ВЫЧИСЛЕНИЕ ЭКСПОНЕНТЫ ОТ АСИМПТОТИЧЕСКИ УСТОЙЧИВОЙ МАТРИЦЫ

А. Я. БУЛГАКОВ

ВВЕДЕНИЕ

Хорошо известна плохая обусловленность задачи вычисления экспоненты от произвольной квадратной матрицы. В этой работе на классе асимптотически устойчивых матриц для широко используемого метода расчета выведены гарантированные оценки точности вычисления матричной экспоненты. Оценки зависят от разрядной сетки используемой ЭВМ, параметра качества устойчивости $\varkappa(A)$ [1], размерности N данной матрицы A и величины ее спектральной нормы. На рассматриваемом классе такие оценки принципиально точнее аналогичных оценок работы [2].

Используемый алгоритм вычисления e^A (см., например, [2—4]) основан на представлении

$$e^{\widetilde{A}} = \sum_{k=0}^{\infty} \frac{1}{k!} \widetilde{A}^k; \quad A = 2^m \widetilde{A},$$

и равенстве

$$e^A = e^{2^m \widetilde{A}} = \left[e^{\widetilde{A}}\right]^{2^m},$$

которое позволяет свести вычисление матричной экспоненты произвольных асимптотически устойчивых матриц к предварительному вычислению экспоненты матрицы малой нормы (норма $\|\widetilde{A}\|$ порядка единицы). Для вычисления экспоненты от матрицы малой нормы с одинаковым успехом используются метод аппроксимации Падэ и тейлоровское разложение матрицы в ряд. В обоих случаях получаемая матрица приближает экспоненту с машинной точностью

$$\|(e^{\widetilde{A}})_{\text{выч}} - e^{\widetilde{A}}\| \leqslant C\varepsilon_1,$$

где $C < 100N$ и ε_1 — машинная константа такая, что $1 + \varepsilon_1$ — наименьшее машинное число, превосходящее единицу. Основным моментом при учете погрешности, накопившейся в указанном методе, является учет погрешностей определения двоичных степеней $e^{\widetilde{A}}$. Так, в работе [2] для получения оценки предлагается использовать рекуррентные соотношения

$$\left\|(e^{2^k\widetilde{A}})_{\text{выч}} - e^{2^k\widetilde{A}}\right\|_1 \leqslant 2\left\|e^{2^{k-1}\widetilde{A}}\right\|_1 \cdot \left\|(e^{2^{k-1}\widetilde{A}})_{\text{выч}} - e^{2^{k-1}\widetilde{A}}\right\|_1 +$$
$$+ \left\|(e^{2^{k-1}\widetilde{A}})_{\text{выч}} - e^{2^{k-1}\widetilde{A}}\right\|_1^2 + N\varepsilon_1 \left\|(e^{2^{k-1}\widetilde{A}})_{\text{выч}}\right\|_1^2,$$

где $\|X\|_1 = \max\limits_{1 \leqslant j \leqslant N} \left(\sum\limits_{i=1}^{N} |X_{ij}|\right)$.

В этом случае оценка будет не точнее, чем

$$\left\|(e^{2^k\widetilde{A}})_{\text{выч}} - e^{2^k\widetilde{A}}\right\|_1 \leqslant 2^k \prod_{i=0}^{k-1} \left\|e^{2^i\widetilde{A}}\right\|_1 \left\|(e^{\widetilde{A}})_{\text{выч}} - e^{\widetilde{A}}\right\|_1.$$

Выведенная в данной работе оценка для не слишком больших k имеет вид

$$\left\|(e^{2^k\widetilde{A}})_{\text{выч}} - e^{2^k\widetilde{A}}\right\| \leqslant \delta_k e^{-2^k\|\widetilde{A}\|/\varkappa(\widetilde{A})},$$

где $\|\widetilde{A}\|$ — спектральная норма матрицы $\widetilde{A}$; δ_k — монотонно растущая числовая последовательность, не превосходящая единицы. Алгоритм вы-

числения δ_k в каждом конкретном случае приводится в статье. Проиллюстрируем на одном примере преимущество выведенной оценки.

Рассмотрим двупараметрическое семейство матриц $A(\alpha, \beta)$ размерности 15×15 $(N=15)$: $A(\alpha,\beta)=C+B+\Lambda+D$. Здесь $C=\{C_{ij}\}$ — двухдиагональная матрица, все ее диагональные элементы равны $-\alpha$, а наддиагональные β. Оставшиеся матрицы $B=\{b_{ij}\}$, $\Lambda=\{\lambda_{ij}\}$, $D=\{d_{ij}\}$ определяются при помощи вектора p $(i=1, 2, \ldots, N)$:

$$p_i=\sin(i/7)\cdot\sqrt{2}\left(\sum_{j=1}^{N}[\sin(j/7)]^2\right)^{-1/2}$$

по формулам

$$\lambda_{ij}=\begin{cases}2p_ip_j\alpha-(p_{i+1}p_j+p_ip_{j+1})\beta & \text{при } j\neq N,\ i\neq N;\\ 2p_ip_j\alpha-p_jp_{i+1}\beta & \text{при } j=N,\ i\neq N;\\ 2p_ip_j\alpha-p_ip_{j+1}\beta & \text{при } j\neq N,\ i=N;\\ 2p_ip_j\alpha & \text{при } j=N,\ i=N;\end{cases}$$

$$B_{ij}=p_ip_j\left[\sum_{m=1}^{N-1}p_mp_{m+1}\beta-2\alpha\right];\quad d_{ij}=-\sum_{m=1}^{2}\beta p_{5m}p_ip_{5m-1}p_j.$$

Положим $\alpha=-16$ и $\beta=107{,}2$. В этом случае $\varkappa(A)=0.8_{10}7$. Допустим, что $\varepsilon_1=10^{-15}$, что соответствует, например, машине ЕС-1050. Погрешность $\|(e^A)_{\text{выч}}-e^A\|$, вычисляемая по алгоритму работы [2], дает величину $1.28_{10}19\varepsilon_1>.1_{10}5$, в то время как погрешность вычисления по алгоритму данной статьи не превосходит $4.4_{10}-2$.

Грубая схема алгоритма приведена в § 1. В § 2 рассмотрены специальные приемы программирования матричных процессов (дополнительные нормировки матриц, вычисления с двойной точностью), и для них проведены аккуратные оценки погрешностей округления при реализации на ЭВМ. Это позволило вывести в § 3 оценки точности вычисления экспоненты от матрицы малой нормы. В § 4, носящем вспомогательный характер, получены неравенства, играющие основную роль при учете влияния ошибок округления на процесс вычисления удвоенных степеней экспоненты, описанный в § 5. В § 6 на основании результатов предыдущих параграфов конкретизирован алгоритм § 1. В § 7 дана сводка численных экспериментов.

Выбор конкретного варианта вычисления e^A сделан с учетом ряда вычислительных экспериментов. Необходимость ограничиться только устойчивыми матрицами показалась автору естественной после продумывания дипломной работы А. М. Исраилова, выполненной в 1980 г. под руководством С. К. Годунова в Новосибирском государственном университете.

Проведенный в данной работе анализ влияния ошибок округления на точность вычисления матричных экспонент позволяет перейти к учету влияния ошибок округления в схеме численного исследования асимптотической устойчивости матриц, предложенной в [5].

Автор выражает свою признательность С. К. Годунову за постановку задач и постоянное внимание к работе.

§ 1. ГРУБАЯ СХЕМА АЛГОРИТМА

Описывается грубая схема алгоритма вычисления экспоненты от асимптотически устойчивой матрицы A $(\varkappa(A)<\infty)$, основанного на представлении $e^{\widetilde{A}}=\sum_{k=0}^{\infty}\frac{1}{k!}\widetilde{A}^k$ и равенстве $e^A=e^{2^m\widetilde{A}}=\left[e^{\widetilde{A}}\right]^{2^m}$, если $A=2^m\widetilde{A}$.

