ЖУРНАЛ
ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ и МАТЕМАТИЧЕСКОЙ ФИЗИКИ

Том 12 Июль 1972 Август № 4

УДК 518 : 517.944/.947

ВЫБОР ИТЕРАЦИОННЫХ ПАРАМЕТРОВ В МЕТОДЕ РИЧАРДСОНА

Е. С. НИКОЛАЕВ, А. А. САМАРСКИЙ

(Москва)

Предлагается упорядочение набора итерационных параметров метода Ричардсона, для которого он становится численно устойчивым. Число параметров произвольно.

В работе рассмотрен вопрос вычислительной устойчивости итерационного метода Ричардсона решения операторного уравнения I рода в гильбертовом пространстве. Этот метод обладает высокой скоростью сходимости, однако для задач с плохо обусловленным оператором на практике была выявлена его численная неустойчивость [1–3].

В [4] показано, что неустойчивость метода связана с порядком использования итерационных параметров и что предлагавшиеся ранее [1–3] способы упорядочения набора параметров не устраняют численную неустойчивость, а только уменьшают ее.

Дальнейшие исследования показали, что существует такой порядок в наборе итерационных параметров, для которого метод становится численно устойчивым. Этот порядок был предложен в [4] и [5], гл. VIII, для случая, когда число параметров $n = 2^p$.

В настоящей статье метод упорядочения параметров, изложенный в [5], обобщается на случай произвольного числа параметров n.

В § 1 дается описание метода Ричардсона и упорядочения набора параметров для случая $n = 2^p$. Там же формулируются теоремы о вычислительной устойчивости. Подробное доказательство теорем будет дано отдельно. § 2 посвящен заданию порядка в наборе параметров для случая произвольного n. В § 3 приведены результаты экспериментального изучения вычислительной устойчивости метода с описанным набором параметров.

§ 1. Постановка задачи

1. Пусть в вещественном гильбертовом пространстве H дано операторное уравнение I рода с самосопряженным оператором $(A = A^* > 0)$

$$Au = f, \tag{1.1}$$

где f — заданный, u — искомый элемент из H.

Для приближенного решения задачи (1.1) рассмотрим неявную двухслойную итерационную схему

$$(1.2)\qquad B\frac{y_{k+1}-y_k}{\tau_{k+1}}+Ay_k=f,\qquad k=0,1,\ldots,n-1,$$

с произвольным начальным приближением $y_0\in H$. Семейство схем (1.2) выделим условием

$$(1.3)\qquad B=B^*\geqslant\beta E,\qquad \beta>0,$$

и будем предполагать, что операторы A и B энергетически эквивалентны [5] с постоянными γ_1 и γ_2:

$$(1.4)\qquad \gamma_1 B\leqslant A\leqslant\gamma_2 B,\qquad \gamma_2>\gamma_1>0.$$

В предположениях (1.3), (1.4) решение задачи об оптимальном наборе итерационных параметров τ_k имеет вид [5]

$$(1.5)\qquad \begin{gathered}\tau_k=\frac{\tau_0}{1+\rho_0\mu_k},\\ \mu_k\in\mathfrak{M}_n=\left\{\cos\frac{2i-1}{2n}\pi,\quad i=1,2,\ldots,n\right\},\qquad k=1,2,\ldots,n,\end{gathered}$$

где $\mathfrak{M}_n$ — множество из n элементов, расположенных по возрастанию i,

$$\tau_0=\frac{2}{\gamma_1+\gamma_2},\qquad \rho_0=\frac{1-\xi}{1+\xi},\qquad \xi=\frac{\gamma_1}{\gamma_2}.$$

При таком наборе параметров справедливы оценки

$$\|y_n-u\|_D\leqslant q_n\|y_0-u\|_D,\qquad D=A \text{ или } B.$$

Здесь норма в энергетическом пространстве H_D определяется следующим образом:

$$\|x\|_D=\sqrt{(Dx,x)}\quad\text{для}\quad D^*=D>0,\qquad x\in H,$$

$$q_n=\frac{2\rho_1{}^n}{1+\rho_1{}^{2n}},\qquad \rho_1=\frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}.$$

Для уменьшения нормы начальной погрешности в пространстве H_A (H_B) в $1/\varepsilon$ раз достаточно n итераций:

$$n=n(\varepsilon,\xi)\approx\frac{\ln 0.5\varepsilon}{\ln\rho_1}\approx\frac{|\ln 0.5\varepsilon|}{2\sqrt{\xi}}.$$

Схему (1.2) с набором параметров (1.5) называют неявным итерационным методом Ричардсона.

2. При изучении сходимости метода (1.2) мы предполагали, что вычислительный процесс является идеальным, т. е. счет ведется с бесконечным числом знаков. Однако процесс округления результатов арифметических операций вносит в решение y_k некоторые погрешности на каждом этапе вычислений. Будем считать, что введение этих погрешностей эквивалентно возмущению входных данных задачи — начального приближения, правой части и операторов A и B.

Тогда реальное решение $\tilde{y}_k$ можно рассматривать, как точное решение задачи

$$\tilde{B}\frac{\tilde{y}_{k+1}-\tilde{y}_k}{\tau_{k+1}}+\tilde{A}\tilde{y}_k=\tilde{f}_{k+1}+\frac{1}{\tau_{k+1}}\tilde{w}_{k+1},\qquad k=0,1,\ldots,n, \tag{1.6}$$

$\tilde{y}_0$ задано.

При таком подходе задача о вычислительной погрешности метода сводится к исследованию устойчивости схемы (1.6) с возмущенными операторами $\tilde{A}$ и $\tilde{B}$ по начальным данным и правой части.

Вторым вопросом численной устойчивости метода (1.2) является вопрос о росте промежуточных решений y_k при различных упорядочениях множества $\mathfrak{M}_n$. Исследование устойчивости схемы (1.6) позволяет получить оценку для величины промежуточного решения.

Некоторые примеры, показывающие влияние порядка использования параметров τ_k на рост решения и точность метода, приведены в [5] и § 3 этой работы.

