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

Препринты ИПМ • Препринт № 71 за 2019 г.

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

Бахвалов П.А., Сурначёв М.Д.

О приведении устойчивых матриц к блочно-диагональному виду

Рекомендуемая форма библиографической ссылки: Бахвалов П.А., Сурначёв М.Д. О приведении устойчивых матриц к блочно-диагональному виду // Препринты ИПМ им. М.В.Келдыша. 2019. № 71. 15 с. doi:10.20948/prepr-2019-71
URL: http://library.keldysh.ru/preprint.asp?id=2019-71

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.КЕЛДЫША
Р о с с и й с к о й а к а д е м и и н а у к

П. А. Бахвалов, М. Д. Сурначёв

О приведении устойчивых матриц к блочно-диагональному виду

Москва — 2019

Бахвалов П. А., Сурначёв М. Д.

О приведении устойчивых матриц к блочно-диагональному виду

Рассматриваются квадратные матрицы A, такие что $\|\exp(tA)\| \leqslant K$ для всех $t \geqslant 0$. Показывается, что любая такая матрица может быть приведена к блочно-диагональному виду, причём числа обусловленности и матрицы перехода, и всех диагональных блоков ограничены величиной, зависящей только от K и размера матрицы. Полученный результат применяется для исследования эффекта повышенной точности разностных схем в длительном счёте.

Ключевые слова: подобие матриц, теорема Крайса, суперсходимость

Pavel Alexeevisch Bakhvalov, Mikhail Dmitrievich Surnachev

On transformation of the stable matrices to a block-diagonal form

We consider square matrices A such that $\|\exp(tA)\| \leqslant K$ for all $t \geqslant 0$. We show that each matrix possessing this property can be transformed to a block-diagonal form such that condition numbers of all the digaonal blocks and the transformation matrix depend only on K and the matrix size. This result is applied to the analysis of long-time simulation accuracy of difference schemes.

Key words: similarity of matrices, Kreiss theorem, superconvergence, long-time simulation accuracy

Оглавление

1. Введение

Рассмотрим модельное уравнение переноса $\partial v/\partial t + \boldsymbol{\omega} \cdot \boldsymbol{\nabla} v = 0$, $\boldsymbol{\omega} \in \mathbb{R}^d$, в d-мерном единичном кубе с периодическими условиями и начальными условиями $v_0(\boldsymbol{r}) = \exp(i\boldsymbol{\alpha} \cdot \boldsymbol{r})$, $\boldsymbol{\alpha}/(2\pi) \in \mathbb{N}^d$. Некоторые разностные схемы для решения на равномерных сетках обладают свойством повышенной точности в длительном счёте, то есть наблюдается оценка ошибки вида

$$\|\varepsilon_h(t)\| \leqslant C_1 h^P |\boldsymbol{\alpha}|^P + C_2 h^Q t |\boldsymbol{\alpha}|^{Q+1}, \quad Q > P > 0, \tag{1}$$

где $\varepsilon_h(t)$ – разность между численным решением на сетке с шагом h в момент времени t и отображением на сетку точного решения на этот же момент времени. Например, для метода Галёркина с разрывными базисными функциями такая оценка наблюдается при $P = k+1$ и $Q = 2k+1$, где k – порядок используемых полиномов (доказательство в одномерном случае см. в [1], в двумерном на декартовой сетке – в [2]).

Обозначим через $\mathbb{C}^{n\times n}$ пространство матриц размера $n \times n$. На равномерной сетке после преобразования Фурье полудискретная схема общего вида сводится к системе из n уравнений, где n – число степеней свободы на одну ячейку, вида

$$Z_{\boldsymbol{\phi}} \frac{d\hat{u}_{\boldsymbol{\phi}}}{d\nu} + L_{\boldsymbol{\phi}} \hat{u}_{\boldsymbol{\phi}} = 0, \tag{2}$$

где $\nu = t/h$, а $\boldsymbol{\phi} = \boldsymbol{\alpha} h$ – волновое число, $Z_{\boldsymbol{\phi}}, L_{\boldsymbol{\phi}} \in \mathbb{C}^{n\times n}$. Если ввести $A_{\boldsymbol{\phi}} = i\boldsymbol{\omega} \cdot \boldsymbol{\phi} - Z_{\boldsymbol{\phi}}^{-1} L_{\boldsymbol{\phi}}$, то численное решение системы (2) примет вид

$$\hat{u}_{\boldsymbol{\phi}}(\nu) = e^{-i\nu\boldsymbol{\omega}\cdot\boldsymbol{\phi}} \exp(\nu A_{\boldsymbol{\phi}})\, \hat{u}_{\boldsymbol{\phi}}(0),$$

а его ошибка относительно точного решения – форму

$$\hat{\varepsilon}_{\boldsymbol{\phi}}(\nu) = e^{-i\nu\boldsymbol{\omega}\cdot\boldsymbol{\phi}} \left(\exp(\nu A_{\boldsymbol{\phi}}) - 1\right) \hat{u}_{\boldsymbol{\phi}}(0).$$