Предлагаемая схема является одним из возможных вариантов вычисления матричной экспоненты удвоениями степеней экспоненты от матрицы малой нормы $(1/2\leqslant\sigma_N(\widetilde{A})=\|\widetilde{A}\|<1)$.

В дальнейшем укажем (§ 2) на специальные приемы программной реализации алгоритма (дополнительные нормировки матриц, вычисления с двойной точностью), которые существенно сократят влияние ошибок округления на окончательные результаты. В § 3 и 5 будут учтены погрешности, возникающие на 4° и 5° этапах работы алгоритма, что позволит дать правило вычисления величины, представляющей гарантированную оценку накопления ошибок округления при вычислении e^A.

Если A — асимптотически устойчивая матрица размерности $N \times N$ и $\varkappa$ — ее параметр качества устойчивости, то для вычисления e^A предлагается следующая схема алгоритма:

1°. Находится $\sigma_N(A) = \|A\|$ — максимальное сингулярное число A.

2°. Вычисляется целое число k_1 такое, что

$$1/2 \leqslant |2^{-k_1}\sigma_N(A)| < 1.$$

3°. Производится нормировка A: $A_1 = 2^{-k_1}A$. Очевидно, что $\left[e^{A_1}\right]^{2^{k_1}} = e^{2^{k_1}A_1} = e^A$.

4°. Определяется матрица

$$B_0 = I + \frac{1}{1!}A_1 + \frac{1}{2!}A_1^2 + \ldots + \frac{1}{k_0!}A_1^{k_0}.$$

Заметим, что величина k_0 выбирается таким образом, чтобы погрешность, накопившаяся при вычислении B_0, была одного порядка с нормой остаточного члена $\|B_0 - e^{A_1}\|$. Правило определения k_0 будет сформулировано в § 3.

5°. Отправляясь от матрицы B_0, полученной в предыдущем пункте, вычислим B_{k_1} по формулам: $B_m = B^2_{m-1}$ $(m = 1, 2, \ldots, k_1)$. При этом гарантируется, что $\|B_{k_1} - e^A\| \leqslant \delta_0$.

§ 2. ПОГРЕШНОСТИ МАШИННОЙ РЕАЛИЗАЦИИ «АРИФМЕТИКИ ВЫНЕСЕННЫХ ПОРЯДКОВ»

Получены оценки машинных погрешностей, возникающих в реализации «арифметики вынесенных порядков» при работе с матрицами (см. [6]).

В наиболее распространенных вычислительных машинах, использующих представление чисел с «плавающей запятой», разрядность машины можно характеризовать двумя машинными постоянными ε_1 и ε_2 такими, что $1 + \varepsilon_1$ — наименьшее машинное число, превосходящее единицу, и $(1 - \varepsilon_1/\gamma)\varepsilon_2/\gamma^2$ — наименьшее по модулю, отличное от нуля машинное число. Здесь γ — основание системы счисления, принятой в машине. Подробнее об этих постоянных см., например, в § 4 книги [7]. Там приведены оценки погрешностей, возникающих при выполнении на ЭВМ арифметических операций с машиннозаданными числами α и β:

$$\begin{aligned} |(\alpha \pm \beta)_{\text{маш}} - (\alpha \pm \beta)| &\leqslant \max\{\varepsilon_1|\alpha \pm \beta|,\ \varepsilon_2/\gamma^2\};\\ |(\alpha \cdot \beta)_{\text{маш}} - (\alpha \cdot \beta)| &\leqslant \max\{\varepsilon_1|\alpha \cdot \beta|,\ \varepsilon_2/\gamma^2\};\\ |(\alpha/\beta)_{\text{маш}} - (\alpha/\beta)| &\leqslant \max\{\varepsilon_1|\alpha/\beta|,\ \varepsilon_2/\gamma^2\}. \end{aligned} \tag{2.1}$$

Для оценки погрешности вычисления скалярного произведения двух машиннозаданных векторов x и y размерности N можно использовать неравенство, приведенное на с. 115 работы [8]: $|[(x, y)]_{\text{выч}} - (x, y)| \leqslant N\varepsilon_1\|x\|\|y\|$. Если скалярное произведение векторов накапливается с удвоенной точностью, то справедливо более сильное неравенство

$$|[(x, y)]_{\text{выч}} - (x, y)| \leqslant \varepsilon_1|(x, y)| + N\varepsilon_1^2/\gamma\|x\|\|y\| + N\varepsilon_2/\gamma^2, \tag{2.2}$$

вывод которого можно найти в § 21 книги [7].

Назовем $[X^0, X^1]$ канонической матричной парой (канонической парой) представления матрицы X, если X^1 — целое число и для матриц X^0, X^1 выполнены условия

$$X = \gamma^{X^1} X^0; \quad 1/\gamma \leqslant \max_{1 \leqslant i,j \leqslant N} |X^0_{ij}| \leqslant 1.$$

Нулевая матрица представляется канонической парой [0, 0]. Аналогично определяется и каноническая пара, задающая число.

Задание матриц и чисел каноническими парами требует специального программирования операций сложения и вычитания двух матриц, а также умножения матрицы на скаляр и определения канонической пары для каждой матрицы и числа. В дальнейшем совокупность таких операций будем называть «арифметикой вынесенных порядков».

Для формального описания этой «арифметики» удобно ввести операторы $\mathfrak{M}$ и $\mathscr{P}$, отображающие пространство матриц размерности $N \times N$ соответственно в себя, и множество целых чисел Z: $\mathfrak{M}(X) = X^0$; $\mathscr{P}(X) = X^1$. Современные ЭВМ позволяют безошибочно вычислять значение $\mathscr{P}(X)$. Ошибка при вычислении $\mathfrak{M}(X)$ может возникнуть лишь в случае выполнения неравенства

$$|\gamma^{-X^1} X_{ij}| < \frac{1}{\gamma^2} \varepsilon_2,$$

что позволяет гарантировать выполнение неравенств ($\|X^0\|_E \geqslant 1/\gamma$, если $X^0 \neq 0$):

$$\|X^0 - (X^0)_{\text{выч}}\|_E \leqslant N/\gamma \varepsilon_2 \|X^0\|_E; \tag{2.3}$$

$$\|X - \gamma^{X^1} (X^0)_{\text{выч}}\|_E \leqslant N/\gamma \varepsilon_2 \|X\|_E. \tag{2.4}$$

Итак, получена оценка ошибки приведения матрицы к ее канонической паре. Перейдем к оценке погрешностей, возникающих при выполнении остальных операций «арифметики вынесенных порядков».

1°. Умножение на скаляр. Для машинной реализации формального равенства $[Y^0, Y^1] = [\alpha^0, \alpha^1][X^0, X^1]$, где Y^0, X^0 — матрицы размерности $N \times N$; α^0 — вещественное число; Y^1, X^1, α^1 — целые числа, предлагается использовать последовательность действий

$$\tilde{Y} = \alpha^0 X^0; \quad Y^0 = \mathfrak{M}(\tilde{Y}); \quad Y^1 = \mathscr{P}(\tilde{Y}) + X^1 + \alpha^1.$$

Для учета погрешностей округления достаточно оценить погрешность вычисления $\alpha^0 X^0$. В силу (2.1) верны неравенства ($i, j = 1, 2, \ldots, N$)

$$|(\alpha^0 X^0_{ij})_{\text{выч}} - \alpha^0 X^0_{ij}| \leqslant \max\{\varepsilon_1 |\alpha^0 X^0_{ij}|, \varepsilon_2/\gamma^2\},$$

позволяющие гарантировать выполнение оценок

$$\|(\alpha^0 X^0)_{\text{выч}} - \alpha^0 X^0\|_E \leqslant \varepsilon_1 \|\alpha^0 X^0\|_E + N\varepsilon_2/\gamma^2 \leqslant \varepsilon_1 \|\alpha^0 X^0\|_E (1 + N\varepsilon_2/\varepsilon_1). \tag{2.5}$$

