Актуальные проблемы современной математики: Оптимизация итерационных методов

Курс лекций А.Н. Коновалова

§1. Двухслойные итерационные методы

Предметом нашего рассмотрения является операторное уравнение

$$Ax = b, \tag{1.1}$$

где $A: \ R^n \longrightarrow R^n$ – линейный, самосопряженный $A = A^*$, положительно определенный $A > 0$, ограниченный оператор; x, b – векторы из R^n с вещественными компонентами. Если в R^n выбран соответствующий базис, то можно считать, что в (1.1) речь идет о численном решении системы линейных алгебраических уравнений с квадратной $n \times n$ матрицей $A = A^T > 0$ с вещественными элементами a_{ij}. В этом смысле и не существует принципиальных различий между операторной и матричной трактовкой излагаемых подходов и результатов.

В итерационных методах решения задачи (1.1) определяется некоторая последовательность векторов x^m (m – номер итерации) такая, что

$$\lim_{m \longrightarrow \infty} \| x^m - x \| = 0. \tag{1.2}$$

Здесь мы ограничимся изучением двухслойных итерационных методов, которые договоримся записывать в следующем каноническом виде

$$B \frac{x^{m+1} - x^m}{\tau_{m+1}} + Ax^m = b, \quad m = 0, 1, ... \tag{1.3}$$

В (1.3) оператор $B: \ R^n \longrightarrow R^n$ и, если не оговорено противное, будем всюду считать, что $B = B^* > 0$. Далее, τ_{m+1} – последовательность итерационных параметров; начальный вектор x^0 предполагается пока произвольным. При $\tau_{m+1} = \tau$ итерационный процесс (1.3) называют стационарным.

Если последовательность векторов, определяемая из (1.3) является сходящейся, то она, очевидно, сходится к решению задачи (1.1), поскольку последняя по предположению имеет единственное решение. Следовательно, B и τ_{m+1} нужно выбирать таким образом, чтобы обеспечить (1.2), т.е. сходимость последовательности x^m из (1.3). С другой стороны, чтобы найти x^{m+1} из (1.3) следует решить операторное уравнение

$$Bx^{m+1} = g^m \equiv (B - \tau_{m+1} A) x^m + \tau_{m+1} b.$$

Эта задача должна быть существенно более простой, чем исходная задача (1.1), иначе и нет смысла для ее решения использовать итерационный процесс (1.3). Например, при $B = A$, $\tau_{m+1} = 1$, $x^0 = 0$ итерационный метод (1.3) сходится за одну итерацию, но задача $Ax^1 = b$ в точности совпадает с исходной задачей (1.1).

Итак, приближенное решение задачи (1.1) отыскивается с помощью (1.3). В принципе достаточно найти такое $m(\varepsilon)$, чтобы

$$\| z^m \| = \| x^m - x \| \le \varepsilon \| z^0 \| = \varepsilon \| x^0 - x \|, \quad \varepsilon > 0. \tag{1.4}$$

Число $m(\varepsilon)$, которое гарантирует выполнение неравенства (1.4), является одной из важнейших характеристик итерационного процесса (1.3). С ним очевидным образом связано $Q(\varepsilon)$ – общее число арифметических действий, достаточное для достижения заданной точности (1.4). Так как

$$Q(\varepsilon) = \sum_{m=1}^{m(\varepsilon)} Q_m,$$

где Q_m – число арифметических действий при вычислении x^m, то задача построения экономичного (оптимального) итерационного процесса (1.3) заключается в минимизации Q_m и $m(\varepsilon)$. Минимизация Q_m может быть осуществлена с помощью подходяще выбранного оператора B, а минимизация $m(\varepsilon)$ связана как с выбором B, так и с выбором τ_{m+1}.

Для вектора погрешности $z^m = x^m - x$ с учетом (1.1), (1.3) получаем

$$B \frac{z^{m+1} - z^m}{\tau_{m+1}} + A z^m = 0 \tag{1.5}$$

и условие сходимости (1.2) итерационного процесса (1.3) можно переписать так:

$$\lim_{m \longrightarrow \infty} \| z^m \| = 0. \tag{1.6}$$

Из (1.5) имеем

$$z^{m+1} = (E - \tau_{m+1} B^{-1} A) z^m = S_{m+1} z^m, \quad \| z^{m+1} \| \le \| S_{m+1} \| \, \| z^m \|. \tag{1.7}$$

Оператор S_{m+1} называют оператором перехода. Далее,

$$z^m = S_m ... S_2 S_1 z^0 = T_{m,0} z^0, \quad \| z^m \| \le \| T_{m,0} \| \, \| z^0 \|. \tag{1.8}$$

Оператор $T_{m,0}$ называют разрешающим оператором.

Если рассматривается стационарный итерационный процесс (1.3), то $S_1 = ... = S_m = S$. Тогда при $S = S^*$

$$\| T_{m,0} \| = \| S^m \| = \| S \|^m.$$

Последнее соотношение позволяет оценить $m(\varepsilon)$. В самом деле, условие (1.4) будет выполнено, если $\| S \|^m \le \varepsilon$. Поэтому

$$m(\varepsilon) \ge \frac{\ln(1/\varepsilon)}{\ln(1/\| S \|)}. \tag{1.9}$$

Величина $\ln(1/\| S \|)$ характеризует скорость сходимости итерационного процесса с оператором перехода S. В (1.7), (1.8) и далее используется операторная норма, подчиненная выбранной векторной норме. Если не оговорено противное, то $\| x \| = \sqrt{(x, x)}$.

Другой характеристикой итерационного процесса (1.3) является вектор невязки r^m:

$$r^m = Ax^m - b, \quad Az^m = A(x^m - x) = Ax^m - b = r^m. \tag{1.10}$$

В отличие от z^m вектор невязки r^m вычисляется при реализации итерационного процесса (1.3). Если итерационный процесс (1.3) сходится, то наряду с (1.6) имеем также и

$$\lim_{m\longrightarrow\infty} \| r^m \| = \lim_{m\longrightarrow\infty} \| z^m \|_A = 0, \quad \| z^m \|_A = \sqrt{(Az^m, z^m)}. \tag{1.11}$$

В связи с (1.11) казалось бы естественным для окончания итерационного процесса (1.3) использовать не (1.9), а условие $\| r^m \| \le \varepsilon$. Однако сама по себе "малость" $\| r^m \|$ не влечет за собою такую же "малость" $\| z^m \|$. Действительно, из (1.10) имеем

$$\| z^m \| \le \| A^{-1} \| \| r^m \|.$$

С другой стороны, в силу (1.1)

$$\| b \| \le \| A \| \| x \|.$$

Следовательно,

$$\frac{\| z^m \|}{\| x \|} \le \nu(A) \frac{\| r^m \|}{\| b \|}, \quad \nu(A) = \| A \| \cdot \| A^{-1} \|. \tag{1.12}$$

В (1.12) $\nu(A)$ – число обусловленности задачи (1.1). Оно является одной из важнейших характеристик задачи (1.1), уже хотя бы в силу справедливости неулучшаемой оценки

$$\frac{\| \delta x \|}{\| x \|} \le \frac{\nu(A)}{1 - \nu(A) \dfrac{\| \delta A \|}{\| A \|}} \Big(\frac{\| \delta b \|}{\| b \|} + \frac{\| \delta A \|}{\| A \|} \Big), \tag{1.13}$$

которая связывает погрешности δb, δA ($\| \delta A \| < \| A^{-1} \|^{-1}$) с погрешностью δx решения задачи (1.1). Из (1.12) вытекает, что лишь для "хорошо" обусловленной задачи (1.1) относительная малость вектора невязки влечет за собою относительную малость вектора погрешности. Именно для таких задач условие окончания итерационного процесса $\| r^m \| \le \varepsilon$ можно считать корректным.

Если дополнительно предположить, что $AB = BA$, то из (1.3), (1.10) получаем

$$B \frac{r^{m+1} - r^m}{\tau_{m+1}} + Ar^m = 0, \tag{1.14}$$

что по форме совпадает с (1.5). В этом случае для сходящегося итерационного процесса (1.3) скорость убывания $\| z^m \|$, $\| r^m \|$ – одна и та же.

Итак, для стационарного итерационного метода (1.3)

$$z^{m+1} = (E - \tau B^{-1} A) z^m = S z^m. \tag{1.15}$$

Теорема 1.1. *Для сходимости стационарного итерационного метода* (1.3) *с оператором перехода* S *из* (1.15) *необходимо и достаточно, чтобы все собственные числа оператора* S *были по модулю меньше единицы.*

Пусть

$$\gamma_1 B \le A \le \gamma_2 B, \quad 0 < \gamma_1 < \gamma_2, \tag{1.16}$$

где γ_1, γ_2 так называемые константы энергетической эквивалентности операторов A и B. Рассмотрим обобщенную спектральную задачу

$$A\psi_i = \mu_i B\psi_i, \quad (B\psi_i, \psi_j) = \delta_{ij}. \tag{1.17}$$

Будем считать, что все собственные числа задачи (1.17) простые, тогда система ψ_i задает базис в R^n. Неравенства (1.16) означают, что $\forall\, v \in R^n$, $v \neq 0$

$$\gamma_1(Bv, v) \le (Av, v) \le \gamma_2(Bv, v).$$

Но для любого собственного вектора ψ_i из (1.17)

$$\gamma_1(B\psi_i, \psi_i) \le (A\psi_i, \psi_i) = \mu_i(B\psi_i, \psi_i) \le \gamma_2(B\psi_i, \psi_i).$$

Поэтому

$$\gamma_1 \le \mu_{\min}(B^{-1}A), \quad \gamma_2 \ge \mu_{\max}(B^{-1}A) \tag{1.18}$$

и, следовательно, $\mu_{\min}(B^{-1}A)$, $\mu_{\max}(B^{-1}A)$ являются самыми точными (неулучшаемыми) константами в (1.16).

Пусть $\rho(S)$ является спектральным радиусом оператора перехода S в (1.15). По определению

$$\rho(S) = \max_i |1 - \tau\mu_i|$$

и в силу теоремы 1.1 условие сходимости стационарного итерационного процесса (1.3) можно записать так

$$\rho(S) < 1, \quad \tau < \frac{2}{\gamma_2} \le \frac{2}{\mu_{\max}}. \tag{1.19}$$

Обратимся теперь к соотношению (1.15), которое перепишем в таком виде

$$Bz^{m+1} = Bz^m - \tau Az^m. \tag{1.20}$$

Если

$$z^m = \sum_{i=1}^{n} c_i^m \psi_i, \quad z^{m+1} = \sum_{i=1}^{n} c_i^{m+1} \psi_i,$$

где B – ортонормированный базис ψ_i определен в (1.17), то из (1.17), (1.20) нетрудно получить

$$\| z^{m+1} \|_B \le \rho(S) \| z^m \|_B, \quad \| v \|_B = \sqrt{(Bv, v)}. \tag{1.21}$$

Спектральный радиус $\rho(S)$ в (1.21) зависит от итерационного параметра τ. Допустимые τ, при которых стационарный итерационный процесс (1.3) сходится, задаются условием (1.19). Выберем теперь такое допустимое τ, которое минимизирует $\rho(S)$. Тем самым мы приходим к задаче

$$\tau: \ \min_{\tau} \max_{\gamma_1 \le \mu_i \le \gamma_2} |1 - \tau\mu_i|, \quad 0 < \tau < \frac{2}{\gamma_2}. \tag{1.22}$$

На интервале $\gamma_1 \le \mu \le \gamma_2$ рассмотрим функцию $g(\tau, \mu) = 1 - \tau\mu$. Поскольку при каждом фиксированном τ эта функция линейна, то $\max |g(\tau, \mu)|$ достигается либо при $\mu = \gamma_1$, либо при $\mu = \gamma_2$. Функция $|g(\tau, \gamma_1)|$ убывает при $0 < \tau < 1/\gamma_1$ и возрастает при $\tau > 1/\gamma_1$. В то же время функция $|g(\tau, \gamma_2)|$ убывает при $0 < \tau < 1/\gamma_2$ и возрастает при $\tau > 1/\gamma_2$. Но $\gamma_2 > \gamma_1$, поэтому существует такое τ_0, что $g(\tau_0, \gamma_1) = -g(\tau_0, \gamma_2)$. Именно это τ_0 и дает решение задачи (1.22). В самом деле, если $\tau < \tau_0$, то $|g(\tau, \gamma_2)| > |g(\tau_0, \gamma_2)|$. Если же $\tau > \tau_0$, то $|g(\tau, \gamma_1)| > |g(\tau_0, \gamma_1)|$. При этом оптимальное значение τ_0 является допустимым

$$\tau_0 = \frac{2}{\gamma_1 + \gamma_2} < \frac{2}{\gamma_2}, \tag{1.23}$$

а для минимального значения $\rho(S)$ будем иметь

$$\rho_0(S) = \rho_0 = \frac{\gamma_2 - \gamma_1}{\gamma_2 + \gamma_1} = \frac{1 - \eta}{1 + \eta}, \quad \eta = \frac{\gamma_1}{\gamma_2}. \tag{1.24}$$

Если в качестве γ_1, γ_2 взяты неулучшаемые константы $\mu_{\min}$, $\mu_{\max}$ из (1.18) то

$$\rho_0 > \rho_* = \frac{1 - \eta_*}{1 + \eta_*}, \quad \eta_* = \frac{\mu_{\min}}{\mu_{\max}}. \tag{1.25}$$

Поэтому ρ_* из (1.25) является точной нижней границей для $\rho(S)$ из (1.21).

Предыдущие рассуждения имеют один недостаток. Дело в том, что оптимизация стационарного итерационного процесса (1.3) связана все-таки с минимизацией $m(\varepsilon)$. В силу (1.9) это приводит к задаче о минимизации $\| S \|$ с самосопряженным оператором перехода S. Без условия $S = S^*$ соотношение (1.9) не имеет места, а формальная замена в нем $\| S \|$ на $\rho(S)$ позволяет говорить о справедливости (1.9) лишь в некотором асимптотическом смысле. Поскольку в (1.15) $S \ne S^*$, то нами решена задача (1.22) о минимизации $\rho(S)$. В общем случае, как хорошо известно, для любой векторной нормы подчиненная ей операторная норма удовлетворяет неравенству $\rho(S) \le \| S \|$.

Приятным исключением в (1.15) является случай $B = E$ (метод простой итерации), а также случай $AB = BA$. Тогда в (1.15) $S = S^*$, $\rho(S)$ можно принять за $\| S \|$, а операторная норма $\rho(S)$ является подчиненной для спектральной векторной нормы.

Далее мы покажем, что задачу (1.15), (1.20) для вектора погрешности z^m с оператором перехода $S \neq S^*$ можно свести к некоторой другой задаче с оператором перехода $\tilde{S} = \tilde{S}^*$. Приведем здесь основные варианты:

$$\begin{gathered} Bz^{m+1} = Bz^m - \tau A^{\frac{1}{2}} A^{\frac{1}{2}} z^m \to y^{m+1} = \\ = y^m - \tau A^{\frac{1}{2}} B^{-1} A^{\frac{1}{2}} y^m \to y^{m+1} = (E - \tau \tilde{C}) y^m, \\ \tilde{C} = A^{\frac{1}{2}} B^{-1} A^{\frac{1}{2}} = \tilde{C}^*, \quad y^m = A^{\frac{1}{2}} z^m, \quad \| y^m \| = \| z^m \|_A. \end{gathered} \tag{1.26}$$

$$\begin{gathered} B^{\frac{1}{2}} B^{\frac{1}{2}} z^{m+1} = B^{\frac{1}{2}} B^{\frac{1}{2}} z^m - \tau A z^m \to y^{m+1} = \\ = y^m - \tau B^{-\frac{1}{2}} A B^{-\frac{1}{2}} y^m \to y^{m+1} = (E - \tau \tilde{C}) y^m, \\ \tilde{C} = B^{-\frac{1}{2}} A B^{-\frac{1}{2}} = \tilde{C}^*, \quad y^m = B^{\frac{1}{2}} z^m, \quad \| y^m \| = \| z^m \|_B. \end{gathered} \tag{1.27}$$

Уместно еще раз напомнить, что в (1.3): $A = A^* > 0$, $B = B^* > 0$. Тогда в (1.26), (1.27)

$$\| z^m \|_A^2 = (Az^m, z^m), \quad \| z^m \|_B^2 = (Bz^m, z^m).$$

Теорема 1.2. *Если $A = A^* > 0$, $B = B^* > 0$, то эквивалентны операторные неравенства*

$$\gamma_1 B \leq A \leq \gamma_2 B, \quad \gamma_1 E \leq \tilde{C} \leq \gamma_2 E, \quad 0 < \gamma_1 < \gamma_2. \tag{1.28}$$

Доказательство. Если $\gamma > 0$, а $\tilde{C} = B^{-\frac{1}{2}} A B^{-\frac{1}{2}}$, то

$$\begin{gathered} (\tilde{C}x, x) - \gamma(x, x) = (B^{-\frac{1}{2}} A B^{-\frac{1}{2}} x, x) - \gamma(x, x) = \\ = (Ay, y) - \gamma(B^{\frac{1}{2}} y, B^{\frac{1}{2}} y) = (Ay, y) - \gamma(By, y). \end{gathered} \tag{1.29}$$

Если же $\tilde{C} = A^{\frac{1}{2}} B^{-1} A^{\frac{1}{2}}$, то

$$\begin{gathered} (\tilde{C}x, x) - \gamma(x, x) = (\tilde{C}^{\frac{1}{2}} x, \tilde{C}^{\frac{1}{2}} x) - \gamma(x, x) = (y, y) - \gamma(C^{-1} y, y) = \\ = (y, y) - \gamma(A^{-\frac{1}{2}} B A^{-\frac{1}{2}} y, y) = (Av, v) - \gamma(Bv, v). \end{gathered} \tag{1.30}$$

Поэтому операторы $(\tilde{C} - \gamma E)$ и $(A - \gamma B)$ имеют одинаковые знаки. Чтобы теперь получить (1.28) остается последовательно положить в (1.29), (1.30) $\gamma = \gamma_1$ и $\gamma = \gamma_2$.

Теорема 1.3. *Если $\tilde{C} = \tilde{C}^* > 0$, $\tau > 0$, $0 < \rho < 1$, то неравенства*

$$\| S \| = \| E - \tau \tilde{C} \| \leq \rho, \quad \frac{1 - \rho}{\tau} E \leq \tilde{C} \leq \frac{1 + \rho}{\tau} E \tag{1.31}$$

эквивалентны.

