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

Препринты ИПМ • Препринт № 73 за 2019 г.

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

Бахвалов П.А., Сурначёв М.Д.

Линейные схемы с несколькими степенями свободы для одномерного уравнения переноса

Рекомендуемая форма библиографической ссылки: Бахвалов П.А., Сурначёв М.Д. Линейные схемы с несколькими степенями свободы для одномерного уравнения переноса // Препринты ИПМ им. М.В.Келдыша. 2019. № 73. 40 с. doi:10.20948/prepr-2019-73
URL: http://library.keldysh.ru/preprint.asp?id=2019-73

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

Линейные схемы с несколькими степенями свободы для одномерного уравнения переноса

Москва — 2019

Бахвалов П. А., Сурначёв М. Д.

Линейные схемы с несколькими степенями свободы для одномерного уравнения переноса

Рассматриваются линейные разностные схемы с несколькими степенями свободы на одну ячейку для одномерного уравнения переноса. Численная ошибка решения таких схем обладет оценкой $O(h^p + th^q)$, причём p совпадает с порядком аппроксимации или превосходит его на единицу, а $q \geqslant p$. В частности, для метода Галёркина с разрывными базисными функциями на основе полиномов порядка k эта оценка справедлива при $p = k + 1$, $q = 2k + 1$. В настоящей работе доказывается, что наличие такой оценки эквивалентно существованию отображения гладких функций на сеточное пространство, отличающееся от обычного (например, L_2-проекции) на величину порядка h^p, в смысле которого схема будет обладать q-м порядком аппроксимации. Это позволяет сформулировать алгоритм определения оптимальных значений p и q.

Ключевые слова: аппроксимация и точность, суперсходимость

Pavel Alexeevisch Bakhvalov, Mikhail Dmitrievich Surnachev

Linear schemes with several degrees of freedom for the 1D transport equation

We consider linear schemes with several degrees of freedom for the 1D transport equation. The solution error possesses the estimate $O(h^p + th^q)$ where p is equal to or greater by one than the truncation error order and $q \geqslant p$ (for the discontinuous Galerkin method $p = k+1$ and $q = 2k+1$ where k is the order of polynomials). We prove that this estimate holds if and only if there exists a mapping of smooth functions on the mesh space providing the q-th order of the truncation error and deviating from the standard mapping (L_2-projection for example) by $O(h^p)$. This fact leads to an algorithm establishing the optimal values p and q for a given scheme.

Key words: consistency and accuracy, superconvergence

Оглавление

1. Введение

Численное решение начальной или начально-краевой задачи для уравнения $\partial u/\partial t + \partial u/\partial x = 0$, полученное при помощи метода Галёркина с разрывными базисными функциями на основе многочленов порядка k, обладает ошибкой

$$\|\varepsilon_h(t)\| \leqslant C_1 h^{k+1} + C_2 h^{2k+1} t. \tag{1.1}$$

Здесь и далее символом $\|\ \cdot\ \|$ обозначается L_2-норма, определённая таким образом, чтобы норма вектора из единиц не зависела от размера вектора. То есть при порядке аппроксимации, равном k (при $k \neq 0$; при $k = 0$ порядок аппроксимации равен 1) формальный порядок точности равен $k+1$, но коэффициент в ошибке решения при h^{k+1} ограничен величиной, не растущей со временем. Этот удивительный факт был обнаружен экспериментально, в частности в [1], и был окончательно доказан для произвольного k лишь почти 20 лет спустя в [2].

На равномерной сетке при постановке периодических граничных условий одним из способов для обоснования оценки (1.1) являлся спектральный анализ. В [3] была получена оценка (1.1) в случае $k = 1$ (кусочно-линейных базисных функций). В [4] авторы доказали её для $k = 2$ и $k = 3$, воспользовавшись системой символьных вычислений. Но для нахождения спектра требуется искать корни алгебраического уравнения порядка $m = k+1$ с коэффициентами, зависящими от волнового числа, которое при $m > 4$, вообще говоря, не разрешимо в радикалах. Поэтому даже использование системы символьных вычислений не позволило получить оценку (1.1) при $k > 3$.