Из неравенств (2.4) — (2.5) следует, что

$$\|([Y^0, Y^1])_{\text{выч}} - [Y^0, Y^1]\|_E \leqslant \varepsilon_1 \|\alpha X\|_E (1 + (N + N^{3/2})\varepsilon_2/\varepsilon_1).$$

Таким образом, показано, что если матрица X и число α заданы своими каноническими парами, то каноническая пара матрицы αX определяется с точностью

$$\|(\alpha X)_{\text{выч}} - \alpha X\|_E \leqslant \varepsilon_1 \|\alpha X\|_E (1 + (N + N^{3/2})\varepsilon_2/\varepsilon_1). \tag{2.6}$$

2°. Сложение и вычитание. Для определения матричной пары $[V^0, V^1]$, задающей матрицу V из формального равенства $[V^0, V^1] = [X^0, X^1] \pm [Y^0, Y^1]$, где пары $[X^0, X^1]$, $[Y^0, Y^1]$ задают матрицы X, Y, предлагается использовать следующую последовательность действий:

$$p = \max\{X^1, Y^1\}; \quad \tilde{V} = \gamma^{X^1 - p} X^0 \pm \gamma^{Y^1 - p} Y^0;$$
$$V^0 = \mathfrak{M}(\tilde{V}); \quad V^1 = \mathscr{P}(\tilde{V}) + p. \tag{2.7}$$

При умножении машиннозаданного числа на неположительную степень числа γ (γ — основание системы счисления используемой ЭВМ), возникает погрешность округления лишь в случае выполнения неравенств

$$|\gamma^{X^1-pX^0_{ij}}| < \frac{1}{\gamma^2}\varepsilon_2; \quad |\gamma^{Y^1-pY^0_{ij}}| < \frac{1}{\gamma^2}\varepsilon_2. \tag{2.8}$$

В то же время из (2.1) следует, что $|(z_{ij})_{\text{маш}} - z_{ij}| \leqslant \max\{\varepsilon_2/\gamma^2; \varepsilon_1|z_{ij}|\}$, где

$$z_{ij} = (\gamma^{X^1-pX^0_{ij}})_{\text{маш}} \pm (\gamma^{Y^1-pY^0_{ij}})_{\text{маш}},$$

и, значит, для матриц $Z=\{z_{ij}\}$ и $Z_{\text{выч}}=\{(z_{ij})_{\text{маш}}\}$ верна оценка $\|Z - Z_{\text{выч}}\|_E \leqslant \varepsilon_1\|Z\|_E + N\varepsilon_2/\gamma^2$. Полученное неравенство вместе с (2.3) и (2.8) позволяет заключить, что

$$\|Z_{\text{выч}} - (\gamma^{X^1-pX^0} \pm \gamma^{Y^1-pY^0})\|_E \leqslant$$
$$\leqslant 3N/\gamma^2\varepsilon_2 + 2N/\gamma^2\varepsilon_2\varepsilon_1 + \varepsilon_1\|\gamma^{X^1-pX^0} \pm \gamma^{Y^1-pY^0}\|_E \leqslant$$
$$\leqslant 3N/\gamma\varepsilon_2\gamma^{-p}(\|X^0\|_E\gamma^{X^1} + \|Y^0\|_E\gamma^{Y^1}) + \varepsilon_1\|\gamma^{X^1-pX^0} \pm \gamma^{Y^1-pY^0}\|_E.$$

Тем самым показано, что если матрицы X и Y заданы своими каноническими парами, то верна оценка

$$\|X \pm Y - (X \pm Y)_{\text{выч}}\|_E \leqslant 2N/\gamma\varepsilon_2(\|X\|_E + \|Y\|_E) + \varepsilon_1\|X \pm Y\|_E. \tag{2.9}$$

Из (2.6) и (2.9) следует, что если матрицы X, Y и число α будут заданы своими каноническими парами, то каноническая пара матрицы $Y + \alpha X$ будет определена с точностью

$$\|Y + \alpha X - (Y + \alpha X)_{\text{выч}}\|_E \leqslant 2N/\gamma\varepsilon_2(\|\alpha X\|_E + \|Y\|_E) +$$
$$+ \varepsilon_1\|Y + \alpha X\|_E + \varepsilon_1\|\alpha X\|_E[1 + (N + N^{3/2})\varepsilon_2/\varepsilon_1] \leqslant$$
$$\leqslant 1{,}01\varepsilon_1\|\alpha X\|_E + 1{,}01\|Y + \alpha X\|_E,$$

если $(2N + 2N^{3/2} + 2N/\gamma)\varepsilon_2/\varepsilon_1 < 0{,}01$.

3°. Произведение матриц. Последовательность операций

$$\bar{V} = X^0Y^0; \quad V = \mathfrak{M}(\bar{V}); \quad V^1 = \mathscr{P}(\bar{V}) + X^1 + Y^1$$

позволяет вычислить каноническую пару $[V^0, V^1]$ матрицы $V = X \cdot Y$. Заметим, что погрешность вычисления произведения X^0Y^0 существенно зависит от того, как накапливаются скалярные произведения векторов — с использованием двойной точности или нет. В [8] (с. 115) приведены оценки

$$\|(X^0Y^0)_{\text{выч}} - X^0Y^0\|_E \leqslant N\varepsilon_1\|X^0\|_E \cdot \|Y^0\|_E; \tag{2.10}$$
$$\|([X^0]^2)_{\text{выч}} - [X^0]^2\|_E \leqslant N\varepsilon_1\|[X^0]^2\|_E, \tag{2.11}$$

выведенные в предположении обычного накопления скалярного произведения векторов. Иногда удобнее использовать оценку

$$\|(X^0Y^0)_{\text{выч}} - X^0Y^0\|_E \leqslant N^{3/2}\varepsilon_1\mu(X^0)\|X^0Y^0\|_E, \tag{2.12}$$

где $\mu(X^0) = \mu(X) = \|X\| \cdot \|X^{-1}\|$ — число обусловленности матрицы X. Неравенство (2.12) является простым следствием неравенства (2.10).

Предположим, что при вычислении X^0Y^0 скалярные произведения векторов накапливаются с удвоенной точностью. В этом случае из (2.2) следует, что при всех $i, j = 1, 2, \ldots, N$ верны неравенства

$$\left|\sum_{k=1}^{N}(X^0_{ik}Y^0_{kj})_{\text{выч}} - \sum_{k=1}^{N}X^0_{ik}Y^0_{kj}\right| \leqslant \varepsilon_1\left|\sum_{k=1}^{N}X^0_{ik}Y^0_{kj}\right| +$$
$$+ \frac{N\varepsilon_1^2}{\gamma}\left(\sum_{k=1}^{N}[X^0_{ik}]^2\right)^{1/2}\left(\sum_{k=1}^{N}[Y^0_{kj}]^2\right)^{1/2} + \frac{N}{\gamma^2}\varepsilon_2,$$

гарантирующие выполнение оценок

$$\|(X^0Y^0)_{\text{выч}} - X^0Y^0\|_E \leqslant \varepsilon_1(1 + N^2\varepsilon_2/\varepsilon_1 + N/\gamma\varepsilon_1)\|X^0\|_E \cdot \|Y^0\|_E;$$
$$\|(X^0Y^0)_{\text{выч}} - X^0Y^0\|_E \leqslant \varepsilon_1\|X^0Y^0\|_E(1 + [N^2\varepsilon_2/\varepsilon_1 + N\varepsilon_1/\gamma]\mu(X)); \quad (2.13)$$
$$\|[X^0]^2 - ([X^0]^2)_{\text{выч}}\|_E \leqslant \varepsilon_1\|[X^0]^2\|_E(1 + N^2\varepsilon_2/\varepsilon_1 + N\varepsilon_1/\gamma).$$

Неравенства (2.10) — (2.13) позволяют заключить, что верны оценки

$$\|(XY)_{\text{выч}} - XY\|_E \leqslant c_1\varepsilon_1\|XY\|_E;$$
$$\|(XY)_{\text{выч}} - XY\|_E \leqslant c_2\varepsilon_1\|X\|_E \cdot \|Y\|_E;$$
$$\|(X^2)_{\text{выч}} - X^2\|_E \leqslant c_3\varepsilon_1\|X^2\|_E.$$

Если выполнены ограничения

$$N^2\varepsilon_2/\varepsilon_1 + N\varepsilon_1/\gamma \leqslant 0{,}01; \quad \mu(X)(N^2\varepsilon_2/\varepsilon_1 + N\varepsilon_1/\gamma) \leqslant 0{,}01,$$

то выбор значений c_1, c_2, c_3 в выведенных оценках зависит от способа вычисления скалярного произведения векторов. Если для вычисления скалярного произведения используется накопление с двойной точностью, то $c_1 = c_2 = c_3 = 1{,}01$. В противном случае $c_1 = N^{3/2}\mu(X)$, $c_2 = c_3 = N$.