Образ Фурье ошибки аппроксимации на $v_0(\boldsymbol{r})$ равен $\hat{\epsilon}(\boldsymbol{\alpha} h)/h$, где

$$\hat{\epsilon}(\boldsymbol{\phi}) = A_{\boldsymbol{\phi}} \hat{u}_{\boldsymbol{\phi}}(0). \tag{3}$$

Будем считать, что схема устойчива с константой K, то есть при всех $\boldsymbol{\phi} \in \mathbb{R}^d$ выполняется

$$\sup_{\nu \geqslant 0} \|\exp(\nu A_{\boldsymbol{\phi}})\| \leqslant K. \tag{4}$$

Таким образом, оценка (1) переписывается в виде

$$\|\left(\exp(\nu A_{\boldsymbol{\phi}}) - 1\right) \hat{u}_{\boldsymbol{\phi}}(0)\| \leqslant \tilde{C}_1 + \tilde{C}_2 \nu, \tag{5}$$

где $\tilde{C}_1 = C_1 |\boldsymbol{\phi}|^P$, $\tilde{C}_2 = C_2 |\boldsymbol{\phi}|^{Q+1}$. Здесь и далее под $\|\cdot\|$ понимается евклидовая векторная норма и порождённая ей матричная норма.

Одним из способов доказательства оценки (1) является метод введения вспомогательного отображения (см. [1, 2] и др.). Его смысл заключается в следующем: если найдутся некоторые другие начальные данные для разностной задачи, отличающиеся от исходных на $O(h^P)$, на которых схема обладает порядком аппроксимации Q, то будет верна оценка (1). Формально он может быть описан следующей теоремой.

Теорема 1. *Пусть $A \in \mathbb{C}^{n\times n}$ удовлетворяет* (4)*, $v \in \mathbb{C}^n$. Тогда при всех $\nu \geqslant 0$ и $w \in \mathbb{C}^n$ выполняется*

$$\|(e^{\nu A} - I)v\| \leqslant (K+1)(\|v-w\| + \nu\|Aw\|).$$

Целью настоящей работы является доказательство обратного утверждения.

Теорема 2. *Пусть $A \in \mathbb{C}^{n\times n}$ удовлетворяет* (4)*, $v \in \mathbb{C}^n$. Пусть существуют такие $C_1, C_2 \geqslant 0$, что при всех $\nu \geqslant 0$ справедлива оценка*

$$\|(e^{\nu A} - I)v\| \leqslant (C_1 + C_2\nu)\|v\|.$$

Тогда для $w = (A^*A + \varepsilon^2)^{-1}\varepsilon^2 v$, где $\varepsilon = C_2/C_1$ и при $C_1 = 0$ или $C_2 = 0$ под w понимается соответствующее предельное значение, справедливы оценки

$$\|v - w\| \leqslant \delta C_1\|v\|, \quad \|Aw\| \leqslant \delta C_2\|v\|, \tag{6}$$

где δ зависит только от n и K.

Из теорем (1) и (2) можно получить следствие.

Следствие 3. *Пусть $A \in \mathbb{C}^{n\times n}$ удовлетворяет* (4)*, $v \in \mathbb{C}^n$. Пусть существуют такие $C_1, C_2 \geqslant 0$, что при всех $\nu \geqslant 0$ справедлива оценка*

$$\|(e^{\nu A} - I)v\| \leqslant (C_1 + C_2\nu)\|v\|. \tag{7}$$

Тогда

$$v^*\left(A^*A + \frac{C_2^2}{C_1^2}\right)^{-1} A^*Av \leqslant \tilde{\delta}^2 C_1^2\|v\|^2, \tag{8}$$

где $\tilde{\delta}$ зависит только от n и K. Обратно, если выполняется (8)*, то*

$$\|(e^{\nu A} - I)v\| \leqslant \tilde{\delta}(K+1)(C_1 + C_2\nu)\|v\|. \tag{9}$$

Доказательство теоремы 2 опирается на следующий результат. Всюду далее через $\varkappa(Y) = \|Y\|\,\|Y^{-1}\|$ будем обозначать число обусловленности матрицы Y.

Теорема 4. *Для любых $n \in \mathbb{N}$, $K \geqslant 1$ существует такое $C = C(n, K)$, что для любой матрицы $A \in \mathbb{C}^{n\times n}$, удовлетворяющей $\sup\limits_{\nu\geqslant 0} \|e^{\nu A}\| \leqslant K$, найдётся представление $A = SMS^{-1}$, где $\varkappa(S) \leqslant C$, а M – блочно-диагональная матрица, для каждого блока M_j которой выполняется либо $M_j = 0$, либо $\varkappa(M_j) \leqslant C$.*

Отметим, что для (косо)эрмитовых матриц A утверждение теоремы 4 вытекает из существования ортонормированного базиса из собственных векторов, при этом все блоки M_j имеют размер 1.

Оставшийся текст работы структурирован следующим образом. Вначале будут доказаны теоремы 1 и 2, вторая – с опорой на теорему 4, и следствие 3 из них. Затем будет доказана сама теорема 4.

