О. П. КИРИЛЮК

ИТЕРАЦИОННОЕ УТОЧНЕНИЕ РЕШЕНИЯ СИСТЕМ ЛИНЕЙНЫХ УРАВНЕНИЙ

Данная работа посвящена итерационному уточнению решения систем линейных уравнений как с квадратными, так и с прямоугольными матрицами коэффициентов полного ранга.

Идея применения итерационного уточнения при решении линейных систем на ЭВМ появилась достаточно давно (см. [1, с. 121—126]). В первых же публикациях на эту тему [1; 2, с. 235—241; 3] проводился анализ погрешностей соответствующих вычислительных процедур (в [1] — для арифметики с фиксированной точкой, в [3] — для арифметики с плавающей точкой). В [1—3] было также замечено, что при итерационном уточнении следует наибольшее внимание уделять вычислению невязки системы (предлагалось использовать на этом этапе арифметику повышенной точности). В [4, 5] были предложены два основных подхода к итерационному уточнению решения систем с прямоугольными матрицами коэффициентов полного ранга: 1) умножение исходной системы на сопряженную матрицу коэффициентов, 2) переход к совместной расширенной системе уравнений. При этом второй подход предусматривал наличие в матрице коэффициентов расширенной системы свободного параметра, позволяющего влиять на обусловленность расширенной системы.

В литературе изложено несколько различных взглядов на итерационное уточнение. При этом различия состоят не только в таких вопросах, как критерий прекращения итераций, но и даже в целях применения итерационного уточнения. Например, в [6] сходимость итерационного уточнения служит критерием хорошей обусловленности решаемой системы, а в [7], напротив, обусловленность решаемой системы предполагается известной и по ней на основании анализа погрешностей априорно определяется количество итераций. При этом в [7] итерационное уточнение используется лишь для улучшения точности решения.

В данной работе изложение ведется в рамках концепции гарантированной точности машинных вычислений, сформулированной в [8, 9]. Результатом вычислений является либо отказ от решения и утверждение, что исходная задача плохо поставлена (очень большая обусловленность или несовместность системы), либо вычисление решения с точностью, близкой к точности представления вектора в ЭВМ.

В работе не описана детальная реализация всех формул уточнения, а сформулированы лишь требования на точность элементарных операций итерационного процесса (см. (1.3) и § 3). Соответствующая реализация дана в [9] и основана на

1) предварительной нормировке матрицы и правой части системы вынесением порядка,

2) использовании арифметики вынесенных порядков в матрично-векторных операциях.

Автор хотел бы выразить благодарность С. К. Годунову и В. И. Костину за помощь при выполнении данной работы.

§ 1. ИТЕРАЦИОННОЕ УТОЧНЕНИЕ РЕШЕНИЯ СИСТЕМ С КВАДРАТНОЙ МАТРИЦЕЙ КОЭФФИЦИЕНТОВ ПОЛНОГО РАНГА

Пусть требуется решить систему

$$Ax = f \tag{1.1}$$

с квадратной $N \times N$-матрицей A ранга N, используя некоторый метод приближенного решения этой системы. Обозначим приближенное решение системы (1.1) через $\tilde{x} = A^{\ominus} f$ (обозначение $A^{\ominus}$ выбрано по аналогии с обозначением A^{-1}). Мы опишем простой итерационный процесс уточнения решения $\tilde{x}$ в случае, когда данный метод гарантирует близость $\tilde{x}$ к x вида

$$\|\tilde{x} - x\| \leqslant \varepsilon \|x\|, \ \|A^{\ominus} f - A^{-1} f\| \leqslant \varepsilon \|A^{-1} f\|, \tag{1.2}$$

причем константа ε не зависит от правой части f системы (1.1). Символами $\odot$, $\ominus$ и $\oplus$ будем обозначать машинные операции умножения матрицы на вектор, вычитания и сложения двух векторов соответственно, допускающие следующие оценки точности:

$$\begin{aligned} \|A \odot x - Ax\| &\leqslant c_1 \varepsilon_1^2 \|A\| \|x\|, \\ \|(x \ominus y) - (x - y)\| &\leqslant \varepsilon_1 \|x - y\| + c_2 \varepsilon_0 (\|x\| + \|y\|), \\ \|(x \oplus y) - (x + y)\| &\leqslant \varepsilon_1 \|x + y\| + c_2 \varepsilon_0 (\|x\| + \|y\|), \end{aligned} \tag{1.3}$$

где c_1, c_2, ε_1, ε_0 — константы, причем c_1, c_2 зависят от порядка N матрицы A и векторов x и y, а ε_1 и ε_0 — от разрядной сетки ЭВМ, на которой проводятся вычисления.

Исследуем сначала модельный процесс итерационного уточнения, состоящий из k шагов и описываемый следующим образом.

Алгоритм I

Шаг 0. Полагаем $\tilde{x}_0 = A^{\ominus} f$.

Шаг i ($i = 1, \ldots, k$). Полагаем

$$\tilde{y}_i = A \odot \tilde{x}_i, \ \tilde{\tilde{r}}_i = f \ominus \tilde{y}_i, \ \tilde{x}_{i+1} = \tilde{x}_i \oplus A^{\ominus} \tilde{\tilde{r}}_i. \tag{1.4}$$

Если $i + 1 < k$, то вычисляем (1.4), заменив i на $i + 1$; если $i + 1 = k$, считаем расчет оконченным. Результатом расчета является $\tilde{x}_k$.

Вектор $\tilde{x}_k$ будем называть *итерационным уточнением решения системы* (1.1) *после k итераций.*

Целью приводимых ниже в данном параграфе рассуждений является:

— Получение оценки относительной близости результата модельного итерационного уточнения к точному решению системы (1.1), т. е. получение константы q_k такой, что

$$\|\tilde{x}_k - x\| \leqslant q_k \|x\|. \tag{1.5}$$

— Выяснение условий на ε, ε_1, ε_0, c_1, c_2, $\mu(A)$, при которых гарантируется, что через некотоое число итераций $k_0 < \infty$ относительная точность решения модельного итерационного уточнения окажется близкой (в данной статье не больше $2\varepsilon_1$) к ε_1. Определение соответствующего числа итераций k_0.

— Формулировка и обоснование алгоритма итерационного уточнения решения системы (1.1) на основании анализа модельного итерационного уточнения.

Начнем анализ модельного итерационного уточнения с получения константы q_k из (1.5). Будем рассуждать по индукции. Согласно (1.2) для вектора $\tilde{x}_0$ верна оценка $\|\tilde{x}_0 - x\| \leqslant q_0\|x\| = \varepsilon\|x\|$.

Лемма 1. *Пусть для некоторого i верна оценка*

$$\|\tilde{x}_i - x\| \leqslant q_i\|x\|. \tag{1.6}$$

Тогда

$$\|\tilde{x}_{i+1} - x\| \leqslant q_{i+1}\|x\|,$$

где

$$q_{i+1} = \alpha q_i + \beta,$$

$$\alpha = \{\varepsilon + [\varepsilon_1 + c_1\varepsilon_1^2(1 + \varepsilon_1 + c_2\varepsilon_0) + c_2\varepsilon_0]\,\mu(A)\,(1 + \varepsilon)\} \times$$
$$\times (1 + \varepsilon_1 + c_2\varepsilon_0) + 2c_2\varepsilon_0, \tag{1.7}$$

$$\beta = \varepsilon_1 + [c_1\varepsilon_1^2(1 + \varepsilon_1 + c_2\varepsilon_0) + 2c_2\varepsilon_0]\,\mu(A)\,(1 + \varepsilon_1 + c_2\varepsilon_0)\,(1 + \varepsilon) + c_2\varepsilon_0.$$

Доказательства. Достаточно подставить в (1.6) формулу (1.4), определяющую векторы $\tilde{x}_{i+1}$, $\tilde{\tilde{r}}_i$ и y_i, перейти в полученных неравенствах с помощью (1.3) от машинных операций $\odot$, $\ominus$, $\oplus$ к обычным арифметическим и воспользоваться линейностью решения систем вида (1.1) по правым частям, а также невырожденностью матрицы A. Проведем выкладки подробно.