§ 3. УЧЕТ ПОГРЕШНОСТЕЙ ОКРУГЛЕНИЯ ПРИ ВЫЧИСЛЕНИИ ЭКСПОНЕНТЫ ОТ МАТРИЦЫ МАЛОЙ НОРМЫ

Пусть A — матрица малой нормы ($1/2 \leqslant \|A\| < 1$). Для вычисления e^A известно много различных алгоритмов (см., например, обзор [3]).

В этом параграфе выведены оценки погрешностей округления, возникающие при вычислении матриц

$$A_m = I + \frac{1}{1!}A + \frac{1}{2!}A^2 + \ldots + \frac{1}{m!}A^m,$$

приближающих e^A с точностью

$$\|e^A - A_m\| \leqslant \|A\|^{m+1}/(m+1)!e^{\|A\|}. \quad (3.1)$$

Аналогичные оценки получены в [9]. Наиболее употребимым способом вычисления матриц A_m является итерационный процесс: $A_0 = I$; $B_0 = I$; $m \geqslant 1$;

$$B_m = 1/mAB_{m-1}; \quad A_m = A_{m-1} + B_m, \quad (3.2)$$

где I — единичная матрица размера $N \times N$.

Для получения оптимального приближения матрицы e^A необходимо остановить процесс (3.2) после k_0 шагов, когда гарантированная погрешность вычисления A_{k_0} станет одного порядка с погрешностью приближения матрицей A_{k_0} экспоненты e^A. В конце параграфа указан алгоритм выбора k_0 в каждом конкретном случае.

Для учета погрешностей, возникающих при реализации процесса (3.2), достаточно предположить, что матрицы, полученные при этом, связаны соотношениями $\bar{A}_0 = \bar{B}_0 = I$; $m \geqslant 1$;

$$\bar{B}_m = 1/mA\bar{B}_{m-1} + \psi_m; \quad \bar{A}_m = \bar{A}_{m-1} + \bar{B}_m + \varphi_m, \quad (3.3)$$

в которых φ_m и ψ_m — квадратные матрицы размерности $N \times N$ — обозначают погрешности. Использование при расчете формул (3.3) «арифметики вынесенных порядков» позволяет гарантировать выполнение неравенств

$$\|\psi_m\| \leqslant d_1\varepsilon_1/m\|A\| \cdot \|\bar{B}_{m-1}\|; \quad (3.4)$$
$$\|\varphi_m\| \leqslant d_2\varepsilon_1(\|\bar{B}_m\| + \|\bar{A}_{m-1}\|), \quad (3.5)$$

где ε_1 — машинная константа; d_1, d_2 — некоторые постоянные, зависящие от N. В силу результатов § 2 в качестве d_1 и d_2 можем взять

$$d_2 = 1{,}01\sqrt{N}; \tag{3.6}$$

$$d_1 = \begin{cases} 1{,}01\sqrt{N} — \text{скалярное произведение векторов накапливается с двойной точностью;} \\ N^{3/2} — \text{в противном случае.} \end{cases} \tag{3.7}$$

Оценим разность $\widetilde{B}_m - 1/m!A^m$. Из (3.3) следует равенство

$$\widetilde{B}_m - 1/m!A^m = 1/mA[\widetilde{B}_{m-1} - 1/(m-1)!A^{m-1}] + \psi_m,$$

позволяющее вывести цепочку неравенств:

$$\|\widetilde{B}_m - 1/m!A^m\| \leqslant \|\widetilde{B}_{m-1} - 1/(m-1)!A^{m-1}\| \cdot \|A\|/m + \|\psi_m\| \leqslant$$
$$\leqslant \|\widetilde{B}_{m-1} - 1/(m-1)!A^{m-1}\| \cdot \|A\|/m + d_1\varepsilon_1\|\widetilde{B}_{m-1}\| \cdot \|A\|/m \leqslant$$
$$\leqslant (1 + d_1\varepsilon_1)\|A\|/m\|\widetilde{B}_{m-1} - 1/(m-1)!A^{m-1}\| +$$
$$+ d_1\varepsilon_1\|A\|/m \cdot \|1/(m-1)!A^{m-1}\| \leqslant d_1\varepsilon_1\|A\|^m/(m-1)!.$$

Опираясь на полученную оценку и используя неравенства (3.4) и (3.5), докажем по индукции, что если

$$\sum_{m=1}^{k} \varepsilon_1\|A\|(d_2m + d_1) \leqslant 0{,}01,$$

то
$$\|\widetilde{A}_k - A_k\| \leqslant 1{,}01\,(d_2k + d_1\|A\|)\,\varepsilon_1 e^{\|A\|}. \tag{3.8}$$

В самом деле, при $k = 1$ неравенство (3.8) справедливо в силу (3.2)—(3.5). Предположим, что оно верно при всех $k < j$, и докажем, что тогда (3.8) верно и при $k = j$. В силу сделанных предположений справедлива цепочка неравенств:

$$\|\widetilde{A}_j - A_j\| \leqslant \left\| \sum_{m=1}^{j} \widetilde{B}_m - 1/m!\,A^m + \varphi_m \right\| \leqslant$$

$$\leqslant \sum_{m=1}^{j} \{\|\widetilde{B}_m - 1/m!\,A^m\| + d_2\varepsilon_1\|\widetilde{B}_m\| + d_2\varepsilon_1\|A_{m-1}\|\} \leqslant$$

$$\leqslant \sum_{m=1}^{j} \{\|\widetilde{B}_m - 1/m!\,A^m\|(1+d_2\varepsilon_1) + d_2\varepsilon_1\|1/m!\,A^m\| +$$

$$+ d_2\varepsilon_1\|A_{m-1}\| + d_2\varepsilon_1\|\widetilde{A}_{m-1} - A_{m-1}\|\} \leqslant$$

$$\leqslant (1 + d_2\varepsilon_1)\,d_1\varepsilon_1\|A\|e^{\|A\|} + jd_2\varepsilon_1 e^{\|A\|} +$$

$$+ \sum_{m=1}^{j} d_2\varepsilon_1 e^{\|A\|} \cdot 1{,}01\,(d_2m + d_1) \leqslant 1{,}01\,(d_2j + d_1\|A\|)\,e^{\|A\|}\varepsilon_1.$$

Оценки (3.1) и (3.8) позволяют заключить, что наиболее подходящим выбором значения k_0 для лучшего приближения матричной экспоненты является минимальное среди всех целых чисел k, удовлетворяющих неравенству $1{,}01\varepsilon_1(d_2k_0 + d_1\|A\|) \cdot e^{\|A\|} \geqslant \|A\|^{k+1}/(k+1)!$, где d_1 и d_2 заданы (3.6), (3.7). Возникающая при этом погрешность оценивается неравенством

$$\|\widetilde{A}_{k_0} - e^A\| \leqslant 2{,}02\varepsilon_1(d_2k_0 + d_1\|A\|)\,e^{\|A\|} =$$
$$= 2{,}02\varepsilon_1(d_2k_0 + d_1\|A\|)\,e^{\|A\|(1+1/\varkappa)}e^{-\|A\|/\varkappa}\varepsilon_1 = \alpha_0\varepsilon_1 e^{-\|A\|/\varkappa},$$

где $\alpha_0 = 2{,}02\,(d_2k_0 + d_1\|A\|)\exp\{\|A\|(1 + 1/\varkappa)\}$.