Доказательство. Для самосопряженного оператора $S = E - \tau\tilde{C}$ операторная норма $\| S \|$, подчиненная векторной норме $\| x \|$ определяется следующим 1образом

$$\| S \| = \sup_{\| x \|=1} |(Sx, x)| = \sup_{\| x \|=1} |((E - \tau\tilde{C})x, x)|.$$

Поэтому $-\| S \| E \le S \le \| S \| E$. Следовательно, при $\tau > 0$, $\rho > 0$ имеет место следующая последовательность эквивалентных операторных неравенств:

$$-\rho E \le S \le \rho E \longleftrightarrow -\rho E \le E - \tau\tilde{C} \le \rho E \longleftrightarrow \frac{1-\rho}{\tau} E \le \tilde{C} \le \frac{1+\rho}{\tau} E. \tag{1.32}$$

Для завершения доказательства остается сравнить (1.31) и (1.32).

Прокомментируем некоторые следствия из теорем 1.2, 1.3. Пусть для неявного ($B \neq E$) стационарного итерационного метода (1.3) известны константы энергетической эквивалентности γ_1, γ_2 из (1.16). Тогда

$$y^{m+1} = (E - \tau\tilde{C})y^m = \tilde{S}y^m, \quad \| y^{m+1} \| \le \| \tilde{S} \| \| y^m \|, \tag{1.33}$$

где $\tilde{C} = \tilde{C}^* > 0$ и y^m определены либо в (1.26) либо в (1.27). При этом $\| y^m \| = \| z^m \|_A$, либо $\| y^m \| = \| z^m \|_B$. Теперь положим в (1.16) $\gamma_1\tau = 1 - \rho$, $\gamma_2\tau = 1 + \rho$. Но тогда при $\tau(\gamma_1 + \gamma_2) = 2$ для $\| \tilde{S} \|$ из (1.33) будем иметь $\| \tilde{S} \| \le \rho$. Итог сказанному подводит следующая

Теорема 1.4. *Пусть в* (1.1), (1.3) $A = A^* > 0$, $B = B^* > 0$, *а в* (1.16) *заданы* γ_1, γ_2. *Тогда при*

$$\tau = \frac{2}{\gamma_1 + \gamma_2}$$

итерационный процесс (1.3) *сходится. Для вектора погрешности* z^m *справедлива оценка:*

$$\| z^m \|_{A,B} \le \rho^m \| z^0 \|_{A,B}; \quad \rho = \frac{1-\eta}{1+\eta}, \quad \eta = \frac{\gamma_1}{\gamma_2}. \tag{1.34}$$

Пусть в (1.3) $B = E$ (метод простой итерации). Рассмотрим спектральную задачу

$$A\varphi_i = \lambda_i\varphi_i, \quad (\varphi_i, \varphi_j) = \delta_{ij}. \tag{1.35}$$

Пусть в (1.16) $\gamma_1 = \delta = \lambda_{\min}(A)$, $\gamma_2 = \Delta = \lambda_{\max}(A)$. Эти константы являются неулучшаемыми. Тогда

$$\tau_0 = \frac{2}{\delta + \Delta}, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \quad \xi = \frac{\delta}{\Delta}. \tag{1.36}$$

В этом случае в (1.36) ξ – величина, обратная числу обусловленности $\nu(A)$ оператора A. Следовательно, если $\nu(A_1) > \nu(A_2)$, то при прочих равных условиях скорость сходимости метода простой итерации для задачи (1.1) с оператором A_2 выше, чем скорость сходимости того же метода для той же задачи (1.1) с оператором A_1. Эта тенденция, как правило, сохраняется и для итерационных методов (1.3), в которых $B \neq E$. Напомним, что скорость сходимости итерационного метода (1.3) с оператором перехода $S = S^*$ определяется в (1.9).

Как уже говорилось, неулучшаемыми константами в (1.16) при $B \neq E$ являются $\gamma_1 = \mu_{\min}(B^{-1}A)$, $\gamma_2 = \mu_{\max}(B^{-1}A)$, а при $B = E$: $\gamma_1 = \lambda_{\min}(A) = \delta$, $\gamma_2 = \lambda_{\max}(A) = \Delta$. Поэтому эффективность неявных итерационных процессов (1.3) связана с возможностью построения достаточно просто обращаемых операторов B, но таких, что

$$\eta_* = \frac{\mu_{\min}(B^{-1}A)}{\mu_{\max}(B^{-1}A)} \gg \frac{\lambda_{\min}(A)}{\lambda_{\max}(A)} = \xi.$$

В этом случае

$$\rho_* = \frac{1-\eta_*}{1+\eta_*} \ll \frac{1-\xi}{1+\xi} = \rho_0,$$

а для $m(\varepsilon)$ из (1.19) имеем $m_*(\varepsilon) \ll m_0(\varepsilon)$.

Итак, если априорно известны (вычислены) константы энергетической эквивалентности γ_1 и γ_2 из (1.16), то это позволяет: во-первых, определить оптимальный параметр τ_0 из (1.23), обеспечивающий сходимость стационарного итерационного метода (1.3), во-вторых, определить ρ в (1.34), т.е. скорость сходимости метода. При неулучшаемых константах γ_1 и γ_2 оценка (1.34) также не может быть улучшена. Если получение априорной информации о γ_1, γ_2 по каким-либо причинам не представляется возможным, то условия сходимости стационарного итерационного метода (1.3) сформулированы в теореме 1.1. Мы приведем здесь другой, эквивалентный, вариант этой теоремы с более просто проверяемым условием сходимости.

Теорема 1.5. *Пусть в* (1.1), (1.3) $A = A^* > 0$, $B = B^* > 0$. *Тогда условие*

$$B > 0,5\,\tau\,A \tag{1.37}$$

является необходимым и достаточным условием сходимости стационарного итерационного процесса (1.3).

Доказательство. По существу достаточно показать эквивалентность условия (1.19): $\tau\mu_{\max} < 2$ и условия (1.37). Последнее условие означает, что $\forall y \in R^n$, $y \neq 0$

$$(By, y) - 0,5\,\tau(Ay, y) > 0.$$

Как и выше, будем предполагать, что собственные векторы ψ_i обобщенной спектральной задачи (1.17) задают B-ортонормированный базис в R^n. Тогда

$$y = \sum_{i=1}^{n} c_i\psi_i, \quad By - 0,5\tau Ay = \sum_{i=1}^{n} c_i(1 - 0,5\tau\mu_i)B\psi_i.$$

Поэтому

$$(By, y) - 0,5\tau(Ay, y) = \sum_{i=1}^{n} c_i^2(1 - 0,5\tau\mu_i).$$

Отсюда и следует эквивалентность условия (1.37) условию (1.19).

§2. Нестационарные двухслойные итерационные методы

Здесь мы рассмотрим нестационарный итерационный процесс (1.3). Начнем с изучения явного метода: $B = E$. Тогда

$$z^{m+1} = (E - \tau_{m+1}A)z^m = S_{m+1}z^m. \tag{2.1}$$

Пусть

$$z^0 = \sum_{i=1}^{n} c_i^0\varphi_i, \quad z^m = \sum_{i=1}^{n} c_i^m\varphi_i, \tag{2.2}$$

где φ_i – ортонормированная система собственных векторов оператора A, определенная в спектральной задаче (1.35) и задающая базис в R^n. Из (2.1), (2.2) вытекает

$$c_i^m = \Big[\prod_{k=1}^{m}(1 - \tau_k\lambda_i)\Big]c_i^0 = P_m(\lambda_i)c_i^0.$$

Поэтому

$$\| z^m \| = \sqrt{\sum_{i=1}^{n}(c_i^m)^2} = \sqrt{\sum_{i=1}^{n}(P_m(\lambda_i)c_i^0)^2} \le$$

$$\le \max_{\gamma_1\le\lambda_i\le\gamma_2}|P_m(\lambda_i)|\sqrt{\sum_{i=1}^{n}(c_i^0)^2} \le \max_{\gamma_1\le\lambda\le\gamma_2}|P_m(\lambda)|\,\| z^0 \|.$$

Поскольку в (2.1) $S_{m+1} = S_{m+1}^*$, то тем самым получена оценка для нормы разрешающего оператора $T_{m,0}$:

$$\| T_{m,0} \| \le \max_{\gamma_1\le\lambda\le\lambda_2}|P_m(\lambda)|. \tag{2.3}$$

Зададимся теперь числом итераций m и выберем последовательность итерационных параметров $\tau_1, ..., \tau_m$ из условия минимальности правой части в (2.3). Понятно, что такой выбор связан с минимизацией нормы разрешающего оператора $T_{m,0}$. Тем самым мы приходим к задаче

$$\tau_1, ..., \tau_m : \quad \min_{\tau_1,...,\tau_m} \max_{\gamma_1 \le \lambda \le \gamma_2} |P_m(\lambda)|. \tag{2.4}$$

Рассмотрим полином Чебышева $T_m(x)$, заданный на отрезке $|x| \le 1$:

$$2^{m-1} T_m(x) \;=\; \cos m(\arccos x), \quad |x| \le 1.$$

Вне $|x| \le 1$ полином $T_m(x)$ доопределяется следующим образом:

$$2^m \tilde{T}_m(x) \;=\; \left(x \,+\, \sqrt{x^2 \,-\, 1}\right)^m \,+\, \left(x \,-\, \sqrt{x^2 \,-\, 1}\right)^m, \quad |x| > 1.$$

Известно, что среди всех полиномов степени m с коэффициентом при x^m равным единице, полином $T_m(x)$ наименее уклоняется от нуля на отрезке $|x| \le 1$. Наибольшее по модулю значение на отрезке $|x| \le 1$ полином $T_m(x)$ принимает в $(m+1)$ точках, среди которых и концы отрезка. Поэтому $\max |T_m(x)| \;=\; 2^{1-m}$, $|x| \le 1$.

Из определения $P_m(\lambda)$ следует, что в задаче (2.4) среди всех полиномов степени m и таких, что $P_m(0) \;=\; 1$, требуется найти полином $P_m^*(\lambda)$, наименее уклоняющийся от нуля на отрезке $[\gamma_1, \gamma_2]$. Сказанное выше означает, что для решения задачи (2.4) следует перевести отрезок $\gamma_1 \le \lambda \le \gamma_2$ в отрезок $-1 \le x \le 1$:

$$x \;=\; \frac{2\lambda \,-\, (\gamma_1 \,+\, \gamma_2)}{\gamma_2 \,-\, \gamma_1}, \tag{2.5}$$

учесть условие нормировки $P_m(0)$ и использовать свойство минимальности полинома Чебышева $T_m(x)$. Поэтому с учетом (2.5) можно записать

$$P_m^*(\lambda) \;=\; \frac{T_m(x)}{\tilde{T}_m(-\frac{1}{\rho})}, \quad \rho \;=\; \frac{1 \,-\, \eta}{1 \,+\, \eta}, \quad \eta \;=\; \frac{\gamma_1}{\gamma_2}. \tag{2.6}$$

Искомый набор итерационных параметров τ_k определяется теперь из условия совпадения множества $\{\lambda_k\}$ нулей полинома $P_m^*(\lambda)$ и множества $\{x_k\}$ нулей полинома $T_m(x)$:

$$\{\lambda_k\} \;=\; \{\frac{1}{\tau_k}\}, \quad \{x_k\} \;=\; \{-\cos\frac{(2k \,-\, 1)\pi}{2m}\}, \quad 1 \le k \le m.$$

Если воспользоваться (2.5), то

$$\tau_k \;=\; \frac{2}{(\gamma_2 + \gamma_1) \,+\, (\gamma_2 - \gamma_1)\cos(2k-1)\pi/2m}, \quad 1 \le k \le m$$

или

$$\tau_k = \frac{\tau}{1 + \rho t_k}, \;\; \tau = \frac{2}{\gamma_1 + \gamma_2}, \;\; t_k \;=\; \cos\frac{(2k-1)\pi}{2m}, \;\; 1 \le k \le m. \tag{2.7}$$

Требуемый, оптимальный в смысле (2.4), набор итерационных параметров τ_k получен.

Из (2.6) вытекает, что $\max |P_m^*(\lambda)| \;=\; 2|T_m^*(-\frac{1}{\rho})|^{-1}$, где

$$T_m^*(-\frac{1}{\rho}) = \Big(-\frac{1}{\rho} + \sqrt{\frac{1}{\rho^2} - 1}\Big)^m + \Big(-\frac{1}{\rho} - \sqrt{\frac{1}{\rho^2} - 1}\Big)^m.$$

Несложные вычисления дают

$$\Big(-\frac{1}{\rho} + \sqrt{\frac{1}{\rho^2} - 1}\Big)^m = (-1)^m \rho_1^m, \;\; \Big(-\frac{1}{\rho} - \sqrt{\frac{1}{\rho^2} - 1}\Big)^m = (-1)^m \frac{1}{\rho_1^m}.$$

Поэтому

$$q_m = \max |P_m^*(\lambda)| = \frac{2\rho_1^m}{1 + \rho_1^{2m}}, \quad \rho_1 = \frac{1 - \sqrt{\eta}}{1 + \sqrt{\eta}} \tag{2.8}$$

и имеет место

Теорема 2.1. *Если $B = E$, $A = A^* > 0$, то при заданном m и τ_k из* (2.7) *нестационарный итерационный метод* (1.3) *сходится и справедлива оценка*

$$\| z^m \|_* \le q_m \| z^0 \|_*, \tag{2.9}$$

где q_m определено в (2.8)*, а $\| \cdot \|_* = \| \cdot \|$, либо $\| \cdot \|_* = \| \cdot \|_{A^2}$.*

Выберем в качестве γ_1, γ_2 неулучшаемые константы $\gamma_1 = \delta = \lambda_{\min}(A)$, $\gamma_2 = \Delta = \lambda_{\max}(A)$. Тогда $\eta = \xi = \dfrac{\delta}{\Delta}$. Если, как обычно потребовать $q_m \le \varepsilon$, то в (1.9)

$$m(\varepsilon) \ge \ln\Big(\frac{1}{\varepsilon} + \sqrt{\frac{1}{\varepsilon^2} - 1}\Big) \Big/ \ln\Big(\frac{1}{\rho_1}\Big),$$

хотя чаще употребляется более простая, но несколько завышенная для $m(\varepsilon)$ формула

$$m_*(\varepsilon) \ge \ln\Big(\frac{2}{\varepsilon}\Big) \Big/ \ln\Big(\frac{1}{\rho_1}\Big), \quad \rho_1 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}. \tag{2.10}$$

Для сравнения приведем аналогичную формулу для метода простой итерации с теми же неулучшаемыми константами δ, Δ и оптимальным итерационным параметром τ:

$$m(\varepsilon) \ge \ln\Big(\frac{1}{\varepsilon}\Big) \Big/ \ln\Big(\frac{1}{\rho_0}\Big), \quad \rho_0 = \frac{1 - \xi}{1 + \xi}. \tag{2.11}$$

Комментарии, как говорится, излишни.

Мы изложили основы теории метода простой итерации с чебышевским набором итерационных параметров. Иногда употребляется также название – метод Ричардсона. Обычно его применяют в циклическом варианте. Длину цикла m определяют из (2.10), затем задаются τ_k из (2.7), проводят m итераций, после чего описанный процесс повторяется.

Следует обратить внимание на то, что в (2.7) не предписан порядок использования τ_k в итерационном процессе (1.3). "Теоретически" любые упорядочивания τ_k в (2.7) приводят к одной и той же оценке (2.8), (2.9) и, следовательно, должны быть равноправны любые оптимальные наборы $\tau_{i_k} : 1 \le i_k \le m$. "Практически" это не так. Возникающую здесь ситуацию поясним в случае естественного упорядочивания. Тогда в (2.7) $k = 1, 2, ..., m$.

Переход от z^{k-1} к z^k осуществляется с помощью оператора перехода $S_k = E - \tau_k A$. Мы доказали, что

$$\| T_{m,0} \| = \| S_m S_{m-1} ... S_k ... S_2 S_1 \| \le q_m < 1.$$

Однако среди операторов S_k могут быть и такие, для которых

$$\| S_k \| > 1. \tag{2.12}$$

Для "теоретического" алгоритма

$$z^m = T_{m,0} z^0 = S_m ... S_k ... S_1 z^0 \tag{2.13}$$

и конечный результат не должен зависеть от той или иной перестановки $\tau_1, ..., \tau_k, ..., \tau_m$ и соответствующей перестановки операторов S_k в (2.13). Но при расчете на ЭВМ с конечным числом разрядов не любая фиксация $\tau_{i_1}, ..., \tau_{i_k}, ..., \tau_{i_m}$ приводит к "правильному" результату.

Действительно, если m "достаточно велико", то может найтись и "достаточно большое" r (оно и на самом деле существует!) такое, что для всех $j + \beta : \beta = 1, ..., r$; $j + r < m$ будем иметь $\| S_{j+\beta} \| > 1$. Но

$$z^{j+r} = T_{j+r,j} z^j = S_{j+r} ... S_{j+1} z^j.$$

Поэтому для любого вектора z^j с ограниченной нормой возможна такая ситуация: $\| z^{j+r} \| > N$, где N – максимально допустимое для данной ЭВМ число. Забавно, что $\| z^{j+r} \|$ при завершении цикла должно уменьшиться до $\| z^m \|$ в точном соответствии с (2.8), (2.9). Однако этого не происходит, ибо авост (аварийный останов) из-за переполнения разрядной сетки, как правило, наступает раньше завершения цикла.

Далее мы убедимся, что множество операторов S_k из (2.12) не является пустым и только что описанная ситуация действительно реализуется. Выберем в качестве базиса в R^n систему собственных векторов спектральной задачи (1.35), а в качестве γ_1 γ_2 неулучшаемые константы δ, Δ. Пусть в разложении

$$z^0 = \sum_{i=1}^{n} c_i^0 \varphi_i, \quad c_n^0 = 1, \quad c_i^0 = 0 \quad i < n. \tag{2.14}$$

Тогда

$$z^m = (1 - \tau_m\Delta)...(1 - \tau_1\Delta)\varphi_n. \tag{2.15}$$

Если учесть (2.7), то в (2.15)

$$1 - \tau_k\Delta = \rho_0\frac{t_k - 1}{1 + \rho_0 t_k}, \quad \rho_0 = \frac{1 - \xi}{1 + \xi}.$$

Следовательно, если $2k < m + 1$, то $t_k > 0$ и

$$|1 - \tau_k\Delta| < 1.$$

Если же $m + 1 < 2k < 2m$, то $t_k < 0$ и при

$$\rho_0(1 + 2|t_k|) > 1 \tag{2.16}$$

будем иметь $|1 - \tau_k\Delta| > 1$. Условие (2.16) является по существу условием на $\nu(A)$. Если $m + 1 < 2k < 2m$, то условие (2.16) выполняется при $\nu(A) > 2$.