В силу (1.3) легко убедиться в справедливости следующей цепочки неравенств:

$$\|\tilde{x}_{i+1} - x\| = \|\tilde{x}_i \oplus A^{\ominus}\tilde{\tilde{r}}_i - x\| \leqslant \|\tilde{x}_i - x + A^{\ominus}\tilde{\tilde{r}}_i\| +$$
$$\|[\tilde{x}_i \oplus A^{\ominus}\tilde{\tilde{r}}_i] - [\tilde{x}_i + A^{\ominus}\tilde{\tilde{r}}_i]\| \leqslant \|\tilde{x}_i - x + A^{\ominus}\tilde{\tilde{r}}_i\| +$$
$$+ \varepsilon_1\|\tilde{x}_i + A^{\ominus}\tilde{\tilde{r}}_i\| + c_2\varepsilon_0(\|\tilde{x}_i\| + \|A^{\ominus}\tilde{\tilde{r}}_i\|) \leqslant (1 + \varepsilon_1)\,\|\tilde{x}_i - x + A^{\ominus}\tilde{\tilde{r}}_i\| +$$
$$+ \varepsilon_1\|x\| + c_2\varepsilon_0(\|x\| + \|\tilde{x}_i - x\| + \|A^{\ominus}\tilde{\tilde{r}}_i\|).$$

Обозначим $r_i = f - A\tilde{x}_i$. Воспользовавшись далее неравенствами (1.6) и (1.2), получим

$$\|\tilde{x}_{i+1} - x\| \leqslant (1 + \varepsilon_1)\,\|\tilde{x}_i - x + A^{-1}\tilde{\tilde{r}}_i\| + [\varepsilon_1 + c_2\varepsilon_0(1 + q_i)]\,\|x\| +$$
$$+ c_2\varepsilon_0\|A^{-1}\tilde{\tilde{r}}_i\| + (1 + \varepsilon_1 + c_2\varepsilon_0)\,\|A^{\ominus}\tilde{\tilde{r}}_i - A^{-1}\tilde{\tilde{r}}_i\| \leqslant (1 + \varepsilon_1)\,\|\tilde{x}_i - x + A^{-1}\tilde{\tilde{r}}_i\| +$$
$$+ [\varepsilon_1 + c_2\varepsilon_0(1 + q_i)]\,\|x\| + [c_2\varepsilon_0 + \varepsilon(1 + \varepsilon_1 + c_2\varepsilon_0)]\,\|A^{-1}\tilde{\tilde{r}}_i\| \leqslant$$
$$\leqslant (1 + \varepsilon_1)\,\|\tilde{x}_i - x + A^{-1}r_i\| + [\varepsilon_1 + c_2\varepsilon_0(1 + q_i)]\,\|x\| +$$
$$+ [c_2\varepsilon_0 + \varepsilon(1 + \varepsilon_1 + c_2\varepsilon_0)]\,\|A^{-1}r_i\| + (1 + \varepsilon_1 + c_2\varepsilon_0)\,(1 + \varepsilon)\,\|A^{-1}\tilde{\tilde{r}}_i - A^{-1}r_i\|.$$

Очевидно, что $A^{-1}r_i - x = A^{-1}r_i - A^{-1}f = A^{-1}(f - A\tilde{x}_i - f) = -\tilde{x}_i$. Поэтому

$$\|\tilde{x}_i - x + A^{-1}r_i\| = \|\tilde{x}_i - \tilde{x}_i\| = 0, \quad \|A^{-1}r_i\| = \|\tilde{x}_i - x\| \leqslant q_i\|x\|$$

и оценку для $\|\tilde{x}_{i+1} - x\|$ можно продолжить так:

$$\|\tilde{x}_{i+1} - x\| \leqslant [\varepsilon_1 + c_2\varepsilon_0(1 + 2q_i) + \varepsilon q_i(1 + \varepsilon_1 + c_2\varepsilon_0)]\,\|x\| +$$
$$+ (1 + \varepsilon_1 + c_2\varepsilon_0)\,(1 + \varepsilon)\,\|A^{-1}\|\|\tilde{\tilde{r}}_i - r_i\|. \tag{1.8}$$

Для завершения доказательства леммы нам осталось оценить $\|\tilde{\tilde{r}}_i - r_i\|$, а именно, показать, что

$$\|\tilde{\tilde{r}}_i - r_i\| \leqslant [q_i(\varepsilon_1 + c_2\varepsilon_0) + c_1\varepsilon_1^2(1 + q_i)\,(1 + \varepsilon_1 + c_2\varepsilon_0) + 2c_2\varepsilon_0]\,\|A\|\|x\|.$$

Введем еще два обозначения $y_i = A\tilde{x}_i$ и $\tilde{r}_i = f - \tilde{y}_i$. Используя опять неравенства (1.3), связывающие машинные операции $\odot$ и $\ominus$ с обычными арифметическими, имеем

$$\|\tilde{\tilde{r}}_i - r_i\| \leqslant \|\tilde{\tilde{r}}_i - \tilde{r}_i\| + \|\tilde{r}_i - r_i\| = \|(f \ominus \tilde{y}_i) - (f - \tilde{y}_i\| +$$
$$+ \|\tilde{r}_i - r_i\| \leqslant \varepsilon_1\|f - \tilde{y}_i\| + \|\tilde{r}_i - r_i\| + c_2\varepsilon_0(\|f\| + \|\tilde{y}_i\|) =$$
$$= \varepsilon_1\|f - \tilde{y}_i\| + \|\tilde{y}_i - y_i\| + c_2\varepsilon_0(\|f\| + \|\tilde{y}_i\|) \leqslant \varepsilon_1\|f - y_i\| +$$
$$+ (1 + \varepsilon_1 + c_2\varepsilon_0)\,\|\tilde{y}_i - y\| + c_2\varepsilon_0(\|f\| + \|y_i\|). \tag{1.9}$$

Следовательно,

$$\|\tilde{r}_i - r_i\| \leqslant \varepsilon_1 \|A(x - \tilde{x}_i)\| + (1 + \varepsilon_1 + c_2\varepsilon_0)\|A \odot \tilde{x}_i - A\tilde{x}_i\| +$$
$$+ c_2\varepsilon_0(\|Ax\| + \|A\tilde{x}_i\|) \leqslant \varepsilon_1\|A(x - \tilde{x}_i)\| + (1 + \varepsilon_1 + c_2\varepsilon_0)c_1\varepsilon_1^2\|A\|\|\tilde{x}_i\| +$$
$$+ c_2\varepsilon_2(\|Ax\| + \|A\tilde{x}_i\|) \leqslant [q_i(\varepsilon_1 + c_2\varepsilon_0) + 2c_2\varepsilon_0]\|A\|\|x\| +$$
$$+ (1 + \varepsilon_1 + c_2\varepsilon_0)c_1\varepsilon_1^2\|A\|\|\tilde{x}_i\| \leqslant [q_i(\varepsilon_1 + c_2\varepsilon_0) + c_1\varepsilon_1^2(1 + q_i)(1 + \varepsilon_1 + c_2\varepsilon_0) +$$
$$+ 2c_2\varepsilon_0]\|A\|\|x\|.$$

Подставляя полученную оценку в (1.8) и собирая члены при q_i, получим утверждение леммы. Лемма доказана.

Итак, $q_0 = \varepsilon$, $q_{i+1} = \alpha q_i + \beta$. Откуда

$$q_i = \alpha q_{i-1} + \beta = \alpha(\alpha q_{i-2} + \beta) + \beta = \alpha^2 q_{i-2} + \beta(1 + \alpha) = \ldots = \alpha^i q_0 +$$
$$+ \beta(1 + \alpha + \ldots + \alpha^{i-1}) = \varepsilon\alpha^i + \beta\frac{1 - \alpha^i}{1 - \alpha} = \alpha^i\left(\varepsilon - \frac{\beta}{1 - \alpha}\right) + \frac{\beta}{1 - \alpha}.$$

Согласно (1.7) верно $\alpha > \varepsilon$. Поэтому при $\alpha < 1$

$$q_i < \varepsilon\alpha^i + \frac{\beta}{1 - \alpha} < \alpha^{i+1} + \frac{\beta}{1 - \alpha}$$

и $q_k < 2\varepsilon_1$ при

$$k \geqslant \frac{\log_2(2\varepsilon_1 - \beta/(1 - \alpha))}{\log_2 \alpha} - 1.$$

Кроме того, условий $\alpha \leqslant 1/2$ и $\beta/(1-\alpha) \leqslant \frac{3}{2}\varepsilon_1$ достаточно, чтобы за $k = \log_2(1/\varepsilon_1)$ шагов модельного итерационного уточнения (для двойной точности ЕС ЭВМ, например, это составляет 52 итерации) получить оценку

$$q_k = \frac{1}{2^{\log_2(1/\varepsilon_1) + 1}} + \frac{3}{2}\varepsilon_1 = \frac{\varepsilon_1}{2} + \frac{3}{2}\varepsilon_1 = 2\varepsilon_1.$$

Таким образом, приходим к следующему выводу.