2. Метод введения вспомогательного отображения

Докажем вначале теорему 1. Для любого $w \in \mathbb{C}^n$ имеем

$$\|(e^A - I)v\| \leqslant \|(e^A - I)w\| + \|(e^A - I)(w - v)\|.$$

Первое слагаемое в правой части оценим с использованием тождества

$$e^A - I = \int\limits_0^1 e^{\tau A} d\tau\ A,$$

а второе – напрямую. С учётом (4) получаем

$$\|(e^A - I)v\| \leqslant K\|Aw\| + (K + 1)\|w - v\| \leqslant (K + 1)(\|v - w\| + \|Aw\|),$$

что и требовалось доказать.

3. Существование вспомогательного отображения

Теперь докажем теорему 2. Воспользуемся разложением $A = SMS^{-1}$, даваемым теоремой 4. Домножив матрицу S на константу, можно считать, что $\|S^{-1}\| = 1$, тогда $\|S\| \leqslant C(n, K)$. Имеем

$$\|(e^{\nu A} - I)v\| = \|S(e^{\nu M} - I)S^{-1}v\| \leqslant (C_1 + C_2\nu)\|v\|,$$

откуда $\|(e^{\nu M} - I)S^{-1}v\| \leqslant (C_1 + C_2\nu)\|v\|$. Положим $y = S^{-1}v$. Обозначим блоки матрицы M через M_j и соответствующие им подвекторы вектора y через y_j.

Поскольку норма подвектора не может превосходить норму вектора, для всех j получаем

$$\|(e^{\nu M_j} - I)y_j\| \leqslant (C_1 + C_2\nu)\|v\|.$$

Матрица M_j либо невырожденная, либо нулевая. Рассмотрим вначале первый случай. Положим $\nu = 1/\|M_j\|$. Поскольку при $\|Y\| \leqslant 1$, $\det Y \neq 0$, справедливо $\|(e^Y - I)^{-1}\| \leqslant 4\|Y^{-1}\|$, подставляя $Y = M_j/\|M_j\|$ и $\|Y^{-1}\| = \varkappa(M_j)$, получаем

$$\|y_j\| \leqslant 4\varkappa(M_j)\left(C_1 + \frac{C_2}{\|M_j\|}\right)\|v\| \leqslant 4C(n,K)\left(C_1 + \frac{C_2}{\|M_j\|}\right)\|v\|.$$

Введём

$$x_j = \begin{cases} 0, & M_j \neq 0,\ \|y_j\| \leqslant 4C(n,K)C_1\|v\|; \\ y_j, & otherwise. \end{cases}$$

Составим из этих компонент вектор x и положим $u = Sx$. По построению для всех j имеем $\|x_j - y_j\| \leqslant 4C(n,K)C_1\|v\|$, откуда

$$\|x - y\| \leqslant 4\sqrt{n}C(n,K)C_1\|v\|$$

и, следовательно,

$$\|u - v\| = \|S(x - y)\| \leqslant 4\sqrt{n}(C(n,K))^2C_1\|v\|. \tag{10}$$

Далее, для таких компонент j, что $x_j \neq 0$, справедливо либо $M_j = 0$, либо $\|x_j\| = \|y_j\| \leqslant 4C(n,K)C_2\|v\|/\|M_j\|$. Отсюда

$$\|M_jx_j\| \leqslant 4C(n,K)C_2\|v\|$$

и, следовательно,

$$\|Au\| \leqslant 4\sqrt{n}(C(n,K))^2C_2\|v\|. \tag{11}$$

Пусть теперь $C_1, C_2 \neq 0$, то есть $\varepsilon = C_2/C_1$ конечно и не равно нулю. Рассмотрим задачу минимизации квадратичного функционала

$$F(w) = \varepsilon^2\|v - w\|^2 + \|Aw\|^2 \to \min$$

по $w \in \mathbb{C}^n$. Если подставить в функционал вектор u, то значение этого функционала будет равно

$$F(u) = \varepsilon^2\|u - v\|^2 + \|Au\|^2 \leqslant 2\left[4\sqrt{n}(C(n,K))^2C_2\|v\|\right]^2.$$

С другой стороны, минимум $F(w)$ достигается на векторе $w = (A^*A+\varepsilon^2)^{-1}\varepsilon^2 v$, поэтому $F(w) \leqslant F(u)$ и, следовательно,

$$\max\{\varepsilon\|w-v\|, \|Aw\|\} \leqslant 4\sqrt{2n}(C(n,K))^2 C_2\|v\|.$$

Полагая $\delta = 4\sqrt{2n}(C(n,K))^2$, получаем неравенства (6).

Остаётся рассмотреть вырожденные случаи. При $C_1 = 0$ имеем $w = u = v$, поэтому из (11) напрямую имеем (6). При $C_2 = 0$ также w совпадает с u и является ближайшим к v вектором из Ker$A$. При этом имеем $\|Aw\| = 0$, а $\|w - v\| \leqslant \delta C_1\|v\|$ следует из (10).