Итак, мы показали, что в рассматриваемом предельном случае (2.14) даже при малых числах обусловленности исходной задачи (1.1) около половины операторов S_k в (2.13) являются операторами типа (2.12).

Обратимся теперь к другому предельному случаю

$$z^0 = \sum_{i=1}^{n} c_i^0\varphi_i, \quad c_1^0 = 1, \quad c_i^0 = 0 \quad i > 1. \tag{2.17}$$

Тогда

$$z^m = (1 - \tau_m\delta)...(1 - \tau_1\delta)\varphi_1 \tag{2.18}$$

и, кроме того,

$$1 - \tau_k\delta = \rho_0\frac{1 + t_k}{1 + \rho_0 t_k}$$

Следовательно, для любого $\nu(A)$ и $1 \le k \le m$

$$|1 - \tau_k\delta| < 1. \tag{2.19}$$

В силу (2.19) может сложиться впечатление, что после завершения цикла для z^m из (2.18) будет справедлива оценка (2.8), (2.9). Однако это не так. Поскольку расчет на ЭВМ ведется с конечным числом разрядов, то ошибки округления хотя и малы, но неизбежны. Поэтому уже после первой итерации мы будем иметь дело не с "теоретическим"

$$z^1 = (1 - \tau_1\delta)\varphi_1,$$

а с "реальным"

$$z^1 = (1 - \tau_1\delta)\varphi_1 + \sum_{i=2}^{n-1} \varepsilon_i\varphi_i + \varepsilon_n\varphi_n. \tag{2.20}$$

За эволюцией последнего члена в (2.20) мы уже проследили, поэтому можно утверждать следующее. Если в предельном случае (2.17) при τ_k из (2.7) $k = 1, ..., m$ в силу малости ε_n и не произойдет авост из-за переполнения разрядной сетки, то после завершения цикла "теоретическое" z^m будет недопустимо искажено.

Следовательно, для явного итерационного метода (1.3) использование в расчетах на ЭВМ параметров τ_k из (2.7) при естественном упорядочивании $k = 1, ..., m$ не представляется возможным. Также очевидно, что подобный вывод справедлив и для упорядочивания $k = m, m-1, ..., 1$.

Эти факты стимулировали исследования, связанные с построением "устойчивого" чебышевского набора итерационных параметров τ_k, $1 \le k \le m$. Основные результаты получены в работах А.А.Самарского, Е.С.Николаева, а также В.И.Лебедева, С.А.Финогенова. Рассмотренные выше предельные случаи (2.14), (2.17) уже позволяют сформулировать основные требования к нужному "устойчивому" упорядочиванию τ_k. Например,

$$\tau_k : \ \| S_k \| > 1 \longrightarrow \tau_{k+1} : \ \| S_{k+1} \| < 1. \tag{2.21}$$

Более детальное рассмотрение ситуации в предельном случае (2.17) позволяет несколько ослабить условие (2.21). Действительно, вследствии неизбежных погрешностей v_k для вектора погрешности z^k, $1 \le k \le m$ вместо $z^k = S_k z^{k-1}$ будем иметь

$$z^k = S_k z^{k-1} + \tau_k v_k.$$

Тогда

$$z^m = T_{m,0} z^0 + \sum_{k=1}^{m} \tau_k T_{m,k} v_k \tag{2.22}$$

и поэтому

$$\| z^m \| \le \| T_{m,0} \| \| z^0 \| + \max_{1 \le k \le m} \| v_k \| \sum_{k=1}^{m} \tau_k \| T_{m,k} \|. \tag{2.23}$$

Оценка $\| T_{m,0} \| \le q_m$ для первого члена в правой части (2.23). не зависит от выбора какого-либо конкретного упорядочивания τ_k из (2.7). В то же время различные упорядочивания приводят и к различным оценкам для $\sum\limits_{k=1}^{m} \tau_k \| T_{m,k} \|$. Поскольку при $B = E$

$$\sum_{k=1}^{m} \tau_k \| T_{m,k} \| \ge \frac{1 - q_m}{\delta}, \quad \delta = \lambda_{\min}(A) \tag{2.24}$$

то оптимальное по отношению к погрешностям округления v_k упорядочивание $\tau_{i_1} ... \tau_{i_m}$ должно приводить к равенству в (2.24). Такие наборы τ_{i_k} построены, именно их имеют в виду, говоря об устойчивых чебышевских параметрах.

Теперь остается заметить, что если в (1.3) $B \neq E$, то теорему 2.1. следует переформулировать так

Теорема 2.2. *Если в* (1.1), (1.3) $A = A^* > 0$, $B = B^* > 0$,

$$\gamma_1 B \leq A \leq \gamma_2 B, \quad 0 < \gamma_1 < \gamma_2,$$

то при заданном m *и устойчивом чебышевском наборе итерационных параметров* τ_k*:*

$$\tau_k = \frac{\tau}{1+\rho t_k}, \quad \rho = \frac{1-\eta}{1+\eta}, \quad \eta = \frac{\gamma_1}{\gamma_2},$$
$$\tau = \frac{2}{\gamma_1+\gamma_2}, \quad t_k = \frac{\cos(2k-1)\pi}{2m}, \quad 1 \leq k \leq m$$

нестационарный итерационный метод (1.3) *является сходящимся, а при* $\|\cdot\|_* = \|\cdot\|_A$*, либо* $\|\cdot\|_* = \|\cdot\|_B$ *справедлива оценка*

$$\| z^m \|_* \leq q_m \| z^0 \|_*, \quad q_m = \frac{2\rho_1^m}{1+\rho_1^{2m}}, \quad \rho_1 = \frac{1-\sqrt{\eta}}{1+\sqrt{\eta}}.$$

Ключевым моментом для оптимизации стационарных и нестационарных итерационных методов (1.3) является *априорная* информация, связанная с константами энергетической эквивалентности γ_1, γ_2. Собственно в определении этих констант и заключается одна из основных задач при построении оптимальных итерационных методов. Сама задача о нахождении неулучшаемых констант γ_1, γ_2 связана либо ($B = E$) с предварительным определением $\lambda_{\min}(A)$, $\lambda_{\max}(A)$, либо ($B \neq E$) с определением $\mu_{\min}(B^{-1}A)$, $\mu_{\max}(B^{-1}A)$. Но эти задачи гораздо сложнее исходной задачи (1.1), а реально речь вообще может идти только о приближенном нахождении γ_1, γ_2. В соответствии с (1.9) реальная скорость сходимости итерационного метода (1.3) будет существенно зависеть от точности вычисления приближенных констант $\tilde{\gamma_1}$, $\tilde{\gamma_2}$. Как правило, $\| S(\tilde{\gamma_1}, \tilde{\gamma_2}) \| > \| S(\gamma_1, \gamma_2) \|$ или же $\| T_{m,0}(\tilde{\gamma_1}, \tilde{\gamma_2}) \| > \| T_{m,0}(\gamma_1, \gamma_2) \|$, т.е. скорость сходимости уменьшается. Именно поэтому столь широкое распространение получили итерационные методы вариационного типа. В этих методах для достаточно содержательных классов "легко" обращаемых операторов B *априорная* информация о γ_1, γ_2 не требуется.

§3. Итерационные методы вариационного типа

В излагаемых здесь методах последовательность итерационных параметров τ_{m+1} в нестационарном итерационном процессе

$$B\frac{x^{m+1} - x^m}{\tau_{m+1}} + Ax^m = b \tag{3.1}$$

выбирается из условия минимума некоторого функционала.

Напомним, что функционалом $F(u)$ в R^n называют числовую функцию, аргументом которой являются векторы из R^n. Функционал $F(u)$ называется линейным, если для $u \in R^n$, $v \in R^n$

$$F(u + v) = F(u) + F(v), \quad F(\alpha u) = \alpha F(u), \quad \alpha = const.$$

Если в R^n ввести некоторым образом скалярное произведение $(.,.)_*$, то любой линейный функционал $F(u)$ представляется в виде

$$F(u) = (u, f)_*,$$

где вектор $f \in R^n$ однозначно определяется функционалом $F(u)$.

Теорема 3.1. *Пусть $A = A^* > 0$ и пусть вектор $x \in R^n$ является решением задачи*

$$A\,x = b. \tag{3.2}$$

Тогда этот же вектор x доставляет минимум функционалу ошибки:

$$J(u) = (Au, u) - 2(u, b). \tag{3.3}$$

Верно и обратное. Если вектор $u_0 \in R^n$ доставляет минимум функционалу ошибки (3.3)*, то этот же вектор u_0 является решением задачи* (3.2).

Доказательство. Пусть v – произвольный вектор из R^n. Положим $v = x + \eta$. Тогда

$$J(v) = J(x + \eta) = (A(x + \eta), x + \eta) - 2(x + \eta, b) = (Ax, x) - 2(x, b) +$$
$$+ 2(Ax - b, \eta) + (A\eta, \eta) = J(x) + (A\eta, \eta) > J(x).$$

Это доказывает первую часть теоремы. Обратно. Образуем вектор $u_0 + \alpha\xi$, где ξ – произвольный вектор из R^n, а $\alpha = const$. Тогда по определению вектора u_0:

$$\frac{dJ(u_0 + \alpha\xi)}{d\alpha}\Big|_{\alpha = 0} = 0, \quad \frac{d^2J}{d\alpha^2}\Big|_{\alpha = 0} > 0.$$

Эти условия равносильны следующим:

$$(Au_0 - b, \xi) = 0, \quad (A\xi, \xi) > 0.$$

Выполнение второго условия связано с предположением $A = A^* > 0$, а из первого условия вытекает, что вектор $(Au_0 - b) \in R^n$ ортогонален любому произвольному вектору $\xi \in R^n$. Это возможно тогда и только тогда, если $Au_0 = b$. Нужное утверждение вытекает теперь из единственности решения задачи (3.2).

Наряду с функционалом ошибки (3.3) задаче (3.2) можно поставить в соответствие *функционал невязки*:

$$\Phi(x) = (Ax - b, Ax - b) = \| Ax - b \|^2 = \| r \|^2. \tag{3.4}$$

Перепишем (3.4) в таком виде

$$\Phi(x) = (A^*Ax, x) - 2(x, A^*b) + \| b \|^2. \tag{3.5}$$

Последнее слагаемое в (3.5) от x не зависит, а первые два определяют функционал ошибки (3.3) для задачи

$$A^*Ax = A^*b. \tag{3.6}$$

В условиях теоремы 3.1 задача (3.6) и задача (3.2) – эквивалентны. Поэтому очевидно, что в тех же условиях минимум функционалам $\Phi(x)$, $J(x)$ доставляет один и тот же вектор $u_0 \in R^n$ такой, что $Au_0 = b$.

Обратимся к итерационному процессу (3.1), где положим $B = E$:

$$\frac{x^{m+1} - x^m}{\tau_{m+1}} + Ax^m = b. \tag{3.7}$$

Те или иные "оптимальные" последовательности итерационных параметров τ_{m+1} в (3.7) будем строить, исходя из условий минимума некоторых функционалов, связанных с такими характеристиками итерационного процесса, как вектор погрешности z^{m+1} или вектор невязки r^{m+1}.

Из (3.7) имеем

$$r^{m+1} = r^m - \tau_{m+1}Ar^m.$$

Тогда

$$\Phi(x^{m+1}) = \| r^{m+1} \|^2 = \| r^m \|^2 - 2\tau_{m+1}(r^m, Ar^m) - \tau_{m+1}^2 \| Ar^m \|^2.$$

Поэтому функционал невязки $\| r^{m+1} \|^2$ достигает минимума при

$$\tau_{m+1} = \frac{(Ar^m, r^m)}{\| Ar^m \|^2} > 0. \tag{3.8}$$

Последовательность итерационных параметров (3.8) вместе с (3.7) определяет *метод минимальных невязок* для решения задачи (3.2).

В методе *скорейшего спуска* последовательность итерационных параметров τ_{m+1} определяется из условия минимальности функционала $\| z^{m+1} \|_A^2$. Поскольку

$$z^{m+1} = z^m - \tau_{m+1}Az^m,$$

то

$$\| z^{m+1} \|_A^2 = \| z^m \|_A^2 - 2\tau_{m+1}(Az^m, Az^m) + \tau_{m+1}^2(A^2 z^m, Az^m).$$

Поэтому функционал $\| z^{m+1} \|_A^2$ достигает минимума при

$$\tau_{m+1} = \frac{(Az^m, Az^m)}{(A^2 z^m, Az^m)} > 0. \tag{3.9}$$

Отметим следующее. Если $Ax = b$, то при $A = A^* > 0$

$$\| z^{m+1} \|_A^2 = (A(x^{m+1} - x), x^{m+1} - x) = (Ax^{m+1}, x^{m+1}) - (Ax, x^{m+1}) - \\ - (Ax^{m+1}, x) + (Ax, x) = J(x^{m+1}) + \| b \|_{A^{-1}}^2.$$

Отсюда вытекает, что в условиях теоремы 3.1 функционал (Az^{m+1}, z^{m+1}), минимизируемый в методе скорейшего спуска, отличается от функционала ошибки (3.3) только постоянным слагаемым $\| b \|_{A^{-1}}^2$. Поэтому последовательность итерационных параметров (3.9) метода скорейшего спуска является минимизирующей и для функционала ошибки $J(x^{m+1})$. И, наконец, поскольку $Az^m = r^m$, то формулу (3.9) следует переписать в таком виде

$$\tau_{m+1} = \frac{(r^m, r^m)}{(Ar^m, r^m)} > 0. \tag{3.10}$$

Теперь о сходимости. В методе минимальных невязок

$$\Phi(x^{m+1}) = \|(E - \tau_{m+1}A)r^m\|^2 \tag{3.11}$$

и последовательность τ_{m+1} из (3.8) минимизирует правую часть в (3.11). Поэтому при любом $\tau \neq \tau_{m+1}$ будем иметь

$$\Phi(x^{m+1}) \leq \|(E - \tau A)r^m\|^2.$$

Выберем в качестве τ оптимальный итерационный параметр стационарного метода простой итерации

$$\tau_* = \frac{2}{\delta + \Delta}, \quad \delta = \lambda_{\min}(A), \quad \Delta = \lambda_{\max}(A).$$

Тогда

$$\| r^{m+1} \| \leq \| (E - \tau_* A)r^m \| \leq \| E - \tau_* A \| \, \| r^m \|.$$

В условиях теоремы 3.1 имеем

$$\| E - \tau_* A \| = \rho_* = \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{\delta}{\Delta}.$$

и тем самым доказана следующая

Теорема 3.2. *Если* $A = A^* > 0$, *то метод минимальных невязок* (3.7), (3.8) *сходится и справедлива оценка*

$$\| r^{m+1} \| = \| Az^{m+1} \| \leq \rho_* \| Az^m \| = \rho_* \| r^m \|. \tag{3.12}$$

Совершенно аналогично устанавливается, что имеет место

Теорема 3.3. *Если* $A = A^* > 0$, *то метод скорейшего спуска* (3.7), (3.10) *сходится и справедлива оценка*

$$\| z^{m+1} \|_A \leq \rho_* \| z^m \|_A. \tag{3.13}$$

Методы минимальных невязок и скорейшего спуска являются некоторыми конкретными реализациями так называемого *градиентного метода* решения задачи (3.2).

Пусть $F(x)$ – некоторый функционал, определенный для $x \in R^n$. Зададимся некоторым вектором $y \in R^n$ таким, что $\| y \| = 1$. Производную от функционала $F(x)$ в "точке" x по направлению y определяют следующим образом

$$\frac{\partial F(x)}{\partial y} = \lim_{h \longrightarrow 0} \frac{F(x+hy) - F(x)}{h} = \frac{d}{dh} F(x+hy)\Big|_{h=0}. \tag{3.14}$$

В (3.14) $h = const$ и поскольку $F(x+hy) = F(x_1 + hy_1, ..., x_n + hy_n)$, то

$$\frac{\partial F(x)}{\partial y} = \frac{\partial F}{\partial x_1} y_1 + ... + \frac{\partial F}{\partial x_n} y_n = (v, y).$$

Здесь вектор $v = \left(\frac{\partial F}{\partial x_1}, ..., \frac{\partial F}{\partial x_n} \right)^T$ – градиент функционала $F(x)$. Итак, по определению

$$\frac{\partial F(x)}{\partial y} = (v, y) = \| v \| \cos(\widehat{v, y}), \quad \| y \| = 1. \tag{3.15}$$

Следовательно, $\frac{\partial F(x)}{\partial y} = \| v \|$, если направление y совпадает с направлением градиента и $\frac{\partial F(x)}{\partial y} = -\| v \|$, если направление y противоположно направлению градиента. Иными словами, направление градиента есть направление наибольшей скорости роста функционала $F(x)$ в "точке" x, а направление антиградиента есть направление наибольшей скорости убывания функционала $F(x)$ в "точке" x.

Обратимся снова к задаче (3.2). Градиентный метод ее решения заключается в следующем. По известному приближению x^m строится приближение

$$x^{m+1} = x^m - t_{m+1} \, grad \, F(x^m). \tag{3.16}$$

Итерационный параметр t_{m+1} в (3.16) выбирается из условия минимума функционала

$$\tilde{F}(x^m - t_{m+1}\, grad\, F(x^m)) = \tilde{F}(x^{m+1}). \tag{3.17}$$

Пусть в (3.16) в качестве функционала $F(x)$ выбран функционал ошибки (3.3). Тогда в силу (3.14)

$$grad\, J(x) = 2(Ax - b).$$

Обозначив затем $2\,t_{m+1} = \tau_{m+1}$, приходим к (3.7):

$$\frac{x^{m+1} - x^m}{\tau_{m+1}} + Ax^m = b.$$

Теперь понятно, что метод скорейшего спуска (3.7), (3.10) соответствует выбору $F(x^m) = J(x^m)$ в (3.16) и выбору $\tilde{F}(x^{m+1}) = J(x^{m+1})$ в (3.17). Последнее утверждение вытекает из уже приводимого тождества

$$\| z^{m+1} \|_A^2 = J(x^{m+1}) + \| b \|_{A^{-1}}^2.$$

Что касается метода минимальных невязок (3.7), (3.8), то здесь $F(x^m) = J(x^m)$, $\tilde{F}(x^{m+1}) = \Phi(x^{m+1})$.

Для задачи (3.2) выбор $F(x^m) = J(x^m)$ в (3.16) является наиболее употребительным. Как нетрудно понять, именно такой выбор предписывается теоремой 3.1.

