ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 128 за 2018 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

Бочев М.А.

Некоторые вопросы матричного анализа методов интегрирования по времени

Рекомендуемая форма библиографической ссылки: Бочев М.А. Некоторые вопросы матричного анализа методов интегрирования по времени // Препринты ИПМ им. М.В.Келдыша. 2018. № 128. 45 с. doi:10.20948/prepr-2018-128
URL: http://library.keldysh.ru/preprint.asp?id=2018-128

ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В. КЕЛДЫША
Российской академии наук

М.А. Бочев

Некоторые вопросы матричного анализа методов интегрирования по времени

Москва — 2018

Михаил Александрович Бочев

Некоторые вопросы матричного анализа методов интегрирования по времени. Препринт Института прикладной математики им. М.В. Келдыша РАН, Москва, 2018.

Данный препринт содержит конспекты лекций, прочитанных в 2016 г. на Римско-Московской школе по матричным методам и прикладной линейной алгебре. Лекции посвящены некоторым задачам матричного анализа, возникающим при разработке и анализе схем интегрирования по времени систем обыкновенных дифференциальных уравнений и дифференциальных уравнений в частных производных. Материал лекций включает некоторые аспекты конечно-разностных пространственных аппроксимаций уравнений конвекции–диффузии (в контексте метода прямых), устойчивости систем обыкновенных дифференциальных уравнений, логарифмической матричной нормы и её применения, явно-неявных схем, методов расщепления, методов Розенброка в сочетании с приближёнными разложениями якобиана и схем с матричной экспонентой на основе подпространств Крылова. Препринт предназначен для аспирантов и студентов, а также для научных работников для ознакомления с указанной тематикой.

Ключевые слова: метод прямых, уравнение конвекции–диффузии, устойчивость дифференциальных уравнений и разностных схем, матричная экспонента, логарифмическая матричная норма, методы Розенброка, подпространства Крылова.

Mikhail A. Botchev

Some topics in matrix analysis for time integration methods. Preprint of Keldysh Institute of Applied Mathematics RAS, Moscow, 2018.

This report contains lecture notes used for the 2016 edition of the Rome-Moscow school of Matrix Methods and Applied Linear Algebra, held in Moscow and Rome (respectively, in August and September 2016). The notes deal with some matrix analysis problems which arise in construction and analysis of time integration methods for solving large systems of ordinary and partial differential equations (ODEs and PDEs). The material treated includes some aspects of finite-difference approximation of convection–diffusion operators (used, following the framework of the methods of lines, to reduce time-dependent convection–diffusion problems to ODE systems), stability of the ODE systems, the logarithmic matrix norm, stability of the implicit–explicit θ-method, splitting methods, Rosenbrock methods with approximate matrix factorizations and Krylov subspace exponential time integration.

Key words: method of lines, convection–diffusion equation, stability of differential equations and difference schemes, matrix exponent, logarithmic matrix norm, Rosenbrock methods, Krylov subspace.

1 Некоторые факты из матричного анализа

В этом разделе мы перечисляем некоторые определения и результаты (как правило, без доказательств), использованные в данных конспектах. Значками $\Diamond$ и $\square$ отмечены конец упражнения и конец доказательства, соответственно.

Под вектором $x \in \mathbb{C}^n$ будем понимать вектор-столбец. Следовательно, обычное скалярное произведение в $\mathbb{C}^n$ можно определить как

$$(x, y) = y^* x, \quad x, y \in \mathbb{C}^n,$$

где y^* — вектор, комплексно-сопряжённый y. Пусть $A \in \mathbb{R}^{n\times n}$. Множество всех собственных чисел матрицы A называется её спектром. Под $\rho(A)$ будем понимать спектральный радиус A:

$$\rho(A) = \max\{|\lambda| \mid \lambda \in \text{спектр } A\}.$$

Определим следующие векторные нормы , называемые 2-й, 1-й и максимальной нормами, соответственно:

$$\|x\|_2 = \sqrt{\sum_{k=1}^{n} |x_k|^2}, \qquad \|x\|_1 = \sum_{k=1}^{n} |x_k|, \qquad \|x\|_\infty = \max_{1\leqslant k\leqslant n} |x_k|. \tag{1.1}$$

Тогда матричные нормы $\|A\| = \max_{x\neq 0}(\|Ax\|/\|x\|)$, подчинённые этим векторным нормам, суть

$$\|A\|_2 = \sqrt{\rho(A^*A)}, \quad \|A\|_1 = \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{ij}|, \quad \|A\|_\infty = \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |a_{ij}|. \tag{1.2}$$

Заметим, что $\|A\|_2$ не может быть вычислена по явной формуле. Вычисление 2-й матричной нормы может оказаться дорогой операцией. Таким образом, если требуется величина какой-либо нормы матрицы, то вычисления 2-й нормы стоит избегать[1].

Для любых матричных норм $\|\cdot\|_\alpha$ и $\|\cdot\|_\beta$ можно найти неулучшаемую константу $C_{\alpha,\beta}$, такую что для любой $A \in \mathbb{C}^{n\times n}$ выполняется $\|A\|_\alpha \leqslant C_{\alpha,\beta}\|A\|_\beta$. Вот эти константы [1, Section 5.6]:

$$\begin{aligned} &C_{1,2} = \sqrt{n}, \quad C_{1,\infty} = n, \\ &C_{2,1} = \sqrt{n}, \quad C_{2,\infty} = \sqrt{n}, \\ &C_{\infty,1} = n, \quad C_{\infty,2} = \sqrt{n}. \end{aligned} \tag{1.3}$$

Квадратная матрица P называется матрицей перестановки, если её столбцы можно переставить так, что получится единичная матрица.

[1] Отметим, что команда `norm(A)` в Матлабе или Октаве вычисляет $\|A\|_2$. Нормы $\|A\|_1$ и $\|A\|_\infty$ считаются командами `norm(A,1)` и `norm(A,'inf')` соответственно.

Квадратная матрица A называется *разложимой*, если существует такая матрица перестановки $\hat{P}$, что

$$\hat{P}A\hat{P}^T = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix},$$

где A_{11} и A_{22} — квадратные матрицы. В противном случае A называется *неразложимой* [2, 3]. Неразложимые матрицы наиболее удобно характеризовать по их направленным графам. Направленный граф матрицы $A \in \mathbb{R}^{n\times n}$ — это граф из n вершин, где имеется стрелка от вершины i к вершине j коль скоро $a_{ij} \neq 0$. Матрица неразложима тогда и только тогда, когда её граф сильно связан, то есть если в графе существует путь по стрелкам между любыми двумя вершинами i и j [4, 2].

Теорема 1.1 *(Теорема Перрона–Фробениуса) Пусть $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ — поэлементно неотрицательная неразложимая матрица ($a_{ij} \geqslant 0$). Тогда*
(1) A имеет положительное собственное значение λ, которое равняется спектральному радиусу матрицы A: $\lambda = \rho(A)$;
(2) собственный вектор $x = (x_i)$, соответствующий λ, может быть выбран поэлементно строго положительным: $x_i > 0$, $i = 1, \ldots, N$;
(3) собственное значение λ является простым.
Если матрица A только поэлементо неотрицательная, то $\rho(A)$ — собственное значение A, и соответствующий собственный вектор может быть выбран поэлементно неотрицательным [1, Theorem 8.3.1].

Матрица $A \in \mathbb{R}^{n\times n}$ называется матрицей со слабым диагональным преобладанием, если

$$|a_{ii}| \geqslant \sum_{j=1, j\neq i}^{n} |a_{ij}|, \qquad i = 1, \ldots, n. \tag{1.4}$$

Если эти неравенства строгие для всех i, то матрица называется матрицей со строгим диагональным преобладанием. Матрица называется неразложимой с диагональным преобладанием, если матрица неразложимая, со слабым диагональным преобладанием и по крайней мере для одного i неравенство диагонального преобладания (1.4) выполняется строго [2, 3].

Теорема 1.2 *[4, 2] Если матрица является матрицей со строгим диагональным преобладанием или неразложимой с диагональным преобладанием, то она невырождена и имеет ненулевые диагональные элементы.*

Упражнение 1.1 Докажите Теорему 1.2 для случая строгого диагонального преобладания. ◊

Матрица A называется M-матрицей [5], если $A = sI - B$, B — поэлементно неотрицательная матрица ($B \geqslant 0$) и $s > \rho(B)$. Если $s = \rho(B)$, то A вырождена и называется вырожденной M-матрицей [6, 7].

Следующая теормема помогает проверить, является ли данная матрица M-матрицей.

Теорема 1.3 *Пусть $A \in \mathbb{R}^{n\times n}$ — матрица со слабым диагональным преобладанием и пусть*

$$a_{ii} \geqslant 0,\ i = 1, \ldots, n, \qquad a_{ij} \leqslant 0,\ i \neq j.$$

Тогда собственные значения матрицы A имеют неотрицательную вещественную часть и A — возможно вырожденная M-матрица.

Доказательство [6, Section 2.5] Определим

$$s = \max_i a_{ii}, \qquad B = sI - A.$$

Заметим, что B — поэлементно неотрицательная, и легко проверить, что

$$\|B\|_\infty \leqslant s. \tag{1.5}$$

Таким образом, из неравенства $\rho(B) \leqslant \|B\|_\infty$ следует $\rho(B) \leqslant s$, а значит, A — по определению возможно вырожденная M-матрица.
Далее, все собственные значения $\lambda(A)$ матрицы A принадлежат множеству $\{z \in \mathbb{C} \mid |z - s| \leqslant \rho(B)\}$. Следовательно, получаем, что $\operatorname{Re}\lambda(A) \geqslant s - \rho(B)$. В силу последней части Теоремы 1.1, $\rho(B)$ является собственным значением B, и, следовательно, $s - \rho(B) \geqslant 0$ является собственным значением A с наименьшей вещественной частью. Значит, если известно, что A — невырожденная (например, в силу Теоремы 1.2), то $s > \rho(B)$. □

Упражнение 1.2 В доказательстве Теоремы 1.3 проверьте выполнение (1.5). ◊

Регулярным расщеплением $A \in \mathbb{R}^{n\times n}$ называется представление [5]

$$A = P - Q,$$

где матрица P невырождена, а P^{-1} и Q — поэлементно неотрицательные.

Теорема 1.4 *[5, 7] Если A — возможно вырожденная M-матрица с регулярным расщеплением $A = P - Q$, то $\rho(P^{-1}Q) \leqslant 1$. Последнее неравенство становится строгим ($\rho(P^{-1}Q) < 1$), коль скоро A невырождена.*

Матрица $A = (a_{ij})$ называется эрмитовой, если A равняется своей комплексно сопряжённой матрице $A^* = (\bar{a}_{ji})$. Если $A = A^T$, то матрица A называется симметричной . Вещественные эрмитовы матрицы являются симметричными. Собственные значения эрмитовой матрицы вещественны. Косоэрмитовы (т.е., матрицы, для которых выполняется $A = -A^*$) и вещественные кососимметричные матрицы ($A = -A^T \in \mathbb{R}^{n\times n}$) имеют чисто мнимые собственные значения. Для каждой матрицы $A \in \mathbb{R}^{n\times n}$ её симметричная и кососимметричная составляющие определяются соответственно как $\frac{1}{2}(A + A^T)$

и $\frac{1}{2}(A - A^T)$. Вещественная квадратная матрица однозначно определяется своими симметричными и кососимметричными составляющими. На плоскости комплексных чисел спектр $\Lambda(A)$ вещественной квадратной матрицы A находится в прямоугольной области,

$$\Lambda(A) \subset [a, b] \times [-ic, ic] \subset \mathbb{C}, \qquad a, b, c \in \mathbb{R},$$

где a и b — соответственно минимальное и максимальное собственные значения $\frac{1}{2}(A + A^T)$, а c — спектральный радиус матрицы $\frac{1}{2}(A - A^T)$.

2 Постановка задачи. Примеры

В данных заметках мы обсуждаем некоторые матричные задачи, возникающие в численном решении задач Коши следующего вида. Для заданных матрицы $A \in \mathbb{R}^{n\times n}$, векторной функции $g(t) : \mathbb{R} \to \mathbb{R}^n$ и вектора $y^0 \in \mathbb{R}^n$, найти такую векторную функцию $y(t) : \mathbb{R} \to \mathbb{R}^n$, что

$$y'(t) = -Ay(t) + g(t), \qquad y(0) = y^0. \tag{2.1}$$

Подобные задачи возникают в разных многочисленных случаях, например, при численном решении начально-краевых задач для параболических или гиперболических уравнений в частных производных (УЧП). Один из распространённых способов решения таких задач состоит в том, что уравнения сначала дискретизируются по пространству (при этом учитываются краевые условия) и, таким образом, сводятся к системе дифференциальных уравнений (ДУ) по времени, например, к системе вида (2.1). Данная система затем решается (интегрируется по времени) численно. Такой подход называется методом прямых (по-английски method of lines) [8, Глава 6.1].

Помимо задач вида (2.1), в главе 5.5 мы также кратко обсуждаем решение задач Коши более общего вида, для нелинейных автономных систем ДУ вида $y'(t) = -Ay - R(y)$, известных как задачи адвекции–диффузии–реакции. Если не оговорено иначе, всюду предполагаем, что A в (2.1) такова, что её симметричная составляющая $\frac{1}{2}(A + A^T)$ положительно определена.

2.1 Пример: нестационарная задача конвекции–диффузии

Пусть $\Omega \subset \mathbb{R}^2$ — область с гладкой границей $\partial\Omega$, а $L[u]$ — линейный дифференциальный оператор, действующий на функции $u(x, y, t)$ из некоторого функционального пространства, для $t > 0$ и $(x, y) \in \Omega$. Рассмотрим задачу

$$\begin{aligned}
&\text{(a)} \quad \frac{\partial u}{\partial t} + L[u] = \tilde{g}(x, y, t), \\
&\qquad u = u(x, y, t), \quad (x, y) \in \Omega, \quad t > 0, \\
&\text{(b)} \quad u(x, y, 0) = u^0(x, y), \\
&\text{(c)} \quad \text{условия на } u(x, y, t)\big|_{(x,y)\in\partial\Omega} \text{ и её производные,}
\end{aligned} \tag{2.2}$$

где функции $\tilde{g}(x,y,t)$, $u^0(x,y)$ заданы, $u(x,y,t)$ — искомая функция. Заметим, что это — начально-краевая задача, поскольку соотношения (2.2)(b), (2.2)(c) представляют собой соответственно начальные и краевые условия на $u(x,y,t)$.