Будем предполагать, что схема (1.6) принадлежит к исходному семейству схем, т. е. выполнены условия

$$\tilde{A}=\tilde{A}^*>0,\qquad \tilde{B}=\tilde{B}^*\geqslant\tilde{\beta}E,\qquad \tilde{\beta}>0. \tag{1.7}$$

Мерой возмущения операторов A и B будем считать относительное изменение их энергии ($0\leqslant\alpha_1,\alpha_2<1$)

$$|((A-\tilde{A})x,x)|\leqslant\alpha_1(Ax,x),\qquad |((B-\tilde{B})x,x)|\leqslant\alpha_2(Bx,x). \tag{1.8}$$

Для исследования роста промежуточных решений $\tilde{y}_k$ перейдем от неявной схемы (1.6) к эквивалентной ей явной

$$\begin{gathered}x_{k+1}=\tilde{S}_{k+1}x_k+\tau_{k+1}\varphi_{k+1}+\psi_{k+1},\qquad k=0,1,\ldots,n-1,\\ \tilde{S}_{k+1}=E-\tau_{k+1}\tilde{C},\end{gathered} \tag{1.9}$$

где

$$x_k=\tilde{B}^{1/2}\tilde{y}_k,\qquad \tilde{C}=\tilde{B}^{-1/2}\tilde{A}\tilde{B}^{-1/2},\qquad \varphi_k=\tilde{B}^{-1/2}\tilde{f}_k,$$
$$\psi_k=\tilde{B}^{-1/2}\tilde{w}_k.$$

Для изучения вычислительной точности метода рассмотрим задачу для погрешности $z_k=\tilde{y}_k-u$:

$$\tilde{B}\frac{z_{k+1}-z_k}{\tau_{k+1}}+\tilde{A}z_k=\tilde{f}_{k+1}-f+\frac{1}{\tau_{k+1}}\tilde{w}_{k+1}+(A-\tilde{A})u,$$
$$z_0=\tilde{y}_0-u.$$

Эквивалентная явная схема имеет вид

$$\begin{gathered}x_{k+1}-\varphi=\tilde{S}_{k+1}(x_k-\varphi)+\tau_{k+1}\varphi_{k+1}+\psi_{k+1},\quad k=0,1,\ldots,n-1,\\ \tilde{S}_{k+1}=E-\tau_{k+1}\tilde{C},\end{gathered} \tag{1.10}$$

где

$$x_k=\tilde{B}^{1/2}z_k,\qquad \tilde{C}=\tilde{B}^{-1/2}\tilde{A}\tilde{B}^{-1/2},\qquad \varphi_k=\tilde{B}^{-1/2}(\tilde{f}_k-f),$$
$$\psi_k=\tilde{B}^{-1/2}\tilde{w}_k,\qquad \varphi=\tilde{B}^{1/2}\tilde{A}^{-1}(A-\tilde{A})u.$$

Из (1.9), (1.10) следует, что для получения необходимых нам оценок достаточно исследовать устойчивость схемы (1.9) по начальным данным и правой части.

Пользуясь рекуррентной формулой (1.9), найдем

$$x_k = \tilde{T}_{k,0}x_0 + \sum_{j=1}^{k} \tau_j \tilde{T}_{k,j}\psi_j + \sum_{j=1}^{k} \tilde{T}_{k,j}\psi_j, \qquad k = 1, 2, \ldots, n$$

(1.11)

$$\tilde{T}_{k,j} = \prod_{i=j+1}^{k} \tilde{S}_i, \qquad \tilde{T}_{k,k} \equiv E.$$

Оператор $\tilde{T}_{k,j}$ называется разрешающим оператором со слоя j на слой k.

Определение. Назовем набор параметров $\{\tau_k\}$ устойчивым, если существуют положительные C_1, C_2, C_3, зависящие, быть может, от γ_1, γ_2, α_1, α_2, но не зависящие от n, такие, что

$$\max_{1\leqslant k\leqslant n}\|\tilde{T}_{k,0}\| \leqslant C_1, \qquad \max_{1\leqslant k\leqslant n}\sum_{j=1}^{k}\tau_j\|\tilde{T}_{k,j}\| \leqslant C_2, \qquad \max_{1\leqslant k\leqslant n}\sum_{j=1}^{k}\|\tilde{T}_{k,j}\| \leqslant C_3.$$

Для устойчивого набора параметров при любом k справедливы оценки

$$\|x_k\| \leqslant C_1\|x_0\| + C_2 \max_{1\leqslant j\leqslant k}\|\varphi_j\| + C_3 \max_{1\leqslant j\leqslant k}\|\psi_j\|$$

для задачи (1.9) и

$$\|x_k\| \leqslant C_1\|x_0\| + C_2 \max_{1\leqslant j\leqslant k}\|\varphi_j\| + C_3 \max_{1\leqslant j\leqslant k}\|\psi_j\| + (1 + C_1)\|\varphi\|$$

для задачи (1.10), из которых следует, что схемы (1.9), (1.10) устойчивы.

3. Перейдем к заданию порядка элементов в множестве $\mathfrak{M}_n$, который порождает устойчивый набор итерационных параметров $\{\tau_k\}$ метода Ричардсона. Рассмотрим здесь случай, когда $n = 2^p$, $p > 0$.

При построении набора параметров $\{\tau_k\}$ будем, следуя [5], исходить из минимального $\beta = \beta_1 = \pi / 2n$ и рекуррентным образом строить множества

$$M_1(\beta) = \{-\cos\beta\},$$

(1.12)

$$M_{2^k}(\beta) = M_{2^{k-1}}(\beta) \cup M_{2^{k-1}}(\pi / 2^{k-1} - \beta), \quad k = 1, 2, \ldots, p.$$

Тогда

(1.13) $$\mathfrak{M}_{2^p} = M_{2^p}(\beta_1).$$

Следующая теорема утверждает, что множество $\mathfrak{M}_{2^p}$, упорядоченное таким образом, порождает устойчивый набор $\{\tau_k\}$ (μ_k в формуле (1.5) есть k-й элемент множества $\mathfrak{M}_{2^p}$).

