Результаты расчета: $\varkappa(A) = 105.76$; $\varkappa_{\text{гр}} = 0.5_{10}4$;

$$\breve{H} = \begin{bmatrix} 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1.5 & 2.0 & 2.5 \\ 0.5 & 2.0 & 4.5 & 7.0 \\ 0.5 & 2.5 & 7.0 & 14.5 \end{bmatrix}.$$

ЛИТЕРАТУРА

1. **Булгаков А. Я.** Эффективно вычисляемый параметр качества устойчивости систем линейных дифференциальных уравнений с постоянными коэффициентами.— Сиб. мат. журн., 1980, т. 21, № 3, с. 32—41.
2. **Godounov S. K., Boulgakov A. J.** Difficultés de calcul dans le probleme de Hurwitz et méthodes pour les surmonter.— In: Analysis and optimization of Systems, Versailles, 1982.— Procedings (Lecture Notes in Control and Information sciences, 44). Springer Verlag, 1982, p. 843—851.
3. **Davison E. J., Man F. T.** The numerical solution of $A^*Q + QA = -C$.— IEEE Trans. Automatic Control., 1968, v. 13, p. 448.
4. **Per Hagander.** Numerical solution of $A^*S + SA + Q = 0$.— Information Sciences, 1972, № 4, p. 45—50.
5. **Ла-Салль Ж., Лефшетц С.** Исследование устойчивости прямым методом Ляпунова.— М.: Мир, 1964.— 168 с.
6. **Карачаров К. А., Пилютик А. Г.** Введение в техническую теорию устойчивости движения.— 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, 1981.— 58 с.— (Препринт/АН СССР, Сиб. отд-ние, ИМ).
9. **Годунов С. К.** Решение систем линейных уравнений.— Новосибирск: Наука, 1980.— 177 с.
10. **Форсайт Дж., Молер К.** Численное решение систем линейных алгебраических уравнений.— М.: Мир, 1969.— 167 с.

УЧЕТ ВЫЧИСЛИТЕЛЬНЫХ ПОГРЕШНОСТЕЙ В ОДНОМ ВАРИАНТЕ МЕТОДА СОПРЯЖЕННЫХ ГРАДИЕНТОВ

А. Я. БУЛГАКОВ, С. К. ГОДУНОВ

ВВЕДЕНИЕ

В работе рассматривается вариант метода сопряженных градиентов для решения операторного уравнения

$$\mathfrak{L}H = C, \qquad (1)$$

где $\mathfrak{L}$ — несамосопряженный линейный оператор, действующий в N-мерном евклидовом 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 от приближенного выполнения арифметических операций на ЭВМ (см., например, [2—5]). Выбор конкретного варианта сделан с учетом ряда проделанных 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 $\mathfrak{L}$ записывается в виде произведения ортогональной и двухдиагональ-

ной матриц. Это обстоятельство позволяет, вычисляя сингулярные числа двухдиагональной матрицы, получить в процессе решения уравнения оценки ее обусловленности $\mu(\mathfrak{L})$. 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) — невозрастание 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 $\mu(\mathfrak{L})$ алгоритм сходится не хуже метода наискорейшего спуска. Для того чтобы процесс давал возможность оценить $\mu(\mathfrak{L})$, вычисляются приближения наибольшего и наименьшего сингулярных чисел $\mathfrak{L}$ 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

$$\mathfrak{L}H = C \tag{1.1}$$

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

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 и числовых параметров c_j, d_j, b_j, η_j, ρ_j, s_j, ξ_j $(j = 1, 2, \ldots)$ по следующим формулам:

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

если $c_j = 0$, то процесс заканчивается;

$$\begin{gathered} b_j = \|V_{j+1/2}\|; \quad Y_j = \mathfrak{L}V_{j+1/2}; \quad \eta_j = (Y_j, G_{j-1}); \\ s_{j-1} = \eta_j / b_j; \quad W_{j-1/2} = V_{j+1/2} - \eta_j W_{j-1}; \quad G_{j-1/2} = \mathfrak{L}W_{j-1/2}; \\ d_j = \|G_{j-1/2}\|; \quad W_j = d_j^{-1} W_{j-1/2}; \quad G_j = d_j^{-1} G_{j-1/2}; \\ \rho_j = d_j / b_j; \quad \xi_j = (V_j, G_j); \quad H_j = H_{j-1} - \xi_j W_j. \end{gathered} \tag{1.2}$$

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

$$\begin{aligned} H_k &= \mathfrak{L}^* \{\alpha_1^{(k)} I + \alpha_2^{(k)} [\mathfrak{L}\mathfrak{L}^*] + \ldots + \alpha_k^{(k)} [\mathfrak{L}\mathfrak{L}^*]^{k-1}\} C; \\ V_{k+1/2} &= \mathfrak{L}^* \{\beta_1^{(k)} I + \beta_2^{(k)} [\mathfrak{L}\mathfrak{L}^*] + \ldots + \beta_k^{(k)} [\mathfrak{L}\mathfrak{L}^*]^{k-1}\} C; \\ G_k &= \mathfrak{L}^* \{\gamma_0^{(k)} I + \gamma_1^{(k)} [\mathfrak{L}\mathfrak{L}^*] + \ldots + \gamma_k^{(k)} [\mathfrak{L}\mathfrak{L}^*]^k\} C, \end{aligned} \tag{1.3}$$

вследствие чего

$$-\mathfrak{L}H_k + C = \{I - \alpha_1^{(k)} [\mathfrak{L}\mathfrak{L}^*] - \ldots - \alpha_k^{(k)} [\mathfrak{L}\mathfrak{L}^*]^k\} C. \tag{1.4}$$

Замечательным и совсем не очевидным свойством процесса (1.2) является выполнение соотношений, обоснование которых составляет содержание следующей леммы:

Лемма 1. *Если процесс* (1.2) *при* $i=1, 2, \ldots, n$ *не прервался, т. е. если* $c_i \neq 0$ *для всех* $i=1, 2, \ldots, n$, *то справедливы соотношения* ($\widehat{V}_i = \|V_{i+1/2}\|^{-1} V_{i+1/2}$):

$$(V_{i+1}, G_j) = 0 \quad \text{при всех } 1 \leqslant i \leqslant j \leqslant n, \tag{1.5}$$

$$(\widehat{V}_i, \widehat{V}_j) = \delta_{ij} \quad \text{при всех } 1 \leqslant i, j \leqslant n, \tag{1.6}$$

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

δ_{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 получено 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:

$$b_j = \|V_{j+1/2}\| = \|\mathfrak{L}^* V_j\| = \|\mathfrak{L}^* (\mathfrak{L} H_{j-1} - C)\| \geqslant \sigma_1(\mathfrak{L}) \|\mathfrak{L} H_{j-1} - C\| = \sigma_1(\mathfrak{L}) c_j,$$

которая вследствие невырожденности оператора $\mathfrak{L}$ 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

$$0 = G_{j-1/2} = \mathfrak{L} V_{j+1/2} - \eta_j G_{j-1}; \quad (\mathfrak{L} V_{j+1/2}, V_j) = \eta_j (G_{j-1}, V_j).$$

Осталось показать, что в силу выбора ξ_j 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:

$$\|\mathfrak{L}(H_k - H)\| = \|\mathfrak{L} H_k - C\| \leqslant$$

$$\leqslant [(\mu(\mathfrak{L}) + 1)^k / (\mu(\mathfrak{L}) - 1)^k + (\mu(\mathfrak{L}) - 1)^k / (\mu(\mathfrak{L}) + 1)^k]^{-1} \cdot 2\|C\|. \tag{1.8}$$

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

$$V = V_1 - \sum_{i=1}^{k} a_i^{(k)} [\mathfrak{L}\mathfrak{L}^*]^i C = \left\{ -I - \sum_{i=1}^{k} a_i^{(k)} [\mathfrak{L}\mathfrak{L}^*]^i \right\} C,$$

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

$$\|\mathfrak{L}H_k - C\| \leqslant \min_{\left(a_1^{(k)},\ldots,a_k^{(k)}\right)} \max_{\lambda} \left|a^{(k)}(\lambda)\right| \cdot \|C\|, \tag{1.9}$$

где максимум модуля полинома

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

берется по всем значениям λ из отрезка

$$\sigma_1^2(\mathfrak{L}) = \sigma_1(\mathfrak{L}\mathfrak{L}^*) \leqslant \lambda \leqslant \sigma_N(\mathfrak{L}\mathfrak{L}^*) = \sigma_N^2(\mathfrak{L}),$$

содержащего весь спектр оператора $\mathfrak{L}$, и выбирается наименьшее возможное значение этого максимума среди всех полиномов $a^{(k)}(\lambda)$ с равным единице коэффициентом $a_0^{(k)} = 1$. Неравенства (1.9) позволяют, воспользовавшись известными свойствами полиномов, наименее уклоняющихся от нуля (см., например, [6, с. 509]), установить неравенства (1.8). Замечательной для процесса (1.2) является возможность получения оценок максимальных и минимальных сингулярных чисел оператора $\mathfrak{L}$ во время счета.