Другим способом, которым, в конечном итоге, и была доказана оценка (1.1) для произвольного k, является метод введения модифицированного отображения функции на сеточное подпространство [1, 2, 5, 6]. Он заключается в следующем. Пусть Π_h – обычный оператор, сопоставляющий функции f её сеточный аналог f_h (в случае метода Галёркина с разрывными базисными функциями – L_2-проекция на пространство многочленов порядка k). Вводится новый оператор $\tilde{\Pi}_h$, такой что для достаточно гладких функций f выполняется $\|\tilde{\Pi}_h f - \Pi_h f\| \leqslant O(h^P)$, в смысле которого схема уже обладает порядком аппроксимации $Q \geqslant P$. Это обеспечивает оценку на точность схемы в смысле Π_h вида

$$\|\varepsilon_h(t)\| \leqslant C_1 h^P + C_2 h^Q t. \tag{1.2}$$

В [7] при помощи проекции Гаусса–Радау для произвольного k были получены значения $P = Q = k+1$, в [6] они были улучшены до $P = k+1$, $Q = k+2$, а в [2] – установлен оптимальный результат $P = k+1$, $Q = 2k+1$. Введённый в [2] оператор $\tilde{\Pi}_h$, действующий на $W^{2k+2}_{2,loc}(\mathbb{R})$, на равномерной сетке можно представить в виде

$$\left(\tilde{\Pi}_h f\right)_\eta = \left(\Pi_h f\right)_\eta + \sum_{m=P}^{Q} h^m \mathfrak{C}^{(m)} \left(\mathcal{P}_h^{(m)} \left(\frac{d^m f}{dx^m}\right)\right)_\eta, \tag{1.3}$$

где $\eta \in \mathbb{Z}$ – номер сеточной ячейки, $(\Pi_h f)_\eta$ и $(\tilde{\Pi}_h f)_\eta$ – наборы значений сеточной функции на ячейке η, $\mathfrak{C}^{(m)}$ – диагональные матрицы, не зависящие от h, а $\mathcal{P}_h^{(m)}$ – некоторым образом определённые отображения.

Возможность переноса оценки (1.1) с равномерной сетки на произвольную неравномерную сетку (полагая в (1.1) $h = h_{\max}$) является специфическим свойством метода Галёркина с разрывными базисными функциями. В то же время на равномерной сетке оценка вида (1.2) при некоторых значениях P и Q, превышающих порядок аппроксимации, оказывается возможной и для других схем, в которых в одной сеточной ячейке определено несколько переменных. Такое же поведение может быть у схем на периодических сетках, измельчаемых гомотетией с коэффициентом h/h_0.

Если задана некоторая полудискретная схема, для получения оценки на ошибку решения вида (1.2) можно тем или иным образом выбрать отображения $\mathcal{P}_h^{(m)}$ (например, положить $\mathcal{P}_h^{(m)} \equiv \Pi_h$) и искать оператор вида (1.3) методом неопределённых коэффициентов. Условием на нахождение этих коэффициентов является Q-й порядок аппроксимации схемы в смысле $\tilde{\Pi}_h$. Если такой оператор удаётся построить, оценка (1.2) оказывается простым следствием устойчивости схемы и неравенства треугольника. Однако при этом возникает вопрос, будет ли полученная оценка оптимальной. То есть означает ли отсутствие отображения в виде (1.3), в смысле которого схема обладает порядком аппроксимации $Q + 1$, неулучшаемость оценки (1.2).

В настоящей работе исследуется связь между спектральным анализом и введением модифицированного отображения. Доказывается, что для произвольной схемы на равномерной сетке, устойчивой в L_2 и обладающей оценкой (1.2), существует такой оператор $\tilde{\Pi}_h$ отображения достаточно гладких функций на пространство сеточных функций вида (1.3), что $\|\tilde{\Pi}_h f - \Pi_h f\| \leqslant C_0 h^P$, а схема обладает Q-м порядком аппроксимации в смысле $\tilde{\Pi}_h$. Таким образом, метод введения модифицированного отображения позволяет по заданным коэффициентам схемы за конечное число операций установить оценку вида (1.2) с *оптимальными* значениями P и Q.

Настоящая работа является продолжением [8, 9]. Все обозначения, используемые в тексте, соответствуют этим работам. Предполагается, что характерный размер сетки h принят равным её шагу, то есть, в принятых в [8] обозначениях, $a = 1$. Координату будем обозначать через x.