Теорема 1. *Если выполнены условия* (1.3), (1.4) *и множество* $\mathfrak{M}_{2^p}$ *упорядочено согласно* (1.13), *то при условии* $\alpha_1 = \alpha_2 = 0$ *справедливы сле-*

дующие оценки, не зависящие от n ($m = j\cdot 2^k$, j — *нечетное число*, $k \geqslant 0$):

$$\|T_{m,0}\| \leqslant 1/\xi, \qquad m = 1, 2, \ldots, 2^p,$$

$$\sum_{i=1}^{m} \tau_i \|T_{m,i}\| \leqslant \frac{1}{\gamma_1}\left[1 + \left(1 - \delta_{1,j}\right)\frac{1}{\xi^2}\right],$$

$$\sum_{i=1}^{m} \|T_{m,i}\| \leqslant \frac{1}{\xi}\left[1 + (1 - \delta_{1,j})\frac{1}{\xi^2}\right],$$

где $\delta_{1,j}$ — *символ Кронекера. Для* $m = 2^p$ *справедливы более точные оценки*

$$\|T_{2^p,0}\| \leqslant q_{2^p} < 1, \qquad \sum_{j=1}^{2^p} \|T_{2^p,j}\| \tau_j \leqslant \frac{1 - q_{2^p}}{\gamma_1} < \frac{1}{\gamma_1}, \tag{1.14}$$

$$\sum_{j=1}^{2^p} \|T_{2^p,j}\| \leqslant \frac{4}{3\sqrt{\xi}}. \tag{1.15}$$

Теорема 1 выражает тот факт, что если в схеме (1.6) пренебречь возмущением операторов A и B, то промежуточные решения ограничены по норме ($m = j\cdot 2^k$)

$$\|\tilde{y}_m\|_B \leqslant \frac{1}{\xi}\|\tilde{y}_0\|_B + \left[1 + (1 - \delta_{1,j})\frac{1}{\xi^2}\right]\left(\frac{1}{\gamma_1}\max_{1\leqslant i\leqslant m}\|f_i\|_{B^{-1}} + \frac{1}{\xi}\max_{1\leqslant i\leqslant m}\|\tilde{w}_i\|_{B^{-1}}\right)$$

и для погрешности решения после n итераций справедлива оценка

$$\|\tilde{y}_n - u\|_B \leqslant q_n\|\tilde{y}_0 - u\|_B + \frac{1 - q_n}{\gamma_1}\max_{1\leqslant i\leqslant n}\|\tilde{f}_i - f\|_{B^{-1}} + \frac{4}{3\sqrt{\xi}}\max_{1\leqslant i\leqslant n}\|\tilde{w}_i\|_{B^{-1}}. \tag{1.16}$$

Теорема 2. *Если выполнены условия* (1.3), (1.4), (1.7), (1.8) *и множество* $\mathfrak{M}_{2^p}$ *упорядочено согласно* (1.13), *то при условии*

$$\alpha = \frac{\alpha_1 + \alpha_2}{1 - \alpha_2} \leqslant \frac{\xi}{2}$$

для погрешности решения задачи (1.6) *после* n *итераций справедлива оценка*

$$\|\tilde{y}_n - u\|_{\tilde{B}} \leqslant \frac{q_n}{\tilde{q}_n}\|\tilde{y}_0 - u\|_{\tilde{B}} + \frac{1}{\tilde{\gamma}_1}\left(1 - \frac{q_n}{\tilde{q}_n}\right)\max_{1\leqslant i\leqslant n}\|\tilde{f}_i - f\|_{\tilde{B}^{-1}} + \tag{1.17}$$
$$+ \frac{4}{3}\frac{1 + \sqrt{\alpha}}{\sqrt{\xi} - \sqrt{\alpha}}\max_{1\leqslant i\leqslant n}\|\tilde{w}_i\|_{\tilde{B}^{-1}} + \frac{\alpha_1}{\tilde{\gamma}_1}\left(1 + \frac{q_n}{\tilde{q}_n}\right)\|f\|_{\tilde{B}^{-1}}.$$

Здесь

$$\tilde{\gamma}_1 = \gamma_1 - \alpha\gamma_2 \geqslant 0.5\,\gamma_1,$$

$$\tilde{q}_n = \frac{2\tilde{\rho}_1^{\,n}}{1 + \tilde{\rho}_1^{\,2n}} > q_n, \qquad \tilde{\rho}_1 = \frac{1 - \sqrt{\xi_0}}{1 + \sqrt{\xi_0}}, \qquad \xi_0 = \frac{\alpha}{1 + \alpha - \xi}.$$

Теорема 2 выражает вычислительную устойчивость итерационного метода Ричардсона для случая, когда $n = 2^p$. Оценка (1.17) переходит при

α_1, $\alpha_2 \to 0$ в полученную ранее (1.16). Оценки (1.14), даваемые теоремой 1, являются неулучшаемыми для любых упорядочений множества $\mathfrak{M}_n$, а (1.15) точна по порядку малости ξ при больших n (ср. с леммой 6 из [4]).

Замечание. Если предположить, что ψ_j в (1.11) имеет вид $\psi_j = T_{j,0} v_j$, то такое возмущение эквивалентно возмущению x_0 на величину $\sum\limits_{j=1}^{k} v_j$.

§ 2. Построение набора параметров для произвольного n*)

1. Идея построения порядка в множестве $\mathfrak{M}_n$ для произвольного n основана на двух соображениях: на переходе от упорядоченного множества $\mathfrak{M}_{2k}$ к заданию порядка в множестве $\mathfrak{M}_{2k+1}$ и на переходе от $\mathfrak{M}_k$ к $\mathfrak{M}_{2k}$, где k — произвольное целое число.