Используя метод приближенного решения систем вида (1.1), *дающий гарантию точности вида* (1.2), *и машинные операции* $\odot$, $\ominus$ *и* $\oplus$, *удовлетворяющие неравенствам* (1.3) *в случае* $\alpha \leqslant 1/2, \beta/(1-\alpha) \leqslant \frac{3}{2}\varepsilon_1$, *где* α *и* β *определяются согласно* (1.7), *можно с помощью модельного итерационного уточнения* (1.4) *довести относительную точность решения системы* (1.1) *до* $2\varepsilon_1$. *При этом достаточно проделать*

$$k = \left[\frac{\log_2(2\varepsilon_1 - \beta/(1 - \alpha))}{\log_2 \alpha} - 1\right]_+ \qquad (1.10)$$

итераций (*всегда* $k \leqslant \log_2(1/\varepsilon_1)$). Символом $[x]_+$ здесь обозначается ближайшее к x целое число, большее x.

Проиллюстрируем полученное утверждение на одном примере. В [8] описан способ вычисления решения системы (1.1), дающий оценку точности решения $\varepsilon = \varepsilon(\mu(A), N)$ вида (1.2). По формулам (1.7), (1.10) для этого метода было просчитано k итераций, достаточных для достижения относительной точности уточненного решения $2\varepsilon_1$. Результаты сведены в следующую таблицу:

$\mu(A) \setminus N$	100	300	500	700	1000	10 000
10^2	1	1	1	2	2	2
10^3	1	2	2	2	2	3
10^4	2	2	2	2	2	3
10^5	2	2	3	3	3	5
10^6	3	3	3	4	4	7
10^7	3	4	5	5	6	15
10^8	5	6	7	8	10	—
10^9	7	12	16	22	38	—
10^{10}	15	—	—	—	—	—

Заметим, что в описанном в [8] методе решения систем вида (1.1) в процессе решения матрица A приводится к двухдиагональному виду. Для этого этапа решения требуется порядка N^3 операций, в то время как на остальные — лишь порядка N^2 арифметических операций. Очевидно, что для многократного решения системы на шаге i ($i = 1, \ldots, k-1$) алгоритма I нет необходимости этап приведения матрицы A к двухдиагональному виду повторять многократно. Следовательно, итерационное уточнение не требует больших временных затрат.

Приведем теперь несложную модификацию алгоритма I, позволяющую во многих случаях уменьшить число итераций, требуемое для достижения относительной точности решения $2\varepsilon_1$. Предлагаемая модификация отличается от модельного итерационного уточнения тем, что вместо априорной оценки точности i-го приближения x_i к решению x используется апостеорная оценка точности, вычисленная по невязке $\tilde{r}_i$.

Алгоритм II

Шаг 1. Приводим матрицу A к двухдиагональному виду (или другому виду, позволяющему быстро вычислять $A^{\ominus}g$ для любого вектора g), вычисляем $\|A\|$, $\|A^{-1}\|$, определяем характеристику точности решения системы ε, величин α и β (по формулам (1.7)), а также $\delta = c_1\varepsilon_1^2(1+\varepsilon_1+c_2\varepsilon_0)+c_2\varepsilon_0$. (Процесс вычисления надо вести так, чтобы приближенно вычисленные величины $\|A\|$, $\|A^{-1}\|$, α, β и δ были не меньше соответствующих точных значений.) Если $\alpha > 1/2$ или $\beta/(1-\alpha) > \frac{3}{2}\varepsilon_1$, то процесс итерационного уточнения заканчивается отказом от вычисления решения, иначе полагаем

$$q_0 = \varepsilon, \qquad k = \left[\frac{\log_2(2\varepsilon_1 - \beta/(1-\alpha))}{\log_2 \alpha} - 1\right]_+$$

и переходим к шагу 2.

Шаг 2. Полагаем $\tilde{x}_0 = A^{\ominus}f$.

Шаг 3. Полагаем

$$\tilde{y}_i = A \odot \tilde{x}_i, \quad \tilde{\tilde{r}}_i = f \ominus \tilde{y}_i, \tag{1.11}$$

$$p_i = \frac{\dfrac{\|A^{-1}\|\,\|\tilde{\tilde{r}}_i\|}{\|\tilde{x}_i\|} + (\delta + c_2\varepsilon_0)\,\mu(A)}{1 - \varepsilon_1 - \dfrac{\|A^{-1}\|\,\|\tilde{\tilde{r}}_i\|}{\|\tilde{x}_i\|} - \delta\mu(A)}.$$

Если $p_i > 2\varepsilon_1$, то переходим к шагу 4, иначе считаем расчет законченным с результатом $\tilde{x}_i$.

Шаг 4. Полагаем

$$\tilde{z}_i = A^{\ominus}\tilde{\tilde{r}}_i,$$

$$\bar{p}_i = \frac{\dfrac{\|\tilde{z}_i\|(1-\varepsilon_1)/(1-\varepsilon) + \varepsilon_1\|A^{-1}\|\,\|\tilde{\tilde{r}}_i\|}{\|\tilde{x}_i\|} + (\delta + c_2\varepsilon_0)\,\mu(A)}{1 - \varepsilon_1 - \dfrac{\|\tilde{z}_i\|(1-\varepsilon_1)/(1-\varepsilon) + \varepsilon_1\|A^{-1}\|\,\|\tilde{\tilde{r}}_i\|}{\|\tilde{x}_i\|} - \delta\mu(A)},$$

$$q_{i+1} = \alpha \min(q_i, p_i, \bar{p}_i) + \beta, \qquad \tilde{x}_{i+1} = \tilde{x}_i \oplus \tilde{z}_i.$$

Если $i+1 < k$ или $q_{i+1} > 2\varepsilon_1$, то переходим к шагу 3, заменив i на $i+1$, иначе считаем расчет законченным с результатом $\tilde{x}_{i+1}$.

Еще раз подчеркнем, что вычисления p_i, $\bar{p}_i$ и всех скалярных величин на шаге 1 алгоритма II необходимо проводить так, чтобы приближенный результат был не меньше точного. Для этого можно использо-

вать так называемые завышенные и заниженные арифметические операции, описанные в гл. IV [9].

Для обоснования приведенного процесса докажем следующую лемму.

Лемма 2. *Пусть* $\tilde{\tilde{r}}_i = f \ominus (A \odot \tilde{x}_i)$ *и* x *— решение уравнения* (1.1). *Тогда верна оценка*

$$\|\tilde{x}_i - x\| \leqslant \hat{p}_i \|x\|,$$

$$\hat{p}_i = \frac{\dfrac{\|A^{-1}\tilde{\tilde{r}}_i\|(1-\varepsilon_1) + \varepsilon_1 \|A^{-1}\| \|\tilde{\tilde{r}}_i\|}{\|\tilde{x}_i\|} + (\delta + c_2\varepsilon_0)\,\mu(A)}{1 - \varepsilon_1 - \dfrac{\|A^{-1}\tilde{\tilde{r}}_i\|(1-\varepsilon_1) + \varepsilon_1 \|A^{-1}\| \|\tilde{\tilde{r}}_i\|}{\|\tilde{x}_i\|} - \delta\mu(A)},$$

$$\delta = c_1\varepsilon_1^2(1 + \varepsilon_1 + c_2\varepsilon_0) + c_2\varepsilon_0.$$

Доказательство. Будем пользоваться введенными ранее обозначениями. По определению, $r_i = f - A\tilde{x}_i$. Поэтому $\tilde{x}_i = A^{-1}(f - r_i)$ и

$$\|\tilde{x}_i - x\| \leqslant \|A^{-1}r_i\| \leqslant \|A^{-1}\tilde{\tilde{r}}_i\| + \|A^{-1}\|\|\tilde{\tilde{r}}_i - r_i\| =$$

$$= \|A^{-1}\tilde{\tilde{r}}_i\|\|\tilde{x}_i\|/\|\tilde{x}_i\| + \|A^{-1}\|\|\tilde{\tilde{r}}_i - r_i\| \leqslant$$

$$\leqslant (\|A^{-1}\tilde{\tilde{r}}_i\|/\|\tilde{x}_i\|)\|x\| + (\|A^{-1}\tilde{\tilde{r}}_i\|/\|\tilde{x}_i\|)\|\tilde{x}_i - x\| + \|A^{-1}\|\|\tilde{\tilde{r}}_i - r_i\|. \qquad (1.12)$$