Применяя к численному решению (2.2) метод прямых, мы сначала дискретизируем УЧП (2.2)(a) по пространству, что приводит к системе ДУ:

$$\frac{\partial u}{\partial t} + L[u] = \tilde{g}(x,y,t) \quad \overset{\text{дискретизация по пространству}}{\longrightarrow} \quad y'(t) = -Ay(t) + g(t), \tag{2.3}$$

где векторная функция $y(t)$, $y : \mathbb{R} \to \mathbb{R}^n$ приближает неизвестную функцию $u(x,y,t)$ в наборе из n дискретных точек $(x_i, y_k) \in \Omega$, матрица $A \in \mathbb{R}^{n\times n}$ аппроксимирует оператор $L[\cdot]$, $Aw \approx L[u]$, а $g(t)$ — векторная функция, $g : \mathbb{R} \to \mathbb{R}^n$, чьи компоненты $g_i(t)$ содержат величины $\tilde{g}(x_i, y_k, t)$ и, возможно, некоторые вклады из краевых условий (2.2)(c). Краевые условия также учитываются структурой матрицы A.

Рассмотрим теперь две простые конечно-разностные дискретизации (2.3) нестационарной задачи конвекции–диффузии. Эта задача задаётся (2.2) с

$$L[u] = -(D_1 u_x)_x - (D_2 u_y)_y + v_1 u_x + v_2 u_y + Du. \tag{2.4}$$

Здесь заданные функции D_i, v_i и D удовлетворяют соотношениям

$$\begin{gathered} D_i = D_i(x,y) \geqslant 0, \quad v_i = v_i(x,y), \qquad i = 1,2, \\ D_1 + D_2 > 0, \quad (v_1)_x + (v_2)_y \equiv 0, \\ D = D(x,y) \geqslant 0, \qquad (x,y) \in \Omega, \end{gathered} \tag{2.5}$$

где индексы $\cdot_x$ и $\cdot_y$ обозначают производные по отношению к x и y соответственно. Функции D_i и v_i называются коэффициентами диффузии и конвекции соответственно. Для простоты изложения будем считать, что область Ω выпуклая и краевые условия (2.2)(c) однородные:

$$u\big|_{\partial\Omega} = 0, \qquad t > 0.$$

Первая из рассматриваемых конечно-разностных дискретизаций приводит к матрице A, такой что $L[u] \approx Ay$ и

$$\begin{gathered} A_0 y = L_{\text{diff}}[u], \qquad A_1 y = L_{\text{conv}}[u], \\ L_{\text{diff}}[u] \equiv -(D_1 u_x)_x - (D_2 u_y)_y + Du, \qquad L_{\text{conv}}[u] \equiv v_1 u_x + v_2 u_y, \end{gathered} \tag{2.6}$$

где A_0 и A_1 соответственно симметричная и кососимметричная составляющие A. Другими словами, дискретизация такова, что симметричная составляющая A аппроксимирует диффузионные члены, а кососимметричная — конвективные.

2.2 Конечно-разностные соотношения. Центральные разности

Введём регулярную декартову сетку, покрывающую область Ω и состоящую из n узлов $(x_i, y_k) \in \Omega$. Сетка имеет шаги $h_1 > 0$ по направлению x и $h_2 > 0$

по y, т.е. $x_{i+1} - x_i = h_1$, $y_{k+1} - y_k = h_2$ для всех возмодных i и k. В каждом узле (x_i, y_k) сетки мы аппроксимируем члены $L_{\text{diff}}[u]$ следующими конечными разностями:

$$\begin{aligned}
(D_1 u_x)_x &\approx \frac{(D_1)_{i+1/2,k}(u_{i+1,k} - u_{i,k}) - (D_1)_{i-1/2,k}(u_{i,k} - u_{i-1,k})}{h_1^2}, \\
(D_2 u_y)_y &\approx \frac{(D_2)_{i,k+1/2}(u_{i,k+1} - u_{i,k}) - (D_2)_{i,k-1/2}(u_{i,k} - u_{i,k-1})}{h_2^2}, \\
Du &\approx D_{i,k} u_{i,k},
\end{aligned} \tag{2.7}$$

где индексы $(\cdot)_{i,k}$ обозначают принадлежность к узлу сетки (x_i, y_k), а индексы $(\cdot)_{i\pm1,k}$, $(\cdot)_{i,k\pm1}$, $(\cdot)_{i\pm1/2,k}$, $(\cdot)_{i,k\pm1/2}$ — к узлам сетки, сдвинутым от (x_i, y_k) соответственно на $\pm h_1$, $\pm h_2$, $\pm h_1/2$ или $\pm h_2/2$ по направлению x или y.

Прежде чем дискретизировать конвективные члены, учитывая соотношение (2.5), перепишем их так [9]:

$$v_1 u_x + v_2 u_y = \frac{1}{2}(v_1 u_x + v_2 u_y) + \frac{1}{2}((v_1 u)_x + (v_2 u)_y). \tag{2.8}$$

Причина, по которой мы это делаем, станет ясна немного позже, в Упражнении 2.2. Для производных в правой части соотношения (2.8) будем использовать такую конечно-разностную схему, относящуюся к точке сетки (x_i, y_k):

$$\begin{aligned}
\frac{1}{2}(v_1 u_x + (v_1 u)_x) &\approx \frac{(v_1)_{i,k}(u_{i+1,k} - u_{i-1,k}) + ((v_1)_{i+1,k} u_{i+1,k} - (v_1)_{i-1,k} u_{i-1,k})}{4h_1}, \\
\frac{1}{2}(v_2 u_y + (v_2 u)_y) &\approx \frac{(v_2)_{i,k}(u_{i,k+1} - u_{i,k-1}) + ((v_2)_{i,k+1} u_{i,k+1} - (v_2)_{i,k-1} u_{i,k-1})}{4h_2}.
\end{aligned} \tag{2.9}$$

Комбинируя соотношения (2.7) и (2.9), получаем следующую аппроксимацию оператора $L[u]$ в каждой точке сетки (x_i, y_k):

$$L[u]\Big|_{(x_i,y_k)} \approx W_{i,k} u_{i-1,k} + S_{i,k} u_{i,k-1} + C_{i,k} u_{i,k} + N_{i,k} u_{i,k+1} + E_{i,k} u_{i+1,k}, \tag{2.10}$$

где

$$\begin{aligned}
W_{i,k} &= -\frac{(D_1)_{i-1/2,k}}{h_1^2} - \frac{(v_1)_{i,k} + (v_1)_{i-1,k}}{4h_1}, \\
S_{i,k} &= -\frac{(D_2)_{i,k-1/2}}{h_2^2} - \frac{(v_2)_{i,k} + (v_2)_{i,k-1}}{4h_2}, \\
C_{i,k} &= \frac{(D_1)_{i-1/2,k} + (D_1)_{i+1/2,k}}{h_1^2} + \frac{(D_2)_{i,k-1/2} + (D_2)_{i,k+1/2}}{h_2^2} + D_{i,k}, \\
N_{i,k} &= -\frac{(D_2)_{i,k+1/2}}{h_2^2} + \frac{(v_2)_{i,k} + (v_2)_{i,k+1}}{4h_2}, \\
E_{i,k} &= -\frac{(D_1)_{i+1/2,k}}{h_1^2} + \frac{(v_1)_{i,k} + (v_1)_{i+1,k}}{4h_1}.
\end{aligned}$$

Обозначения коэффициентов в (2.10) соответствует положению узлов $u_{i\pm1,k\pm1}$ конечно-разностного шаблона по отношению к центральному узлу $u_{i,k}$ (поскольку $u_{i+1,k}$ лежит справа, т.е. к востоку, от $u_{i,k}$, пишем $E_{i,k}u_{i+1,k}$; $u_{i,k+1}$ лежит к северу от центра шаблона, следовательно, мы пишем $N_{i,k}u_{i,k+1}$, и т.д.)

2.3 Структура матрицы

Выписывая соотношения (2.10) для каждого узла сетки в опрелённом порядке, мы можем скомбинировать все эти соотношения в виде матрично-векторного произведения. Действительно, предположим, что $\Omega = [0,1]\times[0,1]$ и сетка определена как

$$x_i = ih_1, \quad y_k = kh_2, \quad \text{где} \quad h_j = 1/n_j, \quad j = 1,2. \tag{2.11}$$

Выписывая соотношения (2.10) последовательно для узлов
(x_1,y_1), (x_1,y_2), …, (x_1,y_{n_2-1}),
(x_2,y_1), (x_2,y_2), …, (x_2,y_{n_2-1}),
…
(x_{n_1-1},y_1), (x_{n_1-1},y_2), …, (x_{n_1-1},y_{n_2-1}), получаем

$$\begin{aligned}
(x_1,y_1): &\quad W_{1,1}u_{0,1} + S_{1,1}u_{1,0} + C_{1,1}u_{1,1} + N_{1,1}u_{1,2} + E_{1,1}u_{2,1},\\
(x_1,y_2): &\quad W_{1,2}u_{0,2} + S_{1,2}u_{1,1} + C_{1,2}u_{1,2} + N_{1,2}u_{1,3} + E_{1,2}u_{2,2},\\
(x_1,y_3): &\quad W_{1,3}u_{0,3} + S_{1,3}u_{1,2} + C_{1,3}u_{1,3} + N_{1,3}u_{1,4} + E_{1,3}u_{2,3},\\
&\quad \dots\\
(x_1,y_{n_2-1}): &\quad W_{1,n_2-1}u_{0,n_2-1} + S_{1,n_2-1}u_{1,n_2-2} + C_{1,n_2-1}u_{1,n_2-1}\\
&\qquad + N_{1,n_2-1}u_{1,n_2} + E_{1,n_2-1}u_{2,n_2-1},\\
(x_2,y_1): &\quad W_{2,1}u_{1,1} + S_{2,1}u_{2,0} + C_{2,1}u_{2,1} + N_{2,1}u_{2,2} + E_{2,1}u_{3,1},\\
&\quad \dots\\
(x_{n_1-1},y_{n_2-1}): &\quad W_{n_1-1,n_2-1}u_{n_1-2,n_2-1} + S_{n_1-1,n_2-1}u_{n_1-1,n_2-2} + C_{n_1-1,n_2-1}u_{n_1-1,n_2-1}\\
&\qquad + N_{n_1-1,n_2-1}u_{n_1-1,n_2} + E_{n_1-1,n_2-1}u_{n_1,n_2-1},\\
&\quad \dots
\end{aligned}$$

Эти соотношения, где в силу однородных граничных условий (2.2)(c) $u_{i,k}=0$ для $i=0$, $i=n_1$, $k=0$ и $k=n_2$, могут быть представлены в виде матрично-векторного произведения так:

$$\begin{bmatrix}
C_{1,1} & N_{1,1} & \dots & \dots & E_{1,1} & & & \\
S_{1,2} & C_{1,2} & N_{1,2} & \dots & & E_{1,2} & & \\
 & \ddots & \ddots & & & & \ddots & \\
 & & S_{1,n_2-1} & C_{1,n_2-1} & 0 & & & \\
W_{2,1} & & & 0 & C_{2,1} & N_{2,1} & & \\
 & W_{2,2} & & & S_{2,2} & C_{2,2} & N_{2,2} & \\
 & & \ddots & & & \ddots & \ddots & \ddots
\end{bmatrix}
\begin{bmatrix} u_{1,1}\\ u_{1,2}\\ \vdots\\ u_{1,n_2-1}\\ u_{2,1}\\ u_{2,2}\\ \vdots \end{bmatrix} = Aw, \tag{2.12}$$

где возникающую матрицу мы называем A, а вектор — w. Матрица A — пятидиагональная, где главная диагональ содержит коэффициенты $C_{i,k}$, соседние с ней диагонали — коэффициенты $S_{i,k}$ и $N_{i,k}$, а две оставшиеся диагонали — $W_{i,k}$ и $E_{i,k}$.

Заметим, что каждый узел сетки соответствует строке матрицы A и что при формировании A мы можем использовать любой порядок узлов. Таким образом, структура A зависит от выбранного порядка узлов (подробности см., например, в [10]).

Координаты вектора w в (2.12) являются функциями от времени t, они аппроксимируют значения искомой функции $u(x, y, t)$ в узлах сетки (x_i, y_k). Следовательно, нам следует заменить $\partial u/\partial t$ в (2.2)(а) вектором производных по времени координат w. Кроме того, заменяя $L[u]$ на Aw, мы получаем систему обыкновенных ДУ (2.3) (где вместо $w(t)$ неизвестная вектор–функция обозначена $y(t)$).

Упражнение 2.1 Сколько нулевых элементов расположено в первой строке A между $N_{1,1}$ и $E_{1,1}$? Предположим теперь, что $n_1 = 20$, $n_2 = 10$. Определите элементы матрицы $a_{(9,10)}$ и $a_{(10,11)}$. Пусть $n_2 = 5$. Выпишите пять первых координат вектор-функции $g(t)$ из (2.3). $\Diamond$

Упражнение 2.2 Покажите, что конечно-разностная аппроксимация (2.7), (2.9) удовлетворяет соотношению (2.6): диффузионные члены определяют симметричную составляющую A, а конвективные члены — её кососимметричную составляющую. $\Diamond$

Предположим теперь, D_1 и D_2 таковы, что на достаточно мелкой сетке в силу свойства $D_1 + D_2 > 0$ (см. (2.4)) имеем

$$(D_1)_{i+1/2,k} + (D_2)_{i,k+1/2} > 0 \qquad \forall (x, y) \in \Omega.$$

Тогда нетрудно показать, что граф матрицы $\frac{1}{2}(A + A^T)$ сильно связан, а сама матрица $\frac{1}{2}(A + A^T)$ — матрица со слабым диагональным преобладанием. В некоторых строках диагональное преобладание выполняется строго. Поэтому приходим к выводу, что симметричная часть A является неразложимой с диагональным преобладанием. Кроме того, по Теоремам 1.2 и 1.3, нетрудно убедиться, что $\frac{1}{2}(A + A^T)$ положительно определена.