Предположим, что множество $\mathfrak{M}_{2k}$ уже упорядочено необходимым нам образом. Тогда представим его в виде следующей суммы:

$$\mathfrak{M}_{2k} = \bigcup_{i=1}^{2k} M_1(\beta_{2k,i}) = M_1(\beta_{2k,1}) \cup M_1(\beta_{2k,2}) \cup \ldots \cup M_1(\beta_{2k,2k}),$$

где $\beta_{2k,i}$ при любом i принадлежит множеству

$$\left\{\frac{2j-1}{4k}\pi, \quad j = 1, 2, \ldots, 2k\right\}.$$

Тогда множество $\mathfrak{M}_{2k+1}$ упорядочим следующим образом:

$$\mathfrak{M}_{2k+1} = \bigcup_{i=1}^{2k} M_1(\beta_{2k+1,i}) \cup M_1\left(\frac{\pi}{2}\right),$$

где $\beta_{2k+1,i}$ — ближайшее к $\beta_{2k,i}$ число из множества

$$\left\{\frac{2j-1}{2(2k+1)}\pi, \; j = 1, 2, \ldots, 2k+1, \; j \neq k+1\right\}.$$

Переход от $\mathfrak{M}_{2k}$ к $\mathfrak{M}_{2k+1}$ осуществлен.

Пусть, далее, упорядочено множество $\mathfrak{M}_k$:

$$\mathfrak{M}_k = \bigcup_{i=1}^{k} M_1(\beta_{k,i}) = M_1(\beta_{k,1}) \cup \ldots \cup M_1(\beta_{k,k}),$$

$$\beta_{k,i} \in \left\{\frac{2j-1}{2k}\pi, \quad j = 1, 2, \ldots, k\right\}.$$

Тогда, используя формулы (1.12), зададим порядок в множестве $\mathfrak{M}_{2k}$ следующим образом:

$$\mathfrak{M}_{2k} = \bigcup_{i=1}^{k} M_2(\beta_{2k,i}) = \bigcup_{i=1}^{k} (M_1(\beta_{2k,i}) \cup M_1(\pi - \beta_{2k,i})),$$

$$\beta_{2k,i} = 0.5\beta_{k,i}, \qquad i = 1, 2, \ldots, k.$$

*) Как сообщил нам В. И. Лебедев, в [6, 7] предложены другие наборы параметров для любого n.

Эти соображения позволяют перейти от упорядоченного множества $\mathfrak{M}_1 = M_1(\pi/2)$, состоящего из одного элемента, к множеству $\mathfrak{M}_n$ с произвольным n, чередуя должным образом переходы от множества с четным числом элементов к множеству с нечетным числом и от множества из k элементов к множеству из $2k$ элементов.

Эта процедура упорядочения множества $\mathfrak{M}_n$ для произвольного n может быть формализована следующим образом.

Представим n в виде разложения в сумму по степеням 2 с целыми показателями k_i:

$$n = 2^{k_1} + 2^{k_2} + \ldots + 2^{k_t}, \qquad k_j \leqslant k_{j-1} - 1, \qquad k_t \geqslant 0.$$

Здесь t — целый индекс. Введем величины

$$n_j = \sum_{i=1}^{j} 2^{k_i - k_j}, \qquad \delta_j = \frac{n_j}{n} 2^{k_j}, \qquad j = 1, 2, \ldots, t, \tag{2.1}$$

и положим $k_{t+1} = -1$. Отметим, что все n_j — нечетные числа.

При построении набора параметров будем исходить из минимального

$$\beta = \beta_1 = \pi / 2n.$$

Образуем упорядоченную сумму множеств

$$M_n(\beta) = \bigcup_{i=1}^{t} M_{2^{k_i}}(n_i\beta) = M_{2^{k_1}}(n_1\beta) \cup \ldots \cup M_{2^{k_t}}(n_t\beta), \tag{2.2}$$

где $M_{2^k}(\beta)$ определяется рекуррентным образом $(j = 1, 2, \ldots, t)$:

$$\begin{gathered} M_1(\beta) = \{-\cos\beta\}, \\ M_{2^k}(\beta) = M_{2^{k-1}}(\beta) \cup M_{2^{k-1}}\left(\frac{\pi}{2^{k-1}}\delta_j - \beta\right), \end{gathered} \tag{2.3}$$

если $k_{j+1} + 2 \leqslant k \leqslant k_j + 1$.

Тогда

$$\mathfrak{M}_n = M_n(\beta_1). \tag{2.4}$$

Заметим, что если $n = 2^p$, то $t = 1$, $n_1 = 1$, $\delta_1 = 1$, $k_1 = p$. При этом формула (2.2) переходит в

$$M_n(\beta) = M_{2^p}(\beta)$$

и рекуррентные соотношения (2.3), определяющие $M_{2^p}(\beta)$, переходят в приведенные выше для $n = 2^p$ формулы (1.12). Следовательно, упорядочение множества $\mathfrak{M}_n$ согласно (2.4) является обобщением построений для случая $n = 2^p$.

Мы изложим сейчас алгоритм, позволяющий упорядочить множество $\mathfrak{M}_n$ в соответствии с (2.4). Построение множества $\mathfrak{M}_n$ непосредственно по формулам (2.2), (2.3) затруднительно; мы используем их для описания упорядочения $\mathfrak{M}_n$ в теоремах 1, 2.

Пусть θ_m — множество из m целочисленных элементов

$$\theta_m = \{\theta_m(1), \theta_m(2), \ldots, \theta_m(m)\}.$$

Положим $n_{t+1} = 2n + 1$. Пусть $j = 1$. Строим множества

$$\theta_{n_j} = \{\theta_{n_j}(i) = \theta_{n_j-1}(i), \quad i = 1, 2, \ldots, n_j - 1; \ \theta_{n_j}(n_j) = n_j\}, \tag{2.5}$$

$$\theta_{2m} = \{\theta_{2m}(2i) = 4m - \theta_m(i), \ \theta_{2m}(2i-1) = \theta_m(i), \ i = 1, 2, \ldots, m\}, \tag{2.6}$$
$$m = n_j, 2n_j, 4n_j, \ldots, 0.25(n_{j+1} - 1).$$

Если $j = t$, то необходимое множество θ_n уже построено, иначе строится множество

$$\theta_{n_{j+1}-1} = \{\theta_{n_{j+1}-1}(2i) = 2n_{j+1} - \theta_{0.5(n_{j+1}-1)}(i), \quad \theta_{n_{j+1}-1}(2i-1) = \tag{2.7}$$
$$= \theta_{0.5(n_{j+1}-1)}(i), \quad i = 1, 2, \ldots, 0.5(n_{j+1} - 1)\}.$$

Затем j увеличивается на 1 и процесс повторяется, начиная с (2.5). В результате будет построено множество θ_n.