§ 4. ОСНОВНЫЕ ТЕОРЕМЫ

На протяжении всего параграфа под матрицей A будем понимать асимптотически устойчивую матрицу размерности $N \times N$ с $\varkappa = \varkappa(A) < \infty$. Все утверждения будут касаться последовательности матриц размерности

$N \times N$: $B_0, B_1, B_2, \ldots, B_k, \ldots$, связанных при помощи квадратных матриц Φ_j $(j = 0, 1, 2, \ldots)$ рекуррентными соотношениями

$$B_0 = e^A + \Phi_0; \quad B_{j+1} = B_j^2 + \Phi_{j+1}. \tag{4.1}$$

На матрицы Φ_j накладываются условия

$$\|\Phi_0\| \leqslant r_0/(2\varkappa)\, e^{-\|A\|/\varkappa}; \quad \|\Phi_j\| \leqslant r_0/(2\varkappa^{3/2})\, \|B_{j-1}^2\|. \tag{4.2}$$

Прежде всего введем некоторые новые обозначения. Из рекуррентных соотношений (4.1) вытекает, что

$$B_{m+1} = (\ldots((e^A + \Phi_0)^2 + \Phi_1)^2 + \ldots + \Phi_m)^2 + \Phi_{m+1}.$$

Раскрывая скобки в полученном выражении и собирая в R_{m+1}^i все слагаемые, содержащие Φ_j в суммарной степени, равной i, можем записать

$$B_{m+1} = e^{2^{m+1}A} + R_{m+1}^1 + R_{m+2}^2 + \ldots + R_{m+1}^{2^{m+1}}. \tag{4.3}$$

Лемма 1. *Пусть при любых* $p, q > 0$ *выполнены для слагаемых в* (4.3) *оценки*: $j = 1, 2, 3, \ldots, 2^k$;

$$\|e^{pA} R_k^j e^{qA}\| \leqslant \rho^j e^{-(2^k+p+q)\|A\|/\varkappa}, \tag{4.4}$$

где ρ *удовлетворяет неравенствам*

$$1 > \rho > 0; \quad 2\rho/(1-\rho) + \rho^2/(1-\rho)^2 < \sqrt{\varkappa}. \tag{4.5}$$

Тогда справедливы оценки: $j = 1, 2, 3, \ldots, 2^{k+1}$;

$$\|e^{pA} R_{k+1}^j e^{qA}\| \leqslant (2\rho + r_0)^j e^{-(2^{k+1}+p+q)\|A\|/\varkappa}.$$

Приступая к доказательству леммы 1, заметим, что из (4.1) и (4.3) следует справедливость соотношений:

$$R_{k+1}^1 = e^{2^k A} R_k^1 + R_k^1 e^{2^k A} + \Phi_{k+1}; \tag{4.6}$$

$$R_{k+1}^j = e^{2^k A} R_k^j + R_k^j e^{2^k A} - \sum_{i=1}^{j-1} 2R_k^i R_k^{j-i} - \delta_{ij} (R_k^i)^2; \tag{4.7}$$

$$R_{k+1}^l = \sum_{i=l-2^k}^{2^k} 2R_k^i R_k^{l-i} - \delta_{l-i,i} (R_k^i)^2, \tag{4.8}$$

где δ_{ij} — символ Кронекера, $j = 2, 3, \ldots, 2^k$; $l = 2^k + 1, 2^k + 2, \ldots, 2^{k+1}$. В свою очередь, из (4.3) и (4.4) нетрудно вывести оценку

$$\|B_k - e^{2^k A}\| \leqslant \rho/(1-\rho)\, e^{-2^k \|A\|/\varkappa},$$

используя которую вместе с неравенствами (4.1) и (4.5), можем получить из

$$B_k^2 = e^{2^{k+1}A} - e^{2^k A}\left[B_k - e^{2^k A}\right] - \left[B_k - e^{2^k A}\right] e^{2^k A} + \left[B_k - e^{2^k A}\right]^2$$

следующую оценку:

$$\|B_k^2\| \leqslant \left\{\sqrt{\varkappa} + \frac{2\rho}{1-\rho} + \frac{\rho^2}{(1-\rho)^2}\right\} e^{-2^{k+1}\|A\|/\varkappa} \leqslant 2\sqrt{\varkappa}\, e^{-2^{k+1}\|A\|/\varkappa}.$$

Найденное неравенство позволяет из (4.2) получить оценку

$$\|\Phi_{k+1}\| \leqslant r_0/\varkappa e^{-2^{k+1}\|A\|/\varkappa}. \tag{4.9}$$

Выведем при помощи (4.4) и (4.9) из (4.6) цепочку неравенств

$$\|e^{pA} R_{k+1}^1 e^{qA}\| \leqslant \|e^{(2^k+p)A} R_k^1 e^{qA}\| + \|e^{pA} R_k^1 e^{(2^k+q)A}\| + \\ + \|e^{pA} \Phi_{k+1} e^{qA}\| \leqslant (2\rho + r_0)\, e^{-(2^{k+1}+p+q)\|A\|/\varkappa}. \tag{4.10}$$

Таким же образом из (4.7) имеем

$$\left\| e^{pA} R_{k+1}^{j} e^{qA} \right\| \leqslant 2\rho^{j} e^{-(2^{k+1}+p+q)\frac{\|A\|}{\varkappa}} + 2(j-1)\rho^{j} e^{-(2^{k+1}+p+q)\frac{\|A\|}{\varkappa}} =$$

$$= \frac{2j}{2^{j}} (2\rho)^{j} e^{-(2^{k+1}+p+q)\frac{\|A\|}{\varkappa}} \leqslant (2\rho + r_0)^{j} e^{-(2^{k+1}+p+q)\frac{\|A\|}{\varkappa}}. \qquad (4.11)$$

Аналогично выведенным неравенствам (4.10) и (4.11) нетрудно получить из (4.8) завершающую доказательство леммы 1 цепочку неравенств $l = 2^k + 1, 2^k + 2, \ldots, 2^{k+1}$;

$$\left\| e^{pA} R_{k+1}^{l} e^{qA} \right\| \leqslant 2l\rho^{l} e^{-(2^{k+1}+p+q)\|A\|/\varkappa} \leqslant (2\rho + r_0)^{l} e^{-(2^{k+1}+p+q)\|A\|/\varkappa}.$$

Лемма 2. *Пусть для матриц* (4.1) *и* (4.2) *при произвольных* $p, q > 0$ *верна оценка*

$$\left\| e^{pA} \left[B_k - e^{2^k A} \right] e^{qA} \right\| \leqslant \rho e^{-(2^k - Q + p + q)\|A\|/\varkappa},$$

где Q *удовлетворяет равенству*

$$(2\exp\{-Q\|A\|/\varkappa\} + \rho)(1 + r_0/\sqrt{\varkappa}) = 1. \qquad (4.12)$$

Тогда при всех $m > 0$ *справедлива оценка*

$$\left\| e^{pA} \left[B_{k+m} - e^{2^{k+m} A} \right] e^{qA} \right\| \leqslant \rho e^{-(2^{k+m} - 2^m Q + p + q)\|A\|/\varkappa}.$$

Для доказательства леммы используем неравенство, получаемое как следствие из (4.2):

$$\|\Phi_{k+1}\| \leqslant r_0/(2\varkappa^{3/2}) \left\| B_k^2 \right\| \leqslant r_0/(2\varkappa^{3/2}) \left\| B_k^2 - e^{2^{k+1} A} + e^{2^{k+1} A} \right\| \leqslant$$

$$\leqslant r_0/(2\varkappa^{3/2}) \left\| B_k^2 - e^{2^{k+1} A} \right\| + r_0/(2\varkappa)\, e^{-2^{k+1}\|A\|/\varkappa},$$