4. Доказательство следствия 3

Пусть выполняется (16). Обозначим $\varepsilon = C_2/C_1$ и $B = A^*A$. Пусть $w = (B+\varepsilon^2)^{-1}\varepsilon^2 v$. Тогда

$$\|v-w\|^2 = \|(I-(B+\varepsilon^2)^{-1}\varepsilon^2)v\|^2 = \|(B+\varepsilon^2)^{-1}Bv\|^2 = v^*(B+\varepsilon^2)^{-1}BB(B+\varepsilon^2)^{-1}v$$

и

$$\|Aw\|^2 = \|A(B+\varepsilon^2)^{-1}\varepsilon^2)v\|^2 = v^T\varepsilon^2(B+\varepsilon^2)^{-1}B(B+\varepsilon^2)^{-1}\varepsilon^2 v.$$

Складывая, получаем

$$\varepsilon^2\|v-w\|^2 + \|Aw\|^2 = \varepsilon^2 v^*(B+\varepsilon^2)^{-1}Bv. \tag{12}$$

С другой стороны, по теореме 2 имеем

$$\varepsilon^2\|v-w\|^2 + \|Aw\|^2 \leqslant 2\delta^2 C_1^2\varepsilon^2\|v\|^2.$$

Таким образом, получаем (8) с $\tilde{\delta} = \sqrt{2}\delta$.

Обратно, пусть выполняется (8), то есть

$$v^*(B+\varepsilon)^{-1}Bv \leqslant \tilde{\delta}C_1^2\|v\|^2.$$

Из (12) получаем

$$\|Aw\|^2 \leqslant \varepsilon^2\tilde{\delta}C_1^2\|v\|^2, \quad \|v-w\|^2 \leqslant \tilde{\delta}C_1^2\|v\|^2,$$

то есть

$$\|Aw\| \leqslant \tilde{\delta}C_2\|v\|, \quad \|v-w\| \leqslant \tilde{\delta}C_1\|v\|.$$

Тогда по теореме 1 имеем (9).

5. Доказательство основной теоремы

Будем говорить, что комплекснозначная матрица A размера $n \times n$ лежит в пространстве $\mathcal{K}_n(\mu)$, если она верхнетреугольная и её элементы a_{jk} удовлетворяют условиям $\operatorname{Re} a_{kk} \leqslant \operatorname{Re} a_{jj} \leqslant 0$ при $j \leqslant k$ и $|a_{jk}| \leqslant \mu|\operatorname{Re} a_{jj}|$ для всех j и k.

Прежде чем перейти непосредственно к доказательству этой теоремы, приведём ряд вспомогательных утверждений.

Утверждение 5. *Для любых $n \in \mathbb{N}$, $K \geqslant 1$ существует такое $f_n(K)$, что для любой матрицы A, удовлетворяющей $\sup\limits_{\nu \geqslant 0} \|e^{\nu A}\| \leqslant K$, найдётся представление $A = \overline{S} M \overline{S}^{-1}$, где $\varkappa(\overline{S}) \leqslant f_n(K)$ и $M \in \mathcal{K}_n(f_n(K))$.*

Это часть теоремы Крайса о матрицах [3]. Очевидно, что функцию $f_n(K)$ можно считать неубывающей с ростом n и K.

Утверждение 6. *Для любых $n \in \mathbb{N}$, $\mu \geqslant 0$ существует такое $g_n(\mu)$, что для любой матрицы $A \in \mathcal{K}_n(\mu)$ выполняется $\sup\limits_{\nu \geqslant 0} \|e^{\nu A}\| \leqslant g_n(\mu)$.*

Это тоже часть теоремы Крайса о матрицах [3]. Поскольку $\mathcal{K}_n(\mu') \subset \mathcal{K}_n(\mu)$ при $\mu' < \mu$, очевидно, что функцию $g_n(\mu)$ можно считать неубывающей с ростом μ.

Утверждение 7. *Для $M \in \mathcal{K}_n(\mu)$, $n > 1$, $\det M \neq 0$, справедливо*

$$\varkappa(M) \leqslant 2^{n-2} \mu^n n^2 \frac{|\lambda(M)|_{\max}}{|\lambda(M)|_{\min}}.$$

Для любой матрицы размера $n \times n$ её норма не превосходит максимального модуля её элемента, умноженного на n. Поскольку M верхнетреугольная, максимальный модуль её элемента не превосходит $\mu|\lambda(M)|_{\max}$, и $\|M\| \leqslant n\mu|\lambda(M)|_{\max}$. Далее вычислим некоторый элемент матрицы M^{-1} по формуле Крамера. В знаменателе стоит произведение диагональных элементов матрицы M. В числителе стоит детерминант некоторого минора порядка $n-1$ матрицы M. Он включает в себя $(n-1)!$ слагаемых, но ввиду того, что матрица M верхнетреугольная, легко показать, что ненулевых слагаемых будет не более 2^{n-2}. В каждое слагаемое в качестве множителя будет входить по одному элементу из каждой (кроме одной) строки матрицы M, поэтому этот минор не превосходит $2^{n-2}\mu^{n-1}\det M/|\lambda(M)|_{\min}$. Таким образом, каждый

элемент матрицы M^{-1} по модулю не превосходит $2^{n-2}\mu^{n-1}/|\lambda(M)|_{\min}$ и, следовательно, $\|M^{-1}\| \leqslant 2^{n-2}\mu^{n-1}n/|\lambda(M)|_{\min}$. Отсюда получаем искомое неравенство.