Согласно (1.9) и (1.3)

$$\|\tilde{\tilde{r}}_i - r_i\| \leqslant \varepsilon_1\|f - A\tilde{x}_i\| + (1 + \varepsilon_1 + c_2\varepsilon_0)\|A \odot \tilde{x}_i - A\tilde{x}_i\| +$$

$$+ c_2\varepsilon_0(\|Ax\| + \|A\tilde{x}_i\|) \leqslant \varepsilon_1\|r_i\| + c_1\varepsilon_1^2(1 + \varepsilon_1 + c_2\varepsilon_0)\|A\|\|\tilde{x}_i\| + 2c_2\varepsilon_0\|Ax\| +$$

$$+ c_2\varepsilon_0\|A(\tilde{x}_i - x)\| \leqslant \varepsilon_1\|r_i\| + \delta\|A\|\|\tilde{x}_i - x\| + (\delta + c_2\varepsilon_0)\|A\|\|x\|,$$

следовательно,

$$\|\tilde{\tilde{r}}_i - r_i\| \leqslant \frac{1}{1-\varepsilon_1}[\varepsilon_1\|\tilde{\tilde{r}}_i\| + \delta\|A\|\|\tilde{x}_i - x\| + (\delta + c_2\varepsilon_0)\|A\|\|x\|] \leqslant$$

$$\leqslant \frac{1}{1-\varepsilon_1}\left[\frac{\varepsilon_1\|\tilde{\tilde{r}}_i\|}{\|\tilde{x}_i\|}\|x\| + \left(\frac{\varepsilon_1\|\tilde{\tilde{r}}_i\|}{\|\tilde{x}_i\|} + \delta\|A\|\right)\|\tilde{x}_i - x\| + (\delta + c_2\varepsilon_0)\|A\|\|x\|\right]$$

и неравенство (1.12) можно продолжить так:

$$\|\tilde{x}_i - x\| \leqslant \left[\frac{\|A^{-1}\tilde{\tilde{r}}_i\|}{\|\tilde{x}_i\|} + \frac{\varepsilon_1\|A^{-1}\|\|\tilde{\tilde{r}}_i\|}{(1-\varepsilon_1)\|\tilde{x}_i\|} + \frac{\delta + c_2\varepsilon_0}{1-\varepsilon_1}\mu(A)\right]\|x\| +$$

$$+ \left[\frac{\|A^{-1}\tilde{\tilde{r}}_i\|}{\|\tilde{x}_i\|} + \frac{\varepsilon_1\|A^{-1}\|\|\tilde{\tilde{r}}_i\|}{(1-\varepsilon_1)\|\tilde{x}_i\|} + \frac{\delta}{1-\varepsilon_1}\mu(A)\right]\|\tilde{x}_i - x\|.$$

Собирая члены при $\|\tilde{x}_i - x\|$ и $\|x\|$, получаем в итоге утверждение леммы.

Итак, лемма 2 доказана. Применение ее к обоснованию алгоритма итерационного уточнения (1.11) и учет очевидных неравенств $\hat{p}_i \leqslant p_i$, $\hat{p}_i \leqslant \bar{p}_i$ делают обоснование этого процесса тривиальным. Таким образом, процесс итерационного уточнения решения систем с квадратной матрицей полного ранга, задаваемый (1.11), полностью обоснован.

§ 2. ИТЕРАЦИОННОЕ УТОЧНЕНИЕ ОБОБЩЕННОГО И НОРМАЛЬНОГО РЕШЕНИЙ СИСТЕМ С ПРЯМОУГОЛЬНОЙ МАТРИЦЕЙ КОЭФФИЦИЕНТОВ ПОЛНОГО РАНГА

Пусть дана система

$$Ax = f \qquad (2.1)$$

с прямоугольной $N \times M$-матрицей ранга $\min(N, M)$. В § 1 был подробно разобран случай $N = M$. Здесь мы остановимся на рассмотрении двух других случаев: $N > M$ и $N < M$.

Прежде всего уточним понятие решения систем вида (2.1). При $N<M$ существует множество векторов x, удовлетворяющих (2.1). Наименьшей по норме из них называется *нормальным решением системы* (2.1). При $N>M$ система (2.1), вообще говоря, несовместна, т. е. не имеет ни одного вектора, удовлетворяющего (2.1). В этом случае под *обобщенным решением системы* (2.1) понимаем такой вектор x, на котором невязка системы минимальна: $\|Ax-f\|=\min\limits_{u}\|Au-f\|$. *Параметром несовместности системы* (2.1) будем называть величину

$$\nu=\nu(A,f)=\|A^{+}\|\frac{\|Ax-f\|}{\|x\|}. \tag{2.2}$$

Как известно (см., например, § 38 [10]), процесс итерационного уточнения, аналогичный описанному в § 2, для решения прямоугольных систем вида (2.1) неприемлем. В случае $N>M$ это прежде всего связано с тем, что оценка точности обобщенного решения системы (2.1) зависит от характеристики несовместности системы (например, от параметра $\nu(A,f)$). При этом в процессе итераций по мере приближения к обобщенному решению пришлось бы решать все менее совместные системы (т. е. с бóльшим ν). Поэтому обосновать сходимость соответствующего итерационного процесса в описанной выше схеме не удается. В случае $N<M$ мы сталкиваемся с другой трудностью. Итерациями можно улучшить невязку системы (2.1). Однако существует множество векторов, удовлетворяющих системе (2.1), и итерационный процесс не различает эти векторы. Поэтому рассчитывать на приближение в процессе итераций к нормальному решению не приходится.

Для преодоления перечисленных трудностей предлагается перейти от системы (2.1) к расширенным системам

$$\begin{bmatrix}\rho I_N & A\\ A^{*} & 0\end{bmatrix}\begin{pmatrix}y\\ x\end{pmatrix}=\begin{pmatrix}f\\ 0\end{pmatrix} \tag{2.3}$$

при $N>M$ и

$$\begin{bmatrix}0 & A\\ A^{*} & \rho I_M\end{bmatrix}\begin{pmatrix}z\\ x\end{pmatrix}=\begin{pmatrix}f\\ 0\end{pmatrix} \tag{2.4}$$

при $N<M$. При любом ненулевом параметре ρ матрицы в (2.3) и (2.4) квадратные, имеют порядок $(N+M)$ и полный ранг, а x-е компоненты решения расширенных систем (2.3) и (2.4) являются соответственно обобщенным и нормальным решениями (2.1).

Переход к расширенной системе при отыскании обобщенного или нормального решения системы (2.1) известен давно (см. [5]). Однако в литературе рассматривался, как правило, лишь случай $\rho=1$. Но как показано в [5, 11], обусловленность расширенной системы равна величине

$$\frac{\frac{1}{2}+\sqrt{\frac{1}{4}+\left(\frac{\sigma_N}{\rho}\right)^2}}{\min\left(1,\sqrt{\frac{1}{4}+\left(\frac{\sigma_1}{\rho}\right)^2}-\frac{1}{2}\right)},$$

где σ_N — операторная норма $A-\|A\|$, а $\sigma_1=1/\|A^{+}\|$ и при $\sigma_N=\|A\|\sim 1$ выбор $\rho=1$ даст обусловленность расширенной системы порядка $\mu^2(A)$, $\mu(A)=\|A\|\,\|A^{+}\|$. В то же время выбор $\rho=\sigma_1/\sqrt{2}$ приводит к величине $\frac{1}{2}+\sqrt{\frac{1}{4}+2\mu^2(A)}\approx\sqrt{2}\,\mu(A)$. При этом доказано, что выбор $\rho=\sigma_1/\sqrt{2}$ оптимален с точки зрения обусловленности расширенной системы.