2. Постановка задачи

Введём пространство $L_{2,per}(\mathbb{R})$ как множество функций $f \in L_{2,loc}(\mathbb{R})$, таких что $\exists N_0 \in \mathbb{N} : f(x+N_0) = f(x)$, с нормой, вычисляемой как среднеквадратичное значение f по периоду. Для $q \in \mathbb{N} \cup \{0\}$ определим пространство $H^q_{per}(\mathbb{R})$ как множество функций $f \in L_{2,per}(\mathbb{R}) \cap W^q_{2,loc}(\mathbb{R})$.

В настоящей работе рассматривается начальная задача для линейного уравнения переноса с постоянным коэффициентом $\omega \in \mathbb{R}$:

$$\frac{\partial v}{\partial t} + \omega \frac{\partial v}{\partial x} = 0, \quad x \in \mathbb{R}, \quad t \geqslant 0, \tag{2.1}$$

$$v(0,x) = v_0(x) \in H^q_{per}(\mathbb{R}). \tag{2.2}$$

Введём следующие обозначения. M^0 – конечное множество степеней свободы в одном сеточном блоке. $M = \mathbb{Z} \times M^0$ – общее множество степеней свободы. Если $f \in \mathbb{C}^M$, то $f_\eta \in \mathbb{C}^{M^0}$ – часть вектора численного решения в блоке $\eta \in \mathbb{Z}$. V^N_{per} – множество последовательностей с периодом N:

$$V^N_{per} = \{f \in \mathbb{C}^M \ : \ \forall \eta, \zeta \in \mathbb{Z} \ \ f_{\eta+N\zeta} = f_\eta\}.$$

$V_{per} = \bigcup\limits_{N \in \mathbb{N}} V^N_{per}$ – множество периодических последовательностей с нормой, определяемой для $f \in V^{N(f)}_{per}$ формулой

$$\|f\|^2 = \frac{1}{N(f)} \sum_{\eta=0}^{N(f)} \|f_\eta\|^2,$$

где $\|f_\eta\|$ – некоторая норма на $\mathbb{C}^{M^0}$. Если функция f имеет период $N(f)$, то она, очевидно, также имеет период $nN(f)$ для любого $n \in \mathbb{N}$, но на значение нормы замена $N(f)$ на $nN(f)$ не влияет.

Для аппроксимации (2.1) рассмотрим полудискретные схемы вида

$$\sum_{\zeta \in \mathcal{S}} Z_\zeta \frac{du_{\eta+\zeta}}{dt}(t) + \frac{1}{h} \sum_{\zeta \in \mathcal{S}} L_\zeta u_{\eta+\zeta}(t) = 0, \quad \eta \in \mathbb{Z}, \quad u_\eta \in \mathbb{C}^{M^0}, \tag{2.3}$$

где $\mathcal{S} \subset \mathbb{Z}$ – конечное множество, а Z_ζ и L_ζ – действительнозначные матрицы. Будем предполагать, что для оператора $Z : V_{per} \to V_{per}$, определённого равенством $(Zu)_\eta = \sum_{\zeta \in \mathcal{S}} Z_\zeta u_{\eta+\zeta}$, существует *ограниченный* обратный. Тогда для любых начальных данных $u_0 \in V_{per}$ система ОДУ (2.3) имеет единственное решение $u \in C^\infty([0,\infty), V_{per})$, $u(0) = u_0$.

Также будем предполагать, что схема (2.3) является устойчивой, то есть существует $K > 0$, такая что для всех $u \in C^\infty([0,\infty), V_{per})$, являющихся решением (2.3), при всех $t \geqslant 0$ выполняется $\|u(t)\| \leqslant K\|u(0)\|$.

Всюду далее будем считать, что $1/h \in \mathbb{N}$. Для отображения данных на пространство сеточных функций используются операторы $\Pi_h, \mathcal{P}_h : L_{2,loc}(\mathbb{R}) \to \mathbb{C}^M$, задаваемые формулами

$$(\Pi_h f)_{\eta,\xi} = \int\limits_G \mu_\xi(x) f(h(x+\eta))dx, \quad (\mathcal{P}_h f)_{\eta,\xi} = \int\limits_G \hat{\mu}_\xi(x) f(h(x+\eta))dx, \quad (2.4)$$

где $\eta \in \mathbb{Z}$ – индекс блока, $\xi \in M^0$ – индекс переменной внутри блока, $G \subset \mathbb{R}$ – некоторый интервал, $\mu_\xi, \hat{\mu}_\xi \in L_2(G)$, $\int_G \mu_\xi(x) = \int_G \hat{\mu}_\xi(x) = 1$. Если $f \in L_{2,per}(\mathbb{R})$ имеет период N_0, то $\Pi_h f, \mathcal{P}_h f \in V_{per}^{(N_0/h)}$. Операторы Π_h и $\mathcal{P}_h$ как операторы из $L_{2,per}(\mathbb{R})$ в V_{per} являются ограниченными равномерно по h.