Утверждение 8. *Существует функция $c_n(\mu, k, m)$, $\mu \in \mathbb{R}$, $k \in \{2, \ldots, n\}$, $m > 1$, не возрастающая с ростом m, такая что выполняется следующее. Пусть $A \in \mathcal{K}_n(\mu)$, $k \in \{2, \ldots, n\}$ и*

$$m = \frac{|\operatorname{Re} a_{kk}|}{|\operatorname{Re} a_{k-1,k-1}|} \neq 1.$$

Введём матрицы F размера $(k-1)\times(k-1)$, G размера $(k-1)\times(n-k+1)$ и H размера $(n-k+1)\times(n-k+1)$, так чтобы

$$A = \begin{pmatrix} F & G \\ 0 & H \end{pmatrix}. \tag{13}$$

Тогда найдётся такая матрица X размера $(k-1)\times(n-k-1)$, что

$$A = \begin{pmatrix} I & X \\ 0 & I \end{pmatrix} \begin{pmatrix} F & 0 \\ 0 & H \end{pmatrix} \begin{pmatrix} I & -X \\ 0 & I \end{pmatrix}, \tag{14}$$

причём $\|X\| \leqslant c_n(\mu, k, m)$.

Приравнивая (13) к (14), получаем уравнение на матрицу X:

$$-FX + XH = G.$$

Поскольку матрицы F и G не имеют общих собственных значений, это уравнение имеет единственное решение [4]. Поскольку интеграл

$$X = \int_0^\infty e^{-Ft} G e^{Ht} dt$$

существует (подынтегральное выражение экспоненциально убывает с ростом t), X является решением этого уравнения [5]. Обозначим $\lambda_- = \operatorname{Re} a_{kk}$ и $\lambda_+ = \operatorname{Re} a_{k-1,k-1}$ (имеем $\lambda_- < \lambda_+ < 0$ и $m = \lambda_-/\lambda_+$). Введём

$$\Lambda_- = (\lambda_+ + 3\lambda_-)/4 = (3+m)\lambda_-/(4m), \quad \Lambda_+ = (3\lambda_+ + \lambda_-)/4 = (3+m)\lambda_+/4.$$

Преобразуем

$$X = \int_0^\infty e^{(\Lambda_+ - F)t}\, G\, e^{(H-\Lambda_-)t}\, e^{t(\lambda_- - \lambda_+)/2}\, dt,$$

откуда

$$\|X\| \leqslant \|G\| \int\limits_0^\infty \|e^{(\Lambda_+ - F)t}\| \, \|e^{(H-\Lambda_-)t}\| \, e^{t(\lambda_- - \lambda_+)/2} \, dt.$$

У матрицы $\Lambda_+ - F$ действительная часть значений на диагонали не превосходит $\Lambda_+ - \lambda_+$, а модули внедиагональных элементов не превосходят $\mu|\mathrm{Re}\,\lambda_+|$. Поэтому

$$\Lambda_+ - F \in \mathcal{K}_{k-1}\left(\mu \frac{\lambda_+}{\Lambda_+ - \lambda_+}\right) = \mathcal{K}_{k-1}\left(\mu \frac{4}{m-1}\right).$$

Аналогично у матрицы $H - \Lambda_-$ действительная часть значений на диагонали не превосходит $\lambda_- - \Lambda_-$, а модули внедиагональных элементов не превосходят $\mu|\mathrm{Re}\,\lambda_-|$. Поэтому

$$H - \Lambda_- \in \mathcal{K}_{n+1-k}\left(\mu \frac{\lambda_-}{\lambda_- - \Lambda_-}\right) = \mathcal{K}_{n+1-k}\left(\mu \frac{4m}{m-1}\right).$$

В силу утверждения 6 имеем

$$\|e^{(\Lambda_+ - F)t}\| \leqslant g_{k-1}\left(\mu \frac{4}{m-1}\right), \quad \|e^{(H-\Lambda_-)t}\| \leqslant g_{n+1-k}\left(\mu \frac{4m}{m-1}\right).$$

Также имеем

$$\int\limits_0^\infty e^{t(\lambda_- - \lambda_+)/2} dt = \frac{2}{\lambda_+ - \lambda_-}$$

и

$$\|G\| \leqslant n\mu|\lambda_+| = n\mu(\lambda_+ - \lambda_-)\frac{1}{m-1}.$$

Следовательно,

$$\|X\| \leqslant c_n(\mu, k, m) := \frac{2n\mu}{m-1} g_{k-1}\left(\mu \frac{4}{m-1}\right) g_{n+1-k}\left(\mu \frac{4m}{m-1}\right).$$

Поскольку $g_k(\mu)$ не убывает с ростом μ, $c_n(\mu, k, m)$ не возрастает с ростом m.

