[ \sigma_5(B_3) = \sigma_5(L) (1 - 0.11 - 10); \quad \sigma_4(B_3) = \sigma_4(L) (1 - 0.0004); \]
\[ \sigma_5(B_3) = \sigma_5(L); \quad \sigma_4(B_3) = (1 - 0.11 - 7) \sigma_1(L); \]
\[ \sigma_2(B_3) = \sigma_2(L) (1 - 0.11 - 11); \quad \sigma_3(B_3) = (1 - 0.001) \sigma_3(L); \]
\[ \sigma_1(B_4) = \sigma_1(L) (1 - 0.11 - 10); \quad \sigma_5(B_3) = (1 - 0.11 - 11) \sigma_4(L); \]
\[ \sigma_6(B_3) = \sigma_4(L) (1 + 0.0001). \]

Сравнивая с результатами примера 1 § 3, можем заключить, что невязка, процесса следует считать совпадающими, но в варианте, использующем специальные приемы программирования, двухдиагональная матрица \( B_m \) оказывается достаточно представительной.
Пример 2. Для примера 2 § 3 получены следующие значения норм невязок:

c_1 = 2; \quad c_2 = 1.732; \quad c_3 = 1.414;
\quad c_4 = 0.974; \quad c_5 = 0.974; \quad c_6 = 0.458;
\quad c_7 = 0.294_{10} - 2; \quad c_8 = 0.155_{10} - 3; \quad c_9 = 0.679_{10} - 5.

Ясно, что невязка постоянно убывает.

ЛИТЕРАТУРА
1. Будагов А. Я., Годунов С. К. Численное определение одного из критериев качества устойчивости систем линейных дифференциальных уравнений с постоянными коэффициентами.— Новосибирск, 1991.— 58 с. (Препринт/АН СССР, Сиб. отделение, ИМ).
2. Годунов С. К., Прокопов Г. П. Вариационный подход к решению больших систем линейных уравнений, возникающих в сильно эллиптических задачах.— М., 1968.— 40 с. (Препринт/АН СССР, ИПМ).
3. Воеводин В. В. О методах сопряженных градиентов.— Журн. вычисл. математики и математической физики, 1979, т. 19, № 5, с. 1315—1317.
5. Федоренко Р. П. Приближенное решение задач оптимального управления.— М.: Наука, 1978.— 486 с.

О СХОДИМОСТИ ОРТОГАНОЛЬНО-СТЕПЕННОГО МЕТОДА РАСЧЕТА СПЕКТРА

В. И. КОТИН, Ш. И. РАЗЗАКОВ

ВВЕДЕНИЕ

В работе обсуждается вопрос об устойчивости и эффективных оценках скорости сходимости ортогонально-степенного метода Воеводина, применяемого для решения полной проблемы собственных значений.

Под полной проблемой собственных значений понимается задача нахождения всех собственных значений матрицы A. Часто при этом пытаются найти и соответствующие им собственные и присоединенные векторы. Достаточно обширная библиография по решению полной проблемы содержится в монографиях и статьях В. В. Воеводина [1, 2], В. Н. Кублановской [3], В. Н. Фаддеевой [4, 5]. Напомним, что собственными значениями матрицы A называются корни ее характеристического полинома, т. е. корни уравнения

$$\det(A - \lambda I) = \left| \begin{array}{cccc} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \\ \end{array} \right| = (-1)^n \left( \lambda^n - p_n \lambda^{n-1} - p_{n-1} \lambda^{n-2} - \cdots - p_1 \right) = 0.$$

Все методы численного решения полной проблемы можно разделить на две группы: прямые и итерационные. Большинство прямых методов