В связи с теоремами 3.2 и 3.3 уместно сделать несколько замечаний. Как мы видели, в условиях теоремы 3.1 скорость сходимости двухслойных градиентных методов минимальных невязок и скорейшего спуска априори не хуже, чем в методе простой итерации с оптимальным итерационным параметром τ_0 из (1.23):

$$\frac{x^{m+1} - x^m}{\tau_*} + Ax^m = b, \quad \tau_* = \frac{2}{\delta + \Delta}. \tag{3.18}$$

Покажем, что если $A = A^* > 0$, то оценки (3.12), (3.13) являются неулучшаемыми. Остановимся для определенности на методе скорейшего спуска (3.7), (3.10). Пусть, как обычно, $\delta = \lambda_1 < \lambda_2 < ...\lambda_n = \Delta$ – собственные числа, а φ_i – собственные векторы спектральной задачи (1.35), так что $A\varphi_i = \lambda_i\varphi_i$, $(\varphi_i, \varphi_j) = \delta_{ij}$. Начальное приближение x^0 в (3.7) выберем таким образом, чтобы

$$z^0 = \lambda_n\varphi_1 + \lambda_1\varphi_n, \quad \| z^0 \|_A^2 = (Az^0, z^0) = \lambda_1\lambda_n(\lambda_1 + \lambda_n). \tag{3.19}$$

Тогда

$$z^1 = (E - \tau_1 A)z^0 = (1 - \tau_1\lambda_1)\lambda_n\varphi_1 + (1 - \tau_1\lambda_n)\lambda_1\varphi_n, \tag{3.20}$$

а для τ_1 имеем

$$(Az^0, Az^0) = 2\lambda_1^2\lambda_n^2, \quad (A^2z^0, Az^0) = \lambda_1^2\lambda_n^2(\lambda_1 + \lambda_n), \quad \tau_1 = \tau_*. \tag{3.21}$$

С учетом (3.21) получим из (3.20)

$$z^1 = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}(\lambda_n\varphi_1 - \lambda_1\varphi_1) = \rho_*(\lambda_n\varphi_1 - \lambda_1\varphi_n).$$

Следовательно,

$$\| z^1 \|_A^2 = (Az^1, z^1) = \rho_*^2\lambda_1\lambda_n(\lambda_1 + \lambda_n) = \rho_*^2 \| z^0 \|_A^2. \tag{3.22}$$

Совершенно аналогично вместо (3.22) получим

$$\tau_2 = \tau_*, \quad \| z^2 \|_A^2 = \rho_*^2 \| z^1 \|_A^2 \quad \text{и т.д.}$$

Поэтому при начальном приближении (3.19) в (3.13) будем иметь $\| z^{m+1} \|_A = \rho_* \| z^m \|_A$. Следовательно, оценка (3.13) является неулучшаемой.

Пусть, однако, начальное приближение x^0 выбрано таким образом, что $z^0 = c_m^0\varphi_m$. Тогда

$$z^1 = (E - \tau_1 A)z^0 = (1 - \tau_1\lambda_m)c_m^0\varphi_m.$$

Далее,

$$(Az^0, Az^0) = \lambda_m^2(c_m^0)^2, \quad (A^2z^0, Az^0) = \lambda_m^3(c_m^0)^2, \quad \tau_1 = \frac{1}{\lambda_m}$$

и поэтому $z^1 \equiv 0$.

Подведем итог сказанному. Оптимизация двухслойных итерационных методов (1.3) в предыдущих параграфах основана либо на минимизации нормы оператора перехода $S(\tau)$, либо на минимизации нормы разрешающего оператора $T_{m,0}(\tau_1, ..., \tau_m)$. Но ни S, ни $T_{m,0}$ не зависят от начального приближения x^0, поэтому найденные $\| S \|_{\min}$, $\| T_{m,0} \|_{\min}$ оптимальны для *всех* начальных приближений x^0, в том числе и для самых "плохих" x^0. Как следствие, скорость сходимости не зависит от выбора начального приближения. В двухслойных градиентных методах (3.16), (3.17) минимизируемые функционалы всегда связаны с предыдущим приближением. Поэтому, как мы только что убедились, качество предыдущего приближения существенным образом сказывается и на качестве следующего приближения, следовательно, и на скорости сходимости двухслойного градиентного метода (3.16), (3.17).

Изложенные выше факты допускают очень простую геометрическую интерпретацию. Пусть

$$A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}, \ \lambda_1 = 1, \ \lambda_2 = 3, \ \varphi_1 = \frac{1}{\sqrt{2}}(1,1)^T, \ \varphi_2 = \frac{1}{\sqrt{2}}(1,-1)^T.$$

Если вектор $x = c_1\varphi_1 + c_2\varphi_2$ является решением задачи $Ax = b$, $b = \beta_1\varphi_1 + \beta_2\varphi_2$, то

$$c_1 = \frac{\beta_1}{\lambda_1}, \quad c_2 = \frac{\beta_2}{\lambda_2}, \quad \beta_1 = (b, \varphi_1), \quad \beta_2 = (b, \varphi_2).$$

Для функционала ошибки $J(x)$ имеем

$$J(x) = \lambda_1(c_1 - \frac{\beta_1}{\lambda_1})^2 + \lambda_2(c_2 - \frac{\beta_2}{\lambda_2})^2 - (\frac{\beta_1^2}{\lambda_1} + \frac{\beta_2^2}{\lambda_2}) =$$
$$= \tilde{F}(c_1, c_2) - \| b \|^2_{A^{-1}} > -\| b \|^2_{A^{-1}}.$$

Следовательно, задача о минимизации функционала $J(x)$ эквивалентна задаче о минимизации функционала $\tilde{F}(c_1, c_2) \equiv \tilde{F}(c)$. Линии уровня $\tilde{F}(c) = \alpha = const. > 0$ функционала $\tilde{F}(c)$ в плоскости c_1, c_2 есть эллипсы с центром в точке $c_1 = \beta_1/\lambda_1$, $c_2 = \beta_2/\lambda_2$. Направления главных осей параллельны координатным осям, а длины полуосей равны $\sqrt{\alpha/\lambda_1}$, $\sqrt{\alpha/\lambda_2}$. Единственный минимум $\tilde{F}(c)$ достигается при $c_1 = \beta_1/\lambda_1$, $c_2 = \beta_2/\lambda_2$, т.е. на решении задачи $Ax = b$.

Обратимся теперь к функционалу $\| z \|^2_A$. Пусть Q – матрица, столбцами которой являются собственные векторы φ_1, φ_2 матрицы A, Λ – диагональная матрица с элементами λ_1, λ_2. Тогда

$$\| z \|^2_A = (Az, z) = (Q^T \Lambda Q z, z) = (\Lambda Q z, Qz) =$$
$$= \frac{\lambda_1}{2}(z_1 + z_2)^2 + \frac{\lambda_2}{2}(z_1 - z_2)^2 = \Big(\frac{z_1 + z_2}{\sqrt{2/\lambda_1}}\Big)^2 + \Big(\frac{z_1 - z_2}{\sqrt{2/\lambda_2}}\Big)^2. \tag{3.23}$$

Линии уровня $\| z \|^2_A = \alpha = const. > 0$ функционала $\| z \|^2_A$ в плоскости z_1, z_2 есть эллипсы с центром в начале координат. Направления главных осей эллипсов совпадают с направлениями соответствующих собственных векторов, а длины полуосей равны $\sqrt{2\alpha/\lambda_1}$, $\sqrt{2\alpha/\lambda_2}$. Единственный минимум достигается при $z_1 = z_2 = 0$.

В методе скорейшего спуска начальное приближение x^0 в (3.16) определяет начальную линию уровня $\| z \|^2_A = \alpha_0$. Из точки $M_0(z_1^0, z_2^0)$ в антиградиентном направлении (оно перпендикулярно линии уровня в точке M_0) проводится спуск. Величина спуска определяется из (3.17) и завершается спуск в некоторой точке $M_1(z_1^1, z_2^1)$. Решение экстремальной задачи (3.17) единственно, поэтому существует такая линия уровня $\| z \|^2_A = \alpha_1 < \alpha_0$, которой принадлежит точка M_1. Иными словами направление спуска является касательным к линии уровня $\| z \|^2_A = \alpha_1 < \alpha_0$. Новое направление спуска из точки M_1 перпендикулярно предыдущему, а задача (3.17) определит точку M_2 и т.д. Тем самым порождается сходящаяся последовательность векторов $z^m = (z_1^m, z_2^m)^T$, для которой справедлива оценка (3.13). Начальная погрешность $z^0 = c_m^0 \varphi_m$ порождает спуск по одной из главных осей эллипса, т.е. по направлению φ_1 или φ_2. Такой спуск независимо от величины c_m^0 сразу же приводит в точку $z_1^1 = z_2^1 = 0$, т.е. в начало координат.

Теперь рассмотрим самые невыгодные направления спуска. В плоскости z_1, z_2 вектору начальной погрешности (3.19): $z^0 = \lambda_1\varphi_2 + \lambda_2\varphi_1$ соответствует точка M_0, лежащая на прямой $z_2 = 2z_1$. Для точек, лежащих на этой прямой, касательной к линии уровня функционала (3.23) является прямая $z_2 = const.$ Поэтому спуск из точки M_0 осуществляется по прямой $z_1 = const.$ и заканчивается в точке M_1. Эта точка соответствует вектору погрешности $z^1 = \rho_*(\lambda_2\varphi_1 - \lambda_1\varphi_2)$ и лежит на прямой $z_2 = \frac{1}{2}z_1$. Для точек, лежащих на этой прямой, касательной к линии уровня

функционала (3.23) является прямая $z_1 = const$. Спуск из точки M_1 осуществляется по прямой $z_2 = const$ и оканчивается в точке M_2, лежащей на прямой $z_2 = 2z_1$ и т.д. Итак, для вектора начальной погрешности $z^0 = \lambda_2\varphi_1 + \lambda_1\varphi_2$ в методе скорейшего спуска реализуется покоординатный спуск. Скорость сходимости при этом – наихудшая.

В связи с покоординатным спуском следует упомянуть об итерационном методе Зейделя. Опишем здесь его матричный вариант: $A = A^T > 0$. Если a_{ij} элементы матрицы A, то формальная схема метода такова:

$$\sum_{j=1}^{i-1} a_{ij}x_j^{m+1} + a_{ii}x_i^{m+1} + \sum_{j=i+1}^{n} a_{ij}x_j^m = bi, \quad i = 1, ..., n. \tag{3.24}$$

Пусть $A = L + D + U$, где $D = diagA$, L – нижняя треугольная матрица $l_{ij} = a_{ij}$, $i > j$; U –верхняя треугольная матрица $u_{ij} = a_{ij}$, $i < j$ и $l_{ii} = u_{ii} = 0$.Тогда (3.24) можно записать в каноническом виде (1.3)

$$(D + L)(x^{m+1} - x^m) + Ax^m = b. \tag{3.25}$$

Поскольку из $A > 0$ следует, что $D > 0$, то в (3.25) имеем $A = A^* > 0$, $B > 0$. Как известно, в этом случае условие (1.37) $B > 0,5\tau A$ теоремы 5.1 является достаточным (но не необходимым!) для утверждения: $\lim\limits_{m\longrightarrow\infty} \| z^m \| = 0$. Для (3.25) условие (1.37) выполнено. Действительно,

$$((D + L - \frac{1}{2}A)x, x) = (Dx, x) + (Lx, x) - \frac{1}{2}(Lx, x) -$$
$$-\frac{1}{2}(Dx, x) - \frac{1}{2}(Ux, x) = \frac{1}{2}(Dx, x) + \frac{1}{2}(Lx, x) - \frac{1}{2}(Ux, x).$$

Остается заметить, что $(Lx, x) = (Ux, x)$.

Пусть при переходе от x^m к x^{m+1} изменяется только i-я компонента (координата) вектора x^m. Тогда $z^{m+1} = z^m - \alpha e_i$, где $\alpha = const$, а e_i – вектор-столбец, i-я компонента которого равна единице, прочие – нулю. Кроме того,

$$\begin{aligned} \| z^{m+1} \|_A^2 &= (Az^{m+1}, z^{m+1}) = (A(z^m - \alpha e_i), z^m - \alpha e_i) = \\ &= \| z^m \|_A^2 + \frac{1}{a_{ii}}(\alpha a_{ii} - r_i^m)^2 - \frac{(r_i^m)^2}{a_{ii}}. \end{aligned} \tag{3.26}$$

Здесь и далее r_i^m – i-я компонента вектора невязки r^m, так что $r_i^m = (r^m, e_i)$. Из (3.26) следует, что минимум функционала $\| z^{m+1} \|_A^2$ достигается при $\alpha = r_i^m/a_{ii}$. При таком значении α

$$r_i^{m+1} = (Az^{m+1}, e_i) = r_i^m - \alpha a_{ii} = 0.$$

Но именно этим свойством i-й компоненты вектора невязки r^{m+1} характеризуется метод Зейделя в форме (3.24). В самом деле, по определению

$$r_i^{m+1} = \sum_{j=1}^{i-1} a_{ij}x_j^{m+1} + a_{ii}x_i^{m+1} + \sum_{j=i+1}^{n} a_{ij}x_j^m - b_i = 0, \quad \text{что}$$

в точности совпадает с i-м уравнением из (3.24). Этот же результат можно получить, используя функционал ошибки (3.3). В самом деле, на переходе $x^m \longrightarrow x^{m+1}$ функционал $J(x)$ можно считать функцией изменяемой компоненты x_i, т.е. положить $J(x) = J(x_i)$, где

$$J(x_i) = J(x_1^{m+1}, ..., x_{i-1}^{m+1}, x_i, x_{i+1}^m, x_{i+2}^m, ..., x_n^m).$$

Новое значение изменяемой компоненты x_i определим из условия минимума функционала $J(x_i)$. Тогда

$$\frac{\partial J}{\partial x_i} = 2\Big(\sum_{j=1}^{i-1} a_{ij}x_j^{m+1} + a_{ii}x_i + \sum_{j=i+1}^{n} a_{ij}x_j^m - b_i\Big) = 0,$$

что опять-таки приводит к i-му уравнению из (3.24). Следовательно, последовательная покоординатная минимизация функционала ошибки также равносильна итерационному методу Зейделя.

В методе скорейшего спуска последовательный покоординатный спуск при отыскании минимума функционала $\| z \|_A^2$ реализуется, как мы видели, только для специально выбранного начального вектора погрешности $z^0 = \lambda_n\varphi_1 + \lambda_1\varphi_n$ (или $z^0 = \lambda_n\varphi_1 - \lambda_1\varphi_n$). При таком выборе z^0 достигается минимально возможная скорость сходимости. Поэтому в условиях теоремы 3.1 скорость сходимости метода Зейделя не может быть выше, чем в методе скорейшего спуска.

§4. Асимптотическое свойство итерационных методов вариационного типа. Неявные методы

В условиях теоремы (3.1) для метода скорейшего спуска:

$$\frac{x^{m+1} - x^m}{\tau_{m+1}} + Ax^m = b, \quad \tau_{m+1} = \frac{(r^m, r^m)}{(Ar^m, r^m)} \tag{4.1}$$

и метода минимальных невязок:

$$\frac{x^{m+1} - x^m}{\tau_{m+1}} + Ax^m = b, \quad \tau_{m+1} = \frac{(Ar^m, r^m)}{(Ar^m, Ar^m)} \tag{4.2}$$

характерно так называемое асимптотическое свойство. Его суть и возможные приложения рассмотрим сначала на примере метода скорейшего спуска (4.1).

Поскольку

$$\|z^{m+1}\|_A^2 = \|z^m\|_A^2 - 2\tau_{m+1}(Az^m, Az^m) + \tau_{m+1}^2(A^2z^m, Az^m),$$

а τ_{m+1} определяется из (4.1),то

$$\|z^{m+1}\|_A^2 = (1 - \tau_{m+1}\frac{\|r^m\|^2}{\|z^m\|_A^2})\|z^m\|_A^2 = \rho_{m+1}^2\|z^m\|_A^2 \tag{4.3}$$

Теорема 4.1. *Если* $\mathrm{A} = \mathrm{A}^* > 0$, *то для последовательности* ρ_{m+1} *из* (4.3) *имеем:* $\rho_{m+2} \geq \rho_{m+1}$.

Доказательство. Прежде всего отметим, что поскольку

$$(r^{m+1}, r^m) = ((E - \tau_{m+1}A)r^m, r^m) = (r^m, r^m) - \tau_{m+1}(Ar^m, r^m),$$

то в силу выбора τ_{m+1} из (4.1)

$$(r^{m+1}, r^m) = (Az^{m+1}, Az^m) = 0 \tag{4.4}$$

Далее, по определению ρ_{m+1}, ρ_{m+2}

$$\|z^{m+2} - \rho_{m+2}\rho_{m+1}z^m\|_A^2 = 2(\|z^{m+2}\|_A^2 - \rho_{m+2}\rho_{m+1}(Az^{m+2}, z^m)). \tag{4.5}$$

Если учесть (4.4), то

$$(Az^{m+2}, z^m) = (Az^{m+1}, z^{m+1}) = \|z^{m+1}\|_A^2 = \frac{1}{\rho_{m+2}^2}\|z^{m+2}\|_A^2. \tag{4.6}$$

Подстановка (4.6) в (4.5) дает

$$\|z^{m+2} - \rho_{m+2}\rho_{m+1}z^m\|_A^2 = 2(1 - \frac{\rho_{m+1}}{\rho_{m+2}})\|z^{m+2}\|_A^2. \tag{4.7}$$

Из (4.7) заключаем, что либо для всех $k \geq m$

$$\rho_{k+2} = \rho_{k+1} = \tilde{\rho}, \quad z^{k+2} = \tilde{\rho}^2 z^k, \tag{4.8}$$

либо $\rho_{m+2} > \rho_{m+1}$. Доказательство завершено.

В силу теоремы (3.3) последовательность ρ_{m+1} ограничена сверху. Поэтому из $\rho_{m+2} > \rho_{m+1}$ вытекает существование

$$\lim_{m\to\infty}\rho_m = \tilde{\rho} \leq \rho_*, \quad \rho_* = \frac{1-\xi}{1+\xi}, \quad \xi = \frac{\delta}{\Delta}. \tag{4.9}$$

Соотношения (4.8), (4.9) определяют асимптотическое свойство метода скорейшего спуска.