2.4 Аппроксимация разностями против потока

Альтернативой аппроксимации центральными разностями (2.9) является аппроксимация разностями против потока. Как мы сейчас увидим это, в этом случае нет необходимости переписывать конвективные члены $v_1 u_x + v_2 u_y$ в виде (2.8). Использование разностей против потока приводит к аппроксимации

вида (2.10) на знакомом пятиточечном шаблоне, где коэффициенты определяются как

$$\begin{aligned}
W_{i,k} &= -\frac{(D_1)_{i-1/2,k}}{h_1^2} - \frac{(v_1)_{i,k} + |v_1|_{i,k}}{2h_1}, \\
S_{i,k} &= -\frac{(D_2)_{i,k-1/2}}{h_2^2} - \frac{(v_2)_{i,k} + |v_2|_{i,k}}{2h_2}, \\
C_{i,k} &= \frac{(D_1)_{i-1/2,k} + (D_1)_{i+1/2,k}}{h_1^2} + \frac{(D_2)_{i,k-1/2} + (D_2)_{i,k+1/2}}{h_2^2} + D_{i,k} + \\
&\quad + \frac{|v_1|_{i,k}}{h_1} + \frac{|v_2|_{i,k}}{h_2}, \\
N_{i,k} &= -\frac{(D_2)_{i,k+1/2}}{h_2^2} + \frac{(v_2)_{i,k} - |v_2|_{i,k}}{2h_2}, \\
E_{i,k} &= -\frac{(D_1)_{i+1/2,k}}{h_1^2} + \frac{(v_1)_{i,k} - |v_1|_{i,k}}{2h_1}.
\end{aligned} \tag{2.13}$$

Как видим, для этой дискретизации конвективные члены дают вклад в коэффициенты $C_{i,k}$ на главной диагонали матрицы A. Следовательно, конвективные вклады не могут давать кососимметричную матрицу (сравните с Упражнением 2.2).

Упражнение 2.3 Покажите, что противопотоковая конечно-разностная аппроксимация (2.10), (2.13) приводит к матрице A, которая является M-матрицей. ◊

2.5 Два других примера

Помимо нестационарной задачи конвекции–диффузии имеется множество других задач и приложений, где возникает необходимость решать задачи Коши вида (2.1). Мы коротко рассмотрим два других таких примера приложений. Оба примера взяты из работы [11]. Первый из них — это типичная задача теории управления: требуется найти такую векторную функцию состояний $y(t)$, что

$$y'(t) = -Ay(t) + Bu(t), \qquad y(0) = y^0, \tag{2.14}$$

где $A \in \mathbb{R}^{n\times n}$ — сопровождающая матрица состояний, $u(t) : \mathbb{R} \to \mathbb{R}^m$ — функция управления, а $B \in \mathbb{R}^{n\times m}$.

Второй пример — это большой класс задач, где задействованы цепи Маркова с непрерывным временем. Как отмечается в [11], *"успешный и широко используемый подход в моделировании поведения разнообразных физических систем состоит в нумерации (взаимно исключающих) состояний, в которых система может находиться в определённый момент времени, и, затем, в описании взаимодействия между этими состояниями"*. При определённых предположениях рассматриваемый физический процесс может быть

описан задачей Чэпмена–Колмогорова:

$$y'(t) = -Ay(t), \qquad y(0) = y^0.$$

Её решение $y(t) = e^{-tA}y^0$ является распределением вероятности перехода цепи Маркова, а матрица коэффициентов $A \in \mathbb{R}^{n\times n}$ называется матрицей интенсивностей, где n — число состояний в цепи Маркова. В силу опредлённых вероятностных предположений, A — вырожденная M-матрица в нулевыми столбцовыми суммами, т.е.,

$$a_{ij} \leqslant 0 \text{ for } i \neq j \quad \text{и} \quad a_{jj} = -\sum_{i\neq j} a_{ij} \geqslant 0.$$

Упражнение 2.4 Основываясь на результатах Главы 1, покажите, что матрица A из последнего примера — действительно вырожденная M-матрица. $\Diamond$

3 Корректность постановки задачи. Оценки устойчивости

Материал, представленный в этой главе, близко следует книге [8, Sect. 2.3].

3.1 Матричная экспонента. Формула вариации постоянных

Для анализа задач Коши (2.1) и численных методов их решения нам потребуется понятие матричной экспоненты, определяемой, для данной матрицы $A \in \mathbb{R}^{n\times n}$, степенными рядами

$$e^A = \sum_{k=0}^{\infty} \frac{1}{k!} A^k, \qquad A^0 = I. \tag{3.1}$$

Данное определение матричной экспоненты — одно из многих возможных определений этого понятия [12].

Теорема 3.1 *Однородная ($g(t) \equiv 0$) задача Коши* (2.1) *имеет решение*

$$y(t) = e^{-tA}y^0. \tag{3.2}$$

Доказательство Рассмотрим степенной ряд (3.1) для e^{-tA} и заметим, что члены ряда ограничены по норме $\frac{1}{k!}t^k\|A\|^k$. Следовательно, ряд сходится и

$$\|e^{-tA}\| \leqslant e^{t\|A\|}. \tag{3.3}$$

Завершение доказательства оставлено читателю в качестве упражнения. $\square$

Упражнение 3.1 Закончите доказательство Теоремы 3.1. $\Diamond$

Теорема 3.2 *Решение задачи Коши* (2.1) *имеет вид*

$$y(t) = e^{-tA}y^0 + \int_0^t e^{(s-t)A}g(s)\,ds. \tag{3.4}$$

Последняя формула называется формулой вариации постоянных.

Доказательство Умножая систему уравнений $y'(t) + Ay(t) = g(t)$ слева на матрицу e^{tA}, мы получаем

$$e^{tA}y'(t) + e^{tA}Ay(t) = e^{tA}g(t) \quad \Leftrightarrow \quad \frac{d}{dt}\left(e^{tA}y(t)\right) = e^{tA}g(t).$$

Выражение (3.4) теперь может быть получено интегрированием последнего равенства:

$$\int_0^t \frac{d}{ds}\left(e^{sA}y(s)\right)\,ds = \int_0^t e^{sA}g(s)\,ds,$$
$$e^{tA}y(t) - \underbrace{e^{0A}y(0)}_{y^0} = \int_0^t e^{sA}g(s)\,ds.$$

□

3.2 Оценки устойчивости

Формула вариации постоянных (3.4) позволяет получить т.н. оценки устойчивости для задачи (2.1), что означает следующее. Рассмотрим, помимо (2.1), возмущённую задачу

$$\tilde{y}'(t) = -A\tilde{y}(t) + g(t) + \delta(t), \qquad \tilde{y}(0) = \tilde{y}^0, \tag{3.5}$$

с заданными $\delta(t)$ и $\tilde{y}^0$. Пусть $\varepsilon(t) = \tilde{y}(t) - y(t)$. Нас интересует получение оценок устойчивости, т.е. оценок, которые показывают зависимость $\|\varepsilon(t)\|$ от $\|\varepsilon(0)\|$ и $\|\delta(t)\|$. Поскольку $\varepsilon(t)$ является решением задачи

$$\varepsilon'(t) = -A\varepsilon(t) + \delta(t), \qquad \varepsilon(0) = \tilde{y}^0 - y^0,$$

используя формулу вариации постоянных, мы получаем

$$\varepsilon(t) = e^{-tA}\varepsilon(0) + \int_0^t e^{(s-t)A}\delta(s)\,ds,$$
$$\|\varepsilon(t)\| \leqslant \|e^{-tA}\|\,\|\varepsilon(0)\| + \int_0^t \|e^{(s-t)A}\|\,ds \max_{s\in[0,t]}\|\delta(s)\|,$$

где мы воспользовались свойством $\|\int_a^b f(x)dx\| \leqslant \int_a^b \|f(x)\|dx$, выполняющимся для всякой непрерывной векторной функции $f : [a,b] \to \mathbb{R}^n$.

Предположим, что существуют константы K и ω, такие что

$$\|e^{-tA}\| \leqslant Ke^{-t\omega}, \qquad t \geqslant 0. \tag{3.6}$$

Тогда

$$\|\varepsilon(t)\| \leqslant Ke^{-t\omega}\|\varepsilon(0)\| + K\frac{1-e^{-t\omega}}{\omega}\max_{s\in[0,t]}\|\delta(s)\|. \tag{3.7}$$

Упражнение 3.2 Проверьте выполнение оценки (3.7). Заметьте, что множитель $(1-e^{-t\omega})/\omega$ не определён для $\omega = 0$. Тем не менее, мы можем формально назначить определённое значение выражению $(e^{-t\omega}-1)/\omega$ для $\omega = 0$. Какое значение? ◊

Для того, чтобы оценка устойчивости (3.7) была полезной, экспоненциальная оценка (3.6) должна быть достаточно точной. Подобные оценки могут быть получены разными способами.

Упражнение 3.3 Заметим, что оценка (3.3) также имеет вид (3.6). Однако эта оценка навряд ли полезна. Объясните почему. Подсказка: рассмотрите скалярный случай $n = 1$ и матрицы $A = 1$, $A = -1$. ◊

Предположим, что A — диагонализируемая как $A = VDV^{-1}$ (где D — диагональная матрица с элементами λ_k, собственными значениями матрицы A, на главной диагонали). Тогда

$$\|e^{-tA}\| = \|Ve^{-tD}V^{-1}\| \leqslant \|V\|\|e^{-tD}\|\|V^{-1}\| = \kappa(V)\max_k|e^{-t\lambda_k}| = \kappa(V)e^{-t\min_k \operatorname{Re}\lambda_k}.$$

Здесь $\kappa(V) = \|V\|\|V^{-1}\|$ — это число обусловленности матрицы собственных векторов V. Мы видим, что (3.6) выполняется с $K = \kappa(V)$ и $\omega = \min_k \operatorname{Re}\lambda_k$. Если A — нормальная, то $\kappa(V) = 1$ в норме $\|\cdot\|_2$. Если же A далека от нормальной, так что большая $\kappa(V)$ делает вышестоящую оценку бесполезной, или если информация о спектре A недоступна, то нам потребуются другие методы оценки. Один из них мы рассмотрим ниже.

3.3 Логарифмическая матричная норма

Чтобы получить более точные экспоненциальные оценки вида (3.6), мы вводим так называемую *логарифмическую норму* матрицы $A \in \mathbb{R}^{n\times n}$, определяемую как [8, Sect. 2.3]

$$\mu(A) = \lim_{\tau\to 0+}\frac{\|I+\tau A\|-1}{\tau}, \tag{3.8}$$

где $\|\cdot\|$ — операторная (т.е. индуцированная некоторой векторной нормой) матричная норма.

Упражнение 3.4 Проверьте, что для $\tau > 0$

$$-\|A\| \leqslant \frac{\|I+\tau A\|-1}{\tau} \leqslant \|A\|.$$

◊

Можно показать, что дробь, стоящая под знаком предела в выражении (3.8), — неубывающая функция от $\tau > 0$. Действительно, для $0 < \sigma < 1$ получаем

$$\|I + \sigma\tau A\| = \|I + \sigma\tau A + \sigma I - \sigma I\| \leqslant \sigma\|I + \tau A\| + 1 - \sigma,$$
$$\frac{\|I + \sigma\tau A\| - 1}{\sigma\tau} \leqslant \frac{\sigma\|I + \tau A\| - \sigma}{\sigma\tau} \leqslant \frac{\|I + \tau A\| - 1}{\tau}.$$

Следовательно, предел в выражении (3.8) существует, и сходимость монотонна.

Упражнение 3.5 Является ли логарифмическая норма нормой? ◊

Определение (3.8) логарифмической матричной нормы показывает, что эту специальную норму можно интерпретировать как одностороннюю производную отображения $\|\cdot\| : \mathbb{R}^{n\times n} \to \mathbb{R}$, вычисленную в точке $I \in \mathbb{R}^{n\times n}$ в направлении $A \in \mathbb{R}^{n\times n}$ [13, Section 1.5]. Название "логарифмическая" становится ясным, если мы заметим, что для $A \in \mathbb{R}^{n\times n}$

$$\mu(A) = \lim_{\tau\to 0+} \frac{\ln\|e^{\tau A}\|}{\tau}.$$

Действительно, для достаточно малых $\tau > 0$ выполняется [13, Section 1.5]

$$\ln\|e^{\tau A}\| = \ln\left(\|I + \tau A\| + O(\tau^2)\right) = \ln\left(1 + \left[\|I + \tau A\| - 1 + O(\tau^2)\right]\right)$$
$$= \|I + \tau A\| - 1 + O(\tau^2).$$

Важность логарифмической нормы проявляется в следующем результате [8].

Теорема 3.3 *Пусть $A \in \mathbb{R}^{n\times n}$ и $\omega \in \mathbb{R}$. Тогда*

$$\mu(-A) \leqslant -\omega \qquad \Leftrightarrow \qquad \|e^{-tA}\| \leqslant e^{-t\omega}\ \forall t \geqslant 0. \tag{3.9}$$

Доказательство Заметим, что последнее выражение может быть переписано в эквивалентной форме:

$$\mu(A) \leqslant \omega \qquad \Leftrightarrow \qquad \|e^{tA}\| \leqslant e^{t\omega}\ \forall t \geqslant 0. \tag{3.10}$$

Результат будет доказан именно в этой форме. Сначала предположим, что $\mu(A) \leqslant \omega$. Тогда для достаточно малых $\tau > 0$, по определению $\mu(A)$, имеем

$$\frac{\|I + \tau A\| - 1}{\tau} - \mu(A) = O(\tau),$$
$$\|I + \tau A\| - 1 - \tau\mu(A) = O(\tau^2),$$
$$\|I + \tau A\| = 1 + \tau\mu(A) + O(\tau^2),$$
$$\|I + \tau A\| \leqslant 1 + \tau\omega + O(\tau^2),$$
$$\|(I + \tau A)^k\| \leqslant (1 + \tau\omega + O(\tau^2))^k,$$

где $t = k\tau$ фиксировано. Переходя к пределу $\tau \to 0+$ в обеих частях неравенства, получаем

$$\|e^{tA}\| \leqslant e^{t\omega}.$$

Здесь для $\tau \to 0+$ и фиксированного $t = k\tau$ имеем $(I + \tau A)^k \to e^{tA}$. Действительно, $I + \tau A$ — это матрица перехода явного метода Эйлера, применённого для решения задачи $y'(t) = Ay(t)$ (смотрите формулу (4.1), где $-A$ заменена на A).

Теперь предположим, что $\|e^{tA}\| \leqslant e^{t\omega}$ для $t > 0$. Тогда

$$\|I + \tau A\| = \|e^{tA} + O(\tau^2)\| \leqslant \|e^{tA}\| + O(\tau^2) \leqslant e^{t\omega} + O(\tau^2) = 1 + \tau\omega + O(\tau^2),$$

откуда следует $\mu(A) \leqslant \omega$. □

Следующий результат перечисляет некоторые важные свойства логарифмической матричной нормы.