Тогда

$$\mathfrak{M}_n = \left\{-\cos\beta_i, \ i = 1, 2, \ldots, n, \ \beta_i = \theta_n(i)\frac{\pi}{2n}\right\} \tag{2.8}$$

и μ_k в формуле (1.5) есть k-й элемент множества $\mathfrak{M}_n$.

Для случая $n = 2^p$ алгоритм (2.5)—(2.7) упрощается:

$$\theta_1 = \{\theta_1(1) = 1\},$$
$$\theta_{2m} = \{\theta_{2m}(2i) = 4m - \theta_m(i), \ \theta_{2m}(2i-1) = \theta_m(i), \ i = 1, 2, \ldots, m\},$$
$$m = 1, 2, 4, \ldots, 2^{p-1},$$

и после нахождения θ_{2^p} множество $\mathfrak{M}_{2^p}$ строится согласно (2.8). Приведем некоторые примеры:

$$\theta_8 = \{1, 15, 7, 9, 3, 13, 5, 11\},$$
$$\theta_9 = \{1, 17, 7, 11, 3, 15, 5, 13, 9\},$$
$$\theta_{12} = \{1, 23, 11, 13, 5, 19, 7, 17, 3, 21, 9, 15\},$$
$$\theta_{16} = \{1, 31, 15, 17, 7, 25, 9, 23, 3, 29, 13, 19, 5, 27, 11, 21\},$$
$$\theta_{18} = \{1, 35, 17, 19, 7, 29, 11, 25, 3, 33, 15, 21, 5, 31, 13, 23, 9, 27\},$$

§ 3. Вычислительная устойчивость метода

1. Для изучения устойчивости набора итерационных параметров $\{\tau_k\}$, построенного в соответствии с упорядочением множества $\mathfrak{M}_n$ по формуле (2.4), используется численный эксперимент.

Исследование характера итерационного метода, очевидно, может быть проведено на простейшей модели, так как характер процесса определяется не конкретным видом оператора A, а его основными функциональными свойствами как оператора в гильбертовом пространстве.

Эксперимент на модельной задаче позволяет сравнить теоретические оценки, полученные для случая $n = 2^p$ в теоремах 1, 2, с численными результатами.

В качестве модельной задачи была выбрана разностная аппроксимация краевой задачи

$$
(3.1)\qquad \begin{aligned}
&\frac{d^4v}{dx^4}=f(x), \qquad 0<x<1, \qquad v(0)=\nu_1,\\
&\frac{d^2v}{dx^2}(0)=\nu_2, \qquad v(1)=\nu_3, \qquad \frac{d^2v}{dx^2}(1)=\nu_4.
\end{aligned}
$$

На равномерной сетке с шагом $h=1/N$ строится разностная схема, аппроксимирующая задачу (3.1) с погрешностью $O(h^2)$:

$$
(3.2)\qquad \begin{aligned}
&u_{\bar{x}x\bar{x}x}=f, \qquad 2h\leqslant x\leqslant 1-2h,\\
&u(0)=\nu_1, \qquad u_{\bar{x}x}-hu_{\bar{x}xx}=\nu_2, \qquad x=h,\\
&u(1)=\nu_3, \qquad u_{\bar{x}x}+hu_{\bar{x}x\bar{x}}=\nu_4, \qquad x=1-h.
\end{aligned}
$$

Оператор A, соответствующий задаче (3.2), является самосопряженным в пространстве H сеточных функций, определенных во внутренних узлах сетки. Скалярное произведение в H задается обычным образом:

$$
(u,v)=\sum_{x=h}^{1-h} u(x)v(x)h.
$$

При этом собственными функциями оператора A будут $\mu_k(x)=\sin k\pi x$ и соответствующими собственными значениями будут

$$
\lambda_k=\frac{16}{h^4}\sin^4\frac{k\pi h}{2}, \qquad k=1,2,\ldots,N-1.
$$

Заметим, что $\|A\|=\lambda_{N-1}\approx 16/h^4=1.6\cdot 10^5$ при $N=10$ и $\|A\|\approx 1.6\cdot 10^9$ при $N=100$.

Выбор задачи (3.2) в качестве объекта эксперимента позволяет моделировать на грубой сетке плохо обусловленные операторы A.

Рассматривался явный итерационный процесс (1.2) ($B\equiv E$). Тогда в условиях (1.4) постоянные энергетической эквивалентности γ_1 и γ_2 есть $\gamma_1=\lambda_1$, $\gamma_2=\lambda_{N-1}$. При этом

$$
\xi=\operatorname{tg}^4\frac{\pi h}{2}.
$$

2. Проводилось несколько серий экспериментов. В первой серии изучалось влияние упорядочения множества $\mathfrak{M}_n$ на рост промежуточных решений и достигаемую после n итераций точность.

В модельной задаче (3.2) бралось $\nu_1=1$, $\nu_2=\nu_3=\nu_4=0$, $f\equiv 0$. При таких входных данных точное решение задачи есть $u(x)=1-x$.