Прокомментируем некоторые следствия из (4.8), (4.9). Пусть $\delta = \lambda_1 < \ldots < \lambda_n = \Delta$ – собственные числа, а φ_i – собственные векторы спектральной задачи (1.35). Пусть также

$$z^0 = \sum_{i=1}^{n} c_i^0 \varphi_i, \quad c_1^0 \neq 0, \quad c_n^0 \neq 0. \tag{4.10}$$

В (4.1) $\tau_{m+1} > 0$, а из анализа функции $g(\tau_{m+1}, \lambda) = |1 - \tau_{m+1}\lambda|$; $\delta \leq \lambda \leq \Delta$ вытекает, что при больших $k \geq m$ с достаточной степенью точности можно считать: $z^k = c_1^k \varphi_1 + c_n^k \varphi_n$. Но тогда

$$z^{k+2} = (1 - \tau_{k+1}\delta)(1 - \tau_{k+2}\delta)c_1^k \varphi_1 + (1 - \tau_{k+1}\Delta)(1 - \tau_{k+2}\Delta)c_n^k \varphi_n,$$

а если при тех же $k \geq m$ выполнено (4.8), то

$$(1 - \tau_{k+1}\delta)(1 - \tau_{k+2}\delta) = \tilde{\rho}^2 = (1 - \tau_{k+1}\Delta)(1 - \tau_{k+2}\Delta). \tag{4.11}$$

Поэтому δ и Δ являются корнями квадратного уравнения:

$$(1 - \tau_{k+1}\lambda)(1 - \tau_{k+2}\lambda) = \tilde{\rho}^2 \tag{4.12}$$

Фактически в (4.12) заложен теоретический алгоритм нахождения $\lambda_{min}(A) = \delta$, $\lambda_{max}(A) = \Delta$. Однако выполнение (4.8) практически маловероятно и реально можно лишь утверждать, что при достаточно больших m

$$\rho_{m+1} \simeq \rho_{m+2}, \quad z^{m+2} \simeq \rho_{m+2}\rho_{m+1} z^m$$

Поэтому вместо (4.12) мы будем иметь дело с уравнением

$$(1 - \tau_{m+1}\lambda)(1 - \tau_{m+2}\lambda) = \rho_{m+1}\rho_{m+2} \tag{4.13}$$

Именно корни этого квадратного уравнения и дают приближенные значения $\tilde{\delta}, \tilde{\Delta}$. Что касается практического вычисления δ, Δ, то следует положить в (4.1) $b = 0$. Тогда исходная задача (1.1) $Ax = 0$ имеет только нулевое решение и поэтому $z^m = x^m$. Следовательно,

$$\rho_{m+1}^2 = \frac{(Az^{m+1}, z^{m+1})}{(Az^m, z^m)} = \frac{(Ax^{m+1}, x^{m+1})}{(Ax^m, x^m)}.$$

Поэтому величины ρ_{m+1}, ρ_{m+2} в уравнении (4.13) могут быть реально вычислены.

Итак, асимптотическое свойство метода скорейшего спуска (4.1) позволяет применить его для решения частичной спектральной задачи (1.35). Для этого следует положить в (4.1) $b = 0$, а затем воспользоваться уравнением (4.13).

Основные вычислительные проблемы связаны здесь с условием $c_1^0 \neq 0, c_n^0 \neq 0$ в (4.10). Если

$$\delta = \lambda_1 < \lambda_2 < \ldots < \lambda_{n-1} < \lambda_n = \Delta,$$

а в (4.10) $c_1^0 = 0, c_2^0 \neq 0$ и $c_{n-1}^0 \neq 0, c_n^0 = 0$, то из уравнения (4.13) будут, вообще говоря, найдены приближения не для δ, Δ, а для λ_2, λ_{n-1}. Такая ситуация является характерной практически для любых модификаций степенных методов решения спектральных задач. Однако именно в подобных ситуациях ошибки округления могут являться не "неизбежным злом", а скорее достоинством. Действительно, уже после нескольких итераций по (4.1) в силу ошибок округления $c_1^m \neq 0$, $c_n^m \neq 0$, хотя сами эти коэффициенты относительно малы. В дальнейшем именно коэффициенты $c_i^k, k \gg m$ с номерами $i = 1, i = n$ являются преобладающими в соответствующем разложении для z^k. Поэтому снова можно считать $z^k = c_1^k\varphi_1 + c_n^k\varphi_n$. Однако нужное k существенно зависит от x^0 и, во всяком случае, требуется тщательный анализ получаемых из (4.13) корней при различных x^0.

Асимптотическое свойство метода скорейшего спуска в форме (4.8) позволяет достаточно просто определить характер стационирования последовательности итерационных параметров τ_{m+1} в (4.1). Из (4.11) получаем при $k > m$

$$\frac{1}{\tau_{k+1}} + \frac{1}{\tau_{k+2}} = \Delta + \delta \tag{4.14}$$

Поскольку правая часть в (4.14) не зависит от номера итерации, то

$$\tau_{k+1} = \frac{2}{(\delta + \Delta) + 2\alpha}, \quad \tau_{k+2} = \frac{2}{(\delta + \Delta) - 2\alpha}. \tag{4.15}$$

Для α в (4.15) с помощью (4.11) нетрудно получить

$$\alpha = \frac{\Delta + \delta}{2}\omega, \quad \omega = \pm\sqrt{\frac{\rho_*^2 - \tilde{\rho}^2}{1 - \tilde{\rho}^2}} \tag{4.16}$$

Величина ρ_*^2 в (4.16) определяется из (4.9), а $\omega = 0$ только для специальных начальных приближений типа (3.19): $z^0 = \lambda_1\varphi_n \pm \lambda_n\varphi_1 = \delta\varphi_n \pm \Delta\varphi_1$. Если $\omega \neq 0$, то выбор знака для ω не имеет существенного значения. Поэтому можно положить

$$\tau_{k+1} = \frac{2}{(\delta + \Delta)(1 + \omega)} = \tilde{\tau}_1, \quad \tau_{k+2} = \frac{2}{(\delta + \Delta)(1 - \omega)} = \tilde{\tau}_2.$$

Следовательно, асимптотическое свойство метода скорейшего спуска в форме (4.8) влечет за собой

$$\tau_{k+1} = \tau_{k+3} = \ldots = \tilde{\tau}_1, \quad \tau_{k+2} = \tau_{k+4} = \ldots = \tilde{\tau}_2. \tag{4.17}$$

При практическом использовании метода (4.1) следует вместо $\tilde{\rho}^2$ использовать $\rho_{k+1}\rho_{k+2}$. Тогда

$$\tau_{k+1} + \tau_{k+2} = \frac{4}{\delta + \Delta} \cdot \frac{1 - \rho_{k+1}\rho_{k+2}}{1 - \rho_*^2},$$

а выход на асимптотический режим можно определять по близости между собой соседних четных и соседних нечетных членов вычисляемой (!) последовательности τ_{m+1}.

Обратимся теперь к явному методу минимальных невязок (4.2). Здесь

$$\|r^{m+1}\|^2 = (1 - \tau_{m+1}\frac{\|r^m\|_A^2}{\|r^m\|^2})\|r^m\|^2 = q_{m+1}^2\|r^m\|^2 \tag{4.18}$$

Теорема 4.2. *Если $A = A^* > 0$, то для последовательности q_{m+1} из* (4.18) *имеем:* $q_{m+2} \geq q_{m+1}$.

Доказательство теоремы (4.2) практически идентично доказательству теоремы (4.1), как и последующие рассуждения, приводящие к (4.12) или (4.13). Следует только заменить ρ_{k+1}, ρ_{k+2} на q_{k+1}, q_{k+2}. Тогда

$$(1 - \tilde{\tau}_1\lambda)(1 - \tilde{\tau}_2\lambda) = q_{k+1}q_{k+2}, \tag{4.19}$$

а корни этого квадратного уравнения, как и прежде, будут служить приближениями для δ и Δ.

В отличие от ρ_{k+1}, ρ_{k+2} величины q_{k+1}, q_{k+2} из (4.18) могут быть реально найдены в процессе вычисления x^{k+1} из (4.2):

$$q_{k+1} = \frac{\|r^{k+1}\|}{\|r^k\|}, \quad q_{k+2} = \frac{\|r^{k+2}\|}{\|r^{k+1}\|}.$$

Поэтому при использовании метода минимальных невязок (4.2) решение частичной спектральной задачи (1.35) может быть найдено *одновременно* с решением исходной прямой задачи (1.1). Возникающие при этом различия между (4.1) и (4.2) с алгоритмической точки зрения понятны. Формально z^{m+1} из (4.1) и r^{m+1} из (4.2) с точностью до определения τ_{m+1} удовлетворяют одному и тому же уравнению

$$z^{m+1} = z^m - \tau_{m+1}Az^m, \quad r^{m+1} = r^m - \tau_{m+1}Ar^m \tag{4.20}$$

Однако реально мы можем вычислить только $r^0 = Ax^0 - b$, а вычисление $z^0 = x^0 - x$ возможно только при заданном $x = A^{-1}b$.

Другое приложение асимптотического свойства итерационных градиентных методов (4.1), (4.2) связано с возможностью существенного ускорения сходимости. Имея в виду указанное выше различие между ρ_{m+1} из (4.3) и q_{m+1} из (4.18), будем на этот раз говорить о явном методе минимальных невязок (4.2). Пусть из (4.2) найдены x^m, x^{m+2}. Образуем линейную комбинацию

$$v = \alpha x^{m+2} + (1-\alpha)x^m. \tag{4.21}$$

Параметр α в (4.21) выберем из условия

$$\alpha: \quad \min \|r\|^2 = \min \|Av - b\|^2. \tag{4.22}$$

Поскольку $r = Av - b = \alpha r^{m+2} + (1-\alpha)r^m$, то

$$\|r\|^2 = \alpha^2 \|r^{m+2}\|^2 + 2\alpha(1-\alpha)(r^{m+2}, r^m) + (1-\alpha)^2 \|r^m\|^2. \tag{4.23}$$

Из (4.2), (4.20) следует, что

$$(r^{m+1}, Ar^m) = (r^m, Ar^m) - \tau_{m+1}(Ar^m, Ar^m) =$$
$$= (r^m, Ar^m) - \frac{(Ar^m, r^m)}{(Ar^m, Ar^m)} \cdot (Ar^m, Ar^m) = 0$$

Поэтому

$$(r^{m+2}, r^m) = (r^{m+1}, r^m) - \tau_{m+2}(Ar^{m+1}, r^m) =$$
$$= (r^{m+1}, r^m) = (r^{m+1}, r^{m+1} + \tau_{m+1} Ar^m) = \|r^{m+1}\|^2.$$

Следовательно, (4.23) можно переписать следующим образом

$$\|r\|^2 = \alpha^2 \|r^{m+2}\|^2 + 2\alpha(1-\alpha)\|r^{m+1}\|^2 + (1-\alpha)^2 \|r^m\|^2 \tag{4.24}$$

В силу асимптотического свойства итерационного метода минимальных невязок (4.2) либо для всех $k \geq m$

$$q_{k+1} = q_{k+2} = \tilde{q},$$

либо

$$q_{k+1} \leq q_{k+2} < \tilde{q} \leq q_* = \rho_*, \ \lim_{m\to\infty} q_m = \tilde{q} \leq q_*.$$

В первом случае для α из (4.22), (4.24) имеем

$$\alpha = \frac{1}{1-\tilde{q}^2}. \tag{4.25}$$

Во втором случае в (4.21) и далее следует положить $\alpha = \alpha_{m+1}$ и тогда

$$\alpha_{m+1} = \frac{1 - q_{m+1}^2}{1 - 2q_{m+1}^2 + q_{m+1}^2 q_{m+2}^2} \tag{4.26}$$

Подстановка (4.25) в (4.21) дает $Av = b$, следовательно, $v = x$, где x-решение исходной задачи $Ax = b$. В качестве следствий из (4.25), (4.26) отметим, что

$$\alpha > 0, \quad 1 - \alpha < 0; \quad \alpha_{m+1} > 0, \quad 1 - \alpha_{m+1} < 0.$$

Поэтому обе формулы

$$v = \alpha x^{m+2} + (1-\alpha)x^m, \quad v = \alpha_{m+1}x^{m+2} + (1-\alpha_{m+1})x^m \tag{4.27}$$

являются экстраполяционными.

Реально процедура ускорения всегда связана с (4.26) и второй формулой из (4.27). Нетрудно оценить эффективность такой процедуры. Подстановка (4.26) в (4.24) дает

$$\|r\|^2 = \frac{q_{m+2}^2 - q_{m+1}^2}{q_{m+2}^2(1 - 2q_{m+1}^2 + q_{m+1}^2 q_{m+2}^2)}\|r^{m+2}\|^2 \tag{4.28}$$

Поскольку

$$1 - 2q_{m+1}^2 + q_{m+1}^2 q_{m+2}^2 \geq (1-\tilde{q}^2)^2,$$

то из (4.28) вытекает, что

$$\|r\|^2 \leq (1 - \frac{q_{m+2}^2}{q_{m+1}^2})\frac{\|r^{m+2}\|^2}{(1-\tilde{q}^2)^2}.$$

С другой стороны $\|r^{m+3}\| = q_{m+3}^2\|r^{m+2}\|^2$ и процедура ускорения эффективна, если $\|r\|^2 < \|r^{m+3}\|^2$. В самом неблагоприятном случае $\tilde{q} = q_*$. Тогда $\tilde{q}^2 \simeq 1 - 4\xi, \quad (1 - \tilde{q}^2)^2 \simeq 16\xi^2$. Поэтому процедуру ускорения (4.26), (4.27) выгодно применять уже при тех m, для которых

$$\frac{q_{m+2}^2}{q_{m+1}^2} - 1 \leq 16\xi^2(1-4\xi).$$

И в заключение этого параграфа рассмотрим некоторые возможные обобщения явного метода скорейшего спуска (4.1) и явного метода минимальных невязок (4.2). Эти обобщения естественным образом связаны с неявным градиентным методом (сравни с (3.16), (3.17)):

$$B\frac{x^{m+1} - x^m}{\tau_{m+1}} + Ax^m = b. \tag{4.29}$$

Как и выше, мы предполагаем, что в (4.29)

$$A = A^* > 0, \quad B = B^* > 0, \quad \gamma_1 B \leq A \leq \gamma_2 B, \quad \gamma_1 > 0. \tag{4.30}$$

Перейдем в (4.29) к вектору погрешности z^m, а затем в соответствии с (1.26) к явной итерационной схеме

$$y^{m+1} = y^m - \tau_{m+1}\tilde{C}y^m, \quad y^m = A^{\frac{1}{2}}z^m, \quad \tilde{C} = A^{\frac{1}{2}}B^{-1}A^{\frac{1}{2}}. \tag{4.31}$$

Тогда

$$\|z^{m+1}\|_A^2 = \|y^{m+1}\|^2 = \|y^m\|^2 - 2\tau_{m+1}(\tilde{C}y^m, y^m) + \tau_{m+1}^2(\tilde{C}y^m, \tilde{C}y^m). \tag{4.32}$$

В методе скорейшего спуска итерационный параметр τ выбирается из условия минимума функционала $\|z^{m+1}\|_A^2$. Поэтому для неявного метода скорейшего спуска из (4.32) получаем

$$\tau_{m+1} = \frac{(\tilde{C}y^m, y^m)}{(\tilde{C}y^m, \tilde{C}y^m)} = \frac{(B^{-1}r^m, r^m)}{(AB^{-1}r^m, B^{-1}r^m)} = \frac{(w^m, r^m)}{(Aw^m, w^m)} \tag{4.33}$$

В (4.33) и далее $w^m = B^{-1}r^m$ – вектор поправки. Расчетные формулы неявного метода скорейшего спуска можно записать следующим образом

$$Bw^m = r^m, \quad \tau_{m+1} = \frac{(w^m, r^m)}{(Aw^m, w^m)}, \quad x^{m+1} = x^m - \tau_{m+1}w^m \tag{4.34}$$

Теорема 4.3. . *Пусть выполнены условия* (4.30). *Тогда неявный метод скорейшего спуска* (4.34) *сходится и справедлива оценка*

$$\|z^m\|_A \le (\frac{1-\eta_*}{1+\eta_*})^m \|z^0\|_A, \quad \eta_* = \frac{\mu_{\min}(B^{-1}A)}{\mu_{max}(B^{-1}A)}. \tag{4.35}$$

Доказательство очевидным образом следует из теорем 1.4 и 3.3. Выбранный способ исследования неявного метода скорейшего спуска является стандартным и связан с переходом от неявной схемы

$$Bz^{m+1} = Bz^m - \tau_{m+1}Az^m \tag{4.36}$$

к явной схеме (4.31). В связи с этим сделаем одно общее замечание. Выбор функционала $F(x^m)$ в (3.16) и $\tilde{F}(x^{m+1})$ в (3.17) сам по себе не гарантирует реализуемость того или иного градиентного метода (3.16). Сказанное в полной мере относится и к неявному методу скорейшего спуска. Действительно, переход от (4.36) к явной итерационной схеме возможен не в соответствии с (1.26), а с (1.27). Но тогда вместо (4.31) получим

$$y^{m+1} = y^m - \tau_{m+1}\tilde{C}y^m, \quad y^m = B^{\frac{1}{2}}z^m, \quad \tilde{C} = B^{-\frac{1}{2}}AB^{-\frac{1}{2}} \tag{4.37}$$

Поскольку в этом случае

$$\|z^{m+1}\|_B^2 = \|y^{m+1}\|^2 = \|y^m\| - 2\tau_{m+1}(\tilde{C}y^m, y^m) + \tau_{m+1}^2(\tilde{C}y^m, \tilde{C}y^m), \tag{4.38}$$

то представляется естественным выбрать τ_{m+1} в (4.38) из условия

$$\tau_{m+1}: \quad \min \|y^{m+1}\|^2 = \min \|z^{m+1}\|_B^2. \tag{4.39}$$

Это дает

$$\tau_{m+1} = \frac{(\tilde{C}y^m, y^m)}{(\tilde{C}y^m, \tilde{C}y^m)} = \frac{(z^m, r^m)}{(w^m, r^m)} \tag{4.40}$$

В терминах y^m, $\tilde{C}$ формулы (4.33), (4.40) совпадают, однако векторы y^m и операторы $\tilde{C}$ в этих формулах различны. И в отличие от (4.33) параметр τ_{m+1} в (4.40) не может быть реально вычислен.

Однако, как нетрудно понять, для неявного градиентного метода

$$Bw^m = r^m, \quad \tau_{m+1} = \frac{(z^m, r^m)}{(w^m, r^m)}, \quad x^{m+1} = x^m - \tau_{m+1} w^m \tag{4.41}$$

справедлива теорема 4.3 с заменой в (4.35) $\|z^m\|_A$ на $\|z^m\|_B$. Кроме того, оба неявных метода (4.34) и (4.41) обладают асимптотическим свойством и при $b = 0$ могут использоваться для решения частичной спектральной задачи (1.17).