Теорема 3.4 *Пусть $A \in \mathbb{R}^{n\times n}$, а $\mu(A)$ определяется в* (3.8)*. Тогда*

$$\mu(sI + A) = s + \mu(A), \qquad \forall s \in \mathbb{R}, \tag{3.11}$$
$$\mu(tA) = |t|\mu(\operatorname{sign}(t)A), \qquad \forall t \in \mathbb{R}, \tag{3.12}$$
$$\mu(A + B) \leqslant \mu(A) + \mu(B), \tag{3.13}$$
$$|\mu(A) - \mu(B)| \leqslant \|A - B\|, \tag{3.14}$$
$$\mu(A) \geqslant -\frac{\|Ax\|}{\|x\|}, \qquad \forall 0 \neq x \in \mathbb{C}^n, \tag{3.15}$$

где $\|\cdot\|$ — норма, участвующая в определении $\mu(\cdot)$, а sign *— сигнум-функция.*

Доказательство Для доказательства (3.11) заметим, что $1 + \tau s \geqslant 0$ для малых $\tau > 0$ и

$$\begin{aligned}\mu(sI + A) &= \lim_{\tau\to 0+} \frac{\|I + \tau(sI + A)\| - 1}{\tau} = \lim_{\tau\to 0+} \frac{(1 + \tau s)\|I + \frac{\tau}{1+\tau s}A\| - 1}{\tau} \\ &= \lim_{\tau\to 0+} \frac{\|I + \frac{\tau}{1+\tau s}A\| - \frac{1}{1+\tau s}}{\frac{\tau}{1+\tau s}} = \lim_{\tau\to 0+} \frac{\|I + \frac{\tau}{1+\tau s}A\| - \frac{1+\tau s-\tau s}{1+\tau s}}{\frac{\tau}{1+\tau s}} = \mu(A) + s.\end{aligned}$$

Доказательство свойства (3.12) оставлено в качестве упражнения.
Далее, свойство (3.13) можно показать, пользуясь оценкой [14]

$$\begin{aligned}\|I + \tau(A + B)\| - 1 &= \|\frac{1}{2}(I + 2\tau A) + \frac{1}{2}(I + 2\tau A)\| - 1 \\ &\leqslant \frac{1}{2}(\|I + 2\tau A\| - 1) + \frac{1}{2}(\|I + 2\tau B\| - 1).\end{aligned}$$

Свойство (3.14) может быть установлено с учётом того, что $-\|A\| \leqslant \mu(A) \leqslant \|A\|$ (см. Упражнение 3.4).

Наконец, чтобы убедиться в выполнении свойства (3.15), напишем

$$\|x\| = \|(I+\tau A)x - \tau Ax\| \leqslant \|(I+\tau A)x\| + \tau\|Ax\|,$$
$$-\|Ax\| \leqslant \frac{\|(I+\tau A)x\| - \|x\|}{\tau} \leqslant \frac{\|I+\tau A\| - 1}{\tau}\|x\|.$$

□

Упражнение 3.6 Закончите детали доказательства Теоремы 3.4. ◊

Поскольку логарифмическая норма вводится на основе любой матричной нормы, на практике можно постараться выбрать норму, наиболее подходящую для конкретной ситуации.

Упражнение 3.7 Проверьте, что для наиболее часто используемых векторных норм (1.1) и индуцированных ими матричных норм (1.2), соответствующие логарифмические нормы определяются так:

$$\begin{aligned}
\mu_2(A) &= \max_{x\neq 0}\frac{\mathrm{Re}(Ax,x)}{(x,x)} = \max\{\lambda \mid \lambda \in \text{spectrum of } \frac{1}{2}(A+A^*)\},\\
\mu_1(A) &= \max_j(\mathrm{Re}\, a_{jj} + \sum_{i\neq j}|a_{ij}|),\\
\mu_\infty(A) &= \max_i(\mathrm{Re}\, a_{ii} + \sum_{j\neq i}|a_{ij}|).
\end{aligned} \tag{3.16}$$

◊

3.4 Примеры

Чтобы увидеть, как результаты этой главы могут быть использованы, рассмотрим несколько примеров, взятых из [8, Sect. 2.3]. Пусть решается задача (2.1) и известно, что матрица $\frac{1}{2}(A+A^T)$ (симметричная часть матрицы A) неотрицательно полуопределена. Как видно из (3.9) и (3.16), $\|e^{-tA}\|_2 \leqslant 1$ выполняется тогда и только тогда, когда $\frac{1}{2}(A+A^T)$ неотрицательно определена. Таким образом, в этом случае оценка устойчивости (3.7) выполняется во 2-й норме. Если же $\frac{1}{2}(A+A^T)$ положительно определена, то мы видим, что (3.6) выполняется для ω, являющегося наименьшим собственным значением матрицы $\frac{1}{2}(A+A^T)$.

Далее, если известно, что диагональные элементы A положительны и A — матрица со строчным (слабым) диагональным преобладанием, то

$$\mu_\infty(-A) = \max_i(-a_{ii} + \sum_{j\neq i}|a_{ij}|) = -\underbrace{\min_i(a_{ii} - \sum_{j\neq i}|a_{ij}|)}_{\text{denote by } \delta} \leqslant 0.$$

Это означает, что оценка устойчивости (3.7) выполняется в максимальной норме с $\omega = \delta$. Таким же образом оценка устойчивости в первой норме может

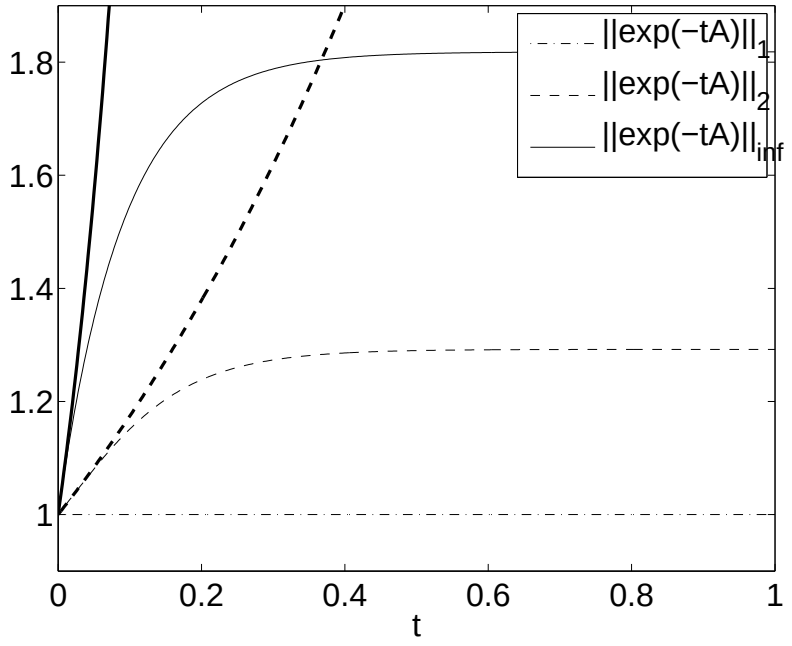

Рис. 1: Зависимость от времени норм $\|e^{-tA}\|_1$ (штрихпунктирная линия), $\|e^{-tA}\|_2$ (штриховая линия) и $\|e^{-tA}\|_\infty$ (сплошная линия). Графики верхних оценок для последних двух норм даны жирными линиями.

быть установлена, если A — матрица со столбцовым (слабым) диагональным преобладанием и положительными диагональными элементами.

Рассмотрим теперь более конкретный пример, взятый из [8, Sect. 2.3]. Задача Коши

$$y'(t) = -Ay(t), \qquad A = \begin{bmatrix} k_1 & -k_2 \\ -k_1 & k_2 \end{bmatrix}, \qquad y(0) = \begin{bmatrix} y_1^0 \\ y_2^0 \end{bmatrix}, \tag{3.17}$$

моделирует химическую реакцию $y_1 \xrightarrow{k_1} y_2 \xrightarrow{k_2} y_1$.

Упражнение 3.8 Проверьте, используя диагонализацию A, что точное решение задачи (3.17) есть

$$\begin{aligned} y_1(t) &= ak_2 + be^{-(k_1+k_2)t}, \\ y_2(t) &= ak_1 - be^{-(k_1+k_2)t}, \qquad a = \frac{y_1^0 + y_2^0}{k_1 + k_2}, \quad b = \frac{k_1 y_1^0 - k_2 y_2^0}{k_1 + k_2}, \end{aligned}$$

где y_1^0 и y_2^0 — заданные начальные значения. $\Diamond$

Рассмотрим решение задачи (3.17) для $0 \leqslant t \leqslant T = 1$, $k_1 = 1$, $y_1^0 = 0.1$ и $y_2^0 = 0.9$. Если $k_2 \gg k_1 = 1$, то $\|A\| \gg 1$, так что оценка (3.7) с учётом (3.3),(3.6) (т.е., для $\omega = -\|A\|$) указывает на неустойчивость (плохую обусловленность) задачи. С другой стороны, учитывая значения логарифмических норм

$$\mu_1(-A) = 0, \quad \mu_2(-A) = -\hat{\omega} = -\frac{k_1 + k_2}{2} + \sqrt{\frac{k_1^2 + k_2^2}{2}} > 0, \quad \mu_\infty(-A) = |k_2 - k_1|$$

и соотношение (3.9), получаем

$$\|e^{-tA}\|_1 \leqslant 1 \qquad \forall t \geqslant 0,$$
$$\|e^{-tA}\|_2 \leqslant e^{-t\hat{\omega}} \qquad \forall t \geqslant 0,$$
$$\|e^{-tA}\|_\infty \leqslant e^{t|k_2-k_1|} \qquad \forall t \geqslant 0.$$

Как видим из Рис. 1, последние две из этих оценок далеки от точных для больших $t > 0$. Для получения более точных оценок заметим, что для любой $n \times n$ матрицы B выполняется (см. (1.3))

$$\|B\|_2 \leqslant \sqrt{n}\|B\|_1, \qquad \|B\|_\infty \leqslant n\|B\|_1.$$

Таким образом, получаем оценки

$$\|e^{-tA}\|_2 \leqslant \sqrt{2} \qquad \forall t \geqslant 0,$$
$$\|e^{-tA}\|_\infty \leqslant 2 \qquad \forall t \geqslant 0,$$

показывающие хорошие свойства устойчивости этой задачи.

4 Основные схемы интегрирования и их устойчивость

Пусть $\tau > 0$ — шаг по времени, а y_k — численное решение задачи (2.1) в $t = k\tau$, $y^k \approx y(k\tau)$. Запишем три стандартные схемы интегрирования по времени задачи (2.1)

$$\frac{y^{k+1} - y^k}{\tau} = -Ay^k + g^k, \tag{4.1}$$
$$\frac{y^{k+1} - y^k}{\tau} = -Ay^{k+1} + g^{k+1}, \tag{4.2}$$
$$\frac{y^{k+1} - y^k}{\tau} = -(1-\theta)Ay^k - \theta Ay^{k+1} + (1-\theta)g^k + \theta g^{k+1}, \quad \theta \in [0,1], \tag{4.3}$$

называемые соответственно явной схемой Эйлера, неявной схемой Эйлера и явно-неявным θ-методом. Для $\theta = \frac{1}{2}$ получаем неявную формулу трапеций, а для значений $\theta = 0$ и $\theta = 1$ — соответственно, явную и неявную схемы Эйлера. Заметим, что неявная формула трапеций также известна как схема Кранка–Николсон, предложенная в 1947 г. Джоном Кранком и Филлис Николсон [15].

Многие схемы интегрирования по времени для задачи (2.1) с $g \equiv 0$ можно записать как

$$y^{k+1} = R(-\tau A)y^k,$$

где $R(z)$ — некоторая рациональная функция. Для каждой схемы интегрирования по времени эту функцию легко получить, выписывая схему для модельной задачи

$$y'(t) = \lambda y(t),\ y(0) = y^0, \qquad \lambda \in \mathbb{C}, \tag{4.4}$$

называмой модельной задачей Дальквиста [8, Sect. I.2]. $R(z)$ называется функцией устойчивости схемы.

Упражнение 4.1 Проверьте, что

$$R(z) = \frac{1 + (1-\theta)z}{1 - \theta z}$$

является функцией устойчивости θ-метода. $\Diamond$

Смысл названия «функция устойчивости» становится ясным, если мы рассмотрим множество

$$\mathcal{S} = \{z \in \mathbb{C} \mid |R(z)| \leqslant 1\}.$$

Заметим, что устойчивость схемы для задачи (4.4) имеет место при условии, что $\tau\lambda = z \in \mathcal{S}$. Поэтому $\mathcal{S}$ называется областью устойчивости схемы.

Следующий результат, используемый нами ниже, называется теоремой максимального модуля.

Теорема 4.1 *[8, Глава I.2] Пусть φ — непостоянная комплексная функция, аналитическая в области $\mathcal{D} \subset \mathbb{C}$ и непрерывная на её замыкании. Тогда максимум φ достигается на границе $\partial\mathcal{D}$ области $\mathcal{D}$. В частности, если φ рациональная и не имеет полюсов в $\mathbb{C}^- = \{z \in \mathbb{C} \mid \operatorname{Re} z \leqslant 0\}$, то* $\max_{z\in\mathbb{C}^-} |\varphi(z)| = \max_{y\in\mathbb{R}} |\varphi(iy)|$.

Схема интегрирования по времени называется A-устойчивой, если её область устойчивости содержит левую комплексную полуплоскость $\mathbb{C}^- = \{z \in \mathbb{C} \mid \operatorname{Re} z \leqslant 0\}$. A-устойчивость означает, что схема, применённая к решению задачи (4.4) с $\lambda \in \mathbb{C}^-$, устойчива безусловно (т.е. для любых $\tau > 0$).

Вопрос состоит в том, возможно ли (и в каком случае) подобные рассуждения об устойчивости для скалярной тестовой задачи (4.4) обобщить для задачи (2.1) с ненормальной матрицей[2]. Следующая теорема даёт ответ на в этот вопрос.