Пусть n кратно 8. Рассматривались следующие упорядочения множества $\mathfrak{M}_n$:

1) $$
\mathfrak{M}_n=\bigcup_{k=1}^{n} M_1\left(\frac{2k-1}{2n}\pi\right)=
$$

$$
=M_1\left(\frac{\pi}{2n}\right)\cup M_1\left(\frac{3\pi}{2n}\right)\cup\ldots\cup M_1\left(\frac{2n-1}{2n}\pi\right);
$$

Этому упорядочению соответствует обычный «обратный» набор $\{\tau_k\}$:

$$\tau_k = \frac{\tau_0}{1+\rho_0\mu_k}, \qquad \mu_k = -\cos\frac{2k-1}{2n}\pi, \qquad k = 1, 2, \ldots, n;$$

2) $$\mathfrak{M}_n = \bigcup_{k=1}^{n} M_1\left(\pi - \frac{2k-1}{2n}\pi\right), \quad \text{здесь} \quad \mu_k = \cos\frac{2k-1}{2n}\pi,$$

что соответствует обычному «прямому» набору параметров $\{\tau_k\}$;

3) $$\mathfrak{M}_n = \bigcup_{k=1}^{n/2} M_2\left(\frac{2k-1}{2n}\pi\right) =$$
$$= \bigcup_{k=1}^{n/2}\left(M_1\left(\frac{2k-1}{2n}\pi\right) \cup M_1\left(\pi - \frac{2k-1}{2n}\pi\right)\right);$$

такому упорядочению соответствует разбиение набора $\{\tau_k\}$ на блоки из двух элементов ($\mu_{2k-1} = -\cos[(2k-1)/2n]\pi$, $\mu_{2k} = \cos[(2k-1)/2n]\pi$); здесь и ниже используются рекуррентные формулы (1.12);

4) $$\mathfrak{M}_n = \bigcup_{k=1}^{n/2} M_2\left(\frac{\pi}{2} - \frac{2k-1}{2n}\pi\right) =$$
$$= \bigcup_{k=1}^{n/2}\left(M_1\left(\frac{\pi}{2} - \frac{2k-1}{2n}\pi\right) \cup M_1\left(\frac{\pi}{2} + \frac{2k-1}{2n}\pi\right)\right).$$

Упорядочение, рекомендованное в [1], в наших обозначениях принимает вид

$$\mathfrak{M}_n = \bigcup_{k=1}^{n/2} M_2\left(\frac{\pi}{2} + \frac{2k-1}{2n}\pi\right);$$

5) $$\mathfrak{M}_n = \bigcup_{k=1}^{n/4} M_4\left(\frac{2k-1}{2n}\pi\right) =$$
$$= \bigcup_{k=1}^{n/4}\left(M_1\left(\frac{2k-1}{2n}\pi\right) \cup M_1\left(\pi - \frac{2k-1}{2n}\pi\right) \cup\right.$$
$$\left.\cup M_1\left(\frac{\pi}{2} - \frac{2k-1}{2n}\pi\right) \cup M_1\left(\frac{\pi}{2} + \frac{2k-1}{2n}\pi\right)\right).$$

Здесь производится организация блоков из 4 элементов:

$$\left(-\cos\beta_k,\ \cos\beta_k,\ -\sin\beta_k,\ \sin\beta_k;\quad \beta_k = \frac{2k-1}{2n}\pi\right);$$

6) $$\mathfrak{M}_n = \bigcup_{k=1}^{n/8} M_8\left(\frac{2k-1}{2n}\pi\right) =$$
$$= \bigcup_{k=1}^{n/8}\left(M_4\left(\frac{2k-1}{2n}\pi\right) \cup M_4\left(\frac{\pi}{4} - \frac{2k-1}{2n}\pi\right)\right);$$

здесь блок состоит из 8 элементов;

7) упорядочение множества $\mathfrak{M}_n$ по формуле (2.4).

Чтобы выявить влияние обусловленности оператора A на вычислительную устойчивость метода, расчеты проводились на последовательности сеток $h = {}^1/_{10}, {}^1/_{12}, {}^1/_{14}$.

Задавалось n с шагом 8 от 8 до 512. Для каждого n и каждого из указанных упорядочений $\mathfrak{M}_n$ проводились итерации по схеме (1.2). Определялось критическое значение $n_{\text{точ}}$, для которого реальная относительная точность еще не превышала теоретическую точность, т. е. выполнялось неравенство

$$\frac{\|y_n - u\|}{\|y_0 - u\|} = \varepsilon_{\text{реал}} \leqslant q_n.$$

Так же определялось $n_{\text{авост}}$, при котором на какой-то промежуточной итерации решение выходило за границу максимально возможного для представления в машине числа (10^{19}).

Расчеты показали, что для упорядочений 1) — 6) при увеличении n сначала происходит потеря точности, когда n становится больше $n_{\text{точ}}$, а затем аварийный останов, вызываемый ростом промежуточных решений.

Использование блоков более сложной структуры приводит к уменьшению численной неустойчивости метода, и чем больше размер блока, тем больше $n_{\text{авост}}$.

При использовании упорядочения 7) реальная относительная точность $\varepsilon_{\text{реал}}$ для всех n не превышала теоретическую, которая для $n = 512$ равна $q_n = 1.4 \cdot 10^{-11}$, $3.9 \cdot 10^{-8}$, $4.5 \cdot 10^{-6}$ для $h = {}^1/_{10}, {}^1/_{12}, {}^1/_{14}$. Промежуточные решения были ограничены и величина

$$R_n = \max_{\substack{0 \leqslant x \leqslant 1 \\ 1 \leqslant k \leqslant n}} |y_k(x)|$$

была монотонно возрастающей функцией n и для больших n не менялась с ростом n, т. е. выходила на асимптотическое значение.

Расчеты показали, что упорядочение $\mathfrak{M}_n$ согласно (2.4) для n, отличных от степени 2, сохраняет те же характеристики вычислительной устойчивости метода, как и в случае $n = 2^p$.

Результаты приведены в табл. 1. Первые строки соответствуют начальному приближению

$$y_0(x) = \begin{cases} 1, & x = 0, \\ 0, & x \neq 0, \end{cases}$$

вторые соответствуют $y_0(x) = \cos(\pi x / 2)$. Для этих приближений при упорядочении 7)

$$\max_{1 \leqslant n \leqslant 512} R_n = 208,\ 427,\ 784 \text{ и } \max_{1 \leqslant n \leqslant 512} R_n = 1.63,\ 2.73,\ 4.00$$

для $h = {}^1/_{10}, {}^1/_{12}, {}^1/_{14}$ соответственно.