Отметим еще одно свойство матриц расширенных систем. Для любых ортогональных матриц P и Q размеров $N\times N$ и $M\times M$ соответствен-

но верны равенства

$$\begin{bmatrix} P & 0 \\ 0 & Q \end{bmatrix}\begin{bmatrix} \rho I_N & A \\ A^* & 0 \end{bmatrix}\begin{bmatrix} P^* & 0 \\ 0 & Q \end{bmatrix} = \begin{bmatrix} \rho I_N & PAQ^* \\ QA^*P^* & 0 \end{bmatrix},$$

$$\begin{bmatrix} P & 0 \\ 0 & Q \end{bmatrix}\begin{bmatrix} 0 & A \\ A^* & \rho I_M \end{bmatrix}\begin{bmatrix} P^* & 0 \\ 0 & Q^* \end{bmatrix} = \begin{bmatrix} 0 & PQ^* \\ QA^*P^* & \rho I_M \end{bmatrix}.$$

Таким образом, зная приведение матрицы A к двухдиагональному виду, мы легко получаем приведение расширенной матрицы к разреженной треугольной форме, что позволяет быстро находить решения расширенных систем (2.3) или (2.4), а значит, выполнять процесс итерационного уточнения расширенных систем (см. § 1).

Итак, имея гарантированный способ решения систем (2.3) и (2.4), можно выполнить алгоритм II и тем самым получить приближенное решение расширенных систем с точностью $2\varepsilon_1$, т. е. получить векторы $\begin{pmatrix} \tilde{y} \\ \tilde{x} \end{pmatrix}$ и $\begin{pmatrix} \tilde{z} \\ \tilde{x} \end{pmatrix}$ такие, что

$$\left\|\begin{pmatrix} \tilde{y} \\ \tilde{x} \end{pmatrix} - \begin{pmatrix} y \\ x \end{pmatrix}\right\| \leqslant 2\varepsilon_1 \left\|\begin{pmatrix} y \\ x \end{pmatrix}\right\|, \quad \left\|\begin{pmatrix} \tilde{z} \\ \tilde{x} \end{pmatrix} - \begin{pmatrix} z \\ x \end{pmatrix}\right\| \leqslant 2\varepsilon_1 \left\|\begin{pmatrix} z \\ x \end{pmatrix}\right\|. \tag{2.5}$$

Используя неравенства (2.5), можно оценить относительную точность полученного нормального (соответственно обобщенного) решения исходной системы (2.1).

При $N < M$ имеем

$$\|\tilde{x} - x\| \leqslant \left\|\begin{pmatrix} \tilde{z} \\ \tilde{x} \end{pmatrix} - \begin{pmatrix} z \\ x \end{pmatrix}\right\| \leqslant 2\varepsilon_1 \left\|\begin{pmatrix} z \\ x \end{pmatrix}\right\| = 2\varepsilon_1 \sqrt{\|x\|^2 + \|z\|^2}.$$

В силу равенства $A^*z = -\rho x$ получим $z = -\rho (A^*)^+ x$ и, следовательно,

$$\|z\| = \rho \|(A^*)^+ x\| \leqslant \rho \|(A^*)^+\| \|x\| = \rho \frac{\|x\|}{\sigma_1(A)} = \frac{\sigma_1(A)}{\sqrt{2}} \frac{\|x\|}{\sigma_1(A)} = \frac{\|x\|}{\sqrt{2}}.$$

Поэтому оценку $\|\tilde{x} - x\|$ при $N < M$ можно продолжить так

$$\|\tilde{x} - x\| \leqslant 2\varepsilon_1 \sqrt{\|x\|^2 + \|z\|^2} \leqslant \sqrt{6}\varepsilon_1 \|x\|. \tag{2.6}$$

Итак, используя алгоритм II, можно получить нормальное решение системы (2.1) с гарантированной оценкой точности (2.6).

Аналогично, при $N > M$ имеем

$$\|\tilde{x} - x\| \leqslant 2\varepsilon_1 \sqrt{\|x\|^2 + \|y\|^2} = 2\varepsilon_1 \sqrt{\|x\|^2 + \frac{\|Ax - f\|^2}{\rho^2}} =$$

$$= 2\varepsilon_1 \sqrt{1 + \frac{2\|Ax - f\|^2}{\sigma_1^2(A)\|x\|^2}} \|x\| = 2\varepsilon_1 \sqrt{1 + 2\nu^2(A, f)} \|x\|. \tag{2.7}$$

Заметим, что при практическом осуществлении итерационного уточнения обобщенного (нормального) решения системы (2.1) удобнее пользоваться вместо (2.6), (2.7) итоговыми оценками точности

$$\begin{aligned} \|\tilde{x} - x\| &\leqslant \frac{2\varepsilon_1}{1 - 2\varepsilon_1} \sqrt{1 + \left(\frac{\|\tilde{y}\|}{\|\tilde{x}\|}\right)^2} \|\tilde{x}\|, \\ \|\tilde{x} - x\| &\leqslant \frac{2\varepsilon_1}{1 - 2\varepsilon_1} \sqrt{1 + \left(\frac{\|\tilde{z}\|}{\|\tilde{x}\|}\right)^2} \|\tilde{x}\|, \end{aligned} \tag{2.8}$$

непосредственно вытекающими из (2.5). В то же время оценки (2.6) и (2.7) полезны для понимания того, каков уровень достигаемой точности решения после итерационного уточнения и от чего этот уровень зависит.

Приведем еще несколько дополнительных замечаний. Пусть $N > M$. Обозначим через Π_0 оператор ортогонального проектирования на ядро матрицы A^*, а через $\Pi_1 = I - \Pi_0$ — оператор ортогонального проектирования на дополнение к ядру A^*. В новых обозначениях $\Pi_0 f = \rho y$, $\Pi_1 f = Ax$. Поэтому в силу неравенств

$$\|\Pi_1 f\| = \|Ax\| \leqslant \|A\| \|x\|,$$

$$\|x\| = \|A^+ f\| = \|A^+ \Pi_1 f\| \leqslant \|A^+\| \|\Pi_1 f\|$$

имеем

$$\nu(A, f) = \|A^+\| \frac{\|\Pi_0 f\|}{\|x\|} \leqslant \mu(A) \frac{\|\Pi_0 f\|}{\|\Pi_1 f\|},$$

$$\nu(A, f) = \|A^+\| \frac{\|\Pi_0 f\|}{\|x\|} \geqslant \frac{\|\Pi_0 f\|}{\|\Pi_1 f\|},$$

т. е.

$$\frac{\|\Pi_0 f\|}{\|\Pi_1 f\|} \leqslant \nu(A, f) \leqslant \mu(A) \frac{\|\Pi_0 f\|}{\|\Pi_1 f\|}. \tag{2.9}$$

Пользуясь (2.9), получим

$$\|\rho\tilde{y} - \Pi_0 f\| = \|\rho\tilde{y} - \rho y\| \leqslant 2\varepsilon_1 \rho \sqrt{\|x\|^2 + \|y\|^2} = 2\varepsilon_1 \sqrt{\rho^2 \|x\|^2 + \|\Pi_0 f\|^2} =$$

$$= 2\varepsilon_1 \sqrt{\frac{\sigma_1^2(A) \|x\|^2}{2\|\Pi_0 f\|^2} + 1} \|\Pi_0 f\| = 2\varepsilon_1 \sqrt{\frac{1}{2\nu^2} + 1} \|\Pi_0 f\| \leqslant$$

$$\leqslant 2\varepsilon_1 \sqrt{\frac{\|\Pi_1 f\|^2}{2\|\Pi_0 f\|^2} + 1} \|\Pi_0 f\|.$$

Следовательно, если f близко к ядру A^*, т. е., если $\|\Pi_1 f\| << \|\Pi_0 f\|$, вектор $\rho\tilde{y}$ будет с хорошей точностью совпадать с проекцией f на ядро A^*. Например, при $\|\Pi_1 f\| \leqslant \|\Pi_0 f\|$ имеем

$$\|\rho\tilde{y} - \Pi_0 f\| \leqslant \sqrt{6}\varepsilon_1 \|\Pi_0 f\|.$$

Таким образом, итерационное уточнение решения расширенной системы может быть использовано для уточнения ядра прямоугольной матрицы A^* полного ранга. Для этого в качестве f необходимо задавать базисные векторы ядра A^*. Тогда y-е компоненты решения расширенной системы, умноженные на ρ, будут давать уточненные базисные векторы ядра A^*. При этом, конечно, нарушается ортогональность векторов базиса. В этом случае указанную процедуру уточнения следует повторить.