Утверждение 9. *Пусть A – некоторая матрица, γ – некоторый замкнутый контур на комплексной плоскости, такой что ни одно из собственных значений матрицы A не лежит на этом контуре. Тогда матрица*

$$P = \frac{1}{2\pi i} \oint\limits_\gamma (zI - A)^{-1} dz \tag{15}$$

является проектором на сумму корневых подпространств матрицы A, соответствующих собственным значениям, лежащим внутри контура. Матрица $I - P$ является проектором на сумму всех корневых подпространств матрицы A, соответствующих остальным собственным значениям.

Доказательство этого утверждения см. в [6], §1.5.3.

Утверждение 10. *Пусть Z_0 – диагональная матрица с элементами λ_j, $j = 1, \ldots, n$, такими что $|\lambda_1| \geqslant \ldots \geqslant |\lambda_k| > \lambda_{k+1} = \ldots = \lambda_n = 0$. Пусть Y – матрица, такая что $\|Y\| \leqslant |\lambda_k|/8$. Тогда матрица $Z = Z_0 + Y$ может быть представлена в виде*

$$Z = S \begin{pmatrix} F & 0 \\ 0 & H \end{pmatrix} S^{-1},$$

где F – матрица размера k, а $\varkappa(S) \leqslant 5$.

При $Y = 0$ утверждение очевидно, поэтому будем считать, что $Y \neq 0$. Определим проекторы

$$P_0(Z_0) = \frac{1}{2\pi i} \int\limits_{|z|=4\|Y\|} (zI - Z_0)^{-1} dz, \quad P_0(Z) = \frac{1}{2\pi i} \int\limits_{|z|=4\|Y\|} (zI - Z)^{-1} dz.$$

Запишем

$$(zI - Z)^{-1} = (zI - Z_0 - Y)^{-1} = \left(I - (zI - Z_0)^{-1} Y\right)^{-1} (zI - Z_0)^{-1}. \qquad (16)$$

Поскольку $(zI - Z_0)^{-1}$ есть диагональная матрица, её норма равна максимальному модулю его элемента и, следовательно, на контуре $|z| = 4\|Y\|$ не превосходит $1/(4\|Y\|)$. Следовательно, на этом контуре $(zI - Z)^{-1}$ конечно, то есть собственные значения Z на нём не лежат. С использованием (16) запишем

$$P_0(Z) - P_0(Z_0) = \frac{1}{2\pi i} \int\limits_{|z|=4\|Y\|} \left[(I - (zI - Z_0)^{-1} Y)^{-1} - I\right] (zI - Z_0)^{-1} dz.$$

Отсюда

$$\|P_0(Z) - P_0(Z_0)\| \leqslant 4\|Y\| \sup_{|z|=4\|Y\|} \left\|(I - (zI - Z_0)^{-1} Y)^{-1} - I\right\| \left\|(zI - Z_0)^{-1}\right\|.$$

Далее, для матрицы M, $\|M\| < 1$, имеем

$$\|(I - M)^{-1} - I\| = \|M + M^2 + \ldots\| \leqslant \|M\| + \|M\|^2 + \ldots = \frac{1}{1 - \|M\|} - 1.$$

Поскольку

$$\|(zI - Z_0)^{-1}Y\| \leqslant \|Y\|/(4\|Y\|) \leqslant 1/4,$$

получаем

$$\|P_0(Z) - P_0(Z_0)\| \leqslant 1/3.$$

Матрицы $P_*(Z_0) = I - P_0(Z_0)$ и $P_*(Z) = I - P_0(Z)$ являются проекторами на сумму корневых подпространств матриц Z_0 и Z, соответствующих остальным собственным значениям. Имеем $\|P_*(Z) - P_*(Z_0)\| = \|P_0(Z) - P_0(Z_0)\| \leqslant 1/3$.

Пусть $\hat{P}_*(Z)$ получена из матрицы $P_*(Z)$ занулением последних $n-k$ столбцов, а матрица $\hat{P}_0(Z)$ получена из матрицы $P_0(Z)$ занулением первых k столбцов. Определим $S = \hat{P}_0(Z) + \hat{P}_*(Z)$. Далее, имеем

$$\begin{gathered}\|S - I\| = \|(\hat{P}_0(Z) + \hat{P}_*(Z)) - (P_0(Z_0) + P_*(Z_*))\| \leqslant \\ \leqslant \|\hat{P}_0(Z) - P_0(Z_0)\| + \|\hat{P}_*(Z) - P_*(Z_0)\|.\end{gathered}$$