Из формул (1.2) следуют равенства

$$d_k G_k = \mathfrak{L}V_{k+1/2} - \eta_k G_{k-1},$$

эквивалентные (ввиду определения $\widehat{V} = \|V_{k+1/2}\|^{-1} V_{k+1/2}$)

$$\mathfrak{L}\widehat{V}_k = \rho_k G_k + s_{k-1} G_{k-1}. \tag{1.10}$$

Равенства (1.10) означают, как уже замечено в [1], что оператор $\mathfrak{L}$ представляется в виде произведения ортогонального преобразования, переводящего ортонормированный базис $\widehat{V}_1, \widehat{V}_2, \ldots, \widehat{V}_k, \ldots$ в ортонормированный базис из векторов $G_1, G_2, \ldots, G_k, \ldots$, на преобразование, задаваемое в этом последнем базисе двухдиагональной матрицей B_N:

$$B_N = \begin{bmatrix} \rho_1 & & & 0 \\ s_1 & \rho_2 & & \\ & s_2 & \rho_3 & \\ 0 & & \ddots & \ddots \\ & & & s_N\rho_N \end{bmatrix}.$$

Поэтому $\sigma_N(\mathfrak{L}) = \sigma_N(B_N)$, $\sigma_1(\mathfrak{L}) = \sigma_1(B_N)$, $\mu(\mathfrak{L}) = \mu(B_N)$. Если провести процесс (1.2) до конца, выполнив N шагов, то, вычисляя сингулярные числа двухдиагональной матрицы B_N, сумели бы определить $\mu(\mathfrak{L})$. На самом же деле сингулярные числа даже урезанных матриц B_k $(k \leqslant N)$, 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:

$$\sigma_1(B_{k-1}) \geqslant \sigma_1(B_k) \geqslant \sigma_1(B_N);$$
$$\sigma_N(B_N) \geqslant \sigma_k(B_k) \geqslant \sigma_{k-1}(B_{k-1});$$
$$\mu(B_{k-1}) \leqslant \mu(B_k) \leqslant \mu(B_N) = \mu(\mathfrak{L}).$$

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

Более того, возможно, используя величины процесса (1.2), вычисленные за первые k шагов $(k = 2, 3, \ldots, N)$, определить векторы, на которых отношение Рэлея оператора $\mathfrak{L}$ достигает значения $\sigma_1(B_k)$ и $\sigma_k(B_k)$. Кратко опишем алгоритм получения этих векторов. Предположим, что σ — сингулярное 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 $X =$

$= x_1\widehat{V}_1 + x_2\widehat{V}_2 + \ldots + x_k\widehat{V}_k$ отношение Рэлея оператора $\mathfrak{L}$ равно σ ($\|\mathfrak{L}X\|/\|X\| = \sigma$). В самом деле, используя (1.10), находим

$$\mathfrak{L}X = \sum_{i=1}^{k} x_i \mathfrak{L}\widehat{V}_i = \sum_{i=1}^{k-1} (x_i\rho_i + x_{i+1}s_i) G_i' + x_k\rho_k G_k.$$

Откуда, воспользовавшись свойством ортонормированности $\{\widehat{V}_i\}_1^k$, $\{G_i\}_1^k$, можно заключить, что

$$\|\mathfrak{L}X\| = \left(\sum_{i=1}^{k-1} (x_i\rho_i + x_{i+1}s_i)^2 + x_k^2\rho_k^2\right)^{1/2} = \|B_k x\|;$$

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

Итак, показано, что $\|\mathfrak{L}X\|/\|X\| = \sigma$. Следовательно, использование $\sigma_1(B_k)$ и $\sigma_k(B_k)$ в качестве оценок сингулярных чисел оператора $\mathfrak{L}$, 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

$$(V_{i+1/2},\ V_{j+1/2}) = (\mathfrak{L}^* V_{i+1},\ W_{j-1/2} + \eta_j W_{j-1}) =$$
$$= (V_{i+1},\ \mathfrak{L}W_{j-1/2} + \eta_j \mathfrak{L}W_{j-1}) = (V_{i+1},\ d_j G_j + \eta_j G_{j-1}),$$

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

$$\eta_k = (\mathfrak{L}\mathfrak{L}^* V_k,\ G_{k-1}); \quad d_k G_k = \mathfrak{L}\mathfrak{L}^* V_k - \eta_k G_{k-1};$$
$$\xi_k = (V_k,\ G_k); \quad V_{k+1} = V_k - \xi_k G_k, \tag{1.11}$$

из которых нетрудно заключить, что выбор ξ_k и η_k обеспечивает выполнение равенств

$$(G_k,\ G_{k-1}) = 0; \quad V_{k+1} = V_k - \xi_k G_k. \tag{1.12}$$

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

$$\xi_k = (V_k, G_k) = d_k^{-1}[(V_k, \mathfrak{L}\mathfrak{L}^* V_k) - \eta_k (V_k, G_k)] =$$
$$= d_k^{-1}(V_k, \mathfrak{L}\mathfrak{L}^* V_k) \geqslant d_k^{-1}\sigma_1(\mathfrak{L}\mathfrak{L}^*)\|V_k\|^2 = d_k^{-1}\sigma_1^2(\mathfrak{L})\, c_k^2 > 0.$$

Из (1.11) следует, что

$$d_{k+1}G_{k+1} = \mathfrak{L}\mathfrak{L}^* V_{k+1} - \eta_{k+1}G_k = \mathfrak{L}\mathfrak{L}^* V_k - \xi_k \mathfrak{L}\mathfrak{L}^* G_k - \eta_{k+1}G_k =$$
$$= d_k G_k + \eta_k G_{k-1} - \xi_k \mathfrak{L}\mathfrak{L}^* V_k - \eta_{k+1}G_k.$$

Так как $\xi_k \neq 0$, то полученное равенство может быть переписано как формула для $\mathfrak{L}\mathfrak{L}^* G_k$:

$$\mathfrak{L}\mathfrak{L}^* G_k = \alpha_k G_{k+1} + \beta_k G_k + \gamma_k G_{k-1} \tag{1.13}$$

с очевидным выражением для коэффициентов α_k, β_k, γ_k. Итак, перейдем к доказательству (1.7), (1.5) методом полной математической индукции.

Для $i = 2$ имеем: $(V_2, G_1) = 0$ в силу выбора ξ_1 на первой итерации, $(G_2, G_1) = 0$, $(V_3, G_2) = 0$ в силу выбора ξ_2, η_2 на второй итерации. Следовательно,

$$(V_3, G_1) = (V_2 - \xi_2 G_2, G_1) = 0,$$

т. е. на начальном шаге утверждение верно.

Предположим, что при некотором k выполнены равенства:

$$1^\circ \quad (G_i,\ G_j) = 0 \quad \text{для } 1 \leqslant j < i \leqslant k;$$
$$2^\circ \quad (V_{i+1},\ G_j) = 0 \quad \text{для } 1 \leqslant j \leqslant i \leqslant k,$$

и покажем, что они останутся справедливыми в случае, если k заменить на $k+1$.

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

$$(G_{k+1}, G_j) = d_k^{-1}(\mathfrak{L}\mathfrak{L}^*V_{k+1} - \eta_{k+1}G_k, G_j) = d_k^{-1}(\mathfrak{L}\mathfrak{L}^*V_{k+1}, G_j),$$

так как $(G_k, G_j)=0$ по предположению индукции. Подставляя сюда $\mathfrak{L}\mathfrak{L}^*G_j$ из формул (1,13):

$$\mathfrak{L}\mathfrak{L}^*G_j = \alpha_jG_{j+1} + \beta_jG_j + \gamma_jG_{j-1}$$

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

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

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 $1\leqslant j\leqslant k+1$. Напомним, что $V_{k+2}=V_{k+1}-\xi_{k+1}G_{k+1}$. Поэтому при $1\leqslant j\leqslant k$ 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 ξ_{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 $\|V_k\|$ за k-й шаг.