§ 3. ОЦЕНКИ ПОГРЕШНОСТЕЙ МАШИННЫХ ВЫЧИСЛЕНИЙ ПРИ РЕШЕНИИ РАСШИРЕННОЙ СИСТЕМЫ УРАВНЕНИЙ

Предложенный в § 2 способ итерационного уточнения обобщенного (нормального) решения прямоугольной системы уравнений предполагает переход к расширенной системе, а затем использование рассмотренного в § 1 итерационного уточнения решения квадратной системы. Однако для выполнения алгоритма II необходимо знать оценку точности решения расширенной системы уравнений. Получением этой оценки мы сейчас и займемся.

Очевидно, что оценки точности решения систем (2.3) и (2.4) отличаются друг от друга лишь заменой N на M. Поэтому выведем оценку лишь для одной из систем, например для (2.3). Ввиду громоздкости выкладок часть из них проведем схематично.

Пусть требуется решить систему

$$\begin{bmatrix} \rho I_N & A \\ A^* & 0 \end{bmatrix} \begin{pmatrix} y \\ x \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}. \tag{3.1}$$

Матрицу системы (3.1) будем обозначать через B, а правую часть — через f. При выводе оценки точности решения (3.1) на ЭВМ мы будем использовать некоторые результаты из гл. IV [9]. Сформулируем их кратко. В [9] описан алгоритм ортогонального приведения прямоугольной $N \times M$-матрицы A к верхнедвухдиагональному виду D, т. е. алгоритм получения по матрице A двухдиагональной матрицы D и ортогональных преобразований P и Q таких, что

$$PAQ^* = \begin{bmatrix} D \\ \cdots \\ 0 \end{bmatrix} + \Delta, \tag{3.2}$$

причем

$$\|\Delta\| \leqslant 2M\sqrt{M}\tau\varepsilon_1\|A\|, \tag{3.3}$$

где $\tau = 34$, если при промежуточных вычислениях в процессе приведения использовалась арифметика удвоенной точности, и $\tau = 4N + 26$ в противном случае. Специфика хранения преобразований P и Q позволяет гарантировать, что точность их применения к произвольным векторам v и w размерностей N и M соответственно оценивается согласно формулам

$$Pv = \tilde{v} + \varphi, \quad Qw = \tilde{w} + \psi, \tag{3.4}$$

$$\|\varphi\| \leqslant M\tau\varepsilon_1\|v\|, \quad \|\psi\| \leqslant M\tau'\varepsilon_1\|w\|, \tag{3.5}$$

где $\tau = 34$ (или $4N + 26$), $\tau' = 34$ (или $4M + 26$), опять в зависимости от использования арифметики удвоенной точности. Векторы $\tilde{v}$ и $\tilde{w}$ являются результатом машинного применения преобразований P и Q к v и w соответственно, а векторы φ и ψ моделируют погрешности машинных вычислений.

Оценим точность решения системы (3.1). Прежде всего, следуя [9], определим по A ортогональные преобразования P и Q и двухдиагональную матрицу D, для которых имеют место (3.2) и (3.3). Затем по матрицам P и Q и векторам f_1, f_2 найдем вектор $g = \begin{pmatrix} g_1 \\ g_2 \end{pmatrix}$ такой, что

$$\begin{bmatrix} P & 0 \\ 0 & Q \end{bmatrix} \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} = \begin{pmatrix} P & f_1 \\ Q & f_2 \end{pmatrix} = \begin{pmatrix} g_1 + \xi_1 \\ g_2 + \xi_2 \end{pmatrix} = g + \xi, \tag{3.6}$$

а для ξ_1 и ξ_2 верны оценки типа (3.5).

В новых обозначениях систему (3.1) можно переписать так:

$$\begin{bmatrix} \rho I_N & \begin{bmatrix} D \\ \cdots \\ 0 \end{bmatrix} + \Delta \\ [D^* \vdots 0] + \Delta^* & 0 \end{bmatrix} u = g + \xi, \quad \text{где} \quad u \begin{pmatrix} Py \\ Qx \end{pmatrix} \tag{3.7}$$

(при реальных вычислениях на ЭВМ величины Δ и ξ, моделирующие машинные погрешности, неизвестны).

Проводя выкладки, аналогичные выкладкам из § 9 гл. IV [9], можно показать, что при решении на ЭВМ системы

$$\tilde{B}u = g, \quad \tilde{B} = \begin{bmatrix} \rho I_N & \begin{bmatrix} D \\ \cdots \\ 0 \end{bmatrix} \\ [D^* \vdots 0 & 0 \end{bmatrix}$$

мы находим вектор $\tilde{u}$, удовлетворяющий системе

$$(\tilde{B} + \Phi)\tilde{u} = g + \eta, \tag{3.8}$$

где

$$\|\Phi\| \leqslant \frac{2\sqrt{3}\,\varepsilon_1}{1-2\varepsilon_1}\|\widetilde{B}\|, \quad \|\eta\| \leqslant 2\varepsilon_1\|g\|. \tag{3.9}$$

Здесь, как и при выводе оценок (3.3), (3.5), предполагается, что матрица A и вектор f предварительно нормированы. Применяя далее ортогональное преобразование $\begin{bmatrix} P^* & 0 \\ 0 & Q^* \end{bmatrix}$ к вектору $\widetilde{u}$, получим

$$\begin{pmatrix} \widetilde{y} \\ \widetilde{x} \end{pmatrix} = \begin{bmatrix} P^* & 0 \\ 0 & Q^* \end{bmatrix} \widetilde{u} + \zeta = \begin{bmatrix} P^* & 0 \\ 0 & Q^* \end{bmatrix} \begin{pmatrix} \widetilde{u}_1 \\ \widetilde{u}_2 \end{pmatrix} + \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix}. \tag{3.10}$$

$$\|\zeta\| \leqslant \sqrt{(M\tau\varepsilon_1)^2\|\widetilde{u}_1\|^2 + (M\tau'\varepsilon_1)^2\|\widetilde{u}_2\|^2} \leqslant$$

$$\leqslant M\varepsilon_1 \max(\tau, \tau')\|\widetilde{u}\| = M\varepsilon_1\tau\|\widetilde{u}\| \leqslant \frac{M\varepsilon_1\tau}{1-M\varepsilon_1\tau}\left\|\begin{pmatrix} \widetilde{y} \\ \widetilde{x} \end{pmatrix}\right\|. \tag{3.11}$$

Собирая (3.2), (3.6), (3.8) и (3.10), приходим к равенству

$$\left(\begin{bmatrix} \rho I_N & A \\ A^* & 0 \end{bmatrix} + \Psi\right)\left(\begin{pmatrix} \widetilde{y} \\ \widetilde{x} \end{pmatrix} - \zeta\right) = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} + \psi, \tag{3.12}$$

где

$$\Psi = \begin{bmatrix} P^* & 0 \\ 0 & Q^* \end{bmatrix}\left(\Phi - \begin{bmatrix} 0 & \Delta \\ \Delta^* & 0 \end{bmatrix}\right)\begin{bmatrix} P & 0 \\ 0 & Q \end{bmatrix}, \quad \psi = \begin{bmatrix} P^* & 0 \\ 0 & Q^* \end{bmatrix}(\eta - \xi).$$

При этом согласно (3.3), (3.9)

$$\|\Psi\| = \left\|\Phi - \begin{bmatrix} 0 & \Delta \\ \Delta^* & 0 \end{bmatrix}\right\| \leqslant \|\Phi\| + \left\|\begin{bmatrix} 0 & \Delta \\ \Delta^* & 0 \end{bmatrix}\right\| = \|\Phi\| + \|\Delta\| \leqslant$$

$$\leqslant \frac{2\sqrt{3}\,\varepsilon_1}{1-2\varepsilon_1}\|\widetilde{B}\| + \|\Delta\| \leqslant \frac{2\sqrt{3}\,\varepsilon_1}{1-2\varepsilon_1}(\|B\| + \|\Delta\|) + \|\Delta\| \leqslant$$

$$\leqslant \frac{2\sqrt{3}\,\varepsilon_1}{1-2\varepsilon_1}\|B\| + 2M\sqrt{M}\,\tau\varepsilon_1\left(1 + \frac{2\sqrt{3}\,\varepsilon_1}{1-2\varepsilon_1}\right)\|A\|,$$

$$\|\psi\| = \|\eta - \xi\| \leqslant \|\eta\| + \|\xi\| \leqslant 2\varepsilon_1\|g\| + \|\xi\| \leqslant 2\varepsilon_1(\|f\| + \|\xi\|) +$$

$$+ \|\xi\| \leqslant 2\varepsilon_1\|f\| + (1+2\varepsilon_1)\sqrt{(M\tau\varepsilon_1)^2\|f_1\|^2 + (M\tau'\varepsilon_1)^2\|f_2\|^2} \leqslant$$

$$\leqslant [2\varepsilon_1 + M\varepsilon_1\tau(1+2\varepsilon_1)]\|f\|,$$

$$\left\|\begin{pmatrix} \widetilde{y} \\ \widetilde{x} \end{pmatrix}\right\| \leqslant \frac{1+\delta_3}{1-\delta_4}\left\|\begin{pmatrix} y \\ x \end{pmatrix}\right\|.$$