$P_0(Z_0)$ есть диагональная матрица, у которой первые k диагональных элементов нулевые, а остальные равны 1, а $P_*(Z_0) = I - P_0(Z_0)$. Значит, матрица $\hat{P}_*(Z) - P_*(Z_0)$ получается из $P_*(Z) - P_*(Z_0)$ занулением последних $n-k$ столбцов, а матрица $\hat{P}_0(Z) - P_0(Z_0)$ получается из матрицы $P_0(Z) - P_0(Z_0)$ занулением первых k столбцов. Зануление столбцов не увеличивает норму матрицы (зануление столбца в матрице можно представить как её домножение справа на матрицу с нормой, равной 1). Поэтому

$$\|S - I\| \leqslant \|P_0(Z) - P_0(Z_0)\| + \|P_*(Z) - P_*(Z_0)\| \leqslant \frac{2}{3}.$$

Следовательно, матрица S обратима и

$$\|S^{-1} - I\| = \|(I + (S - I))^{-1} - I\| \leqslant \frac{1}{1 - \|S - I\|} - 1 \leqslant 2.$$

Таким образом, $\|S\| \leqslant 5/3$, $\|S^{-1}\| \leqslant 3$ и $\varkappa(S) \leqslant 5$. Поскольку первые k и последние $n - k$ столбцов матрицы S по построению лежат в инвариантных подпространствах матрицы Z, матрица $S^{-1}ZS$ имеет блочно-диагональный вид, что и требовалось доказать.

Теперь докажем теорему 4. Доказательство проведём индукцией по размеру матрицы n. При $n = 1$ можно положить $S = 1$ и $C(1, K) \equiv 1$. Предположим, что теорема доказана для матриц размера $n - 1 \geqslant 1$. Докажем её для матриц размера n.

По утверждению 5 справедливо представление $A = \overline{S}M\overline{S}^{-1}$, причём $\varkappa(\overline{S}) \leqslant f_n(K)$, $M \in \mathcal{K}_n(f_n(K))$. Элементы $m_{j,j}$ матрицы M являются собственными значениями матрицы A. Рассмотрим четыре случая.

Случай A. Пусть $\operatorname{Re} m_{11} = 0$, то есть матрица A имеет хотя бы одно чисто мнимое (в том числе нулевое) собственное значение. Тогда в силу $M \in \mathcal{K}_n(f_n(K))$ у матрицы M первая строка и первый столбец содержат только нулевые элементы, кроме, быть может, элемента на их пересечении. Имеем $\|e^{\nu M}\| \leqslant \varkappa(\overline{S})\|e^{\nu A}\| \leqslant K f_n(K)$. Оставшийся блок матрицы M размера $n-1$ (обозначим его через $\hat{M}$) по предположению индукции может быть преобразован к искомому блочно-диагональному виду: $\hat{M} = S' M' (S')^{-1}$, причём все диагональные блоки матрицы M' и матрица преобразования S имеют число обусловленности не более $C_{n-1}(K f_n(K))$. Тогда получаем

$$A = \overline{S} \begin{pmatrix} iy & 0 \\ 0 & S' M' (S')^{-1} \end{pmatrix} \overline{S}^{-1}, \quad y \in \mathbb{R}.$$

Обозначая

$$S = \overline{S} \begin{pmatrix} 1 & 0 \\ 0 & S' \end{pmatrix}, \quad M = \begin{pmatrix} iy & 0 \\ 0 & M' \end{pmatrix},$$

получаем $\varkappa(S) \leqslant \varkappa(\overline{S})\varkappa(S') \leqslant f_n(K) C(n-1, K f_n(K))$. Таким образом, полагая

$$C(n,K) \geqslant C_n^{(A)}(K) = f_n(K) C(n-1, K f_n(K)),$$

получаем утверждение индукции.

Случай B. Пусть $\operatorname{Re} m_{11} \neq 0$ и $p = |\operatorname{Re} m_{nn}|/|\operatorname{Re} m_{11}| \geqslant 2$. Тогда найдётся такое значение $k \in \{2, \ldots, n\}$, что $|\operatorname{Re} m_{k,k}|/|\operatorname{Re} m_{k-1,k-1}| \geqslant 2^{1/n}$. Пользуясь утверждением 8, получаем представление

$$M = U \begin{pmatrix} F & 0 \\ 0 & H \end{pmatrix} U^{-1}, \quad U = \begin{pmatrix} I & X \\ 0 & I \end{pmatrix}, \tag{17}$$

причём $\|X\| \leqslant c_n(\mu, k, 2^{1/n})$. Тогда

$$\|U\|, \|U^{-1}\| \leqslant 1 + \|X\| \leqslant \tilde{c}_n(\mu), \quad \tilde{c}_n(\mu) = 1 + \sup_{k=2,\ldots,n} c_n(\mu, k, 2^{1/n}).$$