Из формул (1.2) получаем соотношения

$$\eta_k = (\mathfrak{L}\mathfrak{L}^*V_k, G_{k-1}); \tag{2.1}$$

$$G_{k-1/2} = \mathfrak{L}\mathfrak{L}^*V_k - \eta_kG_{k-1}; \tag{2.2}$$

$$d_k = \|G_{k-1/2}\|; \tag{2.3}$$

$$G_k = d_k^{-1}G_{k-1/2}; \tag{2.4}$$

$$\xi_k = (V_k, G_k); \tag{2.5}$$

$$V_{k+1} = V_k - \xi_kG_k. \tag{2.6}$$

Мы вправе считать, что $(V_k, G_{k-1})=0$. Это обеспечивается выбором ξ_k. Из (2.6) и (2.5) следует, что

$$\|V_{k+1}\| \leqslant \|V_k\| \tag{2.7}$$

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

$$\|V_{k+1}\|^2 \leqslant \left[\frac{[\mu(\mathfrak{L})]^2-1}{\mu^2(\mathfrak{L})+1}\right]^2\|V_k\|^2 - \eta_k^2\frac{(V_k, \mathfrak{L}\mathfrak{L}^*V_k)^2}{\|\mathfrak{L}\mathfrak{L}^*V_k\|^2}. \tag{2.8}$$

Рассмотрим пучок векторов

$$z = V_k + \alpha\mathfrak{L}\mathfrak{L}^*V_k + \beta G_{k-1}.$$

Тогда для любых α, β выполнено неравенство $\|z\| \geqslant \|V_{k+1}\|$, что следует из равенств

$$\left.\frac{\partial \|z\|^2}{\partial \alpha}\right|_{z=V_{k+1}} = 2(V_{k+1}, \mathfrak{L}\mathfrak{L}^* V_k) = 0;$$

$$\left.\frac{\partial \|z\|^2}{\partial \beta}\right|_{z=V_{k+1}} = 2(V_{k+1}, G_{k-1}) = 0.$$

Кроме того, имеет место цепочка равенств

$$z - \beta G_{k-1} = V_k + \alpha \mathfrak{L}\mathfrak{L}^* V_k;$$

$$\|z\|^2 - 2\beta(z, G_{k-1}) + \beta^2 \|G_{k-1}\|^2 = \|V_k + \alpha \mathfrak{L}\mathfrak{L}^* V_k\|^2; \qquad (2.9)$$

$$(z, G_{k-1}) = (V_k, G_{k-1}) + \alpha(\mathfrak{L}\mathfrak{L}^* V_k, G_{k-1}) + \beta = \alpha(\mathfrak{L}\mathfrak{L}^* V_k, G_{k-1}) + \beta,$$

которая позволяет заключить, что $(z, G_{k-1}) = 0$ при условии $\beta = -\alpha(\mathfrak{L}\mathfrak{L}^* V_k, G_{k-1}) = -\alpha\eta_k$. Тогда при

$$\alpha_0 = (V_k, \mathfrak{L}\mathfrak{L}^* V_k)/\|\mathfrak{L}\mathfrak{L}^* V_k\|^2; \quad \beta_0 = -\alpha_0 \eta_k$$

из (2.9) получаем

$$\|V_{k+1}\|^2 \leqslant \|z_0\|^2 = \|V_k + \alpha_0 \mathfrak{L}\mathfrak{L}^* V_k\|^2 - \beta_0^2 \leqslant$$

$$\leqslant [(\mu^2(\mathfrak{L}) - 1)/(\mu^2(\mathfrak{L}) + 1)]^2 \|V_k\|^2 - \eta_k (V_k, \mathfrak{L}\mathfrak{L}^* V_k)^2 / \|\mathfrak{L}\mathfrak{L}^* V_k\|^4.$$

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

В заключение отметим два несложных утверждения

$$\|\mathfrak{L}\mathfrak{L}^* V_k\|^2 = \|G_{k-1/2}\|^2 + \eta_k^2; \qquad (2.10)$$

$$(\mathfrak{L}\mathfrak{L}^* V_k, V_k)/\|V_k\| \leqslant \|G_{k-1/2}\| \leqslant \|\mathfrak{L}\mathfrak{L}^* V_k\|. \qquad (2.11)$$

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

Выбор ξ_{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

$$(G_{k-1/2}, V_k) = (\mathfrak{L}\mathfrak{L}^* V_k, V_k).$$

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 $d_k = 0$ задавалось d_k равным машинной константе ε_1 ($1 + \varepsilon_1$ — наименьшее машинное число, близкое к 1).

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

$$f^T = (-1.1364325,\ -0.8070552,\ -1.2363081,\ -0.7271348),$$

$$L = \begin{bmatrix} 0.2317832 & -0.2017895 & -0.4012386 & -0.0405040 \\ -0.1206849 & 0.1389269 & 0.0378899 & -0.0242303 \\ -0.9645644 & -0.0906907 & -0.0957022 & 0.0186665 \\ 0.6364325 & -0.6570810 & 0.1444403 & -0,0408918 \end{bmatrix}.$$

Сингулярные числа матрицы L известны: $\sigma_4(L) \approx 1$; $\sigma_3(L) \approx 0.7071067$; $\sigma_2(L) \approx 0.4386913$; $\sigma_1(L) \approx 0.25$.

В процессе расчета получены значения невязки $c_k = \|Lu_{k-1} - f\|$ ($k = 1, 2, 3, 4, 5, 6$) и вычислены сингулярные числа двухдиагональных матриц B_m ($m = 3, 4, 5, 6$), формируемых на m-м шаге процесса.

Получены результаты:

$$c_1 = 2;\ c_2 = 1.2705;\ c_3 = 0.8603;\ c_4 = 0.4862;\ c_5 = 10^{-11};\ c_6 = 2_{10} - 7;$$

$$\sigma_1(B_3) = 0.3618;\ \sigma_2(B_3) = 0.6867;\ \sigma_3(B_3) = \sigma_4(L) \cdot (1 - 3_{10} - 4);$$

$$\sigma_1(B_4) = \sigma_1(L);\ \sigma_2(B_4) = \sigma_2(L);\ \sigma_3(B_4) = \sigma_3(L);\ \sigma_4(B_4) = \sigma_4(L);$$

$$\sigma_1(B_5) = 6_{10} - 2;\ \sigma_2(B_5) = \sigma_1(L)\,(1 + 0.1_{10} - 2);\ \sigma_3(B_5) = \sigma_2(L)\,(1 + 0.1_{10} - 4);$$

$$\sigma_4(B_5) = \sigma_3(L)\,(1 + 0.1_{10} - 6);\ \sigma_5(B_5) = \sigma_4(L)\,(1 + 0.1_{10} - 8);$$

$$\sigma_1(B_6) = 0.6_{10} - 1;\ \sigma_2(B_6) = \sigma_1(L)\,(1 + 0.1_{10} - 2);$$

$$\sigma_3(B_6) = \sigma_2(L)\,(1 + 0.1_{10} - 4);\ \sigma_4(B_6) = \sigma_3(L)\,(1 + 0.1_{10} - 6);$$

$$\sigma_5(B_6) = \sigma_4(L)\,(1 + 0.1_{10} - 8);\ \sigma_6(B_6) = 4209.91.$$

Пример 2. В качестве L взята матрица

$$L = \begin{bmatrix} 0.0331119 & -0.0003639 & -0.3657_{10} - 5 & -0.107_{10} - 6 \\ -0{,}0172407 & 0.0002506 & 0.3450_{10} - 6 & -0.616_{10} - 6 \\ -0.1377948 & -0.0001635 & -0.8720_{10} - 6 & -0.470_{10} - 7 \\ 0.0052046 & -0.0011852 & 0.1317_{10} - 5 & -0.103_{10} - 6 \end{bmatrix}.$$

Правая часть задается вектором f:

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

Известны сингулярные числа матрицы L:

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

В результате расчета получено нарушение одного из основных свойств процесса (1.2) — невозрастание нормы невязки. Приведем значения норм невязок за первые 6 шагов ($c_k = \|Lu_{k-1} - f\|$): $c_1 = 2$, $c_2 = 1.732$, $c_3 = 1.414$, $c_4 = 316.144$, $c_5 = 132.202$, $c_6 = 0.974$. 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 алгоритма метода сопряженных градиентов использовать стандартное программирование формул, то влияние ошибок округления приводит к тому, что теряются два основных свойства алгоритма:

1) невозрастание невязки за шаг процесса (1.2);

2) оценка $\mu(\mathfrak{L})$ числом обусловленности двухдиагональной матрицы B_k, сформированной по результатам первых k шагов процесса.

§ 4. ПОГРЕШНОСТИ МАШИННЫХ ВЫЧИСЛЕНИЙ

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 вынесенных порядков» (см. § 2 из [9]). Здесь, как и в [9], для формального описания арифметики используются операторы $\mathfrak{M}$ и $\mathcal{P}$, заданные на пространстве векторов длины N, и операторы $\mathfrak{m}$, p, заданные на поле вещественных чисел.