Таким образом, $\|\Psi\| \leqslant \delta_1\|B\|$, $\|\psi\| \leqslant \delta_2\|f\|$, где

$$\delta_1 = \frac{2\sqrt{3}\,\varepsilon_1}{1-2\varepsilon_1} + 2M\sqrt{M}\,\tau\varepsilon_1\left(1 + \frac{2\sqrt{3}\,\varepsilon_1}{1-2\varepsilon_1}\right)\frac{\|A\|}{\|B\|},$$

$$\delta_2 = 2\varepsilon_1 + M\tau\varepsilon_1(1+2\varepsilon_1).$$

Используя (3.11), получим

$$\left\|\begin{pmatrix} \widetilde{y} \\ \widetilde{x} \end{pmatrix} - \zeta - \begin{pmatrix} y \\ x \end{pmatrix}\right\| \leqslant \frac{(\delta_1+\delta_2)\,\mu(B)}{1-\delta_1\mu(B)}\left\|\begin{pmatrix} y \\ x \end{pmatrix}\right\|,$$

$$\left\|\begin{pmatrix} \widetilde{y} \\ \widetilde{x} \end{pmatrix} - \begin{pmatrix} y \\ x \end{pmatrix}\right\| \leqslant \frac{(\delta_1+\delta_2)\,\mu(B)}{1-\delta_1\mu(B)}\left\|\begin{pmatrix} y \\ x \end{pmatrix}\right\| + \|\zeta\| \leqslant \delta_3\left\|\begin{pmatrix} y \\ x \end{pmatrix}\right\| + \delta_4\left\|\begin{pmatrix} \widetilde{y} \\ \widetilde{x} \end{pmatrix}\right\|,$$

где

$$\delta_3 = \frac{(\delta_1 + \delta_2)\,\mu(B)}{1 - \delta_1 \mu(B)}, \quad \delta_4 = \frac{M\tau\varepsilon_1}{1 - M\tau\varepsilon_1}.$$

Поэтому

$$\left\|\begin{pmatrix}\tilde{y}\\ \tilde{x}\end{pmatrix}\right\| - \left\|\begin{pmatrix}y\\ x\end{pmatrix}\right\| \leqslant \delta_3 \left\|\begin{pmatrix}y\\ x\end{pmatrix}\right\| + \delta_4 \left\|\begin{pmatrix}\tilde{y}\\ \tilde{x}\end{pmatrix}\right\|, \quad \left\|\begin{pmatrix}\tilde{y}\\ \tilde{x}\end{pmatrix}\right\| \leqslant \frac{1+\delta_3}{1-\delta_4}\left\|\begin{pmatrix}y\\ x\end{pmatrix}\right\|,$$

откуда

$$\left\|\begin{pmatrix}\tilde{y}\\ \tilde{x}\end{pmatrix} - \begin{pmatrix}y\\ x\end{pmatrix}\right\| \leqslant \varepsilon \left\|\begin{pmatrix}y\\ x\end{pmatrix}\right\|, \tag{3.13}$$

$$\varepsilon = \delta_3 + \delta_4 \frac{1+\delta_3}{1-\delta_4} = \frac{\delta_3 + \delta_4}{1 - \delta_4}, \quad \delta_4 = \frac{M\tau\varepsilon_1}{1 - M\tau\varepsilon_1},$$

$$\delta_3 = \frac{(\delta_1 + \delta_2)\,\mu(B)}{1 - \delta_1\mu(B)}, \quad \delta_2 = 2\varepsilon_1 + M\tau\varepsilon_1(1 + 2\varepsilon_1), \tag{3.14}$$

$$\delta_1 = \frac{2\sqrt{3}\,\varepsilon_1}{1 - 2\varepsilon_1} + 2M\sqrt{M}\,\tau\varepsilon_1\left(1 + \frac{2\sqrt{3}\,\varepsilon_1}{1 - 2\varepsilon_1}\right)\frac{\|A\|}{\|B\|}.$$

Для вывода окончательного представления для ε из (3.13) осталось подставить в (3.14) значения величин $\mu(B)$ и $\|A\|/\|B\|$.

Согласно [5, 11]

$$\mu(B) = \frac{\frac{\rho}{2} + \sqrt{\frac{\rho^2}{4} + \sigma_N^2(A)}}{\min\left(\rho, \sqrt{\frac{\rho^2}{4} + \sigma_1^2(A)} - \frac{\rho}{2}\right)} = \frac{\frac{1}{2} + \sqrt{\frac{1}{4} + \left(\frac{\sigma_N(A)}{\rho}\right)^2}}{\min\left(1, \sqrt{\frac{1}{4} + \left(\frac{\sigma_1(A)}{\rho}\right)^2} - \frac{1}{2}\right)}, \tag{3.15}$$

$$\frac{\|A\|}{\|B\|} = \frac{\sigma_N(A)}{\frac{\rho}{2} + \sqrt{\frac{\rho^2}{4} + \sigma_N^2(A)}} = \frac{\sqrt{\frac{\rho^2}{4} + \sigma_N^2(A)} - \frac{\rho}{2}}{\sigma_N(A)} =$$

$$= \sqrt{\left(\frac{\rho}{2\sigma_N(A)}\right)^2 + 1} - \frac{\rho}{2\sigma_N(A)}.$$

Из представления (3.15) видно, что

$$\frac{\partial\mu(B)}{\partial\sigma_N(A)} > 0, \quad \frac{\partial\mu(B)}{\partial\sigma_1(A)} < 0, \quad \frac{\partial\left(\frac{\|A\|}{\|B\|}\right)}{\partial\sigma_N(A)} > 0.$$

Поэтому, подставив в правую часть (3.15) вместо $\sigma_1(A)$ ее оценку снизу, а вместо $\sigma_N(A)$ — оценку сверху, мы тем самым получим оценки сверху для $\mu(B)$ и $\|A\|/\|B\|$.

В [9] описан алгоритм вычисления достаточно узких границ сверху и снизу на любое сингулярное число двухдиагональной матрицы. Вычислим этим алгоритмом величины σ_N и σ_1 такие, что $\sigma_N(D) < \sigma_N$, $\sigma_1(D) \geqslant \sigma_1$. При этом из (3.3) следует, что

$$\sigma_N(A) \leqslant \sigma_N(D) + \|\Delta\| \leqslant \sigma_N + 2M\sqrt{M}\tau\varepsilon_1\|A\|,$$

$$\sigma_N(A) \leqslant \frac{\sigma_N}{1 - 2M\sqrt{M}\,\tau\varepsilon_1}.$$

Аналогично

$$\sigma_1(A) \geqslant \sigma_1(D) - \|\Delta\| \geqslant \sigma_1 - 2M\sqrt{M}\,\tau\varepsilon_1\|A\| \geqslant \sigma_1 - \frac{2M\sqrt{M}\,\tau\varepsilon_1\sigma_N}{1 - 2M\sqrt{M}\,\tau\varepsilon_1}.$$

Таким образом, относительная точность ε решения расширенной системы (3.1) определяется согласно

$$\varepsilon = \frac{\delta_3 + \delta_4}{1 - \delta_4}, \tag{3.16}$$

где $\delta_4 = \dfrac{N_0\tau\varepsilon_1}{1 - N_0\tau\varepsilon_1}$, $\delta_3 = \dfrac{(\delta_1 + \delta_2)\,\mu_B}{1 - \delta_1\mu_B}$, $\delta_2 = 2\varepsilon_1 + N_0\tau\varepsilon_1(1 + 2\varepsilon_1)$,

$$\delta_1 = \frac{2\sqrt{3}\,\varepsilon_1}{1 - 2\varepsilon_1} + 2N_0\sqrt{N_0}\,\tau\varepsilon_1\left(1 + \frac{2\sqrt{3}\,\varepsilon_1}{1 - 2\varepsilon_1}\right)\left(\sqrt{\left(\frac{\rho}{2\tilde{\sigma}_N}\right)^2 + 1} - \frac{\rho}{2\tilde{\sigma}_N}\right),$$

$$\mu_B = \frac{\frac{1}{2} + \sqrt{\frac{1}{4} + \left(\frac{\tilde{\sigma}_N}{\rho}\right)^2}}{\min\left(1,\ \sqrt{\frac{1}{4} + \left(\frac{\tilde{\sigma}_1}{\rho}\right)^2} - \frac{1}{2}\right)}, \quad \tilde{\sigma}_N = \frac{\sigma_N}{1 - 2N_0\sqrt{N_0}\,\tau\varepsilon_1},$$

$$\tilde{\sigma}_1 = \sigma_1 - 2N_0\sqrt{N_0}\tau\varepsilon_1\tilde{\sigma}_N, \quad N_0 = \min(N, M),$$

$\tau = 34$ (соответственно $4\max(N, M) + 26$) при использовании удвоенной (соответственно обычной) точности промежуточных вычислений. Таким образом, итерационное уточнение обобщенного (нормального) решения системы (2.1) проводится по следующему алгоритму.