Обратимся снова к соотношению (4.36), которое перепишем в таком виде

$$r^{m+1} = r^m - \tau_{m+1} A w^m, \quad AB = BA. \tag{4.42}$$

Тогда

$$\|r^{m+1}\|^2 = \|r^m\|^2 - 2\tau_{m+1}(Aw^m, r^m) + \tau_{m+1}^2 (Aw^m, Aw^m).$$

В методе минимальных невязок итерационный параметр τ_{m+1} выбирается из условия минимальности функционала $\|r^{m+1}\|^2$ и поэтому

$$\tau_{m+1} = \frac{(Aw^m, r^m)}{(Aw^m, Aw^m)} \tag{4.43}$$

Вместе с (4.42) это дает расчетные формулы для неявного метода минимальных невязок

$$Bw^m = r^m, \quad \tau_{m+1} = \frac{(Aw^m, r^m)}{(Aw^m, Aw^m)}, \quad x^{m+1} = x^m - \tau_{m+1} w^m \tag{4.44}$$

Теорема 4.4. *. Пусть выполнены условия* (4.30) *и* $AB = BA$. *Тогда для неявного метода минимальных невязок* (4.44) *справедлива оценка*

$$\|r^m\| \le \left(\frac{1-\eta_*}{1+\eta_*}\right)^m \|r^0\|, \quad \eta_* = \frac{\mu_{\min}(B^{-1}A)}{\mu_{\max}(B^{-1}A)} \tag{4.45}$$

Доказательство основано на том, что условия (4.30) вместе с $BA = AB$ позволяют привести (4.42) к виду

$$r^{m+1} = r^m - \tau_{m+1} \tilde{C} r^m, \quad \tilde{C} = A^{\frac{1}{2}} B^{-1} A^{\frac{1}{2}}, \quad \tilde{C} = \tilde{C}^* > 0. \tag{4.46}$$

Теперь можно рассуждать так же, как и при доказательстве теоремы 3.2. Именно,

$$\|r^{m+1}\| = \|(E - \tau_{m+1}\tilde{C}) r^m\| \le \|E - \tau_* \tilde{C}\| \, \|r^m\| = \rho_* \|r^m\| = \frac{1-\eta_*}{1+\eta_*} \|r^m\|$$

Отсюда очевидным образом следует (4.45).

Специально отметим следующее. Возможность использования неявного метода скорейшего спуска (4.34) или неявного метода минимальных невязок (4.44) для решения задачи $Ax = b$ не связана именно с условиями (4.30) или (4.30), $AB = BA$. Эти условия достаточны для справедливости (4.35), (4.45), а также для асимптотического

свойства этих методов. Поэтому в предположениях теоремы 4.4 неявный метод минимальных невязок (4.44) позволяет *одновременно* с решением задачи $Ax = b$ находить приближенные значения $\mu_{\max}(B^{-1}A)$, $\mu_{min}(B^{-1}A)$, а также эффективно использовать процедуру ускорения (4.21), (4.26).

Используемое в теореме 4.4 условие $AB = BA$ все же является достаточно обременительным. Именно поэтому приведем здесь один из вариантов неявного градиентного метода (4.29), который в предположениях (4.30) позволяет *одновременно* найти решение задачи $Ax = b$ и $\mu_{\max}, \mu_{\min}$ в обобщенной спектральной задаче (1.17). Речь пойдет о методе минимальных поправок. Из (4.29) и определения поправки $Bw^m = r^m$ получим

$$Bw^{m+1} = Bw^m - \tau_{m+1}Aw^m$$

или (сравни с (4.37)):

$$y^{m+1} = y^m - \tau_{m+1}\tilde{C}y^m,\ y^m = B^{\frac{1}{2}}w^m,\ \tilde{C} = B^{-\frac{1}{2}}AB^{-\frac{1}{2}}. \tag{4.47}$$

Поэтому (сравни с (4.38)):

$$\|w^{m+1}\|_B^2 = \|y^{m+1}\|^2 = \|y^m\|^2 - 2\tau_{m+1}(\tilde{C}y^m, y^m) + \tau_{m+1}^2(\tilde{C}y^m, \tilde{C}y^m)$$

Выберем, наконец, τ_{m+1} из условия минимальности функционала $\|w^{m+1}\|_B^2$. Тогда (сравни с (4.33), (4.40)):

$$\tau_{m+1} = \frac{(\tilde{C}y^m, y^m)}{(\tilde{C}y^m, \tilde{C}y^m)} = \frac{(Aw^m, w^m)}{(B^{-1}Aw^m, Aw^m)}$$

и для расчетных формул метода минимальных поправок будем иметь

$$Bw^m = r^m,\ \tau_{m+1} = \frac{(Aw^m, w^m)}{(B^{-1}Aw^m, Aw^m)},\ x^{m+1} = x^m - \tau_{m+1}w^m. \tag{4.48}$$

Теорема 4.5. *Пусть выполнены условия* (4.30) *и оператор B является ограниченным в R^n. Тогда для метода минимальных поправок* (4.48) *справедлива оценка*

$$\| r^m \|_{B^{-1}} \le \Big(\frac{1-\eta_*}{1+\eta_*}\Big)^m \| r^0 \|_{B^{-1}}, \quad \eta_* = \frac{\mu_{\min}(B^{-1}A)}{\mu_{\max}(B^{-1}A)}. \tag{4.49}$$

Доказательство. Из (4.30) следует, что в (4.47) $\tilde{C} = \tilde{C}^*$. По предположению $\forall\, v \in R^n$ и $v \ne 0$ $(Bv, v) \le \alpha(v, v)$. Но $B = B^* > 0$, поэтому $(B^{-1}v, v) > 1/\alpha(v, v)$. Далее,

$$(AB^{-1}Av, v) = (B^{-1}Av, Av) > \frac{1}{\alpha}(Av, Av) > \frac{\delta^2}{\alpha}(v, v).$$

Остается заметить, что поскольку

$$\| y^m \|^2 = \| B^{\frac{1}{2}} w^m \|^2 = \| B^{-\frac{1}{2}} r^m \|^2 = \| r^m \|^2_{B^{-1}} = \| z^m \|^2_{AB^{-1}A},$$

то в (4.47) $\tilde{C} > 0$ и эквивалентны следующие неравенства

$$\gamma_1 B \le A \le \gamma_2 B, \quad \gamma_1 E \le \tilde{C} \le \gamma_2 E. \tag{4.50}$$

Для завершения доказательства следует в (4.50) выбрать неулучшаемые константы $\gamma_1 = \mu_{\min}(B^{-1}A)$, $\gamma_2 = \mu_{\max}(B^{-1}A)$ и повторить стандартные рассуждения, связанные с оценкой $\| E - \tau_{m+1}\tilde{C} \|$.

Отметим также, что в условиях теоремы 4.5 метод минимальных поправок (4.48) обладает асимптотическим свойством.

Приведенные выше конкретные явные и неявные итерационные методы вариационного типа очевидным образом могут быть получены с помощью общей схемы построения градиентного метода (3.16), (3.17). Для *всех* рассмотренных методов

$$x^{m+1} = x^m - \tau_{m+1} w^m, \quad Bw^m = r^m$$

и поэтому для вектора поправки w^m имеем

$$w^m = B^{-1} grad\, J(x^m) = B^{-1}(Ax^m - b),$$

где $J(x^m)$ – функционал ошибки. Поэтому все различия методов могут быть связаны только с конкретным определением длины спуска τ_{m+1} из условия

$$\tau_{m+1}: \quad \min \tilde{F}(x^{m+1}) = \min \tilde{F}(x^m - \tau_{m+1} w^m). \tag{4.51}$$

Поскольку $z^{m+1} = z^m - \tau_{m+1} w^m$, то при $D = D^* > 0$ в (4.51) можно положить $\tilde{F}(z^{m+1}) = (Dz^{m+1}, z^{m+1})$, а длину спуска τ_{m+1} выбирать из условия

$$\tau_{m+1}: \quad \min \| z^{m+1} \|^2_D = \min \| y^{m+1} \|^2, \quad y^m = D^{\frac{1}{2}} z^m. \tag{4.52}$$

Для y^{m+1} из (4.52) получим

$$y^{m+1} = (E - \tau_{m+1} C) y^m, \quad C = D^{-\frac{1}{2}} (DB^{-1}A) D^{-\frac{1}{2}}.$$

Тогда

$$\tau_{m+1} = \frac{(Cy^m, y^m)}{(Cy^m, Cy^m)} = \frac{(Dw^m, z^m)}{(Dw^m, w^m)} \tag{4.53}$$

и, кроме того,

$$\| y^{m+1} \| = \rho_{m+1} \| y^m \|, \quad \rho^2_{m+1} = 1 - \frac{(Cy^m, y^m)}{\| Cy^m \|^2 \| y^m \|^2}. \tag{4.54}$$

Итак, каждый из рассмотренных выше двухслойных итерационных методов вариационного типа

$$Bw^m = r^m, \quad x^{m+1} = x^m - \tau_{m+1} w^m,$$
$$\tau_{m+1} : \min(Dz^{m+1}, z^{m+1}),\ D = D^* > 0 \longrightarrow \tau_{m+1} = \frac{(Dw^m, z^m)}{(Dw^m, w^m)} \tag{4.55}$$

можно связать с конкретным выбором оператора D.

Метод минимальных невязок (3.7), (3.8):

$$A = A^* > 0, \quad B = E, \quad D = A^* A = A^2,$$
$$\eta = \xi = \frac{\delta}{\Delta}, \quad \delta E \le A \le \Delta E. \tag{4.56}$$

Метод скорейшего спуска (3.7), (3.10):

$$A = A^* > 0, \quad B = E, \quad D = A,$$
$$\eta = \xi = \frac{\delta}{\Delta}, \quad \delta E \le A \le \Delta E. \tag{4.57}$$

Неявный метод скорейшего спуска (4.34):

$$A = A^* > 0, \quad B = B^* > 0, \quad D = A,$$
$$\eta = \frac{\gamma_1}{\gamma_2}, \quad \gamma_1 B \le A \le \gamma_2 B. \tag{4.58}$$

Неявный метод минимальных невязок (4.44):

$$A = A^* > 0, \quad B = B^* > 0, \quad D = A^* A = A^2, \quad AB = BA,$$
$$\eta = \frac{\gamma_1}{\gamma_2}, \quad \gamma_1 B \le A \le \gamma_2 B. \tag{4.59}$$

Метод минимальных поправок (4.48):

$$A = A^* > 0, \quad B = B^* > 0, \quad B < \alpha E, \quad D = AB^{-1}A,$$
$$\eta = \frac{\gamma_1}{\gamma_2}, \quad \gamma_1 B \le A \le \gamma_2 B. \tag{4.60}$$

Как для (4.55), так и для (4.56)-(4.60) общим является и метод получения оценки скорости сходимости. Поскольку

$$\| y^{m+1} \| = \min_{\tau_{m+1}} \|(E - \tau_{m+1} C) y^m \| = \min_{\tau_{m+1}} \| S_{m+1} y^m \| \le$$
$$\le \min_{\tau_{m+1}} \| S_{m+1} \| \, \| y^m \| \le \min_{\tau} \| S_{m+1} \| \, \| y^m \| = \rho \, \| y^m \|,$$

то достаточно получить оценку сверху для ρ. При этом для ρ_{m+1} из (4.54) будем иметь $\rho_{m+1} \le \rho$, а если в (4.58)-(4.60), как и в (4.56)-(4.57), используются неулучшаемые (точные) константы γ_1, γ_2, то $\rho_{m+1} \le \rho_* < \rho$.

Предположим, что

$$(DB^{-1}A) = (DB^{-1}A)^*,$$
$$\gamma_1 D \le DB^{-1}A \le \gamma_2 D, \quad \gamma_1 > 0, \quad D = D^* > 0. \tag{4.61}$$

Теорема 4.6. *Пусть выполнены условия* (4.61). *Тогда двухслойный итерационный метод вариационного типа* (4.55) *обладает асимптотическим свойством, сходится и справедлива оценка*

$$\| z^m \|_D \le \rho^m \| z^0 \|_D, \quad \rho = \frac{1-\eta}{1+\eta}, \quad \eta = \frac{\gamma_1}{\gamma_2}. \tag{4.62}$$

Практическое использование конкретного метода (4.55) для решения задачи $Ax = b$ всегда связано с возможностью реального вычисления τ_{m+1}. Практическое использование асимптотического свойства конкретного метода (4.55) либо в процедуре ускорения (4.21), (4.26), либо в обобщенной спектральной задаче (1.17) всегда связано с возможностью реального вычисления $\| y^m \|^2 = (Dz^m, z^m)$. Поэтому условия (4.61) не в полной мере определяют возможность применимости метода (4.55) в тех или иных реальных задачах. В этом смысле весьма показательным является неявный метод (4.41). Здесь $D = B$ и при $A = A^* > 0$, $B = B^* > 0$ условия теоремы 4.6 выполнены. Однако, поскольку

$$\tau_{m+1} = \frac{(Dw^m, z^m)}{(Dw^m, w^m)} = \frac{(r^m, z^m)}{(r^m, w^m)}, \quad \| y^m \|^2 = (Bz^m, z^m),$$

то какое-либо практическое применение метода (4.41) возможно только при $b = 0$.

Для неявного метода минимальных невязок ($D = A^*A$) условие коммутативности $AB = BA$ в (4.59) не является необходимым ни для сходимости, ни для асимптотического свойства. Это условие позволяет лишь вместо (4.61) использовать более простое неравенство: $\gamma_1 B \le A \le \gamma_2 B$. Далее, если в методе минимальных поправок ($D = AB^{-1}A$), $A \ne A^*$, $B = B^* > 0$, то при

$$\gamma_1 B \le A, \quad \gamma_1 > 0, \quad (Av, B^{-1}Av) \le \gamma_2 (Av, v)$$

итерационный метод (4.55) сходится, однако $\rho = \sqrt{1-\eta}$ и не гарантируется асимптотическое свойство. Примеры подобного рода легко продолжить. Поэтому следует отчетливо представлять, что как условия (4.56)-(4.60), так и условия (4.61) не являются необходимыми ни для сходимости какого-либо конкретного метода, ни для его применимости. Выбор достаточных условий в форме (4.61) обусловлен тем, что наряду с теоремой 4.6 справедлива также и

Теорема 4.7. *Пусть выполнены условия* (4.61). *Тогда нестационарный двухслойный итерационный метод*

$$x^k = x^{k-1} - \tau_k w^k, \quad 1 \le k \le m$$
$$\tau_k = \frac{\tau}{1+\rho t_k}, \ \tau = \frac{2}{\gamma_1+\gamma_2}, \ \rho = \frac{1-\eta}{1+\eta}, \ t_k = \frac{\cos(2k-1)\pi}{2m} \tag{4.63}$$

сходится и справедлива оценка

$$\| z^m \|_D \le \frac{2\rho_1^m}{1 + \rho_1^{2m}} \| z^0 \|_D, \quad \rho_1 = \frac{1 - \sqrt{\eta}}{1 + \sqrt{\eta}}, \quad \eta = \frac{\gamma_1}{\gamma_2}. \tag{4.64}$$

Итак, в условиях (4.61) скорость сходимости двухслойного итерационного метода вариационного типа (4.55) не хуже, чем у двухслойного стационарного метода с оптимальным τ:

$$x^k = x^{k-1} - \tau w^k, \quad \tau = \frac{2}{\gamma_1 + \gamma_2},$$

но хуже, чем у нестационарного двухслойного метода (4.63) с чебышевским набором итерационных параметров. Дальнейшая задача заключается в построении итерационных методов вариационного типа, для которых выполняется оценка (4.64). Такие методы существуют.

§5. Методы сопряженных направлений

В рассмотренных выше явных и неявных градиентных методах реализуется стратегия поэтапного уменьшения функционала ошибки $J(x^m)$ (обобщенного функционала ошибки) при последовательных переходах $x^0 \longrightarrow x^1 \longrightarrow ... \longrightarrow x^m$. Однако, можно поставить задачу и о глобальной минимизации этого функционала при переходе $x^0 \longrightarrow x^m$. При изучении нестационарных двухслойных итерационных методов мы убедились, что такая стратегия минимизации (оптимизации) является более выгодной. А именно, для нестационарного двухслойного итерационного метода $x^k = x^{k-1} - \tau_k w^k$ выбор $\tau_1, ..., \tau_m$ из условия минимизации нормы разрешающего оператора $T_{m,0}: \ D^{\frac{1}{2}} z^m = T_{m,0} D^{\frac{1}{2}} z^0$,

$$T_{m,0} = S_m ... S_1 = (E - \tau_m C) ... (E - \tau_1 C), \quad C = D^{\frac{1}{2}} B^{-1} A D^{-\frac{1}{2}}$$

обеспечивает более высокую скорость сходимости, чем выбор оптимального τ в стационарном двухслойном методе $x^k = x^{k-1} - \tau w^k$:

$$D^{\frac{1}{2}} z^k = S D^{\frac{1}{2}} z^{k-1} = (E - \tau C) D^{\frac{1}{2}} z^{k-1}$$

из условия минимума нормы оператора перехода $S = E - \tau C$.

Пусть $A = A^* > 0$. Векторы $\varphi \in R^n$, $\psi \in R^n$ такие, что $(A\psi, \varphi) = 0$ называют сопряженными по отношению к заданному оператору (матрице) A. Точно так же называют и направления в R^n, которые определяются этими векторами. Система векторов $\psi_i \in R^n, \ i = 1, ..., n$ образует A-базис в R^n (сопряженный базис), если $(A\psi_i, \psi_j) = 0, \ i \ne j$.

Теорема 5.1. *Если $A = A^* > 0$, то векторы сопряженного базиса $\psi_i, \ i = 1, ..., n$ являются линейно независимыми в R^n.*

Действительно, если $\psi_1 = \alpha_2\psi_2 + ... + \alpha_n\psi_n$, то $(A\psi_1, \psi_1) = 0$, что противоречит условию $A > 0$.