В § 2 [9] показано, что использование «арифметики вынесенных порядков» позволяет вычислить нормировку вектора x и разность векторов x и αy (α — скаляр) с точностью:

$$\|(\|x\|^{-1}x)_{\text{выч}} - \|x\|^{-1}x\| \leqslant 3\varepsilon_1; \tag{4.1}$$

$$\|(x - \alpha y)_{\text{выч}} - (x - \alpha y)\| \leqslant 1{,}01\varepsilon_1\|\alpha y\| + 1{,}01\varepsilon_1\|x - \alpha y\|, \tag{4.2}$$

если $(2N^{1/2}+2N^{3/2}+2\sqrt{N}/\gamma)\varepsilon_2/\varepsilon_1<0{,}01$, где ε_1 — машинная постоянная $(1+\varepsilon_1$ — наименьшее машинное число, превосходящее 1). Пусть γ — основание системы счисления используемой ЭВМ и ε_2 — машинная постоянная такая, что $1/\gamma^2(1-\varepsilon_1/\gamma)\varepsilon_2$ — наименьшее по модулю отличное от нуля машинное число.

В § 2 работы [9] и в § 21 работы [7] показано, что, используя «арифметику вынесенных порядков», можно гарантировать выполнение неравенства

$$|(x,\ y)-[(x,\ y)]_{\text{выч}}|\leqslant 1{,}01\varepsilon_1\|x\|\cdot\|y\|. \tag{4.3}$$

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

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

полученного для случая накопления скалярного произведения векторов с двойной точностью.

Итак, осталось оценить погрешности выполнения двух операций: вычисления образа оператора $\mathfrak{L}$ и вычисления невязки.

Предположим, что погрешности этих операций оцениваются неравенствами

$$\|(\mathfrak{L}H)_{\text{выч}}-\mathfrak{L}H\|\leqslant 2\varepsilon_1\|\mathfrak{L}H\| \tag{4.5}$$

и

$$\|(\mathfrak{L}H-C)_{\text{выч}}-(\mathfrak{L}H-C)\|\leqslant 2\varepsilon_1\|\mathfrak{L}H-C\|. \tag{4.6}$$

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 в случае, если оператор $\mathfrak{L}$ задается матрицей L размерности $N\times N$.

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

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

Следовательно,

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

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

$$\|(LH)_{\text{выч}}-LH\|\leqslant\varepsilon_1\|LH\|\{1+N^{3/2}\varepsilon_1^2/\gamma\cdot\mu(L)+N^{3/2}/(\gamma\sigma_1(L)\varepsilon_1)\cdot\varepsilon_2\}.$$

Итак, если выполнены условия

$$2N^{3/2}\varepsilon_1^2/\gamma\mu(L)<1;\ 2N^{3/2}\varepsilon_2/(\gamma\sigma_1(L)\varepsilon_1)<1, \tag{4.7}$$

то выполнено неравенство

$$\|(LH)_{\text{выч}}-LH\|\leqslant 2\varepsilon_1\|LH\|. \tag{4.8}$$

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 $(j=1,\ 2,\ \ldots,\ N)$:

$$|(V_j)_{\text{выч}}-V_j|\leqslant\varepsilon_1|V_j|+(N+1)\varepsilon_1^2/\gamma\sqrt{\|L_j\|^2+C_j^2}\sqrt{\|H\|^2+1}+$$
$$+(N+1)/\gamma^2\varepsilon_2$$

и, значит,

$$\|(V)_{\text{выч}} - V\| \leqslant \varepsilon_1 \|V\| + (N+1)\varepsilon_1^2/\gamma \sqrt{N\|L\|^2 + \|C\|^2}\sqrt{\|H\|^2 + 1} + \\ + (N+1)^{3/2}/\gamma^2 \varepsilon_2.$$

Заметим, что поскольку $1/\gamma < \|C\| < \sqrt{N}$ и $(N+1)^{3/2}/\gamma^2\varepsilon_2 < (N+1)^{3/2}/\gamma\varepsilon_2\|C\|$, то, предположив, что

$$\|C\|/\|V\| < \delta_1, \tag{4.9}$$

где δ_1 выбирается удовлетворяющим неравенству $2(N+1)^{3/2}\varepsilon_2/(\gamma\varepsilon_1)\delta_1 < 1$, можно гарантировать выполнение оценки $(N+1)^{3/2}/\gamma^2\varepsilon_2 < 1/2\varepsilon_1\|V\|$.

Аналогично, воспользовавшись очевидной цепочкой неравенств

$$\sqrt{1+\|H\|^2}\sqrt{N\|L\|^2+\|C\|^2} \leqslant \sqrt{N}\|H\|\cdot\|L\| + \|C\|\cdot\|H\| + \sqrt{N}\|L\| + \\ + \|C\| \leqslant \left[\mu(L)\sqrt{N}\left(1+\frac{\|C\|}{\|V\|}\right) + \frac{\sqrt{N}}{\sigma_1(L)}\left(1+\frac{\|C\|}{|V|}\right) + (1+\sqrt{N}\|L\|\gamma)\frac{\|C\|}{\|V\|}\right]\|V\|,$$

можем, взяв в (4.9) δ_1 удовлетворяющим неравенству

$$2(N+1)\varepsilon_1/\gamma\{[\mu(L)\sqrt{N} + \sqrt{N}/\sigma_1(L)](1+\delta_1) + (1+\sqrt{N}\|L\|\gamma)\delta_1\} < 1,$$

записать, что $\|(LH-C)_{\text{выч}} - (LH-C)\| \leqslant 2\varepsilon_1\|LH-C\|$.

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

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 собою соотношениями

$$V_k = \mathfrak{L}H_{k-1} - C + \varphi_1; \quad c_k = \|V_k\| + \alpha_1; \quad V_{k+1/2} = \mathfrak{L}^* V_k + \varphi_2;$$
$$b_k = \|V_{k+1/2}\| + \alpha_2; \quad Y_k = \mathfrak{L}V_{k+1/2} + \varphi_3; \quad \eta_k = (Y_k, G_{k-1}) + \alpha_3;$$
$$s_{k-1} = \eta_k/b_k + \alpha_4; \quad W_{k-1/2} = V_{k+1/2} - \eta_k W_{k-1} + \varphi_4; \tag{5.1}$$
$$G_{k-1/2} = \mathfrak{L}W_{k-1/2} + \varphi_5; \quad d_k = \|G_{k-1/2}\| + \alpha_5; \quad W_k = d_k^{-1}W_{k-1/2} + \varphi_6;$$
$$G_k = d_k^{-1}G_{k-1/2} + \varphi_7; \quad \rho_k = d_k/b_k + \alpha_6; \quad \xi_k = (V_k, G_k) + \alpha_7;$$
$$H_k = H_{k-1} - \xi_k W_k + \varphi_8,$$

в которых φ_i, α_j $(i = 1, 2, \ldots, 8;\ j = 1, 2, \ldots, 7)$ обозначают погрешности, допущенные при вычислениях по формулам (1.2). Оценки $\|\varphi_i\|$ и $|\alpha_j|$ 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 произведений с двойной точностью. Расчет формул (1.2) подробно представлен в § 4. Считая выполненными условия (4.5)—(4.6), приведем оценки из § 4 для некоторых $\|\varphi_i\|$, $|\alpha_j|$ процесса (5.1):

$$\|\varphi_1\| \leqslant 2\varepsilon_1\|V_k\|; \quad \|\varphi_2\| \leqslant 2\varepsilon_1\|\mathfrak{L}^* V_k\|;$$
$$\|\varphi_3\| \leqslant 2\varepsilon_1\|\mathfrak{L}V_{k+1/2}\|; \quad |\alpha_3| \leqslant 2\varepsilon_1\|Y_k\| = 2\varepsilon_1\|\mathfrak{L}V_{k+1/2}\|;$$
$$\|\varphi_4\| \leqslant 1{,}01\varepsilon_1\|\eta_k W_{k-1}\| + 1{,}01\varepsilon_1\|V_{k+1/2} - \eta_k W_{k-1}\|; \tag{5.2}$$
$$\|\varphi_5\| \leqslant 2\varepsilon_1\|\mathfrak{L}W_{k-1/2}\|; \quad |\alpha_5| \leqslant 2\varepsilon_1\|G_{k-1/2}\|;$$
$$\|\varphi_6\| \leqslant 2\varepsilon_1|d_k^{-1}|\cdot\|W_{k-1/2}\|; \quad \|\varphi_7\| \leqslant 2\varepsilon_1;$$
$$|\alpha_7| \leqslant 2\varepsilon_1\|V_k\|; \quad \|\varphi_8\| \leqslant 1{,}01\varepsilon_1\|\xi_k W_k\| + 1{,}01\varepsilon_1\|H_{k-1} - \xi_k W_k\|.$$