Алгоритм III

Шаг 1. Приводим матрицу A к двухдиагональному виду D; вычисляем оценку снизу σ_1 на наименьшее сингулярное число $\sigma_1(D)$ и σ_N — оценки сверху на $\sigma_N(D)$; определяем ε, μ_B, $\tilde{\sigma}_1$, $\tilde{\sigma}_N$ по формулам (3.16) и оценку сверху $\sigma_{B^{-1}}$ на $\sigma_1(B^{-1})$ по формуле

$$\sigma_{B^{-1}} = \rho\min\left(1,\ \sqrt{\frac{1}{4} + \left(\frac{\tilde{\sigma}_1}{\rho}\right)^2} - \frac{1}{2}\right),$$

где $\rho = \sigma_1/\sqrt{2}$; вычисляем α и β по формулам (см. (1.7))

$$\alpha = \{\varepsilon + [\varepsilon_1 + c_1\varepsilon_1^2(1 + \varepsilon_1 + c_2\varepsilon_0) + c_2\varepsilon_0]\,\mu_B(1 + \varepsilon)\}(1 + \varepsilon_1 + c_2\varepsilon_0) + 2c_2\varepsilon_0,$$

$$\beta = \varepsilon_1 + [c_1\varepsilon_1^2(1 + \varepsilon_1 + c_2\varepsilon_0) + 2c_2\varepsilon_0]\,\mu_B(1 + \varepsilon_1 + c_2\varepsilon_0)(1 + \varepsilon) + c_2\varepsilon_0,$$

а также

$$\delta = c_1\varepsilon_1^2(1 + \varepsilon_1 + c_2\varepsilon_0) + c_2\varepsilon_0.$$

Если $\alpha > 1/2$ или $\beta/(1 - \alpha) > \dfrac{3}{2}\varepsilon_1$, то процесс итерационного уточнения заканчивается отказом от вычисления решения, иначе полагаем

$$q_0 = \varepsilon, \qquad k = \left[\frac{\log_2(2\varepsilon_1 - \beta/(1 - \alpha))}{\log_2\alpha} - 1\right]_+$$

и переходим к шагу 2.

Шаг 2. Полагаем $\tilde{u}_0 = B^{\ominus}\begin{pmatrix} f \\ 0 \end{pmatrix}$.

Шаг 3. Полагаем

$$\tilde{v}_i = B \odot \tilde{u}_i, \quad \tilde{\tilde{r}}_i = \begin{pmatrix} f \\ 0 \end{pmatrix} \ominus \tilde{v}_i,$$

$$p_i = \frac{\dfrac{\|\tilde{\tilde{r}}_i\|}{\|\tilde{u}_i\|\sigma_{B^{-1}}} + (\delta + c_2\varepsilon_0)\,\mu_B}{1 - \varepsilon_1 - \dfrac{\|\tilde{\tilde{r}}_i\|}{\|\tilde{u}_i\|\sigma_{B^{-1}}} - \delta\mu_B}.$$

Если $p_i > 2\varepsilon_1$, то переходим к шагу 4, иначе заменяем q_i на $\min(q_i, p)$ и переходим к шагу 5.

Шаг 4. Полагаем $\widetilde{w}_i = B^{\ominus}\widetilde{\widetilde{r}}_i$,

$$\bar{p}_i = \frac{\dfrac{\|\widetilde{w}_i\|(1-\varepsilon_1)/(1-\varepsilon) + \varepsilon_1\|\widetilde{\widetilde{r}}_i\|/\sigma_{B^{-1}}}{\|\widetilde{u}_i\|} + (\delta + c_2\varepsilon_0)\mu_B}{1 - \varepsilon_1 - \dfrac{\|\widetilde{w}_i\|(1-\varepsilon_1)/(1-\varepsilon) + \varepsilon_1\|\widetilde{\widetilde{r}}_i\|/\sigma_{B^{-1}}}{\|\widetilde{u}_i\|} - \delta\mu_B},$$

$$q_{i+1} = \alpha\min(q_i,\ p_i,\ \bar{p}_i) + \beta,\ \widetilde{u}_{i+1} = \widetilde{u}_i \oplus \widetilde{w}_i.$$

Заменяем $i+1$ на i. Если $i < k$ или $q_i > 2\varepsilon_1$, то переходим к шагу 2, иначе — к шагу 5.

Шаг 5. Возьмем в качестве $\tilde{z}$ первые M компонент, а в качестве $\tilde{x}$ последние N компонент вектора $\widetilde{u}_i$. Считаем расчет законченным с результатом $\tilde{x}$ и оценкой точности вида

$$\|\widetilde{x} - x\| \leqslant q\|\widetilde{x}\|, \quad q = \frac{\min(q_i, 2\varepsilon_1)}{1 - \min(q_i, 2\varepsilon_1)}\sqrt{1 + \left(\frac{\|\widetilde{z}\|}{\|\widetilde{x}\|}\right)^2}.$$

Величину $\mu = \widetilde{\sigma}_N/\widetilde{\sigma}_1$ — оценку сверху на обусловленность системы — считаем дополнительной выходной информацией. Кроме того, при $N < M$ в качестве дополнительной выходной информации выдаем также вектор невязки системы $\tilde{y} = \tilde{z}/\rho$ и оценку сверху на параметр несовместности

$$\widetilde{\nu} = \frac{\rho(\|\widetilde{z}\| + q\|\widetilde{x}\|)}{\widetilde{\sigma}_1(1-q)\|\widetilde{x}\|}.$$

Здесь, как и выше вычисление всех скалярных величин следует вести с помощью арифметических операций с направленным округлением (см. § 4 гл. IV [9]).

СПИСОК ЛИТЕРАТУРЫ

1. **Wilkinson J. H.** Rounding errors in algebraic processes.— Prentice — Hall, 1963.
2. **Уилкинсон Дж. Х.** Алгебраическая проблема собственных значений.— М.: Наука, 1970.
3. **Moler C. B.** Iterative refinement in floating point // J. Assoc. Com. Math.— 1967.— V. 14, N 2.— P. 316—321.
4. **Golub G. H., Wilkinson J. H.** Note on the iterative refinement of least squares solution // Numer. Math.— 1966.— V. 9, N 2.— P. 139—148.
5. **Björck A.** Iterative refinement of linear least squares solutions. I // BIT.— 1967.— V. 7, N 4.— P. 257—278.
6. **Форсайт Дж., Молер К.** Численное решение систем линейных алгебраических уравнений.— М.: Мир, 1969.
7. **Jankowski M., Wozniakowski H.** Iterative refinement implies numerical stability // BIT.— 1977.— V. 17, N 5.— P. 303—311.
8. **Годунов С. К.** Решение систем линейных уравнений.— Новосибирск: Наука. Сиб. отд-ние, 1980.
9. **Гарантированная** точность решения систем линейных уравнений в евклидовых пространствах/Годунов С. К., Антонов А. Г., Кирилюк О. П., Костин В. И.— Новосибирск: Наука. Сиб. отд-ние, 1988.
10. **Воеводин В. В.** Вычислительные основы линейной алгебры.— М.: Наука, 1977.
11. **Годунов С. К.** О решении однородных и линейных уравнений // Актуальные проблемы вычислительной математики и математического моделирования.— Новосибирск: Наука. Сиб. отд-ние, 1985.— С. 179—188.