Теорема 4.2 *[16, Глава IV.11] (Теорема Джона фон Неймана) Пусть $R(z)$ — рациональная функция, ограниченная в $\mathbb{C}^- = \{z \in \mathbb{C} \mid \operatorname{Re} z \leqslant 0\}$, и пусть $A \in \mathbb{C}^{n\times n}$ — такая, что*

$$\operatorname{Re}(y, Ay) \geqslant 0, \qquad \forall y \in \mathbb{C}^n.$$

Тогда для матричной нормы, соответствующей скалярному произведению в вышестоящем неравенстве, выполняется

$$\|R(-\tau A)\| \leqslant \max_{z\in\mathbb{C}^-} |R(z)|. \tag{4.5}$$

Доказательство [16, Глава IV.11] Для простоты обозначений в неравенстве, которое надо доказать, заменим $-\tau A$ на A. В этих новых обозначениях требуется показать, что $\|R(A)\| \leqslant \max_{z\in\mathbb{C}^-} |R(z)|$. Имеем

$$\operatorname{Re}(y, Ay) \leqslant 0, \qquad \forall y \in \mathbb{C}^n.$$

[2]Если A близка к нормальной или если имеется информация о (псевдо)спектре A, то могут быть использованы оценки на $\|R(-\tau A)\|$, подобные показанным в конце Секции 3.2.

Предположим, что A — ненормальная (иначе доказательство оставляется читателю, см. Упражнение 4.2). Для $\alpha \in \mathbb{C}$ введём

$$A(\alpha) = \frac{\alpha}{2}(A + A^*) + \frac{1}{2}(A - A^*).$$

Заметим, что $A(1) = A$. Нетрудно увидеть, что

$$(v, A(\alpha)v) = \bar{\alpha}\,\mathrm{Re}(v, Av) + i\,\mathrm{Im}(v, Av). \tag{4.6}$$

Тогда

$$\mathrm{Re}(y, A(\alpha)y) \leqslant 0, \qquad \forall y \in \mathbb{C}^n,$$

выполняется, коль скоро $\mathrm{Re}\,\alpha \geqslant 0$. Следовательно, для $\mathrm{Re}\,\alpha \geqslant 0$ собственные значения $A(\alpha)$ также имеют неположительную вещественную часть. Поэтому рациональная функция

$$\varphi(\alpha) = \|R(A(\alpha))v\|^2,$$

где v фиксирован, не имеет полюсов в $\mathrm{Re}\,\alpha \geqslant 0$. По Теореме 4.1 отсюда вытекает, что

$$\begin{aligned}\|R(A)v\|^2 = \varphi(1) &\leqslant \max_{y\in\mathbb{R}} \varphi(iy) = \max_{y\in\mathbb{R}} \|R(A(iy))v\|^2 \\ &\leqslant \max_{y\in\mathbb{R}} \|R(A(iy))\|^2 \|v\|^2.\end{aligned}$$

Нетрудно проверить, что матрица $A(iy)$ — нормальная. Поскольку доказываемый результат выполняется для нормальных матриц (см. Упражнение 4.2), имеем

$$\|R(A(iy))\| \leqslant \max_{z\in\mathbb{C}^-} |R(z)| \qquad \forall y \in \mathbb{R},$$

откуда следует

$$\|R(A)v\|^2 \leqslant \left(\max_{z\in\mathbb{C}^-} |R(z)|\right)^2 \|v\|^2.$$

□

Упражнение 4.2 Докажите 4.2 для нормальных матриц A. ◊

Упражнение 4.3 Проверьте соотношение (4.6) и нормальность матрицы $A(iy)$. ◊

Чтобы оценить силу Теоремы 4.2 и элегантность её доказательства, попробуем установить подобный результат для одной конкретной схемы интегрирования по времени, а именно, для θ-метода. Нетрудно увидеть, что θ-метод, применённый к задаче (2.1) с $g \equiv 0$, можно записать как

$$y^{k+1} = R(-\tau A)y^k, \qquad R(Z) = (I - \theta Z)^{-1}(I + (1-\theta)Z),$$

где введено обозначение $Z = -\tau A$.

Упражнение 4.4 Проверьте, что для любой квадратной матрицы Z выполняется

$$(I-\theta Z)^{-1}(I+(1-\theta)Z)=(I+(1-\theta)Z)(I-\theta Z)^{-1}.$$

$\Diamond$

Для выбранного скалярного произведения и соответствующих ему векторной и операторной матричной норм можем записать

$$\frac{\|y^{k+1}\|^2}{\|y^k\|^2}=\frac{\|R(Z)y^k\|^2}{\|y^k\|^2}=\frac{\|(I+(1-\theta)Z)(I-\theta Z)^{-1}y^k\|^2}{\|y^k\|^2}=\frac{\|(I+(1-\theta)Z)u\|^2}{\|(I-\theta Z)u\|^2},$$

где $u=(I-\theta Z)^{-1}y^k$. Последнее соотношение можно представить в виде

$$\frac{\|y^{k+1}\|^2}{\|y^k\|^2}=\frac{1+2(1-\theta)\operatorname{Re}(v,Zv)+(1-\theta)^2\|Zv\|^2}{1-2\theta\operatorname{Re}(v,Zv)+\theta^2\|Zv\|^2}=|R(\zeta)|^2, \tag{4.7}$$

где $v=u/\|u\|$ и $\zeta=\operatorname{Re}(v,Zv)+i\sqrt{\|Zv\|^2-(\operatorname{Re}(v,Zv))^2}$.

Упражнение 4.5 Покажите, что (4.7) выполняется для данной ζ. $\Diamond$

Таким образом, мы получаем

$$\|R(-\tau A)\|=|R(\zeta)|,$$

где ζ определена выше. Получим оценку для $\|R(-\tau A)\|$ локализацией ζ. Учитывая (3.9), естественно предположить, что

$$\mu_2(-A)\leqslant-\omega, \tag{4.8}$$

где $\mu_2(\cdot)$ задана для той же самой операторной матричной нормы.

Упражнение 4.6 Предложите условие на ω, θ и τ, достаточное для того, чтобы все собственные значения матрицы $I+\theta\tau A$ имели бы положительную вещественную часть. Заметьте, что при выполнении этого условия матрица $I+\theta\tau A$ невырождена. $\Diamond$

Используя (4.8), в оценке, полученной выше, имеем

$$\operatorname{Re}\zeta=\operatorname{Re}(v,Zv)=\tau\operatorname{Re}(v,-Av)\leqslant\tau\mu_2(-A)\leqslant-\tau\omega,$$

так что, в силу теоремы максимального модуля, выполняется

$$\|R(-\tau A)\|\leqslant\max_{\operatorname{Re}\zeta\leqslant-\tau\omega}|R(\zeta)|=\max\big\{|R(-\tau\omega)|,\underbrace{\lim_{z\to\infty}|R(z)|}_{1-1/\theta}\big\}.$$

Мы доказали следующий результат.

Теорема 4.3 *[8, Глава I.2] Пусть $\|\cdot\|$ — векторная или операторная матричная норма, соответствующая выбранному скалярному произведению, и*

пусть $A \in \mathbb{C}^{n\times n}$ *такая, что* $\mu_2(-A) \leqslant -\omega$ *для* ω*, удовлетворяющей условию, полученному в Упражнении 4.6. Тогда для функции устойчивости* θ*-метода* (4.3)

$$R(z) = \frac{1+(1-\theta)z}{1-\theta z}$$

выполняется

$$\|R(-\tau A)\| \leqslant \max_{\mathrm{Re}\, z \leqslant -\tau\omega} |R(z)| = \max\bigl\{|R(-\tau\omega)|, 1-\frac{1}{\theta}\bigr\}.$$

Упражнение 4.7 Покажите, что оценка (4.8) для логарифмической нормы выполняется тогда и только тогда, когда

$$\mathrm{Re}(v, Av) \geqslant \omega\|v\|^2, \qquad \forall v \in \mathbb{C}^n.$$

$\Diamond$

Заметим, что результаты устойчивости в этом разделе получены для θ-метода для норм, заданных скалярным произведением. Получение результатов устойчивости для этой схемы в других нормах является достаточно сложной задачей, за исключением случая $\theta = 1$ [8, Глава I.2]. Для неявной схемы Эйлера ($\theta = 1$) результаты устойчивости получаются легче, и, кроме того, некоторые результаты выполняются только для $\theta = 1$. Например, требуя для функции устойчивости θ-метода условие

$$\|R(-\tau A)\|_* \leqslant 1, \qquad \text{где } * = 1 \text{ or } * = \infty,$$

получаем, что должно выполняться $\theta = 1$ [16, Глава IV.11].

5 Операторное расщепление

5.1 Краткое введение в методы расщепления

В изложении материала этого раздела мы следовали [8, Глава IV.1]. Одним из очень полезных подходов в численных методах являются так называемые методы операторного расщепления. Для ознакомления с этим подходом предположим, что решается задача (2.1) без источника ($g \equiv 0$), и пусть[3]

$$A = A_1 + A_2.$$

Если системы уравнений $y' = -A_1 y(t)$ и $y' = -A_1 y(t)$ могут быть решены легче, чем $y' = -Ay(t)$, то решение задачи (2.1) за один шаг по времени, а именно

$$y^1 \approx y(\tau) = e^{-\tau A} y^0, \tag{5.1}$$

[3]Здесь матрицу A_1 не следует путать с кососимметрической частью матрицы A.

мы можем получить приближённо, делая шаг по времени для системы с матрицей A_1, а затем — для системы с A_2. Фактически мы последовательно решаем две задачи Коши

$$\begin{aligned} \tilde{y}' &= -A_1\tilde{y}(t), \quad \text{for } t \in [0,\tau] \quad \text{где } \tilde{y}(0) = y^0, \\ \hat{y}' &= -A_2\hat{y}(t), \quad \text{for } t \in [0,\tau] \quad \text{где } \hat{y}(0) = \tilde{y}(\tau), \end{aligned} \tag{5.2}$$

где результат $\tilde{y}(\tau)$ первой задачи является входными данными $\hat{y}(0)$ для второй задачи. Такая процедура расщепления (5.2), повторяемая на каждом шаге по времени $k = 2, 3, \ldots$, является простейшим методом расщепления и называется последовательным расщеплением.

Предположим теперь, что y^1 в (5.1) и решения задач (5.2) вычисляются точно, что означает

$$y^1 = e^{-\tau A}y^0, \qquad y^1_{\text{split}} = e^{-\tau A_2}e^{-\tau A_1}y^0.$$

Сравнивая точное решение y^1 с решением метода расщепления y^1_{split}, мы видим, что

$$\begin{aligned} e^{-\tau A} &= I + \tau(-A_1 - A_2) + \frac{\tau^2}{2}(A_1 + A_2)^2 + \ldots, \\ e^{-\tau A_2}e^{-\tau A_1} &= I + \tau(-A_1 - A_2) + \frac{\tau^2}{2}(A_1^2 + 2A_2A_1 + A_2^2) + \ldots. \end{aligned}$$

Следовательно, расщепление привносит дополнительную ошибку на каждом шаге. Если шаг стартует с точными исходными данными, то ошибка, привнесённая за один шаг по времени, называется локальной ошибкой. Для локальной ошибки в последовательном расщеплении получаем

$$\left(e^{-\tau A} - e^{-\tau A_2}e^{-\tau A_1}\right) y^0 = \frac{\tau^2}{2}(A_1A_2 - A_2A_1) + O(\tau^3).$$

Поскольку локальная ошибка — $O(\tau^2)$, глобальная ошибка, т.е. ошибка, накопленная после всех сделанных шагов по времени, — на порядок ниже, $O(\tau)$. Таким образом, последовательное расщепление (5.2) имеет первый порядок точности. Введём обозначение

$$[A_1, A_2] = A_1A_2 - A_2A_1,$$

называемое коммутатором A_1 и A_2.

Упражнение 5.1 Покажите, что если матрицы A_1 и A_2 диагонализируемые и коммутируют, то

$$e^{-\tau A_2}e^{-\tau A_1} = e^{-\tau A_2 - \tau A_1} = e^{-\tau A}. \tag{5.3}$$

Таким образом, последовательное расщепление в этом случае является точным. Если же A_1 и A_2 коммутируют, но необязательно диагонализируемые, то (5.3) по-прежнему выполняется, что можно установить степенным разложением матричных экспонент. $\Diamond$

Иногда для двух некоммутирующих матриц бывает полезно выразить произведение их экспонент как матричную экспоненту какой-то одной матрицы. Другими словами, надо для заданных $A_{1,2}$ найти такую матрицу $\tilde{A}$, что

$$e^{\tau A_2} e^{\tau A_1} = e^{\tau \tilde{A}}.$$

Такая $\tilde{A}$ определяется формулой Бейкера–Кемпбелла–Хаусдорфа:

$$\begin{aligned}\tilde{A} =& (A_1 + A_2) + \frac{\tau}{2}[A_2, A_1] + \frac{\tau^2}{12}\left([A_2, [A_2, A_1]] + [A_1, [A_1, A_2]]\right) + \\ &+ \frac{\tau^3}{24}[A_2, [A_1, [A_1, A_2]]] + O(\tau^4).\end{aligned}$$

Здесь выражения для членов высших порядков достаточно громоздки, но могут быть получены рекурсивно [17].

5.2 Расщепления второго порядка

Точность последовательного расщепления (5.2) можно улучшить, если повторить шаги расщепления в обратном порядке:
(1) шаг для задачи с A_1;
(2) шаг для задачи с A_2;
(3) шаг для задачи с A_2;
(4) шаг для задачи с A_1.
Предполагая, что расщеплённые задачи решаются точно, мы можем представить решение такого метода расщепления на шаге $k = 1$ в виде

$$y^1_{\text{split}} = e^{-\frac{\tau}{2}A_1} e^{-\frac{\tau}{2}A_2} e^{-\frac{\tau}{2}A_2} e^{-\frac{\tau}{2}A_1} y^0 = e^{-\frac{\tau}{2}A_1} e^{-\tau A_2} e^{-\frac{\tau}{2}A_1} y^0. \tag{5.4}$$

После некоторых преобразований можно показать, что

$$\left(e^{-\tau A} - e^{-\frac{\tau}{2}A_1} e^{-\tau A_2} e^{-\frac{\tau}{2}A_1}\right) y^0 = \frac{\tau^3}{24}\left([A_1, [A_1, A_2]] + 2[A_2, [A_1, A_2]]\right) y(\tau/2) + O(\tau^5),$$

что показывает второй порядок точности. Этот метод расщепления был предложен в 1968 г. независимо Г.И. Марчуком [18] и Г. Стренгом [19]. Поэтому назовём (5.4) расщеплением Марчука–Стренга.