Любую линейно независимую в R^n систему векторов φ_i, $i = 1, ..., n$ можно превратить в A-базис ψ_i. Для этого, например, можно воспользоваться методом ортогонализации Грама-Шмидта. Положим $\psi_1 = \varphi_1$ и определим ψ_2 следующим образом: $\psi_2 = \varphi_2 + a_{12}\psi_1$, где параметр a_{12} выберем из условия $(A\psi_2, \psi_1) = 0$. Тогда $a_{12} = -(A\psi_1, \varphi_2)/(A\psi_1, \psi_1)$. Затем

$$\psi_3 = \varphi_3 + a_{13}\psi_1 + a_{23}\psi_2, \quad (A\psi_3, \psi_1) = 0, \quad (A\psi_3, \psi_2) = 0.$$

Это дает

$$a_{13} = -\frac{(A\psi_1, \varphi_3)}{(A\psi_1, \psi_1)}, \quad a_{23} = -\frac{(A\psi_2, \varphi_3)}{(A\psi_2, \psi_2)}$$

Дальнейшее комментариев не требует:

$$\begin{aligned} &\psi_m = \varphi_m + a_{1,m}\psi_1 + ... + a_{m-1,m}\psi_{m-1} \\ &\psi_1 = \varphi_1, \quad a_{i,j} = -\frac{(A\psi_i, \varphi_j)}{(A\psi_i, \psi_i)}, \quad i < j. \end{aligned} \tag{5.1}$$

Хотя, как известно, алгоритм (5.1) является одним из наименее устойчивых по отношению к погрешностям округления, но важно, что такой алгоритм теоретически существует.

Обратимся к задаче (1.1):

$$Ax = b. \tag{5.2}$$

Пусть x^0 – начальное приближение в каком-либо итерационном методе решения задачи (5.2). Пусть также система векторов ψ_i задает A-базис в R^n. В силу теоремы 5.1 существуют константы c_i такие, что

$$x = x^0 - c_1\psi_1 - c_2\psi_2 - ... - c_n\psi_n \tag{5.3}$$

и, кроме того

$$c_i = \frac{(\psi_i, r^0)}{(A\psi_i, \psi_i)}, \quad r^0 = Ax^0 - b, \quad i = 1, ..., n. \tag{5.4}$$

Итерационный метод, осуществляющий переход $x^0 \longrightarrow x^m$ определим следующим образом:

$$x^m = x^0 - \tau_1\psi_1 - ... - \tau_m\psi_m, \quad m = 1, 2, ... \tag{5.5}$$

Тогда для функционала ошибки (3.3) получим

$$J(x^m) = (Ax^m, x^m) - 2(x^m, b) = \\ = J(x^0) + \sum_{i=1}^{m} \left[\tau_i^2 (A\psi_i, \psi_i) - 2\tau_i (\psi_i, r^0)\right]. \tag{5.6}$$

Как это следует из (5.6), глобальная (переход $x^0 \longrightarrow x^m$) минимизация функционала $J(x^m)$ равносильна минимизации каждого отдельного члена суммы в правой части (5.6). Поэтому решение задачи о глобальной минимизации функционала $J(x^m)$ дается формулами

$$\tau_i = \frac{(\psi_i, r^0)}{(A\psi_i, \psi_i)}, \quad i = 1, 2, ..., m. \tag{5.7}$$

С точностью до обозначений формулы (5.4) и (5.7) совпадают.

Движение (спуск) по какому-либо из сопряженных направлений ψ_i изменяет один и только один член в правой части (5.5). Если такой спуск связан также и с выбором длины спуска в соответствии с (5.7), то глобальная минимизация в R^n функционала ошибки (3.3) достигается после выполнения не более чем n сопряженных спусков. Этот принципиальный вывод следует также из сравнения (5.3) и (5.5): для некоторого $m \leq n$ будем иметь $x^m = x$. Следовательно, итерационный метод (5.5), (5.7) при отсутствии ошибок округления приводит к точному решению задачи (5.2) не более, чем за n итераций. Такова в самых общих чертах одна из возможных интерпретаций *метода сопряженных направлений.*

Выпишем некоторые следствия из (5.5):

$$x^m = x^{m-1} - \tau_m \psi_m \longrightarrow z^m = z^{m-1} - \tau_m \psi_m \longrightarrow r^m = r^{m-1} - \tau_m A\psi_m.$$

Далее заметим, что минимизация функционала ошибки $J(x^m)$ при $A = A^* > 0$ эквивалентна минимизации функционала $\| z^m \|_A^2$. Это позволяет переписать (5.5), (5.7) следующим образом:

$$x^m = x^{m-1} - \tau_m \psi_m, \quad r^m = r^{m-1} - \tau_m A\psi_m \\ \tau_m = \frac{(\psi_m, r^{m-1})}{(A\psi_m, \psi_m)}, \quad m = 1, 2, ... \tag{5.8}$$

Конкретная реализация метода сопряженных направлений связана с конкретным A-базисом. Это, в свою очередь, связано с определенным выбором системы линейно независимых векторов из R^n, которую следует превратить в A-базис.

Выберем в качестве линейно независимой системы векторов $\varphi_1, ..., \varphi_n$ систему невязок $r^m : \ m = 0, 1, ..., n-1$ итерационного метода (5.8). Такая система r^m заранее, естественно, не может быть задана, в соответствии с (5.8) она вычисляется на последовательных переходах $x^{m-1} \longrightarrow x^m$.

Теорема 5.2. *Ненулевые невязки r^m из* (5.8) *взаимно ортогональны, т.е.*

$$(r^m, r^k) = 0, \quad k < m. \tag{5.9}$$

Доказательство. Поскольку в (5.8) ψ_m является A-базисом, то разложение (5.3) единственно. По определению вектора невязки и в силу (5.3), (5.5)

$$r^m = Ax^m - b = Az^m = A(x^m - x) = \sum_{j=m+1}^{n} \tau_j A\psi_j.$$

Поэтому

$$(r^m, \psi_k) = \sum_{j=m+1}^{n} \tau_j (A\psi_j, \psi_k).$$

Следовательно,

$$(r^m, \psi_k) = 0, \quad k \le m. \tag{5.10}$$

Пусть система векторов $\psi_1, ..., \psi_n$ вычисляется вместе с системой $r^0, ..., r^{n-1}$ посредством A-ортогонализации последней. Положим $\psi_1 = r^0$. Тогда (5.1) можно переписать следующим образом

$$\psi_{m+1} = r^m + \sum_{k=1}^{m} a_{k,m+1}\psi_k. \tag{5.11}$$

Из (5.11) вытекает, что r^k является линейной комбинацией не всех векторов системы $\psi_1, ..., \psi_n$, а только $\psi_1, ..., \psi_{k+1}$. Следовательно, и (r^m, r^k) является линейной комбинацией скалярных произведений $(r^m, \psi_1), ..., (r^m, \psi_{k+1})$. Вместе с (5.10) это и приводит к (5.9).

Линейная независимость системы $r^0, ..., r^{n-1}$ является простым следствием ее ортогональности. Поэтому в отсутствии ошибок округления единственной причиной остановки процесса (5.8) может быть только условие $r^m = 0$. Ортогональность системы невязок в (5.8) позволяет существенно упростить процесс построения A-базиса.

Пусть уже известна система $\psi_1 = r^0, \psi_2, ..., \psi_m$ такая, что $(A\psi_m, \psi_k) = 0$ для $k < m$. Вместо (5.11) определим ψ_{m+1} следующим образом

$$\psi_{m+1} = r^m + b_m\psi_m, \quad b_m = -\frac{(A\psi_m, r^m)}{(A\psi_m, \psi_m)}. \tag{5.12}$$

Коэффициент b_m в (5.12) выбран из условия A-ортогональности векторов ψ_m, ψ_{m+1}. Тогда

$$(A\psi_{m+1}, \psi_k) = (Ar^m, \psi_k) + b_m(A\psi_m, \psi_k) = (Ar^m, \psi_k) = (r^m, A\psi_k).$$

Но из (5.8) следует, что

$$A\psi_k = -\frac{r^k - r^{k-1}}{\tau_k}.$$

Поэтому

$$(r^m, A\psi_k) = -\frac{1}{\tau_k}(r^m, r^k) + \frac{1}{\tau_k}(r^m, r^{k-1}).$$

Теперь $(A\psi_{m+1}, \psi_k) = 0$ для $k = 1, ..., m-1$ вытекает из (5.9), а $(A\psi_{m+1}, \psi_m) = 0$ следует из (5.12). Для обоснования (5.12) остается заметить, что система векторов $\psi_1 = r^0$, $\psi_2 = r^0 + b_1\psi_1$ является A-ортогональной.

Итак, мы описали *метод сопряженных градиентов* в его классической реализации:

$$\begin{aligned} &r^0 = Ax^0 - b, \quad \psi_1 = r^0, \quad x^m = x^{m-1} - \tau_m\psi_m, \\ &r^m = r^{m-1} - \tau_m A\psi_m, \quad \psi_{m+1} = r^m + b_m\psi_m, \\ &\tau_m = \frac{(\psi_m, r^{m-1})}{(A\psi_m, \varphi_m)}, \quad b_m = -\frac{(A\psi_m, r^m)}{(A\psi_m, \psi_m)}, \quad m = 1, 2, ... \end{aligned} \tag{5.13}$$

В (5.13) система $r^0, r^1, ..., r^{m-1}$ – ортогональна, система $\psi_1, ..., \psi_m$ – A-ортогональна. Этап $\psi_m \longrightarrow \psi_{m+1}$ в (5.13) соответствует ортогонализации Грама-Шмидта (5.1), а этап $x^{m-1} \longrightarrow x^m$ соответствует разложению (5.5) по векторам сопряженного базиса ψ_m. Это разложение порождает и конкретный минимизируемый функционал, так что в методе сопряженных градиентов (5.13) при каждом $m \geq 1$ осуществляется как глобальная $(x^0 \longrightarrow x^m)$, так и локальная $(x^{m-1} \longrightarrow x^m)$ минимизация функционала ошибки $J(x^m)$ и функционала $\| z^m \|_A^2$.

Теоретически метод сопряженных градиентов (5.13) является прямым методом и при отсутствии ошибок округления приводит к точному решению задачи (5.2) не более, чем за n итераций. Однако ошибки округления неизбежны. Как следствие, даже в самом благоприятном случае $m > n$ вместо $z^m = x^m - x = 0$ будем иметь $z^m \neq 0$. Под "самым благоприятным" можно, например, понимать (5.13) в том случае, когда ведется контроль, связанный с вычислением скалярных произведений $(\psi_k, A\psi_m)$, (r^k, r^m) (теоретически эти скалярные произведения равны нулю при $k < m$), а при "существенном" нарушении ортогональности проводится в какой-либо форме процесс переортогонализации.

Поэтому в любом случае (особенно, если $m < n$) полезно иметь представление о характере убывания z^m. В (5.13) при *каждом* m осуществляется глобальная минимизация функционала ошибки $J(x^m)$. Следовательно, при *каждом* m осуществляется и глобальная минимизация функционала $\| z^m \|_A^2$. Поэтому если z^m соответствует методу сопряженных градиентов (5.13), а $\tilde{z}^m$ – любому другому итерационному методу, в котором минимизация функционала $\| \tilde{z}^m \|_A^2$ осуществляется при *фиксированном* m, то $\| z^m \|_A^2 \leq \| \tilde{z}^m \|_A^2$.

Пусть $\tilde{z}^m$ соответствует нестационарному методу простой итерации с чебышевским набором итерационных параметров τ_j:

$$\tilde{z}^m = \tilde{z}^{m-1} - \tau_m A\tilde{z}^m, \quad A = A^* > 0, \quad \delta E \le A \le \Delta E$$
$$\tau_j = \frac{2}{(\Delta + \delta) + (\Delta - \delta)t_j}, \quad \tau_j = \cos\frac{(2j-1)\pi}{2m}, \quad 1 \le j \le m. \tag{5.14}$$

Этот метод достаточно подробно рассматривался в §2. Мы сохраним обозначения этого параграфа и, опуская детали, еще раз приведем нужные для дальнейшего оценки. Итак,

$$(A\tilde{z}^m, \tilde{z}^m) = \sum_{i=1}^n \lambda_i (c_i^m)^2 = \sum_{i=1}^n \lambda_i [P_m(\lambda_i)]^2 (c_i^0)^2 \le$$
$$\le \max_{\delta\le\lambda\le\Delta} |P_m(\lambda)|^2 \sum_{i=1}^n \lambda_i (c_i^0)^2 = \max_{\delta\le\lambda\le\Delta} |P_m(\lambda)|^2 (A\tilde{z}^0, \tilde{z}^0).$$

При τ_j из (5.14)

$$\max_{\delta\le\lambda\le\Delta} |P_m(\lambda)| = q_m = \frac{2\rho_1^m}{1+\rho_1^{2m}}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad \xi = \frac{\delta}{\Delta}. \tag{5.15}$$

Теперь заметим, что для z^m из (5.13) справедливо представление

$$z^m = z^0 + \sum_{i=1}^m a_i A^i z^0, \quad z^0 = x^0 - x \tag{5.16}$$

и метод сопряженных градиентов для любого $m \ge 1$ минимизирует $\| z^m \|_A^2$ на классе всех полиномиальных приближений вида (5.16). Но к этому же классу принадлежит и $\tilde{z}^m$ из (5.14), поскольку справедливо представление $\tilde{z}^m = P_m(A)\tilde{z}^0$ с той же нормировкой $P_m(0) = E$ операторного (матричного) полинома $P_m(A)$, что и в (5.16). Итог вышесказанному подводит следующая

Теорема 5.3. *Если $A = A^* > 0$, то метод сопряженных градиентов* (5.13) *сходится и справедлива оценка*

$$\| z^m \|_A \le q_m \| z^0 \|_A, \quad m \ge 2, \tag{5.17}$$

где q_m определено в (5.15).

Доказательство уже очевидно, поскольку в (5.13) и в (5.14) можно выбрать одно и то же начальное приближение x^0, а тогда

$$\| z^m \|_A^2 \le \| \tilde{z}^m \|_A^2 \le q_m^2 \| z^0 \|_A^2.$$

Сделаем некоторые замечания относительно вычислительных аспектов метода сопряженных градиентов (5.13). Часть этих замечаний носит общий характер и связана с вычислением скалярных произведений (x, y), т.е. с реализацией на ЭВМ макрооперации $S = \sum\limits_i x_i y_i$. Общепринятая рекомендация здесь такова: арифметические

операции следует проводить с двойной точностью и округлять не каждое слагаемое $x_i y_i$, а лишь окончательный результат S. Для указания на такой режим вычислений используется обозначение $fl_2(S) = fl_2(\sum\limits_i x_i y_i)$. Другая особенность при реализации на ЭВМ макрооперации S связана с возможной потерей точности для S близких к "машинному нулю" ω. Пусть $x_i = y_i$. Тогда сферическая (евклидова) норма вектора x вычисляется по формуле

$$\| x \| = \sqrt{(x,x)} = \sqrt{S} = \sqrt{\sum_i x_i^2}. \tag{5.18}$$

Если для всех i имеем $|x_i| < \omega^{\frac{1}{2}}$, то $fl(S) = 0$, хотя $S \neq 0$. Более того, возможно, что $\sqrt{S} > \omega$. Как видим, ни о какой точности формулы (5.18) вблизи ω говорить не приходится. Поэтому другая общепринятая рекомендация заключается в замене формулы (5.18) на эквивалентную:

$$\| x \| = \begin{cases} 0, \ \| x \|_1 = 0, \ \| x \|_1 = \max\limits_i |x_i|, \ i = 1, ..., n \\ \| x \|_1 \sqrt{\sum\limits_i (x_i / \| x \|_1)^2}, \ \| x \|_1 \neq 0. \end{cases} \tag{5.19}$$

Анализ (5.19) приводит к соотношению

$$fl(\| x \|) \equiv \sqrt{\sum_i [x_i(1+\varepsilon_i)]^2},$$

где оценка сверху для эквивалентного возмущения ε_i не зависит от ω и тем меньше, чем выше разрядность используемой ЭВМ.

Что касается непосредственно (5.13), то соотношения ортогональности (5.9), (5.10) для систем векторов $r^0, r^1, ..., r^{m-1}$ и $\psi_1, ..., \psi_m$ позволяют получить различные эквивалентные формулы для параметров τ_m, b_m. Например,

$$\begin{aligned} \tau_m &= \frac{(\psi_m, r^{m-1})}{(A\psi_m, \psi_m)} = \frac{(r^{m-1}, r^{m-1})}{(A\psi_m, \psi_m)}, \\ b_m &= -\frac{(A\psi_m, r^m)}{(A\psi_m, \psi_m)} = \frac{(r_m, r^m)}{(r^{m-1}, r^{m-1})}. \end{aligned} \tag{5.20}$$

Доказательство справедливости (5.20) особых сложностей не представляет, если учесть, что из (5.9), (5.10) вытекает еще одно соотношение ортогональности

$$(r^m, A\psi_k) = 0, \quad m \neq k, \quad m \neq k+1. \tag{5.21}$$

Отметим, что из (5.20) при $A = A^* > 0$ следует и $\tau_m > 0$, $b_m > 0$. Использование второй формулы в (5.20) для b_m "теоретически" дает и реальное представление о

скорости убывания $\| r^m \|^2$ в процессе реализации метода сопряженных градиентов (5.13).

Потеря точности в методе сопряженных градиентов обычно связана с тем, что с ростом m система векторов $r^0, r^1, ..., r^{m-1}$ перестает быть ортогональной. Поэтому какой-либо контроль в (5.13), безусловно, необходим. "Идеальный", следовательно и "дорогой", контроль заключается, как уже говорилось, в проверке условий ортогональности (5.9), (5.10). Для контроля могут быть использованы и формулы (5.20). Одна из общепринятых рекомендаций заключается в том, чтобы вычислять τ_m по второй из формул (5.20), а первую – использовать для контроля. Другая рекомендация связана с одновременным вычислением невязок

$$r^m = r^{m-1} - \tau_m A\psi_m, \quad \tilde{r}^m = Ax^m - b. \tag{5.22}$$

В случае "существенного" расхождения результатов следует в (5.13) перейти на вторую формулу (5.22). Подобная рекомендация по существу означает переход на циклический вариант метода: $x^m, \tilde{r}^m$ принимаются за новое начальное приближение в классической реализации (5.13). Но тогда мы будем вынуждены считаться с тем, что для циклического варианта метода (5.13) уже не будет справедлива оценка (5.17).