Предположим, что при некоторых δ_1, δ_2 имеет место

$$\|C\|/\|V_k\| \leqslant \delta_1; \quad |\eta_k|/d_k \leqslant \delta_2, \tag{5.3}$$

тогда справедлива оценка

$$\|V_{k+1}\| \leqslant [\mu^2(\mathfrak{L})-1]/[\mu^2(\mathfrak{L})+1]\|V_k\| + (50\delta_1 + 7\delta_2)\,\mu(\mathfrak{L})\,\varepsilon_1\|V_k\|, \qquad (5.4)$$

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 наискорейшего спуска. Если же нарушено одно из условий (5.3) (что легко проверяется в процессе расчета), то предлагается останавливать процесс, предварительно запомнив приближенное решение H_{k-1}, а затем переходить к его уточнению, а именно для решения уравнения $\mathfrak{L}\bar{H} = \bar{C}$, где $\bar{C} = \mathfrak{L}H_{k-1} - C$, 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:

$$\eta_k = (\mathfrak{L}\mathfrak{L}^*V_k,\ G_{k-1}) + \beta_1;\quad G_{k-1/2} = \mathfrak{L}\mathfrak{L}^*V_k - \eta_k G_{k-1} + \psi_1;$$
$$d_k = \|G_{k-1/2}\| + \beta_2;\ G_k = d_k^{-1}G_{k-1/2} + \psi_2; \qquad (5.5)$$
$$\xi_k = (V_k,\ G_k) + \beta_3;\quad V_{k+1} = V_k - \xi_k G_k + \psi_3$$

с явными выражениями для β_i, ψ_i:

$$\beta_1 = (\mathfrak{L}\varphi_2,\ G_{k-1}) + (\varphi_3,\ G_{k-1}) + \alpha_3;$$
$$\beta_2 = \alpha_5;\quad \beta_3 = \alpha_7;\quad \psi_2 = \varphi_7; \qquad (5.6)$$
$$\psi_1 = \mathfrak{L}\varphi_2 + \eta_k(G_{k-1} - \mathfrak{L}W_{k-1}) + \mathfrak{L}\varphi_4 + \varphi_5;$$
$$\psi_3 = \tilde{\varphi}_1 - \varphi_1 - \xi_k(\mathfrak{L}W_k - G_k) + \mathfrak{L}\varphi_8.$$

Здесь $\tilde{\varphi}_1$ — погрешность вычисления невязки, полученной в результате работы $(k-1)$-го шага процесса (5.1).

Отметим, что, производя оценки, будем пренебрегать членами малости порядка ε_1^2 по сравнению с членами порядка ε_1.

Оценки для ψ_i, β_i $(i = 1, 2, 3)$ нетрудно получить из (5.6), (5.2), зная оценки $\|G_k - \mathfrak{L}W_k\|$, $|\eta_k|$, $\|G_{k-1} - \mathfrak{L}W_{k-1}\|$, $|\xi_k|$, к выведению которых сейчас и перейдем.

Из (5.5) следует:

$$|\eta_k| \leqslant \|\mathfrak{L}\mathfrak{L}^*V_k\| + |\beta_1|; \qquad (5.7)$$
$$|\xi_k| \leqslant \|V_k\| + |\beta_3|. \qquad (5.8)$$

Наконец, покажем, что

$$\|\mathfrak{L}W_k - G_k\| \leqslant 3\mu(\mathfrak{L})\,\varepsilon_1. \qquad (5.9)$$

В самом деле, из (5.1) имеем цепочку равенств

$$G_k = d_k^{-1}G_{k-1/2} + \varphi_7 = d_k^{-1}(\mathfrak{L}W_{k-1/2} + \varphi_5) + \varphi_7 = \mathfrak{L}W_k - \mathfrak{L}\varphi_6 + d_k^{-1}\varphi_5 + \varphi_7,$$

позволяющую заключить, следуя (5.2), что

$$\|\mathfrak{L}W_k - G_k\| \leqslant \|d_k^{-1}\varphi_5\| + \|\varphi_7\| + \|\mathfrak{L}\varphi_6\| \leqslant 2\varepsilon_1/(1-\varepsilon_1) + 2\varepsilon_1 + 2\mu(\mathfrak{L})\,\varepsilon_1.$$

Полученное неравенство гарантирует справедливость (5.9) при условии $\mu(\mathfrak{L}) > 4$. Если $\mu(\mathfrak{L}) < 4$, 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

$$\|\psi_1\| \leqslant 4\varepsilon_1\|\mathfrak{L}\| \cdot \|\mathfrak{L}^*V_k\| + 2\mu(\mathfrak{L})\,\varepsilon_1|\eta_k| + 2\mu(\mathfrak{L})\,\varepsilon_1\|G_{k-1/2}\|;$$
$$|\beta_1| \leqslant 2\varepsilon_1\|\mathfrak{L}\| \cdot \|\mathfrak{L}^*V_k\| + 4\varepsilon_1\|\mathfrak{L}\mathfrak{L}^*V_k\|; \qquad (5.10)$$
$$|\beta_2| \leqslant 2\varepsilon_1\|G_{k-1/2}\|;\quad |\beta_3| \leqslant 2\varepsilon_1\|V_k\|;$$
$$\|\psi_2\| \leqslant 2\varepsilon_1;\quad \|\psi_3\| \leqslant 5\mu(\mathfrak{L})\,\varepsilon_1\|V_k\| + \mu(\mathfrak{L})\,\varepsilon_1\|C\|.$$

Таким образом, вопрос уменьшения нормы невязки процесса (5.1), (5.2) сведен 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 выяснена ниже) соответствующим величинам, вычисляемым по известным

$$\overline{V}_k = V_k; \cdot \overline{G}_{k-1} = G_{k-1} - (G_{k-1},\ V_k)/\|V_k\|^2 \cdot V_k \tag{5.11}$$

по формулам

$$\begin{gathered}\overline{\eta}_k = (\mathfrak{L}\mathfrak{L}^*\overline{V}_k,\ \overline{G}_{k-1});\quad \overline{G}_{k-1/2} = \mathfrak{L}\mathfrak{L}^*\overline{V}_k - \overline{\eta}_k\overline{G}_{k-1};\\ \overline{d}_k = \|\overline{G}_{k-1/2}\|;\ \overline{G}_k = \overline{d}_k^{-1}\overline{G}_{k-1/2};\\ \overline{\xi}_k = (\overline{V}_k,\ \overline{G}_k);\quad \overline{V}_{k+1} = \overline{V}_k - \overline{\xi}_k\overline{G}_k.\end{gathered} \tag{5.12}$$

Отметим, что $(\overline{V}_k,\ \overline{G}_{k-1}) = 0$ в силу (5.11). Следовательно, из сказанного в § 2 имеем оценку

$$\|\overline{V}_{k+1}\| \leqslant [\mu^2(\mathfrak{L}) - 1]/[\mu^2(\mathfrak{L}) + 1] \cdot \|\overline{V}_k\|,$$

и для доказательства справедливости (5.4) при условии (5.3) осталось показать, что если

$$\|C\|/\|V_k\| \leqslant \delta_1;\quad |\eta_k|/d_k \leqslant \delta_2,$$

то

$$\|\overline{V}_{k+1} - V_{k+1}\| \leqslant \mu(\mathfrak{L})\,\varepsilon_1(50\delta_1 + 7\delta_2)\|V_k\|. \tag{5.13}$$

Для выяснения степени близости векторов V_{k+1}, $\overline{V}_{k+1}$ (оценка $\|\overline{V}_{k+1} - V_{k+1}\|$) требуется оценить близость векторов и скаляров в (5.5) и (5.12).