Другое расщепление второго порядка точности, предложенное в 1963 г. Стренгом [20], имеет для шага $k = 1$ вид

$$y^1_{\text{split}} = \frac{1}{2}\left(e^{-\tau A_1} e^{-\tau A_2} + e^{-\tau A_2} e^{-\tau A_1}\right) y^0. \tag{5.5}$$

Здесь снова для простоты изложения мы предполагаем, что расщеплённые задачи решаются точно. Заметим, что вычисления $e^{-\tau A_1} e^{-\tau A_2} y^0$ и $e^{-\tau A_2} e^{-\tau A_1} y^0$ могут быть выполнены параллельно. По этой причине расщепление (5.5) называют параллельным или симметрично взвешенным .

5.3 Примеры расщепления

В предыдущей главе мы использовали матричные экспоненты e^{-tA_j}, $j = 1, 2$, только для того, чтобы представить разные методы расщепления в удобной компактной форме. На практике каждый шаг расщепления может быть выполнен любой подходящей схемой интегрирования. Разумеется, методы расщепления могут быть применены (и широко применяются) к любым системам дифференциальных уравнений

$$y'(t) = f(t, y(t)), \qquad f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n,$$

при условии, что определено расщепление $f(t, y) = f_1(t, y) + f_2(t, y)$. Благодаря своей алгоритмической простоте методы расщепления достаточно популярны. В сложных математических моделях они позволяют заниматься различными процессами модели независимо, в разных модулях программного обеспечения.

Чтобы подчеркнуть многосторонний характер и эффективность подхода расщепления, кратко перечислим несколько возможных методов расщепления, применяемых для решения по времени задач адвекции–диффузии (2.3).

1. θ-метод. См. Упражнение 5.2.
2. Методы расщепления по направлениям, где A_1 содержит все вклады частных производных по x, а $A_2 = A - A_1$. Хорошо известны методы расщепления этого типа, называемые методами переменных направлений [21] и локально-одномерными схемами [22].
3. Методы расщепления по физическим процессам, где, например, A_1 содержит вклады диффузионных, а A_2 — адвективных членов.
4. Специальные схемы расщепления, где для шагов расщепления, выполняемых неявными схемами, A_j выбираются так, что линейные системы с $I + \tau A_j$ легко решаются.

Упражнение 5.2 Покажите, что θ-метод можно рассматривать как метод последовательного расщепления, где шаг для задачи с $A_1 = (1 - \theta)A$ выполняется явной схемой Эйлера, а шаг для $A_2 = \theta A$ — неявной схемой Эйлера. $\Diamond$

5.4 Расщепления M-матрицами

Пусть $A \in \mathbb{R}^{n \times n}$ — матрица со слабым диагональным преобладанием, неотрицательными диагональными и неположительными внедиагональными элементами. По Теореме 1.3 A — (возможно вырожденная) M-матрица. В некоторых ситуациях, рассмотренных ниже, может быть полезным расщепление

$$A = M + N, \tag{5.6}$$

где M и N имеют те же свойства, что и A, т.е. M и N — матрицы со слабым диагональным преобладанием, неотрицательными диагональными и неположительными внедиагональными элементами. Подобные расщепления иногда называют репликативными [23]. Рассмотрим следующую явно-неявную схему численного решения (2.1):

$$\frac{y^{k+1}-y^k}{\tau} = -My^{k+1} - Ny^k + g^{k+1/2}. \tag{5.7}$$

Эту схему можно записать в виде

$$y^{k+1} = (I+\tau M)^{-1}(I-\tau N)y^k + \tau(I+\tau M)^{-1}g^{k+1/2}. \tag{5.8}$$

Рассмотрим следующий результат об устойчивости этой схемы.

Теорема 5.1 *[23] Пусть M и N — матрицы со слабым диагональным преобладанием, неотрицательными диагональными и неположительными внедиагональными элементами. Если $N \neq 0$, то схема интегрирования по времени* (5.8),(5.6) *с шагом по времени*

$$\tau \leqslant \frac{1}{\max_i n_{ii}} \tag{5.9}$$

устойчива, т.е.

$$\|(I+\tau M)^{-1}(I-\tau N)\|_\infty \leqslant \|(I+\tau M)^{-1}\|\|(I-\tau N)\|_\infty \leqslant 1, \tag{5.10}$$

$$\|y^{k+1}\|_\infty \leqslant \|y_k\|_\infty + \tau\|g^{k+1/2}\|_\infty, \tag{5.11}$$

и монотонна, т.е.

$$g^{k+1/2} \geqslant 0,\ l \leqslant m \quad \Rightarrow \quad y^k \geqslant 0,$$

где векторные неравенства понимаются поэлементно. Кроме того, если $A = M + N$ невырождена, то

$$\rho((I+\tau M)^{-1}(I-\tau N)) < 1.$$

Если $N = 0$, то устойчивость и монотонность схемы выполняются для всех $\tau > 0$.

Доказательство Для краткости записи, в доказательстве будем опускать индекс $\cdot_\infty$ в нормах. По Теореме 1.3, M — возможно вырожденная M-матрица, и, значит, $M = sI - B$, где $s = \max_i m_{ii} \geqslant \|B\| \geqslant \rho(B)$, а B — поэлементно

неотрицательная (см. доказательство Теоремы 1.3). Тогда

$$\begin{aligned}\|(I+\tau M)y\| &= \|(I+\tau sI-\tau B)y\| \geqslant |\|(1+\tau s)y\| - \|\tau By\|| \geqslant \\ &\geqslant \|(1+\tau s)y\| - \|\tau By\| = (1+\tau s)\|y\| - \tau\|By\| \geqslant \\ &\geqslant \|y\| + \tau s\|y\| - \tau\|B\|\|y\| = \|y\| + \tau(s-\|B\|)\|y\|, \\ \|(I+\tau M)^{-1}\| &= \max_{x\neq 0}\frac{\|(I+\tau M)^{-1}x\|}{\|x\|} = \max_{y\neq 0}\frac{\|y\|}{\|(I+\tau M)y\|} \leqslant \\ &\leqslant \max_{y\neq 0}\frac{\|y\|}{\|y\| + \tau(s-\|B\|)\|y\|} \leqslant 1.\end{aligned}$$

Нетрудно увидеть, что $\|I-\tau N\| \leqslant 1$ при условии, что выполняется (5.9). Таким образом, справедливо соотношение (5.10), и, следовательно, выполняется и (5.11). Монотонность следует из поэлементной неотрицательности матриц $(I+\tau M)^{-1}$ и $I-\tau N$. Наконец, оценка для спектрального радиуса следует из наблюдения, что $P-Q$, с $P = I+\tau M$ и $Q = I-\tau N$, является регулярным расщеплением M-матрицы τA (см. Теорему 1.4). □

Упражнение 5.3 (a) Покажите, что из (5.9) следует $\|(I-\tau N)\|_\infty \leqslant 1$.

(b) Проверьте, что если $g(t) \equiv 0$, то из (5.8), (5.6), (5.9) следует

$$\|y^{k+1}\|_\infty \leqslant \frac{1-\tau\min_i\sum_j n_{ij}}{1+\tau(s-\|B\|_\infty)}\|y^k\|_\infty.$$

◊

5.5 Уменьшаем ошибку расщепления: методы Розенброка

В некоторых прикладных задачах ошибка расщепления может быть достаточно велика, иногда неприемлемо велика. Это зачастую так, если собственные значения матриц A_1 и A_2 сильно отличаются по величине, как, например, в жёстких системах, где A_1 and A_2 могут иметь собственные значения, отличающиеся на порядки [24, 25]. В этом случае методам расщепления имеется альтернатива — так называемые методы Розенброка [26, 27]. Рассмотрим задачу Коши для системы линейных автономных дифференциальных уравнений

$$y'(t) = f(y), \qquad y(0) = y^0, \tag{5.12}$$

где $f : \mathbb{R}^n \to \mathbb{R}^n$ и $y^0 \in \mathbb{R}^n$ заданы. Двухстадийная схема Розенброка, называемая РОЗ2, имеет вид

$$\begin{aligned} y^{k+1} &= y^k + \frac{3}{2}k_1 + \frac{1}{2}k_2, \\ (I+\gamma\tau\hat{A})k_1 &= \tau f(y^k), \\ (I+\gamma\tau\hat{A})k_2 &= \tau f(y^k+k_1) - 2k_1, \end{aligned} \tag{5.13}$$

где $-\hat{A}$ — матрица, являющаяся приближением якобиана $f'(y^k)$, $\tau > 0$ — шаг по времени, а параметр $\gamma > 0$ определён ниже. Методы Розенброка в целом и схема РОЗ2 (5.13) в частности обладают примечательным свойством: их порядок точности не зависит от того, насколько хорошо $-\hat{A}$ приближает якобиан $f'(y^k)$. Применяя схему РОЗ2 к тестовой задаче Дальквиста, нетрудно получить функцию устойчивости схемы РОЗ2. Тогда для $-\hat{A} = f'(y^k)$ можно показать, что схема РОЗ2 A-устойчива для $\gamma \geqslant \frac{1}{4}$. Кроме того, можно показать, что схема РОЗ2 имеет второй порядок точности для любой матрицы $\hat{A}$ [26], [8, Глава IV.5.2]. Если $-\hat{A} = f'(y^k) + O(\tau)$, то двухстадийную схему РОЗ2 можно легко модифицировать так, что она будет иметь третий порядок точности для определённых значений γ [8, Глава IV.5.2], [28]. Заметим, что методы Розенброка, допускающие произвольные приближения к якобиану, также называются W-методами, см., например, [16, Глава IV.7].

Как отмечено выше, важным и привлекательным свойством методов Розенброка является то, что они могут использованы вместо схем расщепления. Действительно, поскольку $\hat{A}$ может быть выбрана произвольно, мы можем определить $\hat{A}$ как

$$I + \gamma\tau\hat{A} = (I + \gamma\tau A_1)(I + \gamma\tau A_2), \qquad \text{где} \quad A_1 + A_2 = -f'(y^k). \tag{5.14}$$

Для малых τ этот выбор означает приближение

$$I + \gamma\tau\hat{A} = I + \gamma\tau(A_1 + A_2) + (\gamma\tau)^2 A_1 A_2 = I - \gamma\tau f'(y^k) + O(\tau^2), \tag{5.15}$$

а, следовательно, соотношение (5.14) будем называть *приближённой матричной факторизацией* (ПМФ). Идея ПМФ предложена в работах [29, 30], она близка к классическим методам расщепления переменных направлений [21]. Комбинация схемы РОЗ2 с ПМФ, называемая РОЗ2-ПМФ, может быть применена вместо схемы расщепления с A_1 и A_2 [26, 31]. Можно образно сказать, что расщепление в схеме РОЗ2-ПМФ убрано с уровня схемы интегрирования по времени на уровень линейной алгебры (расщепление присутствует только при решении систем с $I + \gamma\tau A$). Схема РОЗ2-ПМФ с успехом применяется в решении задач адвекции–диффузии–реакции [26, 32, 25, 27].

Если A_1 и A_2 в (5.14) не коммутируют, то строгий анализ методов Розенброка в комбинации с ПМФ — сложная задача со многими нерешёнными вопросами (см., например, [26, 33, 34]). В частности, рассмотрим задачу адвекции–диффузии–реакции (5.12), где

$$f(y) = -A_{\text{adv}}y - A_{\text{diff}}y - R(y), \tag{5.16}$$

A_{adv} и A_{diff} — матрицы дискретных операторов адвекции и диффузии, соответственно (см. (2.6)), а R — оператор реакции. В численном решении таких задач обычным является предположение, что вещества из разных клеток вычислительной сетки химически не реагируют между собой. При таком предположении якобиан $R'(y)$ — блочно-диагональная матрица, где количество блоков равняется количеству клеток сетки. Применяя схему РОЗ2 (5.13)

с ПМФ (5.14) для решения (5.12),(5.16), разумно не включать адвективные члены в матрицу $\hat{A}$, поскольку эти члены обычно невелики (и называются «нежёсткими»). Следовательно, они не приводят к жёсткому ограничению на шаг по времени в явных схемах. Невключение адвективных членов в $\hat{A}$ и означает, что эти члены считаются схемой РОЗ2 явно. Расчёт адвективных членов явными схемами является стандартной практикой и, как правило, приводит к лучшей точности, чем расчёт неявными схемами.

Напротив, члены диффузии и реакции обычно «жёсткие», их необходимо считать неявными схемами. Следовательно, определим члены в (5.14) так:

$$\begin{gathered} I + \gamma\tau\hat{A} = (I + \gamma\tau A_1)(I + \gamma\tau A_2), \qquad \text{где} \\ A_1 = A_{\text{diff}},\ A_2 = R'(y^k),\ A_1 + A_2 = -f'(y^k) - A_{\text{adv}}. \end{gathered} \tag{5.17}$$

Для анализа устойчивости схемы РОЗ2-ПМФ (5.13),(5.17) можно было бы рассмотреть тестовую задачу, подобную задаче Дальквиста (4.4), т.е.

$$y'(t) = \lambda_{\text{adv}} y + \lambda_{\text{diff}} y + \lambda_{\text{react}} y.$$

Нетрудно заметить, что решаемая задача (5.12),(5.16) может быть сведена к этой тестовой задаче, если предположить, что члены реакции линейны, а все три оператора задачи (адвекции, диффузии, реакции) коммутируют. Разумеется, такое предположение нереалистично.

В качестве альтернативы можно провести анализ устойчивости для более простой схемы первого порядка, называемой РОЗ1 (одностадийная схема Розенброка)

$$y^{k+1} = y^k + k_1, \qquad (I + \tau\hat{A})k_1 = \tau f(y^k), \tag{5.18}$$

где $\tau > 0$ — шаг по времени. Легко проверить, что для $f(y) = -Ay$ схема РОЗ1 принимает вид

$$B\frac{y^{k+1} - y^k}{\tau} + Ay^k = 0, \qquad B = I + \tau\hat{A}. \tag{5.19}$$

Эта формула известна в русско-язычной научной литературе как каноническая двухуровневая разностная схема, см., например, [35, 36, 37]. Если $A = A^T$ и A — положительно определена и определены нормы

$$\|y\|_A = \sqrt{(Ay, y)}, \quad \|S\|_A^2 = \inf\left\{M \mid (ASy, Sy) \leqslant M(Ay, y)\ \forall y \in \mathbb{R}^n\right\},$$

то выполняется следующий результат устойчивости А.А. Самарского.