Возникающую здесь ситуацию поясним на примере чебышевского метода (5.14), в котором при фиксированном m для $\| z^m \|$ также имеет место оценка (5.17). В соответствии с (1.9) и (5.15) число итераций в (5.14), требуемое для достижения точности ε дается формулой

$$m(\varepsilon) = \frac{1}{2}\sqrt{\nu(A)}\ln\frac{2}{\varepsilon} = \frac{1}{2}\sqrt{\frac{1}{\xi}}\ln\frac{2}{\varepsilon}.$$

Как правило, реально приходится иметь дело с плохо обусловленными задачами (1.1): число обусловленности ν – велико, ξ – мало. Поэтому и достигаемую точность ε естественно связывать с ξ. Положим $\nu = N^2$, $\varepsilon = 1/N^2$. Тогда для уменьшения нормы начальной погрешности в N^2 раз требуется

$$M\big(\frac{1}{N^2}\big) = \frac{1}{2}\sqrt{\nu}(\ln 2 + 2\ln N)$$

итераций. Циклический вариант метода (5.14) организуем следующим образом. Зададимся длиной цикла m и потребуем $q_m < 1/e$, где e – основание натурального логарифма. Такой выбор m соответствует уменьшению $\| z^0 \|$ в e-раз. Тогда

$$m\big(\frac{1}{e}\big) = \frac{1}{2}\sqrt{\nu}(1 + \ln 2). \tag{5.23}$$

Чтобы теперь уменьшить $\| z^0 \|$ в N^2-раз потребуется p-циклов, где $p = 2\ln N$. Общее число итераций в циклическом варианте метода (5.14):

$$\tilde{M}\Big(\frac{1}{N^2}\Big) \;=\; \sqrt{\nu}(1 \,+\, \ln 2)\ln N.$$

Отметим, что

$$\tilde{M}\Big(\frac{1}{N^2}\Big) \;=\; (1 \,+\, \ln 2)M\Big(\frac{1}{N^2}\Big).$$

и общее число итераций в только что построенном циклическом варианте незначительно превосходит число итераций для (5.14) при обычной реализации. В этом смысле указанный циклический вариант метода (5.14) является оптимальным.

В качестве следствия из вышесказанного отметим, что очень невыгодно использовать "короткие" длины цикла, когда $2m \ll \sqrt{\nu}$. В частности, при $m \,=\, 1$ (5.14) вырождается в метод простой итерации с оптимальным значением итерационного параметра. Это приводит к увеличению $\tilde{M}$ в $\sqrt{\nu}$ - раз.

В циклическом варианте метода сопряженных градиентов длину цикла заранее выбрать невозможно: при переходе на расчет r^m по второй из формул (5.22) m будет определяться из условия

$$r^{m-1} \,-\, \tau_m A\psi_m \neq Ax^m \,-\, b. \tag{5.24}$$

Поэтому возможны как оптимальные длины цикла типа (5.23): $m \simeq \sqrt{\nu}$, так и очень короткие. Следовательно, нарушение условий ортогональности (5.9), (5.10) в лучшем случае может приводить к существенному уменьшению скорости сходимости. В том числе и по этой причине широкое распространение получили трехчленные реализации метода сопряженных градиентов (5.13).

§6. Трехслойные итерационные методы

Начальный этап в методе сопряженных градиентов (5.13) совпадает с аналогичным этапом в явном методе скорейшего спуска (3.7), (3.10):

$$x^1 \;=\; x^0 \,-\, \tau_1 r^0, \quad \tau_1 \;=\; \frac{(r^0, r^0)}{(Ar^0, r^0)}. \tag{6.1}$$

Однако следующие приближения x^m в (5.13) будут связаны между собой уже трехчленными соотношениями. Действительно,

$$x^{m+1} \;=\; x^m \,-\, \tau_{m+1}\psi_{m+1} \;=\; x^m \,-\, \tau_{m+1}(r^m \,+\, b_m\psi_m),$$

откуда

$$\frac{x^{m+1} \,-\, x^m}{\tau_{m+1}} \,+\, b_m\psi_m \,+\, r^m \;=\; 0.$$

Кроме того,

$$\psi_m \;=\; -\frac{x^m \,-\, x^{m-1}}{\tau_m}. \tag{6.2}$$

Следовательно,

$$\frac{x^{m+1} - x^m}{\tau_{m+1}} - b_m \frac{x^m - x^{m-1}}{\tau_m} + r^m = 0. \tag{6.3}$$

Строго говоря, теперь в (6.3) следует положить $r^m = r^{m-1} - \tau_m A\psi_{m-1}$, затем с помощью (6.2) исключить ψ_{m-1} и т.д. В результате, как и в (5.16), получим

$$x^{m+1} = x^0 + \sum_{i=1}^{m} a_i A^i r^0, \quad r^0 = Ax^0 - b, \tag{6.4}$$

что равносильно представлению x^{m+1} в виде линейной комбинации *всех* предшествующих приближений $x^0, x^1, \ldots, x^m$.
Если же считать (в отсутствии ошибок округления – это так!)

$$r^m = r^{m-1} - \tau_m A\psi_{m-1} \equiv Ax^m - b, \tag{6.5}$$

то метод сопряженных градиентов (5.13) можно интерпретировать как явный трехслойный метод (6.1), (6.3), (6.5), итерационные параметры которого τ_m и b_m выбираются из условия минимальности $\|z^m\|_A^2$ при любом $m \geq 1$.

Итак, переход к трехчленному варианту метода сопряженных градиентов существенно связан с предположением (6.5). В этом случае двухчленные соотношения (5.13) обычно используют в таком виде:

$$\begin{gathered} r^0 = \psi_0, \quad r^m = Ax^m - b, \quad m = 0, 1, \ldots \\ x^{m+1} = x^m - \tau_m \psi_m, \ \tau_m = \frac{(\psi_m, r^m)}{(\psi_m, A\psi_m)}, \ m = 0, 1, \ldots \\ \psi_m = r^m - b_m \psi_{m-1}, \ b_m = -\frac{(r^m, A\psi_{m-1})}{(\psi_{m-1}, A\psi_{m-1})}, \ m = 1, 2, \ldots. \end{gathered} \tag{6.6}$$

Несущественные различия в (5.13) и (6.6) для итерационных параметров τ_m, b_m связаны с иной нумерацией векторов A–базиса ψ_m. В (5.13) $\psi_1 = r^0$, а в (6.6) $\psi_0 = r^0$. Такая нумерация позволяет несколько упростить формулы трехчленного варианта метода сопряженных градиентов.

В силу (6.5) для последовательных приближений x^m из (6.6) имеем

$$x^{m+1} = \Big(1 + b_m \frac{\tau_m}{\tau_{m-1}}\Big) x^m - b_m \frac{\tau_m}{\tau_{m-1}} x^{m-1} - \tau_m r^m$$

или

$$\begin{gathered} x^{m+1} = \alpha_{m+1} x^m + (1 - \alpha_{m+1}) x^{m-1} - \alpha_{m+1}\beta_{m+1} r^m \\ \alpha_{m+1} = 1 + b_m \frac{\tau_m}{\tau_{m-1}}, \quad \tau_m = \alpha_{m+1}\beta_{m+1}. \end{gathered} \tag{6.7}$$

Сразу же отметим, что поскольку $A = A^* > 0$, то в силу (5.20) для всех ненулевых невязок r^m будем иметь $\alpha_{m+1} > 1$. Отметим также и такие очевидные следствия из (6.7):

$$r^{m+1} = \alpha_{m+1} r^m + (1 - \alpha_{m+1}) r^{m-1} - \alpha_{m+1}\beta_{m+1} A r^m, \tag{6.8}$$

$$r^m = \alpha_m r^{m-1} + (1 - \alpha_m) r^{m-2} - \alpha_m \beta_m A r^{m-1}. \tag{6.9}$$

Теперь воспользуемся условием (5.9) ортогональности невязок: $(r^m, r^k) = 0$, $k < m$. Из (6.8) и условий $(r^{m+1}, r^m) = 0$, $(r^m, r^{m-1}) = 0$, $\alpha_{m+1} \neq 0$ сразу же получаем

$$\beta_{m+1} = \frac{(r^m, r^m)}{(Ar^m, r^m)}.$$

Опять-таки из (6.8) и условий $(r^{m+1}, r^{m-1}) = 0$, $(r^m, r^{m-1}) = 0$ будем иметь

$$(1 - \alpha_{m+1})(r^{m-1}, r^{m-1}) = \alpha_{m+1}\beta_{m+1}(Ar^m, r^{m-1}). \tag{6.10}$$

Наконец, (6.9) и условия $(r^m, r^{m-1}) = 0$, $(r^m, r^{m-2}) = 0$ дают

$$(r^m, r^m) = -\alpha_m \beta_m (Ar^m, r^{m-1}). \tag{6.11}$$

Из (6.10), (6.11) вытекает, что

$$\alpha_{m+1} = \Big(1 - \frac{\beta_{m+1}(r^m, r^m)}{\beta_m (r^{m-1}, r^{m-1})\alpha_m}\Big)^{-1}.$$

Тем самым мы приходим к эквивалентному трехчленному варианту метода сопряженных градиентов (6.6):

$$\begin{gathered}
r^0 = Ax^0 - b, \quad \tau_1 = \frac{(r^0, r^0)}{(Ar^0, r^0)}, \quad x^1 = x^0 - \tau_1 r^0, \\
x^{m+1} = \alpha_{m+1} x^m + (1 - \alpha_{m+1}) x^{m-1} - \alpha_{m+1}\beta_{m+1} r^m, \\
r^m = Ax^m - b, \quad \beta_{m+1} = \frac{(r^m, r^m)}{(Ar^m, r^m)}, \quad m \geq 1, \\
\alpha_1 = 1, \quad \alpha_{m+1} = \Big(1 - \frac{\beta_{m+1}(r^m, r^m)}{\beta_m (r^{m-1}, r^{m-1})\alpha_m}\Big)^{-1}, \quad m \geq 1.
\end{gathered} \tag{6.12}$$

Отметим, что в процессе реализации (6.12), как и в (5.13), (5.20) имеется возможность следить за поведением $\|r^m\|^2/\|r^{m-1}\|^2$, а для $\|z^m\|_A^2$ из (6.12) для любого $m \geq 2$ справедлива теорема 5.3.

Выбранная в (6.12) форма записи начального этапа $x^0 \to x^1$ лишний раз подчеркивает совпадение с тем же этапом в явном методе скорейшего спуска (3.7), (3.10). Сравним эти два метода. Очевидно, что параметры β_{m+1} в (6.12) совпадают с итерационными параметрами τ_{m+1} в (3.7), (3.10). Определение α_{m+1} в (6.12) не требует вычисления новых скалярных произведений. Поэтому дополнительные вычислительные затраты в (6.12) относительно невелики, а скорость сходимости, как это следует из сравнения (5.17) и (3.13), существенно выше, чем в методе (3.7), (3.10).

Попытаемся выяснить причину такого различия. Пусть x^m, x^{m+1}, x^{m+2} – последовательные приближения, полученные с помощью явного метода скорейшего спуска (3.7), (3.10). Как и в явном методе минимальных невязок, рассмотрим следующую процедуру ускорения (сравни с (4.21)):

$$v = \alpha x^{m+2} + (1 - \alpha)x^m. \tag{6.13}$$

Если $x = A^{-1}b$, то для $z = v - x$ получим

$$\begin{gathered} z = \alpha z^{m+2} + (1 - \alpha)z^m, \quad Az = \alpha Az^{m+2} + (1 - \alpha)Az^m, \\ z^{m+2} = z^{m+1} - \beta_{m+2}Az^{m+1}, \quad z^{m+1} = z^m - \beta_{m+1}Az^m. \end{gathered} \tag{6.14}$$

В (6.14) учтено, что параметры τ_k из (3.7), (3.10) совпадают с параметрами β_k из (6.12). Параметр α ускоряющей процедуры (6.13) выберем из условия минимальности функционала $\|z\|_A^2$. Тогда

$$\alpha = \frac{(z^m, Az^m) - (z^{m+2}, Az^m)}{(z^{m+2}, Az^{m+2}) - 2(z^{m+2}, Az^m) + (z^m, Az^m)}. \tag{6.15}$$

Дальнейшее связано с тем, что в методе скорейшего спуска (3.7), (3.10) ортогональны невязки *последовательных* приближений, так что

$$(Az^{m+1}, Az^m) = 0, \quad (Az^{m+1}, Az^{m+2}) = 0$$

и поэтому, например,

$$(z^{m+2}, Az^m) = (z^{m+1}, Az^m) = (z^{m+1}, Az^{m+1}).$$

Эти соотношения вместе со второй группой формул из (6.14) позволяют переписать (6.15) в таком виде

$$\alpha = \left(1 - \frac{\beta_{m+2}}{\beta_{m+1}} \cdot \frac{(r^{m+1}, r^{m+1})}{(r^m, r^m)}\right)^{-1}. \tag{6.16}$$

Теперь заметим, что по определению x^m, x^{m+1}, x^{m+2} ускоряющую процедуру (6.13) можно представить следующим образом:

$$\begin{gathered} x^m \to r^m = Ax^m - b, \ \ \beta_{m+1} = \frac{(r^m, r^m)}{(Ar^m, r^m)}, \ \ x^{m+1} = x^m - \beta_{m+1}r^m \\ v = \alpha(E - \beta_{m+2}A)x^{m+1} + (1 - \alpha)x^m + \alpha\beta_{m+2}b. \end{gathered} \tag{6.17}$$

С другой стороны, для трехчленного соотношения метода сопряженных градиентов (6.12) имеем

$$x^{m+2} = \alpha_{m+2}(E - \beta_{m+2}A)x^{m+1} + (1 - \alpha_{m+2})x^m + \alpha_{m+2}\beta_{m+2}b. \tag{6.18}$$

Остается сравнить α из (6.16) с α_{m+2} из (6.12), чтобы сделать следующий вывод. Если приближение x^m – задано, а x^{m+1} в (6.17) вычисляется с помощью метода скорейшего спуска (3.7), (3.10), то именно метод сопряженных градиентов (6.12) определяет ускоряющую процедуру (6.13), (6.16).

Очевидно, что метод (6.16), (6.17) с вычислением β_{m+1}, β_{m+2} из (3.10) можно интерпретировать как циклический вариант метода сопряженных градиентов (6.12) с "очень короткой" (следовательно, не оптимальной) длиной цикла. Но тем не менее уже для одного такого цикла: $x^m \longrightarrow x^{m+1} \longrightarrow v$ справедлива оценка (5.17) вместо оценки (3.13) для метода (3.7), (3.10) без процедуры ускорения $x^m \longrightarrow x^{m+1} \longrightarrow x^{m+2}$.

Трехчленный вариант (6.12) метода сопряженных градиентов обладает тем же принципиальным недостатком, что и двучленный вариант (5.13): ошибки округления приводят к невыполнению условий (5.9) ортогональности невязок $(r^k, r^m) = 0$ для $k < m$. Но именно эти условия при любом $m \geq 1$ обеспечивают глобальную минимизацию функционала $\|z^m\|_A^2$ для вычисляемых приближений x^m. Можно также показать, что (5.9) являются и необходимыми условиями глобальной минимизации.

Циклическая модификация (6.16), (6.17) трехчленного варианта (6.12)

$$v^m = x^m \to r^m \to x^{m+1} \to v^{m+2} = x^{m+2} \to r^{m+2} \to x^{m+3} \to v^{m+4} \dots \qquad (6.19)$$

не использует условие (5.9), она, как мы видели, построена на ортогональности невязок *только для последовательных приближений*. Поэтому ошибки округления в цикле $x^{m+2} \longrightarrow v^{m+4}$, приводящие к приближенному соотношению ортогональности $(r^{m+2}, r^{m+3}) \simeq 0$, можно считать независимыми от ошибок округления предыдущего цикла $x^m \longrightarrow v^{m+2}$, приводящих к $(r^m, r^{m+1}) \simeq 0$. Более того, степень влияния ошибок округления предыдущих циклов на ошибки округления вычисляемого цикла уменьшаются с ростом m. Модификация (6.16), (6.17) приобретает новое ценное качество: итерационный процесс становится "самоисправляющимся". Однако, как известно, за все хорошее приходится платить. Для данного случая такой платой является короткая длина цикла в (6.16), (6.17). Далее мы построим аналогичную "самоисправляющуюся" модификацию трехчленного варианта метода сопряженных градиентов (6.12), в которой длина цикла не лимитируется.

Еще раз напомним, что предписанный в (5.13) или в (6.12) выбор параметров τ_m, b_m или α_m, β_m обеспечивает глобальную минимизацию функционала $\|z^{m+1}\|_A^2$. Этот же функционал минимизируется и на каждом шаге вычислительных процессов (5.13), (6.12). Иными словами, глобально оптимальный итерационный процесс: $\min \|z^{m+1}\|_A^2$, $x^0 \longrightarrow x^{m+1}$ является и локально оптимальным: $\min \|z^{m+1}\|_A^2$, $x^m \longrightarrow x^{m+1}$. Говоря о локальности в трехчленном варианте метода, следует иметь в виду переход $x^{m-1}, x^m \longrightarrow x^{m+1}$. При этом весьма существенно, что для всех приближений x^k, $k < m+1$ выполнены условия ортогональности невязок (5.9). Поэтому формулы трехчленного варианта (6.12) метода сопряженных градиентов можно переписать следующим образом:

$$x^{m+1} = \alpha_{m+1}(E - \beta_{m+1}A)x^m + (1 - \alpha_{m+1})x^{m-1} + \alpha_{m+1}\beta_{m+1}b,$$
$$x^1 = (E - \tau_1 A)x^0 + \tau_1 b, \quad \tau_1 = \frac{(r^0, r^0)}{(Ar^0, r^0)}, \quad r^0 = Ax^0 - b, \tag{6.20}$$
$$\alpha_{m+1}\beta_{m+1} : (r^k, r^{m+1}) = 0, \quad k < m+1 \longleftrightarrow \min \|z^{m+1}\|_A^2.$$

Но именно в силу условия $(r^k, r^{m+1}) = 0$, $k < m+1$ погрешности округления на вычисляемом этапе x^{m-1}, $x^m \longrightarrow x^{m+1}$ нельзя считать независимыми от погрешностей округления на предшествующих этапах. В этом и заключается одна из главных причин существенного накопления вычислительных погрешностей в трехчленном варианте метода сопряженных градиентов (6.12).

Только что сказанное приводит к следующей модификации трехчленного варианта:

$$y^{m+1} = \gamma_{m+1}(E - \delta_{m+1}A)y^m + (1 - \gamma_{m+1})y^{m-1} + \gamma_{m+1}\delta_{m+1}b,$$
$$y^1 = (E - \tau_1 A)y^0 + \tau_1 b, \quad \tau_1 = \frac{(r^0, r^0)}{(Ar^0, r^0)}, \quad r^0 = Ay^0 - b, \tag{6.21}$$
$$\gamma_{m+1}, \delta_{m+1} : \min \|z^{m+1}\|_A^2 = \min \|y^{m+1} - x\|_A^2, \quad y^{m-1}, y^m \longrightarrow y^{m+1}.$$