Предварительно заметим, что из (5.5) и (5.10) следует

$$\begin{gathered}|(V_{k+1},\ G_k)| = \|-\beta_3 G_k + \psi_3\| \leqslant 2\varepsilon_1\|V_k\| + 5\mu(\mathfrak{L})\,\varepsilon_1\|V_k\| +\\ + \mu(\mathfrak{L})\,\varepsilon_1\|C\| \leqslant 5{,}5\mu(\mathfrak{L})\,\varepsilon_1\|V_k\| + \mu(\mathfrak{L})\,\varepsilon_1\|C\|,\end{gathered}$$

откуда

$$\begin{gathered}|(V_{k+1},\ G_k)| \leqslant 5{,}5\mu(\mathfrak{L})\,\varepsilon_1\|V_k\| + \mu(\mathfrak{L})\,\varepsilon_1\|C\|;\\ |(V_k,\ G_{k-1})| \leqslant 5{,}5\mu(\mathfrak{L})\,\varepsilon_1\|V_{k-1}\| + \mu(\mathfrak{L})\,\varepsilon_1\|C\|.\end{gathered} \tag{5.14}$$

Полученные неравенства позволяют, используя (5.11), заключить, что

$$\|\overline{G}_{k-1} - G_k\| \leqslant \|V_k\|^{-1}\{5{,}5\mu(\mathfrak{L})\,\varepsilon_1\|V_{k-1}\| + \mu(\mathfrak{L})\,\varepsilon_1\|C\|\}. \tag{5.15}$$

Далее, опираясь на (5.10), (5.11), (5.15) и равенство

$$\overline{\eta}_k - \eta_k = \beta_1 + (\mathfrak{L}\mathfrak{L}^*V_k,\ \overline{G}_{k-1} - G_{k-1}),$$

нетрудно вывести оценку

$$\begin{gathered}|\overline{\eta}_k - \eta_k| \leqslant 2\varepsilon_1\|\mathfrak{L}\| \cdot \|\mathfrak{L}^*V_k\| + 4\varepsilon_1\|\mathfrak{L}\mathfrak{L}^*V_k\| + |(G_{k-1},\ V_k)| \cdot \|\mathfrak{L}^*V_k\|^2/\|V_k\|^2 \leqslant\\ \leqslant 2\varepsilon_1\|\mathfrak{L}\| \cdot \|\mathfrak{L}^*V_k\| + 4\varepsilon_1\|\mathfrak{L}\mathfrak{L}^*V_k\| + \|\mathfrak{L}\mathfrak{L}^*V_k\|^2/\|V_k\|^2[5{,}5\mu(\mathfrak{L})\,\varepsilon_1\|V_{k-1}\| +\\ + \mu(\mathfrak{L})\,\varepsilon_1\|C\| \leqslant 2\varepsilon_1\|\mathfrak{L}\| \cdot \|\mathfrak{L}^*V_k\| + 4\varepsilon_1\|\mathfrak{L}\mathfrak{L}^*V_k\| +\\ + 6{,}5\mu(\mathfrak{L})\,\varepsilon_1\|C\| \cdot \|\mathfrak{L}^*V_k\|^2/\|V_k\|^2.\end{gathered} \tag{5.16}$$

Аналогично, используя равенство

$$\begin{gathered}G_{k-1/2} - \overline{G}_{k-1/2} = \overline{\eta}_k\overline{G}_{k-1} - \eta_k G_{k-1} - \psi_1 =\\ = (\eta_k - \overline{\eta}_k)G_{k-1/2} + \overline{\eta}_k(\overline{G}_{k-1} - G_{k-1}) - \psi_1,\end{gathered}$$

вытекающее из (5.5) и (5.12), и опираясь на неравенства (5.6), (5.15)—(5.16), легко найти оценку

$$\begin{gathered}\|G_{k-1/2} - \overline{G}_{k-1/2}\| \leqslant 2\varepsilon_1\|\mathfrak{L}\| \cdot \|\mathfrak{L}^*V_k\| + 4\varepsilon_1\|\mathfrak{L}\mathfrak{L}^*V_k\| +\\ + 6{,}5\mu(\mathfrak{L})\,\varepsilon_1\{\|\mathfrak{L}^*V_k\|^2/\|V_k\|^2 + |\eta_k|/\|V_k\|\},\end{gathered}$$

из которой непосредственно получаем

$$\|G_{k-1/2}-\overline{G}_{k-1/2}\|/\max\{\|G_{k-1/2}\|,\ \|\overline{G}_{k-1/2}\|\}\leqslant$$
$$\leqslant 6{,}5\mu(\mathfrak{L})\varepsilon_1\|C\|/\|V_k\|+\{2\varepsilon_1\|\mathfrak{L}\|\cdot\|\mathfrak{L}^*V_k\|+4\varepsilon_1\|\mathfrak{L}\mathfrak{L}^*V_k\|+$$
$$+|\eta_k|/\|V_k\|\cdot 6{,}5\mu(\mathfrak{L})\varepsilon_1\|C\|\}/\max\{\|G_{k-1/2}\|,\ \|\overline{G}_{k-1/2}\|\}. \qquad (5.17)$$

При выводе (5.17) использовалось неравенство

$$\|\mathfrak{L}^*V_k\|^2/\|V_k\|\leqslant\|G_{k-1/2}\|,$$

которое было выведено в § 2.

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

$$\overline{G}_k-G_k=\overline{d}_k^{-1}\overline{G}_{k-1/2}-d_k^{-1}G_{k-1/2}+\psi_2=$$
$$=\|\overline{G}_{k-1/2}\|^{-1}\overline{G}_{k-1/2}-\|G_{k-1/2}\|^{-1}G_{k-1/2}-(d_k^{-1}-\|G_{k-1/2}\|^{-1})\cdot G_{k-1/2}+\psi_2,$$

откуда следует, что

$$\|\overline{G}_k-G_k\|\leqslant 2\|\overline{G}_{k-1/2}-G_{k-1/2}\|/\max\{\|\overline{G}_{k-1/2}\|,\ \|G_{k-1/2}\|\}+$$
$$+|\alpha_5|/(\|G_{k-1/2}\|-|\alpha_5|)+\|\psi_2\|\leqslant 2\varepsilon_1/(1-\varepsilon_1)+2\varepsilon_1+$$
$$+2\|\overline{G}_{k-1/2}-G_{k-1/2}\|/\max\{\|\overline{G}_{k-1/2}\|,\ \|G_{k-1/2}\|\}. \qquad (5.18)$$

Тем самым получена оценка близости векторов G_k и $\overline{G}_k$.

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

$$\overline{V}_{k+1}-V_{k+1}=\xi_k\overline{G}_k-\overline{\xi}_k\overline{G}_k-\psi_3=(\xi_k-\overline{\xi}_k)G_k+$$
$$+\overline{\xi}_k(G_k-\overline{G}_k)-\psi_3=[(V_k,\ G_k-\overline{G}_k)G_k+\beta_3G_k]+\overline{\xi}_k(G_k-\overline{G}_k)-\psi_3,$$

откуда

$$\|\overline{V}_{k+1}-V_{k+1}\|\leqslant\|\psi_3\|+|\beta_3|+2\|V_k\|\cdot\|G_k-\overline{G}_k\|.$$

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

$$\|\overline{V}_{k+1}-V_{k+1}\|\leqslant 6{,}2\varepsilon_1\|V_k\|+21\mu(\mathfrak{L})\varepsilon_1\|V_k\|+27\mu(\mathfrak{L})\varepsilon_1\|C\|+$$
$$+6{,}5\mu(\mathfrak{L})\varepsilon_1\|C\|\,|\eta_k|/\max\{\|\overline{G}_{k-1/2}\|,\ \|G_{k-1/2}\|\}.$$

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 $\mathfrak{L}H=C$

Опишем общую схему решения уравнения

$$\mathfrak{L}H=C, \qquad (6.1)$$

все этапы которой рассматривались в предыдущих параграфах. При построении алгоритма особое внимание будем уделять арифметике процесса. В § 4 отмечено, что для реализации алгоритма (1.2) предлагается использовать «арифметику вынесенных порядков». Предположим, что в нашем распоряжении имеются два оператора: $\mathfrak{m}$ и p, позволяющие каждому действительному числу α поставить в соответствие целое число $p(\alpha)$ — порядок числа α и $\mathfrak{m}(\alpha)$ — мантиссу числа α ($1/\gamma\leqslant|\mathfrak{m}(\alpha)|<1$ и $\alpha=\mathfrak{m}(\alpha)\gamma^{p(\alpha)}$, где γ — основание системы счисления используемой ЭВМ. Назовем канонической парой представления числа α пару $[\alpha^0,\ \alpha^1]$, если α^1 — целое число, $1/\gamma\leqslant|\alpha^0|<1$ и $\alpha=\alpha^0\gamma^{\alpha^1}$ (α, равное нулю, представляется парой $[0,\ 0]$). Используя операторы $\mathfrak{m}$ и p, каждому действительному числу α поставим в соответствие каноническую пару $[\mathfrak{m}(\alpha),\ p(\alpha)]$. Аналогично для вектора введены операторы $\mathfrak{M}$ и $\mathcal{P}$, позволяющие вычислить его каноническую пару. Подробно эти операторы рассмотрены в работе [9].

Операторы $\mathfrak{M}$, $\mathscr{P}$, $\mathfrak{m}$, p дают возможность формально описывать «арифметику вынесенных порядков». Ниже перечислены основные этапы алгоритма, выполняемые последовательно один за другим. Скалярное произведение (x, y) в пространстве векторов считается подобранным так, чтобы возможно более простой вид имел оператор $\mathfrak{L}^*$ — сопряженный к оператору $\mathfrak{L}$.