Теорема 5.2 *[35, 36, 37] Пусть $A = A^T \in \mathbb{R}^{n\times n}$ положительно определена и $B \in \mathbb{R}^{n\times n}$ — такая, что $B + B^T$ положительно определена. Схема интегрирования по времени* (5.19) *для решения системы $y'(t) = -Ay$ устойчива, а именно*

$$\|S\|_A \leqslant 1, \qquad S = B^{-1}(I - \tau A),$$

тогда и только тогда, когда

$$(Bx, x) \geqslant \frac{\tau}{2}(Ax, x), \quad \forall x \in \mathbb{R}^n.$$

Полезно рассмотреть несколько примеров, иллюстрирующих этот результат. Прежде всего, при $B = I$ схема РОЗ1 превращается в неявную схему Эйлера. По Теореме 5.2, схема устойчива тогда и только тогда, когда

$$(x, x) \geqslant \frac{\tau}{2}(Ax, x), \quad \forall x \in \mathbb{R}^n,$$

что, как нетрудно убедиться, эквивалентно

$$\tau \leqslant \frac{2}{\|A\|_2}.$$

Далее, РОЗ1 с $B = I + \tau A$ — это неявная схема Эйлера, для которой условие устойчивости

$$((I + \tau A)x, x) \geqslant \frac{\tau}{2}(Ax, x), \quad \forall x \in \mathbb{R}^n,$$

тривиально выполняется для всех $\tau > 0$. Наконец, θ-метод (4.3) соответствует выбору $B = I + \tau\theta A$ с условием устойчивости

$$((I + \tau\theta A)x, x) \geqslant \frac{\tau}{2}(Ax, x), \quad \forall x \in \mathbb{R}^n.$$

Его можно переписать в виде

$$1 + (\theta - \frac{1}{2})\tau\frac{(Ax, x)}{(x, x)} \geqslant 0, \quad \forall x \in \mathbb{R}^n,$$

выполняющемся для всех $\tau > 0$ при условии $\theta \geqslant \frac{1}{2}$ (что легко проверить, используя результаты Главы 4). Для $\theta < \frac{1}{2}$ условие устойчивости можно переписать в виде

$$\tau \geqslant \frac{2}{(1 - 2\theta)\|A\|_2}.$$

5.6 ПМФ+ — улучшаем матричную факторизацию

Для задач адвекции–диффузии–реакции схема РОЗ2-ПМФ обычно даёт более точные результаты, чем схемы расщепления [25, 27]. Однако, в некоторых случаях и ошибка РОЗ2-ПМФ может быть значительна. Это становится ясным, если рассмотреть соотношение (5.15). Обозначая якобиан диффузии–реакции $A_1 + A_2$ в (5.17) как A, получаем

$$\text{ошибка ПМФ} = I + \gamma\tau\hat{A} - (I + \gamma\tau A) = (\gamma\tau)^2 A_1 A_2.$$

Как видим, ошибка ПМФ мала *асимптотически* для $\tau \to 0$. В глобальной модели загрязнения воздуха ТМ5 [38] для типичных шагов по времени τ собственные значения τA_1 (диффузия) меняются в диапазоне от 10^{-5} до 10, а

собственные значения τA_2 (реакция) — от 10^{-5} до 10^6 [34]. Таким образом, хотя ПМФ и работает в данной задачи, нет оснований ожидать, что ПМФ даёт хорошое приближение к якобиану.

Как показано в [34], ПМФ можно улучшить в случае, если матрица диффузии A_1 имеет слабое столбцовое диагональное преобладание (т.е. элементы A_1^T удовлетворяют (1.4)) и существует LU разложение матрицы $I+\gamma\tau A_1$ (без перестановки строк или столбцов). Действительно, пусть $I + \gamma\tau A_1 = LU$, и, не нарушая общности, выберем LU разложение так, что диагональные элементы в нижней треугольной матрице L равны 1 (почему это возможно?). Рассмотрим следующую улучшенную приближённую матричную факторизацию, которую будем называть ПМФ+:

$$I + \gamma\tau\hat{A} = L(U + \gamma\tau A_2), \tag{5.20}$$

где $A_2 = R'(y^k)$ якобиан членов реакции. Тогда

$$\text{ошибка ПМФ+} = I + \gamma\tau\hat{A} - (I + \gamma\tau A) = \gamma\tau(L - I)A_2.$$

Применяя результат Упражнения 5.4 (приведённого ниже) к LU факторизации матрицы $I + \gamma\tau A_1$, мы видим, что при малых τ внедиагональные элементы в L суть $O(\tau)$. Следовательно, ошибка ПМФ+ — $O(\tau^2)$. Рассмотрим теперь, как ведёт себя ошибка ПМФ+ для реалистичных, бóльших значениях τ. Как мы увидим сейчас, L наследует от $I + \gamma\tau A_1$ слабое столбцовое диагональное преобладание, так что

$$\|L - I\|_1 \leqslant 1.$$

Таким образом, получаем

$$\|\text{ошибка ПМФ+}\|_1 \leqslant \gamma\tau\|A_2\|_1.$$

Таким образом, ошибка ПМФ+ имеет порядок τ^2 для малых τ, но, в отличие от ПМФ, растёт с τ не быстрее, чем линейно.

Упражнение 5.4 Предположим, что матрица $A \in \mathbb{R}^{n\times n}$ такова, что существует LU разложение матрицы $I + \tau A = LU$, где L и U — соответственно, нижняя и верхняя треугольная матрицы. Пусть $l_{ii} = 1$, $i = 1, \ldots, n$. Показать, что для всех внедиагональных элементов L выполняется $l_{ij} = O(\tau)$. Подсказка: используйте математическую индукцию по размерности матрицы n. $\Diamond$

Сформулируем и докажем следующую теорему.

Теорема 5.3 *Пусть $A \in \mathbb{R}^{n\times n}$ имеет слабое столбцовое диагональное преобладание и пусть имеется LU разложение $A = LU$, где L и U — соответственно нижняя и верхняя треугольные матрицы. Пусть $l_{ii} = 1$,*

$i = 1, \ldots, n$. Тогда L также имеет слабое столбцовое диагональное преобладание:

$$\sum_{i=j+1}^{n} |l_{ij}| \leqslant |l_{jj}| = 1.$$

Доказательство Доказательство проведём индукцией по размерности n. Нетрудно проверить, что утверждение справедливо для $n = 2$. Предполагая справедливость для $n - 1$, представим $A \in \mathbb{R}^{n\times n}$ в виде

$$A = \begin{bmatrix} a_{11} & a_U^T \\ a_L & A_{n-1} \end{bmatrix}, \qquad a_L, a_U \in \mathbb{R}^{n-1},\ A_{n-1} \in \mathbb{R}^{(n-1)\times(n-1)}.$$

Аналогично для матриц LU разложения A имеем

$$L = \begin{bmatrix} 1 & 0 \\ l & L_{n-1} \end{bmatrix}, \qquad U = \begin{bmatrix} a_{11} & a_U^T \\ 0 & U_{n-1} \end{bmatrix}.$$

Первый столбец L имеет элементы $l_{i1} = a_{i1}/a_{11}$, $i = 2, \ldots, n$, а, следовательно,

$$\sum_{i=2}^{n} |l_{i1}| = \frac{1}{|a_{11}|} \sum_{i=2}^{n} |a_{i1}| \leqslant \frac{1}{|a_{11}|} |a_{11}| = 1.$$

Таким образом, имеется диагональное преобладание в первом столбце L. Кроме того,

$$L_{n-1}U_{n-1} = A_{n-1} - la_U^T,$$

так что, по предположению индукции, L_{n-1} имеет слабое диагональное преобладание по столбцам при условии, что слабое диагональное преобладание по столбцам имеется у $A_{n-1} - la_U^T$. Проверка того, что $A_{n-1} - la_U^T$ действительно обладает этим свойством, оставлено читателю в качестве упражнения. □

Упражнение 5.5 Завершите доказательство Теоремы 5.3. ◊

Предположение о диагональном преобладании *по столбцам* в этом разделе рассматривается потому, что этим свойством матрица диффузии A_1 в глобальной модели TM5 [27]. Если же A_1 имеет слабое диагональное преобладание *по строкам*, то следует изменить определение ПМФ+ в (5.20) так:

$$I + \gamma\tau\hat{A} = (L + \gamma\tau A_2)U,$$

где $LU = I + \gamma\tau A_1$ — LU разложение с $u_{kk} = 1$, $k = 1, \ldots, n$. Для этой модифицированной факторизации выполняются соотношения $\|U - I\|_\infty \leqslant 1$ и $\|\text{ошибка ПМФ+}\|_\infty \leqslant \gamma\tau\|A_2\|_\infty$.

6 Методы подпространства Крылова для вычисления действия матричной экспоненты на вектор

6.1 Подпространство Крылова и полиномы на матрице

В случае, если источник g равен нулю, решение (2.1) даётся

$$y(t) = e^{-tA}y^0. \tag{6.1}$$

Приближённое действие оператора матричной экспоненты на вектор y_0 может быть вычислено, используя методику подпространств А.Н. Крылова [39, 40, 41, 42, 43] следующим образом. Используя так называемый модифицированный процесс Грама–Шмидта, нетрудно вычислить такие матрицы $V_{k+1} \in \mathbb{R}^{n\times(k+1)}$ и верхнюю хессенбергову[4] $H_{k+1,k} \in \mathbb{R}^{(k+1)\times k}$, что [10, 44]

$$V_{k+1} = \begin{bmatrix} v_1 & \dots & v_{k+1} \end{bmatrix}, \quad V_{k+1}^T V_{k+1} = I \in \mathbb{R}^{(k+1)\times(k+1)},$$
$$\operatorname{colspan}(V_{k+1}) = \operatorname{span}(y^0, Ay^0, \dots, A^k y^0)$$

и

$$AV_k = V_{k+1}H_{k+1,k} = V_k H_{k,k} + h_{k+1,k} v_{k+1} e_k^T, \tag{6.2}$$

где v_i — i-й столбец матрицы V_k, $v_1 = y^0/\|y^0\|_2$, $H_{k,k}$ — это матрица $H_{k+1,k}$ без последней строки, а $e_k = (0, \dots, 0, 1)^T \in \mathbb{R}^k$. Подпространство, натянутое на столбцы V_k, называется подпространством А.Н. Крылова и обозначается $\mathcal{K}_k(A, y^0)$:

$$\mathcal{K}_k(A, y^0) \equiv \operatorname{span}(y^0, Ay^0, \dots, A^{k-1}y^0).$$

Используя построенные V_k и $H_{k,k}$, мы можем вычислить приближённо (6.1) как

$$y(t) = e^{-tA}y^0 = \beta e^{-tA} V_k e_1 \approx \beta V_k e^{-tH_{k,k}} e_1, \tag{6.3}$$

где $e_1 = (1, 0, \dots, 0)^T \in \mathbb{R}^k$, а $\beta = \|y^0\|_2$. Следуя [45], приведём несколько аргументов, объясняющих, почему это может быть хорошей аппроксимацией произведения матричной экспоненты на вектор.

Во-первых, выполняется следующее утверждение.

Лемма 6.1 [45] *Пусть матрица $V_k \in \mathbb{R}^{n\times k}$ и верхняя хессенбергова матрица $H_{k,k} \in \mathbb{R}^{k\times k}$ — матрицы, определённые выше. Тогда для любого многочлена p_j степени $j \leqslant k-1$ справедливо*

$$p_j(A)v_1 = V_k p_j(H_k)e_1,$$

где v_1 и e_1 определены выше.

Доказательство [45] Пусть $\pi_k = V_k V_k^T$. Докажем по индукции, что $A^j v_1 = V_k H_k^j e_1$, $j = 0, 1, \dots, k-1$. Для $j = 0$ имеем $v_1 = V_k e_1$, и, следовательно, соотношение для $j = 0$ верно. Предполагая теперь, что оно выполняется

[4] Матрица $H = (h_{ij})$ называется верхней хессенберговой, если $h_{ij} = 0$ для $i > j + 1$.

для некоторого $j \leqslant k-2$, рассмотрим соотношение для $j+1$. Заметим, что $A^{j+1}v_1, A^j v_1 \in \mathcal{K}_k(A, y_0)$. Тогда получаем

$$A^{j+1}v_1 = \pi_k A^{j+1}v_1 = \pi_k A A^j v_1 = \pi_k A \pi_k A^j v_1 = V_k H_k V_k^T A^j v_1 = \\ = V_k H_k V_k^T V_k H_k^j e_1 = V_k H_k^{j+1} e_1.$$

□

Во-вторых, напомним, что если ν — степень минимального многочлена A, то любая степень A — это многочлен от A степени не выше $\nu - 1$.

В-третьих, справедлив следующий фундаментальный результат (доказательство см. в [46]).

Теорема 6.1 [46] *Пусть $A \in \mathbb{R}^{n\times n}$ имеет минимальный многочлен степени ν. Тогда для любой функции f аналитической в открытой области, содержащей спектр $\Lambda(A)$ матрицы A, выполняется*

$$f(A) = p_{\nu-1}(A),$$

где $p_{\nu-1}$ интерполирует f на $\Lambda(A)$ в эрмитовом смысле с собственными значениями, взятыми столько раз, какова их кратность[5].

Предположим теперь, что все поддиагональные элементы $H_{k,k}$ ненулевые, т.е., $h_{j+1,j} \neq 0$, $j = 1, \ldots, k-1$. (Иначе, если для какого-то k $h_{k+1,k} = 0$, то столбцы V_k являются базисом инвариантного подпространства A.) Следовательно, геометрическая кратность всех собственных значений матрицы $H_{k,k}$ равна 1, а её минимальный многочлен совпадает с характеристическим многочленом. Поэтому

$$e^{H_{k,k}} = p_{k-1}(H_{k,k}), \tag{6.4}$$

где p_{k-1} — единственно определённый многочлен степени $k-1$, интерполирующий функцию экспоненты на спектре $\Lambda(H_{k,k})$ в эрмитовом смысле с собственными значениями, взятыми столько раз, какова их кратность.

Все эти рассуждения приводят нас к следующему результату.

Теорема 6.2 *[45] Для аппроксимации* (6.3) *выполняется*

$$\beta V_k e^{-tH_{k,k}} e_1 = p_{k-1}(-tA)y^0,$$

где p_{k-1} — многочлен, определённый в (6.4).