3. Во второй серии находились оценки для

$$\|T_{n,0}\|, \qquad \sum_{j=1}^{n} \tau_j \|T_{n,j}\|, \qquad \sum_{j=1}^{n} \|T_{n,j}\|,$$

Таблица 1

	$h = 1/10$, $\xi = 6.29\cdot10^{-4}$				$h = 1/12$, $\xi = 3.004\cdot10^{-4}$				$h = 1/14$, $\xi = 1.61\cdot10^{-4}$			
	$n_{\text{точ}}$	q_n	$\varepsilon_{\text{реал}}$	$n_{\text{авост}}$	$n_{\text{точ}}$	q_n	$\varepsilon_{\text{реал}}$	$n_{\text{авост}}$	$n_{\text{точ}}$	q_n	$\varepsilon_{\text{реал}}$	$n_{\text{авост}}$
1	24	0.55	0.495	48	24	0.732	0.729	48	24	0.839	0.773	40
	24	0.55	0.548	48	24	0.732	0.731	48	24	0.839	0.837	48
2	24	0.55	0.481	72	24	0.732	0.713	64	24	0.839	0.765	64
	16	0.746	0.746	64	16	0.864	0.86	64	16	0.923	0.918	64
3	48	0.178	0.17	88	48	0.366	0.364	80	48	0.544	0.376	80
	48	0.178	0.177	96	48	0.366	0.366	88	48	0.544	0.542	88
4	48	0.178	0.171	136	40	0.47	0.42	136	48	0.544	0.482	128
	40	0.264	0.263	136	40	0.47	0.469	128	40	0.64	0.639	128
5	88	$2.4\cdot10^{-2}$	$2.2\cdot10^{-2}$	176	96	$7.2\cdot10^{-2}$	$5.8\cdot10^{-2}$	168	104	0.142	0.141	160
	104	$1.08\cdot10^{-2}$	$8.68\cdot10^{-3}$	192	104	$5.4\cdot10^{-2}$	$4.9\cdot10^{-2}$	184	104	0.142	0.105	176
6	184	$1.95\cdot10^{-4}$	$1.24\cdot10^{-4}$	416	176	$4.48\cdot10^{-3}$	$3.74\cdot10^{-3}$	408	176	$2.29\cdot10^{-2}$	$1.98\cdot10^{-2}$	336
	184	$1.95\cdot10^{-4}$	$1.9\cdot10^{-4}$	456	176	$4.48\cdot10^{-3}$	$4.44\cdot10^{-3}$	448	200	$1.25\cdot10^{-2}$	$9\cdot10^{-3}$	376

когда $\mathfrak{M}_n$ упорядочено согласно (2.4). Для этого вычислялись величины

$$I_1 = \|T_{n,0}y\| / \|y\| \leqslant \|T_{n,0}\|, \tag{3.3}$$

$$I_2 = \sum_{j=1}^{n} \tau_j \frac{\|T_{n,j}y\|}{\|y\|} \leqslant \sum_{j=1}^{n} \tau_j \|T_{n,j}\|, \tag{3.4}$$

$$I_3 = \sum_{j=1}^{n} \frac{\|T_{n,j}y\|}{\|y\|} \leqslant \sum_{j=1}^{n} \|T_{n,j}\|. \tag{3.5}$$

Для случая $n = 2^p$ в теореме 1 доказано, что равенства в (3.3)—(3.5) достигаются, если y является собственной функцией, соответствующей минимальному собственному значению задачи

$$Ay - \lambda By = 0.$$

При этом справедливы равенства

$$I_1 = \|T_{2^p,0}\| = q_{2^p}, \qquad I_2 = \sum_{j=1}^{2^p} \tau_j \|T_{2^p,j}\| = \frac{1 - q_{2^p}}{\gamma_1} \tag{3.6}$$

и оценка

$$I_3 = \sum_{j=1}^{2^p} \|T_{2^p,j}\| \leqslant \frac{4}{3\sqrt{\xi}}. \tag{3.7}$$

Для модельной задачи (3.2) имеем $y = \sin \pi x$.

Кроме того в теореме 1 доказано, что для произвольного n при любом упорядочении множества $\mathfrak{M}_n$ справедлива оценка

$$\sum_{j=1}^{n} \tau_j \|T_{n,j}\| \geqslant \frac{1 - q_n}{\gamma_1}.$$

Таблица 2

$h = 1/10, \quad \xi = 6.29 \cdot 10^{-4}$

n	I_1	q_n	I_2	$(1-q_n)/\gamma_1$	I_3	$I_3\sqrt{\xi}$
64	$8.0451 \cdot 10^{-2}$	$8.0451 \cdot 10^{-2}$	$9.5968 \cdot 10^{-3}$ $3.5085 \cdot 10^{-4}$	$9.5968 \cdot 10^{-3}$	42.726 27.171	1.072 0.6816
96	$1.6174 \cdot 10^{-2}$	$1.6174 \cdot 10^{-2}$	$1.0268 \cdot 10^{-2}$ $3.6973 \cdot 10^{-4}$	$1.0268 \cdot 10^{-2}$	45.034 28.641	1.1297 0.718
128	$3.2467 \cdot 10^{-3}$	$3.2467 \cdot 10^{-3}$	$1.0403 \cdot 10^{-2}$ $3.8662 \cdot 10^{-4}$	$1.0403 \cdot 10^{-2}$	47.072 29.933	1.181 0.7509
192	$1.308 \cdot 10^{-4}$	$1.308 \cdot 10^{-4}$	$1.0435 \cdot 10^{-2}$ $3.8184 \cdot 10^{-4}$	$1.0435 \cdot 10^{-2}$	46.5 29.57	1.167 0.7418
256	$5.27 \cdot 10^{-6}$	$5.27 \cdot 10^{-6}$	$1.044 \cdot 10^{-2}$ $3.868 \cdot 10^{-4}$	$1.044 \cdot 10^{-2}$	47.098 29.95	1.182 0.7513
344	$6.37 \cdot 10^{-8}$	$6.37 \cdot 10^{-8}$	$1.044 \cdot 10^{-2}$ $4.3697 \cdot 10^{-4}$	$1.044 \cdot 10^{-2}$	53.143 33.768	1.333 0.8471
384	$8.55 \cdot 10^{-9}$	$8.55 \cdot 10^{-9}$	$1.044 \cdot 10^{-2}$ $3.8787 \cdot 10^{-4}$	$1.044 \cdot 10^{-2}$	47.225 30.03	1.185 0.753