а также представление

$$Z(k) = B_k^2 - e^{2^{k+1} A} = \left[B_k - e^{2^k A} \right]^2 + e^{2^k A} \left[B_k - e^{2^k A} \right] + \left[B_k - e^{2^k A} \right] e^{2^k A}.$$

Они позволяют вывести из (4.1) неравенство

$$\left\| e^{pA} \left[B_{k+1} - e^{2^{k+1} A} \right] e^{qA} \right\| \leqslant \left\| e^{pA} Z(k)\, e^{qA} \right\| (1 + r_0/\sqrt{\varkappa}). \qquad (4.13)$$

Наконец, по условию леммы верно неравенство

$$\left\| e^{pA} Z(k)\, e^{qA} \right\| \leqslant (2\rho e^{-Q\|A\|/\varkappa} + \rho^2)\, e^{-(2^{k+1} - 2Q + p + q)\|A\|/\varkappa},$$

которое вместе с (4.12) и (4.13) гарантирует справедливость оценки

$$\left\| e^{pA} \left[B_{k+1} - e^{2^{k+1} A} \right] e^{qA} \right\| \leqslant$$

$$\leqslant (2\rho e^{-Q\|A\|/\varkappa} + \rho^2)(1 + r_0/\sqrt{\varkappa})\, e^{-(2^{k+1} - 2Q + p + q)\|A\|/\varkappa} =$$

$$= \rho e^{-(2^{k+1} - 2Q + p + q)\|A\|/\varkappa}.$$

Неравенства леммы 2, выполненные при $m = 1$, аналогично выводятся и при $m > 1$. Для этого достаточно заметить, что $\exp\{-Q\|A\|/\varkappa\} > \exp\{-2^m Q\|A\|/\varkappa\}$ при всех $m > 1$ в силу выбора Q.

Основываясь на леммах 1 и 2, докажем теорему 1, но для удобства ее формулировки введем некоторые новые обозначения. Пусть на интервале $(0, -\log_2 r_0)$ задана функция $\rho(t) = (2^t - 1) r_0/(1 - (2^t - 1) r_0)$ и t_0 — минимальный положительный корень уравнения

$$2\rho(t_0)/(1 - \rho(t_0)) + [\rho(t_0)]^2/[1 - \rho(t_0)]^2 = \sqrt{\varkappa}.$$

Тогда, взяв k_0, где $k_0 = [t_0]$ — целая часть числа t_0, определим Q_0 как корень уравнения

$$(2e^{-Q_0\|A\|/\varkappa} + \rho(k_0))(1 + r_0/\sqrt{\varkappa}) = 1. \qquad (4.14)$$

Применяя введенные числа и функцию $\rho(t)$, определим

$$\delta(t)=\begin{cases}\rho(t) & \text{при } t\leqslant k_0;\\ \rho(k_0)\exp\{2^{t-k_0}Q_0\|A\|/\varkappa\} & \text{при } t>k_0.\end{cases}$$

В дальнейшем функцию $\delta(t)$ будем использовать как мажорантную при оценке точности вычисления степеней матричной экспоненты.

Теорема 1. *Пусть* $\varkappa=\varkappa(A)<\infty$, *где* A — *некоторая матрица, и пусть квадратные матрицы* Φ_j $(j=0,\ 1,\ 2,\ \ldots,\ k+1)$ *той же размерности* $N\times N$ *удовлетворяют неравенствам*

$$\|\Phi_j\|\leqslant r_0/(2\varkappa^{3/2})\|B_{j-1}^2\|;\quad \|\Phi_0\|\leqslant r_0/(2\varkappa)\,e^{-\|A\|/\varkappa}$$

с достаточно малой постоянной r_0. *Если далее предполагать, что матрицы* $B_0,\ B_1,\ \ldots,\ B_{k+1}$ *определены рекуррентными соотношениями* $B_0=e^A+\Phi_0$; $B_{m+1}=B_m^2+\Phi_{m+1}$, *то имеет место следующая оценка:*

$$\|e^{2^kA}-B_k\|\leqslant\delta(k)\,e^{-2^k\|A\|/\varkappa}. \tag{4.15}$$

Переходя к доказательству теоремы, прежде всего заметим, что для $R_0^1=B_0-e^A$ при любых $p,\ q\geqslant 0$ верна оценка

$$\|e^{pA}R_0^1e^{qA}\|=\|e^{pA}\Phi_0e^{qA}\|\leqslant$$

$$\leqslant\varkappa e^{-(p+q)\|A\|/\varkappa}r_0/(2\varkappa)\,e^{-\|A\|/\varkappa}=r_0/2e^{-(2^0+p+q)\|A\|/\varkappa}.$$

Тем самым показана справедливость неравенства (4.15) при $k=0$. Это позволяет сделать индуктивное предположение, что для некоторого k $(k<k_0)$ выполнены оценки

$$\|e^{pA}R_k^je^{qA}\|\leqslant[(2^{k+1}-1)\,r_0]^je^{-(2^k+p+q)\|A\|/\varkappa}.$$

Тогда в силу неравенств леммы 1 (где в качестве ρ взято $(2^{k+1}-1)r_0$, неравенство (4.5) для такого ρ выполнено ввиду выбора k_0) верны также неравенства

$$\|e^{pA}R_{k+1}^je^{qA}\|\leqslant[(2^{k+2}-1)\,r_0]^je^{-(2^{k+1}+p+q)\|A\|/\varkappa}.$$

Итак, мы доказали, что при всех $k\leqslant k_0$ верна оценка

$$\|e^{2^kA}-B_k\|\leqslant(2^{k+1}-1)\,r_0/[1-(2^{k+1}-1)\,r_0]\,e^{-2^k\|A\|/\varkappa},$$

при этом для всех $p,\ q>0$ справедливы неравенства

$$\|e^{pA}[e^{2^kA}-B_k]\,e^{qA}\|\leqslant\frac{(2^{k+1}-1)\,r_0}{1-(2^{k+1}-1)\,r_0}e^{-(2^k+p+q)\|A\|/\varkappa}.$$

Для завершения доказательства осталось воспользоваться леммой 2, взяв $\rho=(2^{k_0+1}-1)\,r_0/[1-(2^{k_0+1}-1)\,r_0]$ и $Q=Q_0$, где Q_0 определено как решение уравнения (4.14). Теорема 1 доказана.

§ 5. ОЦЕНКА ПОГРЕШНОСТЕЙ ОКРУГЛЕНИЯ ПРИ ВЫЧИСЛЕНИИ ЭКСПОНЕНТЫ ОТ АСИМПТОТИЧЕСКИ УСТОЙЧИВОЙ МАТРИЦЫ

На основании теоремы 1 § 4 выведена оценка погрешностей округления, возникающих на этапе 5^0 схемы § 1. Эта оценка позволила сформулировать правило вычисления δ_0 — гарантированной погрешности вычисления e^A.