1°. Входные данные. В машину вводятся следующие величины: N — целое число; $C = \{C_j\}$ — вектор правой части размерности N; $H_0 = \{(H_0)_j\}$ — вектор начального приближения размерности N; M — целое число, ограничитель итераций; δ_1, δ_2 — параметры, задаваемые для уменьшения влияния ошибок округления (см. § 5); ρ — требуемая точность.

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

$$(F_1)_j = (\mathfrak{L}H_0)_j + C_j;\ F_1^0 = \mathfrak{M}(F_1);\ F_1^1 = \mathscr{P}(F_1).$$

2°.1. Предположим, что для некоторого $i > 0$ вычислен вектор H_{i-1} такой, что каноническая пара $[F_i^0, F_i^1]$ задает вектор-невязку $\mathfrak{L}H_{i-1} - \gamma^{F_{i-1}^1} F_{i-1}$. Далее, начиная с пункта 3°, описывается итерационный процесс получения приближенного решения операторного уравнения (6.1) с вектором F_i^0 в правой части.

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

$$[U_0^0, U_0^1] = [0, 0];\ [W_0^0, W_0^1] = [0, 0];\ [G_0^0, G_0^1] = [0, 0].$$

Напомним, что означают эти формальные равенства:

$$(U_0^0)_j = (G_0^0)_j = (W_0^0)_j = 0 \quad (j = 1, 2, \ldots, N),$$
$$U_0^1 = G_0^1 = W_0^1 = 0.$$

4°. Опишем стандартный шаг (k-й шаг, $k \geqslant 1$) итерационного алгоритма (1.2). Пусть имеются полученные в предыдущих вычислениях:

$[U_{k-1}^0, U_{k-1}^1]$ — $(k-1)$-е приближение;

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

Стандартный шаг состоит в переходе к каноническим парам $[U_k^0, U_k^1]$, $[G_k^0, G_k^1]$, $[W_k^0, W_k^1]$.

4°.1. Вычисляется V_k — невязка $(k-1)$-го шага по схеме § 4. С двойной точностью накапливаются суммы ($j = 1, 2, \ldots, N$):

$$(\tilde{V}_k)_j = (\mathfrak{L}U_{k-1}^0)_j + \gamma^{-U_{k-1}^1}(F_i^0)_j\, V_k^1 = \mathscr{P}(\tilde{V}_k) + U_{k-1}^1;\ V_k^0 = \mathfrak{M}(\tilde{V}_k).$$

4°.2. Определяется c_k — норма невязки на k-м шаге:

$$\tilde{c}_k = \left(\sum_{j=1}^{N} (V_k^0)_j^2\right)^{1/2};\ c_k = \tilde{c}_k \gamma^{V_{k+1}^1}.$$

4°.3. Проверяется справедливость неравенства $\delta_1 c_k < \gamma^{-1}$. Если оно выполнено, то управление передается п. 5°, в противном случае начинает работу п. 4°.4. В случае выполнения проверяемого неравенства тем более верна оценка $\delta_1 \|V\| < \|F_i^0\|$, и, значит, для уменьшения влияния ошибок округления (см. § 5) итерации по формулам (1.2) прекращаются.

4°.4. Вычисляется вектор $V_{k+1/2}$:

$$\tilde{V}_{k+1/2} = \mathfrak{L}^* V_k;\ V_{k+1/2}^0 = \mathfrak{M}(\tilde{V}_{k+1/2});\ V_{k+1/2}^1 = \mathscr{P}(\tilde{V}_{k+1/2}) + V_k^1.$$

4°.5. Находится b_k (норма $V_{k+1/2}$):

$$\widetilde{b}_k = \left(\sum_{j=1}^{N} (V^0_{k+1/2})^2_j\right)^{1/2};\ b^1_k = p(\widetilde{b}_k) + V^1_{k+1/2};\ \ b^0_k = \mathfrak{m}(\widetilde{b}_k).$$

$$\widetilde{Y}_k = \mathfrak{L}V^0_{k+1/2};\ Y^1_k = \mathscr{P}(\widetilde{Y}_k) + V^1_{k+1/2};\ Y^0_k = \mathfrak{M}(\widetilde{Y}_k);$$
$$\widetilde{\eta}_k = (Y^0_k, G^0_{k-1});\ \eta^0_k = \mathfrak{m}(\widetilde{\eta}_k);\ \eta^1_k = p(\widetilde{\eta}_k) + Y^1_k + G^1_{k-1}.$$

4°.6. Определяется η_k:

4°.7. Вычисляем s_{k-1} — $(k-1)$-й элемент побочной диагонали матрицы B_k (см. § 1). Если $k \geqslant 2$, то полагаем $\widetilde{s}_{k-1} = \eta^0_k / b^0_k$, затем задается $s_{k-1} = \widetilde{s}_k \gamma^{\eta^1_k - b^1_k}$.

4°.8. Определяется вектор $W_{k-1/2}$ (ищется линейная комбинация канонических пар):

$$q = \max\{V^1_{k+1/2}, \eta^1_k + W^1_{k-1}\}\quad (j = 1, 2, \ldots, N);$$
$$(\widetilde{W}_{k-1/2})_j = \gamma^{V^1_{k+1/2} - q}(V^0_{k+1/2})_j - \gamma^{\eta^1_k + W^1_{k-1} - q}\eta^0_k (W^0_{k-1})_j;$$
$$W^0_{k-1/2} = \mathfrak{M}(\widetilde{W}_{k-1/2});\ W^1_{k-1/2} = q + \mathscr{P}(\widetilde{W}_{k-1/2}).$$

4°.9. Вычисляются $G_{k-1/2}$ и $d_k = \|G_{k-1/2}\|$:

$$\widetilde{G}_{k-1/2} = \mathfrak{L}W^0_{k-1/2};\ G^0_{k-1/2} = \mathfrak{M}(\widetilde{G}_{k-1/2});\ G^1_{k-1/2} = \mathscr{P}(\widetilde{G}_{k-1/2}) + W^1_{k-1/2};$$
$$\widetilde{d}_k = \left(\sum_{j=1}^{N} (G^0_{k-1/2})^2_j\right)^{1/2};\ d^1_k = p(\widetilde{d}_k) + G^1_{k-1/2};\ d^0_k = \mathfrak{m}(\widetilde{d}_k).$$

4°.10. Проверяем неравенство $\eta^1_k - d^1_k < p(\delta_2) - 1$. Если оно выполнено, то управление передается п. 5°, иначе начинает работу 4°.11. Заметим, что проверяемое неравенство в случае выполнения влечет за собой справедливость неравенства $\eta_k / d_k < \delta_2$. Значит, для уменьшения влияния ошибок округления (см. § 5) необходимо прекратить итерации.

4°.11. Вычисляются векторы G_k, W_k $(j = 1, 2, \ldots, N)$:

$$(\widetilde{W}_k)_j = (W^0_{k-1/2})_j / d^0_k;\ W^0_k = \mathfrak{M}(\widetilde{W}_k);$$
$$(\widetilde{G}_k)_j = (G^0_{k-1/2})_j / d^0_k;\ G^0_k = \mathfrak{M}(\widetilde{G}_k);$$
$$W^1_k = \mathscr{P}(\widetilde{W}_k) + W^1_{k-1/2} - d^1_k;\ G^1_k = \mathscr{P}(\widetilde{G}_k) + G^1_{k-1/2} - d^1_k.$$

4°.12. Определяем ρ_k (k-й элемент главной диагонали матрицы B_k) (см. § 1):

$$\widetilde{\rho}_k = d^0_k / b^0_k;\ \rho_k = \gamma^{d^1_k - b^1_k}\widetilde{\rho}_k.$$

4°.13. Находим ξ_k:

$$\widetilde{\xi}_k = (V^0_k, G^0_k);\ \xi^1_k = p(\widetilde{\xi}_k) + V^1_k + G^1_k;\ \ \mathfrak{m}(\widetilde{\xi}_k) = \xi^0_k.$$

4°.14. Вычисляем k-е приближение $U_k = U_{k-1} - \xi_k W_k$ (аналогично п. 4°.8). Полагаем $q = \max\{U^1_{k-1}, \xi^1_k + W^1_k\}$;

$$(\widetilde{U}_k)_j = \gamma^{U^1_{k-1} - q}(U^0_{k-1})_j - \gamma^{\xi^1_k + W^1_k - q}\xi_k (W^0_k)_j,$$
$$(j = 1, 2, \ldots, N);\ U^0_k = \mathfrak{M}(\widetilde{U}_k);\ U^1_k = q + \mathscr{P}(\widetilde{U}_k).$$

На этом основании k-й шаг итерации процесса завершен. Заметим, что k меняется от 1 до M_1, где $M_1 \leqslant M$, так как процесс может оборваться в п. 4°.3, 4°10 еще до того, как k станет равным M. После завершения описанного процесса определена следующая выходная информация: M_1, ρ_1, $\rho_2, \ldots, \rho_{M_1}$, $s_1, s_2, \ldots, s_{M_1 - 1}$, $U^0_{M_1}$, $U^1_{M_1}$, F^0_i, F^1_i.