Поскольку U имеет вид (17), матрицы $e^{\nu F}$ и $e^{\nu H}$ являются подматрицами матрицы $e^{\nu M}$, и их нормы не превосходят $\|e^{\nu M}\| \leqslant \varkappa(\overline{S})K$. Матрицы F и H можно, применив утверждение индукции, преобразовать к блочно-диагональному виду: $F = S_F M_F S_F^{-1}$, $H = S_H M_H S_H^{-1}$. В результате получаем $A = S M S^{-1}$, где

$$S = \overline{S} U \begin{pmatrix} S_F & 0 \\ 0 & S_H \end{pmatrix}, \quad M = \begin{pmatrix} M_F & 0 \\ 0 & M_H \end{pmatrix}.$$

Поскольку матрицы M_F и M_H блочно-диагональные и их блоки имеют числа обусловленности, не превосходящие $\max\limits_{n'<n} C(n', K f_n(K))$, этим же свойством

обладает и матрица M. Далее,

$$\varkappa(S) \leqslant \varkappa(\overline{S})\varkappa(U)\max\{\varkappa(S_F), \varkappa(S_H)\} \leqslant C_n^{(B)}(K),$$

где

$$C_n^{(B)}(K) = f_n(K)\tilde{c}_n^2(\mu)\max_{n'<n} C\left(n', K f_n(K)\right).$$

Таким образом, полагая $C(n,K) \geqslant C_n^{(B)}(K)$, получаем утверждение индукции.

Случай C. Пусть теперь $\mathrm{Re}\, m_{11} \neq 0$, $p = |\mathrm{Re}\, m_{nn}|/|\mathrm{Re}\, m_{11}| \leqslant 2$ и $|\lambda(A)|_{\max}/|\lambda(A)|_{\min} \leqslant (8(2n+1))^{n-1}$. Тогда то по утверждению 7 справедливо

$$\varkappa(A) \leqslant C_n^{(C)}(K) = 2^{n-2}n^2(8(2n+1))^{n-1}.$$

Полагая $S = I$ и $C(n,K) \geqslant C_n^{(C)}(K)$, получаем утверждение индукции.

Случай D. Пусть теперь у матрицы A действительная часть всех собственных значений лежит на отрезке $[-2\kappa, -\kappa]$, где $\kappa > 0$, и $|\lambda(A)|_{\max}/|\lambda(A)|_{\min} > Q^{n-1}$, где $Q = 8(2n+1)$. Тогда всё множество собственных значений можно разбить на две группы, так чтобы минимальное по модулю собственное значение в первой группе (обозначим его модуль через Λ_+) не менее чем в Q раз превосходило максимум модуля собственных значений во второй группе (обозначим его через Λ_-).

Введя матрицу перестановок U, можно представить A в виде $A = UZU^{-1}$, где на диагонали Z стоят её собственные значения, упорядоченные по убыванию их модулей. Положим $Z = Z_0 + Y$, где Z_0 – диагональная матрица, элементы которой совпадают с диагональными элементами Z для собственных значений из первой группы и равны нулю для собственных значений из второй группы. Поскольку все внедиагональные элементы в матрице Y по модулю не превосходят 2κ, а диагональные не превосходят Λ_-, имеем $\|Y\| \leqslant \Lambda_- + 2n\kappa \leqslant (1+2n)\Lambda_- \leqslant \Lambda_+/8$. Применяя утверждение 8, получаем

$$A = UZU^{-1} = US\begin{pmatrix} F & 0 \\ 0 & H \end{pmatrix}S^{-1}U^{-1},$$

причём $\varkappa(S) \leqslant 5$. Поскольку U – матрица перестановок, $\varkappa(U) = 1$. Следовательно, $\|e^{\nu F}\|$ и $\|e^{\nu H}\|$ не превосходят $5K$ и к ним можно применить утверждение индукции. Таким образом, полагая

$$C(n,K) \geqslant C_n^{(D)}(K) = 5\max_{n'<n} C(n', 5K),$$

получаем утверждение индукции.

Таким образом, если задать $C(n,K)$ рекуррентной формулой $C(1,K) = 1$, $C(n,K) = \max\{C_n^{(A)}(K), C_n^{(B)}(K), C_n^{(C)}(K), C_n^{(D)}(K)\}$ при $n \geqslant 2$, то во всех возможных случаях утверждение индукции будет выполняться. Это заканчивает доказательство теоремы.

Список литературы

1. Cao W., Zhang Z., Zou Q. Superconvergence of discontinuous Galerkin methods for linear hyperbolic equations // SIAM Journal on Numerical Analysis. 2014. Vol. 52, no. 5. P. 2555–2573.

2. Superconvergence of discontinuous Galerkin methods for two-dimensional hyperbolic equations / Cao W., Shu C.-W., Yang Y. et al. // SIAM Journal on Numerical Analysis. 2015. Vol. 53, no. 4. P. 1651–1671.

3. Kreiss H. O. Über Matrizen die beshränkte Halbgruppen erzeugen // Mathematica Scandinavica. 1959. P. 71–80.

4. Гантмахер Ф. Р. Теория матриц. М.: Наука, 1967. с. 576.

5. Беллман Р. Введение в теорию матриц. М.: Наука, 1976. с. 367.

6. Kato T. Perturbation theory for linear operators. Grund. math. Wiss., B. 132, Springer, 1966. P. XIX, 592.