Предположим, что возникающие в машинной реализации этапа 5^0 матрицы связаны между собою соотношениями

$$\widetilde{B}_m=\widetilde{B}_{m-1}^2+\Phi_m, \tag{5.1}$$

в которых квадратные матрицы Φ_m $(m=1, 2, \ldots, k_1)$ обозначают погрешности, допущенные при вычислении. Величина погрешностей зависит от конкретно выбранного способа расчета этапа 5^0. Из сказанного в § 2 следует, что если использовать «арифметику вынесенных порядков», то для матриц погрешностей выполнены оценки

$$\|\Phi_m\| \leqslant \alpha\varepsilon_1 \|\widetilde{B}_{m-1}^2\|, \tag{5.2}$$

где

$$\alpha = \sqrt{N}\widetilde{\alpha};\ \widetilde{\alpha} = \begin{cases} 1{,}01 — \text{если скалярное произведение векторов} \\ \quad \text{накапливается с двойной точностью;} \\ N — \text{в противном случае.} \end{cases} \tag{5.3}$$

Напомним, что в § 3 для погрешностей, возникающих при расчете этапа 4^0, выведена оценка

$$\|\widetilde{B}_0 - e^{A_1}\| \leqslant \alpha_0\varepsilon_1 e^{-\|A_1\|/\varkappa}, \tag{5.4}$$

где

$$\alpha_0 = 2{,}02\left(2k_0\sqrt{N} + d_1\|A_1\|\right)e^{\|A_1\|(1+1/\varkappa)}, \tag{5.5}$$

а k_0 определяется как минимальное целое k, при котором справедлива оценка $1{,}01\left(\sqrt{N}k + d_1\|A_1\|\right)\varepsilon_1 \geqslant \|A_1\|^{k+1}/(k+1)!$, где d_1 задана формулой (3.7). Так как $\Phi_0 = \widetilde{B}_0 - e^{A_1}$, то, используя неравенства (5.1)—(5.5) и взяв $r_0 = 2\varkappa(A)\varepsilon_1 \max\{\alpha_0, \sqrt{\varkappa(A)}\alpha\}$, можем воспользоваться теоремой 1 для оценки близости матриц $e^A = e^{2^{k_1}A_1}$ и $\widetilde{B}_{k_1}$. В этом случае верна оценка

$$\left\|\widetilde{B}_{k_1} - e^{2^{k_1}A_1}\right\| \leqslant \delta(k_1)\, e^{-2^{k_1}\|A_1\|/\varkappa} = \delta(k_1)\, e^{-\|A\|/\varkappa}, \tag{5.6}$$

где

$$\delta(k_1) = \begin{cases} \rho(k_1) & \text{при } k_1 \leqslant k_2; \\ \rho(k_2)\exp\{2^{k_1-k_2}Q_0\|A\|/\varkappa\} & \text{при } k_1 > k_2. \end{cases} \tag{5.7}$$

При вычислении $\delta(k_1)$ участвует функция $\rho(t)$, заданная на интервале $(0, -\log_2 r_0)$:

$$\rho(t) = (2^t - 1)r_0/[1 - (2^t - 1)r_0]. \tag{5.8}$$

Целое число k_2 находится как целая часть минимального положительного корня t_0 $(k_2 = [t_0])$ уравнения

$$2\rho(t_0)/(1-\rho(t_0)) + [\rho(t_0)]^2/[1-\rho(t_0)]^2 = \sqrt{\varkappa}. \tag{5.9}$$

Очевидно, что при этом $1 - \rho(t_0) > 0$. Аналогично Q_0 определяется как корень уравнения

$$[2\exp\{-Q_0\|A_1\|/\varkappa\} + \rho(k_2)](1 + r_0/\sqrt{\varkappa}) = 1. \tag{5.10}$$

Неравенства и уравнения (5.6)—(5.10) позволяют найти интересующее нас значение $\delta_0 = \delta(k_1)\exp\{-\|A\|/\varkappa\}$.

§ 6. ОПИСАНИЕ ОБЩЕГО АЛГОРИТМА ВЫЧИСЛЕНИЯ ЭКСПОНЕНТЫ ОТ АСИМПТОТИЧЕСКИ УСТОЙЧИВОЙ МАТРИЦЫ С ОЦЕНКОЙ ТОЧНОСТИ

Опишем общую схему вычисления экспоненты от асимптотически устойчивой матрицы A, все этапы которой подробно рассматривались в предыдущих параграфах. При построении алгоритма особое внимание будем уделять арифметике матричных процессов. При этом используем «арифметику вынесенных порядков», формальное описание которой позволяют сделать операторы $\mathfrak{M}$ и $\mathscr{P}$ (см. § 2). Величины некоторых постоянных зависят от способа накопления скалярных произведений век-

торов. Ниже перечислены основные этапы алгоритма вычисления e^A, выполняемые последовательно один за другим.

1°. Входные данные. В машину вводятся следующие величины: N — целое число; $A=\{A_{ij}\}$ — исходная матрица размерности $N\times N$; $\varkappa$ — вещественное число — параметр качества устойчивости A ($\varkappa<\infty$); γ, ε_1 — машинные постоянные (см. § 2).

2°. Находится вещественное число σ_N — максимальное сингулярное число матрицы A ($\sigma_N=\|A\|$). Алгоритм его вычисления с оценкой гарантированной точности подробно описан в [7].

3°. Определяется целое число k_1 как целая часть числа $z=\log_2(\sigma_N)$ ($k_1=[z]+1$). Полагаем $\sigma=2^{-k_1}\sigma_N$.

4°. Приводится алгоритм получения δ_0 — числового значения гарантированной точности определения матрицы e^A.

4°.1. Вычисляется вещественное число r_0. Для этого отыскивается целое число k_2 — минимальное среди целых k, удовлетворяющих неравенству $\varepsilon_1(2{,}02k\sqrt{N}+2{,}02d_1)\geqslant 1/(k+1)!$, где d_1 и α удовлетворяют соответственно (3.7) и (5.3), и

$$\alpha_0=2{,}02(k_2\sqrt{N}+d_1)e^{1+1/\varkappa};\quad r_0=2\varkappa\varepsilon_1\max\{\alpha_0,\alpha\sqrt{\varkappa}\}.$$

4°.2. Вычисляются числа

$$x=1-1/(\sqrt{\varkappa}+1)^{1/2};\quad t_0=\log_2\{[x/(x+1)+r_0]/r_0\}.$$

Нетрудно проверить, что полученное число t_0 является минимальным положительным корнем системы уравнений

$$2\rho(t_0)(1-\rho(t_0))+[\rho(t_0)]^2/[1-\rho(t_0)]^2=\sqrt{\varkappa};$$
$$\rho(t_0)=\left(2^{t_0}-1\right)r_0\Big/\left[1-\left(2^{t_0}-1\right)r_0\right].$$

4°.3. Возьмем в качестве целого числа k_0 целую часть вещественного числа t_0 ($k_0=[t_0]$).

4°.4. Найдем вещественное число Q_0 по формулам

$$x_1=(1+r_0/\sqrt{\varkappa})^{-1}-\left(2^{k_0}-1\right)r_0\Big/\left[1-\left(2^{k_0}-1\right)r_0\right];$$
$$Q_0=-\ln(x_1/2)\,2^{k_1}\varkappa/\sigma,$$

где k_1 и σ получены на предыдущих этапах.

4°.5. Вычисляется число δ:

$$\delta=\begin{cases}\left(2^{k_1}-1\right)r_0\Big/\left[1-\left(2^{k_1}-1\right)r_0\right] & \text{при } k_1\leqslant k_0;\\ \left(2^{k_0}-1\right)r_0\Big/\left[1-\left(2^{k_0}-1\right)r_0\right]\exp\left\{2^{k_1-k_0}Q_0\|A\|/\varkappa\right\} & \text{при } k_1>k_0.\end{cases}$$

4°.6. Находится значение δ_0 — гарантированная погрешность вычисления матричной экспоненты: $\delta_0=\delta\exp\{-\sigma/\varkappa\}$.

5°. Нормируется матрица A: $A_1=2^{-k_1}A$.

6°. Опишем алгоритм вычисления матрицы

$$B_0=I+\frac{1}{1!}A_1+\frac{1}{2!}A_1^2+\ldots+\frac{1}{k_2!}A_1^{k_2},$$

где k_2 определено в п. 4°.1.

6°.1. Определение матрицы B_0 основано на итерационном процессе, стандартный шаг которого описан в п. 6°.2. Для начала процесса задаются канонические матричные пары: $[U_1^0, U_1^1]=[I+A_1, 0]$; $[V_1^0, V_1^1]=[A_1, 0]$. Напомним, что означают эти формальные равенства:

$$(U_1^0)_{ij}=\begin{cases}(A_1)_{ij}, & \text{если } i\neq j;\\ 1+(A_1)_{ii}, & \text{если } i=j;\end{cases}$$
$$(V_1^0)_{ij}=(A_1)_{ij};\quad U_1^1=V_1^1=0.$$

6°.2. Стандартный k-й шаг ($k=2, 3, \ldots, k_2$) вычисления матрицы $[U^0_{k_2}, U^1_{k_2}]$. Пусть канонические матричные пары $[U^0_{k-1}, U^1_{k-1}]$, $[V^0_{k-1}, V^1_{k-1}]$ получены на предыдущем шаге итерационного процесса. Стандартный шаг состоит в переходе к каноническим матричным парам $[U^0_k, U^1_k]$, $[V^0_k, V^1_k]$.