5°. Вычисляется H_i — приближенное решение уравнения (6.1), полученное после завершения четырех этапов процесса ($j=1, 2, \ldots, N$):

$$(H_i)_j = (H_{i-1})_j + \gamma^{U_{M_1}+F_i^1}\left(U^0_{M_1}\right)_j;$$

$\varepsilon_5 = c_{M_1}\gamma^{F_i^1}$ — норма погрешности найденного приближения.

5°.1. Проверяется неравенство $\varepsilon_5 < \rho$. Если оно выполнено, то полагается $\widehat{H} = H_i$ и управление 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 обозначим $\bar{x}$ и $\underline{x}$ ($\bar{x}$ соответствует $\sigma_1(B_{M_1})$, $\underline{x}$ — $\sigma_{M_1}(B_{M_1})$). Выходная информация: $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{M_1}, \underline{x}_1, \underline{x}_2, \ldots, \underline{x}_{M_1}$, $\sigma_1(B_{M_1})$, $\sigma_{M_1}(B_{M_1})$, $\rho_1, \rho_2, \ldots, \rho_{M_1}, s_1, s_2, \ldots, s_{M_1-1}$.

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

7°.1. Совпадает с п. 2°.

7°.2. Задаются $\underline{X}_0$ и $\bar{X}_0$: $\underline{X}_0 = 0$; $\bar{X}_0 = 0$.

7°.3. Начинается описание k-го шага ($1 \leqslant k \leqslant M_1$) итерационного повторного расчета формул (1.2). Считаются вычисленными векторы $\bar{X}_{k-1}$, $\underline{X}_{k-1}$ и канонические пары $[U^0_{k-1}, U^1_{k-1}]$, $[G^0_{k-1}, G^1_{k-1}]$, $[W^0_{k-1}, W^1_{k-1}]$ Последовательно работают формулы, совпадающие с формулами п. 4°.1, 4°.4, 4°.5.

7°.4. Вычисляем векторы $\bar{X}_k$ и $\underline{X}_k$ ($j=1, 2, \ldots, N$):

$$(\bar{X}_k)_j = \bar{x}_k b^0_k (V^0_{k+1/2})_j \gamma^{b^1_k - V^1_{k+1/2}} + (\bar{X}_{k-1})_j;$$

$$(\underline{X}_k)_j = \underline{x}_k b^0_k (V^0_{k+1/2})_j \gamma^{b^1_k - V^1_{k+1/2}} + (\underline{X}_{k-1})_j,$$

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 $\bar{X}$ и $\underline{X}$, каноническую пару $[V^0_{M_1+1}, V^1_{M_1+1}]$.

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

8°.1. Вычисляем $\sigma(\overline{X})$, $\sigma(\underline{X})$:

$$z = \mathfrak{L}\overline{X};\ \sigma(\overline{X}) = \|z\|/\|\overline{X}\|;$$
$$\tilde{z} = \mathfrak{L}\underline{X};\ \sigma(\underline{X}) = \|\tilde{z}\|/\|\underline{X}\|.$$

8°.2. Проверяется неравенство

$$\frac{\sigma^2(\underline{X}) - \sigma^2(\overline{X})}{\sigma^2(\underline{X}) + \sigma^2(\overline{X})}\sigma(\overline{X}) + (50\delta_1 + 7\delta_2)\sigma(\underline{X})\varepsilon_1 < \sigma(\overline{X}),$$

где ε_1 — машинная постоянная (см. § 4). Если оно выполнено, то процесс заканчивается без указания вычисленного приближенного решения и его результатом является утверждение, что оператор $\mathfrak{L}$ плохо обусловлен. Оценка обусловленности гарантируется указанием векторов $\overline{X}$, $\underline{X}$ и чисел $\sigma(\overline{X})$, $\sigma(\underline{X})$.

8°.3. Вычисляется новая правая часть для итерационного уточнения полученного приближенного решения:

$$F^0_{i+1} = V^0_{M_1+1};\quad F^1_{i+1} = V^1_{M_1+1} + F^1_i.$$

Затем в качестве i берется $i+1$, и управление передается п. 2°.1.

9°. Выходная информация. В результате работы алгоритма выдаются следующие данные:

1) $\underline{\sigma}$ — оценка снизу максимального сингулярного числа оператора $\mathfrak{L}$ и $\overline{\sigma}$ — оценка сверху минимального сингулярного числа оператора $\mathfrak{L}$;

2) векторы $\overline{X}$, $\underline{X}$, на которых отношение Рэлея оператора $\mathfrak{L}$ достигает соответственно значений $\overline{\sigma}$ и $\underline{\sigma}$;

3) $\widehat{H} = \{\widehat{H}_j\}_{j=1}^N$ — полученное приближенное решение (6.1);

4) ε_5 — норма невязки полученного приближения ($\varepsilon_5 \leqslant \rho$).

§ 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:

$$c_1 = 2;\ c_2 = 1.2705;\ c_3 = 0.86029;\ c_4 = 0.48627;$$
$$c_5 = 0.2_{10} - 9;\ c_6 = 0.2_{10} - 10;$$
$$\sigma_1(B_3) = \sigma_2(L)\,(1 - 0.15);\ \sigma_2(B_3) = \sigma_3(L)\,(1 - 0.014);$$
$$\sigma_3(B_3) = \sigma_4(L)\,(1 - 0.0002);\ \sigma_1(B_4) = \sigma_1(L);$$
$$\sigma_2(B_4) = \sigma_2(L);\ \sigma_3(B_4) = \sigma_3(L);\ \sigma_4(B_4) = \sigma_4(L);$$
$$\sigma_1(B_5) = \sigma_1(L)\,(1 - 0.1_{10} - 7);\ \sigma_2(B_5) = \sigma_2(L)\,(1 - 0.1_{10} - 8);$$
$$\sigma_3(B_5) = \sigma_3(L)\,(1 - 0.1_{10} - 10);\ \sigma_4(B_5) = \sigma_4(L)\,(1 - 0.0004);$$
$$\sigma_5(B_5) = \sigma_4(L);\ \sigma_1(B_6) = (1 - 0.1_{10} - 7)\sigma_1(L);$$
$$\sigma_2(B_6) = \sigma_2(L)\,(1 - 0.1_{10} - 11);\ \sigma_3(B_6) = (1 - 0.001)\sigma_3(L);$$
$$\sigma_4(B_6) = \sigma_3(L)\,(1 - 0.1_{10} - 10);\ \sigma_5(B_6) = (1 - 0.1_{10} - 11)\sigma_4(L);$$
$$\sigma_6(B_6) = \sigma_4(L)\,(1 + 0.0001).$$

Сравнивая с результатами примера 1 § 3, можем заключить, что невязки 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:

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

Ясно, что невязка постоянно убывает.

ЛИТЕРАТУРА

1. **Булгаков А. Я., Годунов С. К.** Численное определение одного из критериев качества устойчивости систем линейных дифференциальных уравнений с постоянными коэффициентами.— Новосибирск, 1981.— 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 мат. физики, 1979, т. 19, № 5, с. 1313—1317.
4. **Hestenes M. R., Stiefel E.** Methods of conjugate gradients for solving linear systems.— Nat. Bur. Standards. J. Res., 1952, v. 49, p. 409—436.
5. **Федоренко Р. П.** Приближенное решение задач оптимального управления.— М.: Наука, 1978.— 488 с.
6. **Фаддеев Д. К., Фаддеева В. Н.** Вычислительные методы линейной алгебры.— М.—Л.: Физматгиз, 1963.— 734 с.
7. **Годунов С. К.** Решение систем линейных уравнений.— Новосибирск: Наука, Сиб. отд-ние, 1980.— 177 с.
8. **Форсайт Дж., Моулер К.** Численное решение систем линейных алгебраических уравнений.— М.: Мир, 1969.— 167 с.
9. **Булгаков А. Я.** Вычисление экспоненты от асимптотически устойчивой матрицы.— В кн.: Вычислительные методы линейной алгебры. Новосибирск: Наука, 1985, с. 4—17.

О СХОДИМОСТИ ОРТОГОНАЛЬНО-СТЕПЕННОГО МЕТОДА РАСЧЕТА СПЕКТРА

В. И. КОСТИН, 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

$$\det(A - \lambda I) = \begin{vmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & & a_{nn} - \lambda \end{vmatrix} =$$
$$= (-1)^n (\lambda^n - p_1\lambda^{n-1} - p_2\lambda^{n-2} - \dots - p_n) = 0.$$

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