Таблица 3

$h = 1/20, \quad \xi = 3.84 \cdot 10^{-5}$

n	I_1	q_n	I_2	$(1-q_n)/\gamma_1$	I_3	$I_3\sqrt{\xi}$
64	0.75125	0.75125	$2.5641 \cdot 10^{-3}$ $3.115 \cdot 10^{-5}$	$2.5641 \cdot 10^{-3}$	62.066 39.506	0.384 0.245
96	0.55725	0.55725	$4.564 \cdot 10^{-3}$ $4.48 \cdot 10^{-5}$	$4.564 \cdot 10^{-3}$	89.331 56.863	0.553 0.352
128	0.39313	0.39313	$6.2558 \cdot 10^{-3}$ $5.708 \cdot 10^{-5}$	$6.2558 \cdot 10^{-3}$	113.86 72.474	0.705 0.449
192	0.1838	0.1838	$8.4137 \cdot 10^{-3}$ $7.42 \cdot 10^{-5}$	$8.4137 \cdot 10^{-3}$	148.04 94.234	0.917 0.584
256	$8.3747 \cdot 10^{-2}$	$8.3747 \cdot 10^{-2}$	$9.445 \cdot 10^{-3}$ $8.64 \cdot 10^{-5}$	$9.445 \cdot 10^{-3}$	172.26 109.65	1.067 0.679
344	$2.8197 \cdot 10^{-2}$	$2.8197 \cdot 10^{-2}$	$1.002 \cdot 10^{-2}$ $9.88 \cdot 10^{-5}$	$1.002 \cdot 10^{-2}$	197.03 125.4	1.22 0.777
384	$1.7181 \cdot 10^{-2}$	$1.7181 \cdot 10^{-2}$	$1.013 \cdot 10^{-2}$ $9.136 \cdot 10^{-5}$	$1.013 \cdot 10^{-2}$	182.23 116.0	1.129 0.718
512	$3.5191 \cdot 10^{-3}$	$3.5191 \cdot 10^{-3}$	$1.027 \cdot 10^{-2}$ $9.56 \cdot 10^{-5}$	$1.027 \cdot 10^{-2}$	190.66 121.37	1.181 0.752
768	$1.4762 \cdot 10^{-4}$	$1.4762 \cdot 10^{-4}$	$1.0307 \cdot 10^{-2}$ $9.43 \cdot 10^{-5}$	$1.0307 \cdot 10^{-2}$	188.18 119.78	1.166 0.742
1024	$6.192 \cdot 10^{-6}$	$6.192 \cdot 10^{-6}$	$1.0308 \cdot 10^{-2}$ $9.56 \cdot 10^{-5}$	$1.0308 \cdot 10^{-2}$	190.72 121.4	1.181 0.752

Чтобы исследовать, как меняются все три нормы I_1, I_2, I_3 в окрестности $n = 2^p$, используется численный эксперимент. Для вычисления указанных выше норм достаточно найти

$$\max_y I_1, \qquad \max_y I_2, \qquad \max_y I_3,$$

когда y пробегает множество собственных функций оператора A.

Расчеты проводились на последовательности сеток $h = {}^1/_{10}, {}^1/_{12}, {}^1/_{14}, {}^1/_{20}$. Число итераций n задавалось, как и в первой серии. При этом оказалось, что в (3.3) — (3.5) равенство достигалось на первой собственной функции оператора A, как и для случая $n = 2^p$, и были справедливы равенства (3.6).

Попутно мы убедились в том, что теоретические оценки (3.6), (3.7) при $n = 2^p$ достигаются.

Некоторые результаты приведены в табл. 2, 3. В первых строках значения I_1, I_2, I_3 даны для $y = \sin \pi x$, во вторых строках I_2, I_3 даны для сравнения при $y = \sin (N - 1)\pi x$.

Рассмотренные выше эксперименты на модельной задаче показывают, что для n, отличного от степени 2, при упорядочении множества $\mathfrak{M}_n$ согласно (2.4) сохраняются те же характеристики вычислительной устойчивости метода, как и для случая $n = 2^p$, для которого в теоремах 1, 2 получены теоретические оценки.

Использование упорядочения $\mathfrak{M}_n$ согласно (2.4) необходимо при решении задач с плохо обусловленным оператором, например разностных задач, возникающих из аппроксимации эллиптических уравнений высокого порядка. В то же время, как показывает пример 1, если n не слишком велико и задача не очень плохо обусловлена, то можно пользоваться более простым упорядочением с помощью блоков малой длины.

Поступила в редакцию 24.03.1972

Цитированная литература

1. D. Young. On Richardson's method for solving linear system with positive definite matrices. J. Math. and Phys., 1954, **32**, № 4, 243—255.
2. В. Вазов, Дж. Форсайт. Разностные методы решения дифференциальных уравнений в частных производных. М., Изд-во ин. лит., 1963.
3. Д. К. Фаддеев, В. Н. Фаддеева. Вычислительные методы линейной алгебры. М., Физматгиз, 1963.
4. В. Н. Лебедев, С. А. Финогенов. О порядке выбора итерационных параметров в чебышевском циклическом итерационном методе. Ж. вычисл. матем. и матем. физ. 1971, **11**, № 2, 425—438.
5. А. А. Самарский. Введение в теорию разностных схем. М., «Наука», 1971.
6. Г. И. Марчук, В. И. Лебедев. Численные методы в теории переноса нейтронов. М., Атомиздат, 1971.
7. В. И. Лебедев, С. А. Финогенов. Об одном алгоритме выбора параметров в чебышевских циклических методах. В сб. «Вычисл. методы линейной алгебры». Новосибирск, ВЦ СО АН СССР, 1972, 21—27.