6°.2.1. Вычисляется матрица V_k:

$$(\widetilde{V}_k)_{ij} = \sum_{l=1}^{N} (A_1)_{il}(V^0_k)_{lj}/k; \quad i, j = 1, 2, \ldots, N;$$

$$V^0_k = \mathfrak{M}(\widetilde{V}_k); \quad V^1_k = \mathscr{P}(\widetilde{V}_k) + V^1_{k-1}.$$

6°2.2. Определяется матрица U_k (ищется линейная комбинация канонических матричных пар). Полагаем

$$q = \max\{U^1_{k-1}, V^1_k\}; \quad i, j = 1, 2, \ldots, N;$$

$$(\widetilde{U}_k)_{ij} = \gamma^{U^1_{k-1}-q}(U^0_{k-1})_{ij} + \gamma^{V^1_k-q}(V^0_k)_{ij};$$

$$U^0_k = \mathfrak{M}(\widetilde{U}_k); \quad U^1_k = q + \mathscr{P}(\widetilde{U}_k).$$

После завершения описанного процесса получим каноническую матричную пару $[B^0_0, B^1_0] = [U^0_{k_2}, U^1_{k_2}]$, определяющую искомую матрицу B_0.

7°. Входная информация совпадает с выходной информацией пункта 6°, а также служит в качестве начальных значений для итерационного процесса вычисления e^A. Стандартный шаг (k-й шаг, $k=1, 2, \ldots, k_1$) процесса описан здесь. Пусть имеется полученная в предыдущих вычислениях каноническая матричная пара $[B^0_{k-1}, B^1_{k-1}]$. Стандартный шаг состоит в переходе к канонической паре $[B^0_k, B^1_k]$.

7°.1. Вычисляется матрица B_k:

$$(\widetilde{B}_k)_{ij} = \sum_{l=1}^{N} (B^0_{k-1})_{il}(B^0_{k-1})_{lj}, \quad i, j = 1, 2, \ldots, N;$$

$$B^0_k = \mathfrak{M}(\widetilde{B}_k); \quad B^1_k = 2B^1_{k-1} + \mathscr{P}(\widetilde{B}_k).$$

Таким образом, искомое приближение матричной экспоненты e^A получено в виде канонической матричной пары $[B^0_{k_1}, B^1_{k_1}]$.

7°.2. Каноническую матричную пару приводим к нормальному виду ($i, j = 1, 2, \ldots, N$):

$$B_{ij} = \gamma^{B^1_{k_1}}(B^0_{k_1})_{ij}.$$

8°. Выходная информация. В результате работы алгоритма выдаются следующие данные: 1) δ_0 — гарантированная точность полученного приближения матричной экспоненты e^A; 2) $B = \{B_{ij}\}$, $i, j = 1, 2, \ldots, N$, — матрица полученного приближения матричной экспоненты ($\|B - e^A\| < \delta_0$).

§ 7. ПРИМЕР ИСПОЛЬЗОВАНИЯ

Значения входных параметров: $N = 15$; $\gamma = 16$; $\varepsilon_1 = 0.2_{10}-14$. В качестве входной матрицы A использовано двупараметрическое семейство матриц $A(\alpha, \beta)$, описанное во введении. Известны значения параметра качества устойчивости $\varkappa_i = \varkappa(A(16, \beta_i))$ ($i = 1, 2, 3, 4$), $\varkappa_1 = 0.8_{10}7$, $\varkappa_2 = 3.8_{10}6$, $\varkappa_3 = 1.2_{10}6$, $\varkappa_4 = 4_{10}5$, где $\beta_1 = 107.2$; $\beta_2 = 97.6$; $\beta_3 = 84.8$; $\beta_4 = 75.2$. В этом случае, если обозначить через δ_i ($i = 1, 2, 3, 4$) гарантированные погрешности вычисления матричных экспонент от матриц $A(16, \beta_i)$, то $\delta_1 = 4.4_{10}-2$, $\delta_2 = 1.2_{10}-2$, $\delta_3 = 0.2_{10}-2$, $\delta_4 = 4_{10}-4$. Отме-

тим, что выбранные в качестве характеристик ЭВМ числа γ и ε_i соответствуют характеристикам машины ЕС-1050, если для представления вещественных переменных используются двойные слова. Оценки δ_i получены при условии, что в алгоритме скалярные произведения векторов накапливаются с обычной точностью.

ЛИТЕРАТУРА

1. **Булгаков А. Я.** Эффективно вычисляемый параметр качества устойчивости систем линейных дифференциальных уравнений с постоянными коэффициентами.— Сиб. мат. журн., 1980, т. XXI, № 3, с. 32—41.
2. **Ward R. C.** Numerical computation of the matrix exponential with accuracy estimate.— SIAM J. Numerical Anal., 1977, v. 14, N 4, p. 600—610.
3. **Moler C., Van Loan C.** Nineteen dubious ways to compute the exponential of a matrix.— SIAM Review, 1978, v. 20, N 4, p. 801—836.
4. **Повзнер А. Я., Павлов Б. В.** Об одном методе численного интегрирования систем обыкновенных дифференциальных уравнений.— Журн. вычисл. математики и мат. физики, 1973, т. 13, № 4, с. 256—259.
5. **Godounov S. K., Boulgakov A. J.** Difficultés de calcul dans le probleme de Hurwitz et méthodes pour les surmonter.— In: Analysis and optimization of Systems, Versailles, 1982.— Proceedings (Lecture Notes in Control and Information Sciences, 44). Springer Verlag, 1982, p. 843—851.
6. **Булгаков А. Я., Годунов С. К.** Численное определение одного из критериев качества устойчивости систем линейных дифференциальных уравнений с постоянными коэффициентами.— Новосибирск, 1981.— 58 с. (Препринт/АН СССР, Сиб. отд-ние, ИМ).
7. **Годунов С. К.** Решение систем линейных уравнений.— Новосибирск: Наука. Сиб. отд-ние, 1980.— 177 с.
8. **Уилкинсон Дж. Х.** Алгебраическая проблема собственных значений.— М.: Наука, 1970.— 564 с.
9. **Levis A. H.** Some Computational Aspects of the Matrix Exponential.— IEEE Trans. Automatic Control, 1969, v. AC-14, N 4, p. 410—411.

РАСЧЕТ ПОЛОЖИТЕЛЬНО ОПРЕДЕЛЕННЫХ РЕШЕНИЙ УРАВНЕНИЯ ЛЯПУНОВА

А. Я. БУЛГАКОВ, С. К. ГОДУНОВ

ВВЕДЕНИЕ

В большинстве приложений представляют интерес лишь положительно определенные решения H матричного уравнения Ляпунова

$$A^*H + HA + I = 0, \tag{1}$$

где I — единичная матрица размерности $N \times N$. Положительная определенность имеет место только тогда, когда A гурвицева, т. е. если нулевое решение системы обыкновенных уравнений

$$\dot{x} = Ax \tag{2}$$

асимптотически устойчиво.

В работе предлагается детальный алгоритм расчета H, который сопровождается анализом гурвицевости A. Если оказалось, что A не гурвицева, то процесс завершается без указания H. В противном случае указывается положительно определенное, приближенное с машинной точностью решение уравнения (1).

Численный анализ гурвицевости использует числовую характеристику устойчивости

$$\varkappa(A) = 2\|A\| \max_{\dot{x}=Ax} \left\{ \int_0^\infty \|x(t)\|^2 \, dt / \|x(0)\|^2 \right\}, \tag{3}$$