Доказательство

$$\beta V_k e^{-tH_{k,k}} e_1 = \beta V_k p_{k-1}(-tH_{k,k}) e_1 \overset{\text{(Лемма 6.1)}}{=} \beta p_{k-1}(-tA)v_1 = p_{k-1}(-tA)y^0.$$

□

Заметим, что собственные значения $H_{k,k}$ называются числами Ритца матрицы A и имеется большое количество результатов, объясняющих, почему и насколько хорошо собственные значения A аппроксимируются числами Ритца при растущем k (см., например, [10, 44]).

[5] Многочлен p интерполирует функцию f в эрмитовом смысле в точке x, взятой l раз, если $f^{(j)}(x) = p^{(j)}(x)$, $j = 0, \ldots, l-1$.

6.2 Альтернативный вывод приближения

Следуя работам [42, 47, 48], закончим этот раздел альтернативным обоснованием аппроксимации (6.3). Предположим, мы приближённо решаем (2.1) с нулевым источником g, проецируя задачу (2.1) в галёркинском смысле на крыловское подпространство $\mathrm{colspan}V_k$. Это означает, что мы ищем такое приближённое решение $y_k(t) \approx y(t)$, что

$$y_k(t) = V_k u(t) \quad \text{и} \quad r_k(t) \perp \mathrm{colspan}V_k, \tag{6.5}$$

где $r_k(t)$ — невязка решения $y_k(t)$, определяемая как [42, 47, 48]

$$r_k(t) = -y_k'(t) - Ay_k(t).$$

Подставляя $y_k(t) = V_k u(t)$ в $y' = -Ay(t)$ и учитывая, что $V_k^T V_k$ — единичная матрица, получаем спроецированную задачу Коши для функции $u(t)$:

$$u'(t) = -\underbrace{V_k A V_k}_{H_{k,k}} u(t), \qquad u(0) = \beta e_1. \tag{6.6}$$

Здесь все обозначения такие же, как в начале главы. Заметим, что $u(t) = \beta e^{-tH_{k,k}} e_1$. Далее, используя (6.2), мы можем получить выражение для невязки $r_k(t)$, позволяющее контролировать качество приближённого решения $y_k(t)$. Действительно [42, 47],

$$\begin{aligned} r_k(t) &= -y_k'(t) - Ay_k(t) = -V_k u'(t) - AV_k u(t) = (V_k H_{k,k} - AV_k)u(t) \\ &= (V_k H_{k,k} - V_{k+1}H_{k+1,k})u(t) = -h_{k+1,k} v_{k+1} e_k^T u(t) = -h_{k+1,k} v_{k+1} e_k^T e^{-tH_{k,k}} u(0) = \\ &= -h_{k+1,k} v_{k+1} e_k^T e^{-tH_{k,k}} \beta e_1 = \underbrace{-h_{k+1,k} e_k^T e^{-tH_{k,k}} \beta e_1}_{\text{скалярная функция от } t} v_{k+1} \perp \mathrm{colspan}V_k. \end{aligned}$$

Эта невязка может быть использована для разных целей, см., например, [48, 49].

Для неоднородных задач (2.1), т.е., для ненулевого источника $g(t)$, аппроксимации матричной экспоненты на подпространствах Крылова могут быть применены в рамках так называемых экспоненциальных схем интегрирования по времени, см., например, [50]. Для неоднородных задач, можно также использовать проекцию на единственное *блочное* подпространство Крылова [51], подобно проекции (6.5),(6.6).

7 Благодарности

Благодарю организаторов Римско-Московской школы за приглашение прочитать лекции в школе, студентов школы за их интерес к лекциям и за указание мне ряда моих ошибок в первых вариантах этих конспектов (все оставшиеся ошибки — мои), Николая Замарашкина за ряд полезных замечаний по содержанию лекций. Особую благодарность выражаю бывшим коллегам Япу ван

дер Вегту (Jaap van der Vegt) и Герриту Звиру (Gerrit Zwier) по Университету Твенте (Нидерланды), сделавших возможным моё участие в школе.

Список литературы

[1] Horn R. A., Johnson C. R. Matrix Analysis. — Cambridge University Press, 1986. — Russian translation: Р. Хорн, Ч. Джонсон. Матричный анализ.— М.: Мир, 1989.

[2] Ortega J. M. Matrix theory. A second course. The University Series in Mathematics. — Plenum Press, New York, 1987. — P. xii+262. — ISBN: 0-306-42433-9. — URL: http://dx.doi.org/10.1007/978-1-4899-0471-3.

[3] Ortega J. M. Introduction to Parallel and Vector Solution of Linear Systems. — Plenum Press, 1988. — Russian translation: Дж. Ортега. Введение в параллельные и векторные методы решения линейных систем.—М.: Мир, 1991.

[4] Young D. M. Iterative Solution of Large Linear Systems. — Academic Press, 1971.

[5] Varga R. S. Matrix Iterative Analysis. — Prentice-Hall, 1962.

[6] Horn R. A., Johnson C. R. Topics in Matrix Analysis. — Cambridge University Press, 1991.

[7] Rose D. J. Convergent regular splittings for singular M-matrices // SIAM J. Algebraic Discrete Methods. — 1984. — Vol. 5, no. 1. — P. 133–144. — URL: http://dx.doi.org/10.1137/0605015.

[8] Hundsdorfer W., Verwer J. G. Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations. — Springer Verlag, 2003.

[9] Krukier L. A. Implicit difference schemes and an iterative method for solving them for a certain class of systems of quasi-linear equations // Sov. Math. — 1979. — Vol. 23, no. 7. — P. 43–55. — Translation from Izv. Vyssh. Uchebn. Zaved., Mat. 1979, No. 7(206), 41–52 (1979).

[10] Saad Y. Iterative Methods for Sparse Linear Systems. — 2d edition. — SIAM, 2003. — Available from http://www-users.cs.umn.edu/~saad/books.html.

[11] Sidje R. B. Expokit. A software package for computing matrix exponentials // ACM Trans. Math. Softw. — 1998. — Vol. 24, no. 1. — P. 130–156. — www.maths.uq.edu.au/expokit/.

[12] Higham N. J. Functions of Matrices: Theory and Computation. — Philadelphia, PA, USA : Society for Industrial and Applied Mathematics, 2008.

[13] Dekker K., Verwer J. G. Stability of Runge–Kutta methods for stiff non-linear differential equations. — North-Holland Elsevier Science Publishers, 1984. — Russian translation: К. Деккер, Я. Вервер. Устойчивость методов Рунге–Кутты для жёстких нелинейных дифференциальных уравнений.—М.: Мир, 1988.

[14] Lozinskiĭ S. M. Error estimate for numerical integration of ordinary differential equations. I // Izv. Vysš. Učebn. Zaved. Matematika 1958, no. 5 (6), 52-90; Translated as: Izvestija Vysših Učebnyh Zavedeniĭ Matematika. — 1959. — Vol. 1959, no. 5 (12). — P. 222.

[15] Crank J., Nicolson P. A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type // Proc. Camb. Philos. Soc. — 1947. — Vol. 43. — P. 50–67.

[16] Hairer E., Wanner G. Solving Ordinary Differential Equations II. Stiff and Differential–Algebraic Problems. Springer Series in Computational Mathematics 14. — 2 edition. — Springer–Verlag, 1996.

[17] Sanz-Serna J. M., Calvo M. P. Numerical Hamiltonian Problems. — Chapman & Hall, 1994.

[18] Marčuk G. I. Some application of splitting-up methods to the solution of mathematical physics problems // Apl. Mat. — 1968. — Vol. 13. — P. 103–132.

[19] Strang G. On the construction and comparison of difference schemes // SIAM J. Numer. Anal. — 1968. — Vol. 5, no. 3. — P. 506–517.

[20] Strang G. Accurate partial difference methods I: linear Cauchy problems // Archive for Rational Mechanics and Analysis. — 1963. — Vol. 12. — P. 392–402.

[21] Peaceman D. W., Rachford Jr. H. H. The numerical solution of parabolic and elliptic differential equations // J. Soc. Indust. Appl. Math. — 1955. — Vol. 3. — P. 28–41.

[22] Yanenko N. N. The method of fractional steps. The solution of problems of mathematical physics in several variables. — New York : Springer-Verlag, 1971. — P. viii+160.

[23] Bochev (Botchev) M. A. On the stability of nonselfadjoint difference schemes with M-matrices for evolution boundary value problems with an elliptic operator with respect to space // Izv. Vyssh. Uchebn. Zaved. Mat. — 1995. — Vol. 9. — P. 15–22.

[24] Sportisse B. An analysis of operator splitting techniques in the stiff case // J. Comput. Phys. — 2000. — Vol. 161, no. 1. — P. 140–168.

[25] Verwer J. G., Hundsdorfer W., Blom J. G. Numerical time integration for air pollution models // Surveys for Mathematics in Industry. — 2002. — Vol. 10. — P. 107–174.

[26] A second order Rosenbrock method applied to photochemical dispersion problems / Jan G. Verwer, Edwin J. Spee, Joke G. Blom, Willem Hundsdorfer // SIAM J. Sci. Comput. — 1999. — Vol. 20. — P. 456–480.

[27] Solving vertical transport and chemistry in air pollution models / P. J. F. Berkvens, M. A. Botchev, M. C. Krol et al. // Atmospheric Modeling / Ed. by D.P. Chock, G.R. Carmichael. — Springer, 2002. — Vol. 130 of IMA Volumes in Mathematics and its Applications. — P. 1–20.

[28] Lastdrager B., Koren B., Verwer J. G. Solution of time-dependent advection-diffusion problems with the sparse-grid combination technique and a Rosenbrock solver // Comput. Methods Appl. Math. — 2001. — Vol. 1, no. 1. — P. 86–99.

[29] D'yakonov E. G. Difference systems of second order accuracy with a divided operator for parabolic equations without mixed derivatives // USSR Comput. Math. Math. Phys. — 1964. — Vol. 4, no. 5. — P. 206–216.

[30] Beam R. M., Warming R. F. An implicit finite-difference algorithm for hyperbolic systems in conservation-law form // J. Comput. Phys. — 1976. — Vol. 22. — P. 87–110.

[31] van der Houwen P. J., Sommeijer B. P. Approximate factorization for time-dependent partial differential equations // J. Comput. Appl. Math. — 2001. — Vol. 128, no. 1-2. — P. 447–466. — Numerical analysis 2000, Vol. VII, Partial differential equations.

[32] Gerisch A., Verwer J. G. Operator splitting and approximate factorization for taxis-diffusion-reaction models // Appl. Numer. Math. — 2002. — Vol. 42. — P. 159–176.

[33] Ostermann A. Stability of W-methods with applications to operator splitting and to geometric theory // Appl. Numer. Math. — 2002. — Vol. 42, no. 1–3. — P. 353–366. — http://dx.doi.org/10.1016/S0168-9274(01)00160-X.

[34] Botchev M. A., Verwer J. G. A new approximate matrix factorization for implicit time integration in air pollution modeling // J. Comp. Appl. Math. — 2003. — Vol. 157. — P. 309–327. — http://dx.doi.org/10.1016/S0377-0427(03)00414-X.

[35] Samarskii A. A. Regularization of difference schemes // USSR Comput. Math. and Math. Phys. — 1967. — Vol. 7. — P. 62–93.

[36] Samarskii A. A. Theorie der Differenzenverfahren. — Leipzig : Akademische Verlagsgesellschaft Geest & Portig K.-G., 1984. — P. 356. — Translated from the Russian by Gisbert Stoyan.

[37] Samarskii A. A., Nikolaev E. S. Numerical methods for grid equations. Vol. I&II. — Basel : Birkhäuser Verlag, 1989. — P. xvi+502. — ISBN: 3-7643-2277-2.

[38] TM5: global chemistry transport model. — Wageningen University, the Netherlands. — 2016. — http://tm5.sourceforge.net/.

[39] Park T. J., Light J. C. Unitary quantum time evolution by iterative Lanczos reduction // J. Chem. Phys. — 1986. — Vol. 85. — P. 5870–5876.

[40] Druskin V. L., Knizhnerman L. A. Two polynomial methods of calculating functions of symmetric matrices // U.S.S.R. Comput. Maths. Math. Phys. — 1989. — Vol. 29, no. 6. — P. 112–121.

[41] Druskin V. L., Knizhnerman L. A. Krylov subspace approximations of eigenpairs and matrix functions in exact and computer arithmetic // Numer. Lin. Alg. Appl. — 1995. — Vol. 2. — P. 205–217.

[42] Celledoni E., Moret I. A Krylov projection method for systems of ODEs // Appl. Numer. Math. — 1997. — Vol. 24, no. 2-3. — P. 365–378.

[43] Hochbruck M., Lubich C. On Krylov subspace approximations to the matrix exponential operator // SIAM J. Numer. Anal. — 1997. — Oct. — Vol. 34, no. 5. — P. 1911–1925.

[44] van der Vorst H. A. Iterative Krylov methods for large linear systems. — Cambridge University Press, 2003.

[45] Saad Y. Analysis of some Krylov subspace approximations to the matrix exponential operator // SIAM J. Numer. Anal. — 1992. — Vol. 29, no. 1. — P. 209–228.

[46] Gantmacher F. R. The Theory of Matrices. Vol. 1. — AMS Chelsea Publishing, Providence, RI, 1998. — Translated from the Russian by K. A. Hirsch, Reprint of the 1959 translation.

[47] Druskin V. L., Greenbaum A., Knizhnerman L. A. Using nonorthogonal Lanczos vectors in the computation of matrix functions // SIAM J. Sci. Comput. — 1998. — Vol. 19, no. 1. — P. 38–54.

[48] Botchev M. A., Grimm V., Hochbruck M. Residual, restarting and Richardson iteration for the matrix exponential // SIAM J. Sci. Comput. — 2013. — Vol. 35, no. 3. — P. A1376–A1397. — http://dx.doi.org/10.1137/110820191.

[49] Botchev M. A., Oseledets I. V., Tyrtyshnikov E. E. Iterative across-time solution of linear differential equations: Krylov subspace versus waveform relaxation // Computers & Mathematics with Applications. — 2014. — Vol. 67, no. 12. — P. 2088–2098. — http://dx.doi.org/10.1016/j.camwa.2014.03.002.

[50] Hochbruck M., Ostermann A. Exponential integrators // Acta Numer. — 2010. — Vol. 19. — P. 209–286.

[51] Botchev M. A. A block Krylov subspace time-exact solution method for linear ordinary differential equation systems // Numer. Linear Algebra Appl. — 2013. — Vol. 20, no. 4. — P. 557–574.

Предметный указатель

Содержание