Также для $q \in \mathbb{N} \cup \{0\}$ и $p \in \mathbb{N}$ будем рассматривать оператор $\tilde{\Pi}_h^{(p,q)} : W_{2,loc}^q(\mathbb{R}) \to \mathbb{C}^M$, задаваемый формулой

$$\left(\tilde{\Pi}_h^{(p,q)} f\right)_\eta = (\Pi_h f)_\eta + \sum_{m=p}^{q} h^m \mathfrak{C}^{(m)} \left(\mathcal{P}_h \left(\nabla^m x\right)\right)_\eta, \quad (2.5)$$

где Π_h и $\mathcal{P}_h$ определены (2.4), а $\mathfrak{C}^{(m)}$ – некоторые действительнозначные диагональные матрицы размера $|M^0| \times |M^0|$. Здесь и далее под ∇ понимается производная по x. При $p > q$ будем считать, что $\tilde{\Pi}_h^{(p,q)} = \Pi_h$ и, таким образом, Π_h является частным случаем оператора вида (2.5). Далее там, где индексы p и q у $\tilde{\Pi}_h^{(p,q)}$ несущественны, будем их опускать, полагая, что q фиксировано, а p без ограничения общности можно считать равным 1.

Определение 1. *Пусть $\tilde{\Pi}_h$ – некоторый оператор вида* (2.5)*. Ошибкой решения в смысле $\tilde{\Pi}_h$ с начальными данными $v_0 \in H_{per}^q(\mathbb{R})$ будем называть величину*

$$\varepsilon_h(t, v_0, \tilde{\Pi}_h) = u(t) - \tilde{\Pi}_h v(t, \cdot), \quad (2.6)$$

где $u(t)$ – решение (2.3) *с условием $u(0) = \tilde{\Pi}_h v_0$, а $v(t,x) = v_0(x - \omega t)$.*

Определение 2. *Пусть $\mathbf{\Pi}$ – оператор из $H_{loc}^q(\mathbb{R})$ или $C^q(\mathbb{R})$ в $\mathbb{C}^M$. Ошибкой аппроксимации на функции f в смысле $\mathbf{\Pi}$ будем называть величину $\epsilon_h(f, \mathbf{\Pi}) \in \mathbb{C}^M$ с компонентами*

$$(\epsilon_h(f, \mathbf{\Pi}))_\eta = \sum_{\zeta \in \mathcal{S}} \left[-\omega Z_\zeta (\mathbf{\Pi}(\nabla f))_{\eta+\zeta} + \frac{1}{h} L_\zeta (\mathbf{\Pi} f)_{\eta+\zeta} \right]. \quad (2.7)$$

Если $\tilde{\Pi}_h$ – оператор вида (2.5), то для функций $f \in H_{per}^{q+1}(\mathbb{R})$, имеющих период N_0, выполняется $\epsilon_h(f, \tilde{\Pi}_h) \in V_{per}^{(N_0/h)}$.

Определение 3. *Предположим, что существуют такие константы C_1 и C_2, что при всех начальных условиях $v_0 \in H^r_{per}(\mathbb{R})$, $r \geqslant Q+1$, и при всех t и h схема обладает оценкой ошибки*

$$\|\varepsilon_h(t, v_0, \tilde{\Pi}_h)\| \leqslant C_1 h^P \|\nabla^P v_0\| + C_2 (t+h) h^Q \|\nabla^{Q+1} v_0\|, \tag{2.8}$$

где $+\infty \geqslant Q \geqslant P > 0$. Тогда будем говорить, что схема обладает формальным порядком точности P и порядком точности в длительном счёте Q в смысле $\tilde{\Pi}_h$ на $H^r_{per}(\mathbb{R})$. Если выполняется оценка $\|\varepsilon_h(t, v_0, \tilde{\Pi}_h)\| \leqslant C_1 h^P \|\nabla^P v_0\|$ при $r \geqslant P > 0$, будем говорить, что $Q = \infty$, а если $\varepsilon_h(t, v_0, \tilde{\Pi}_h) \equiv 0$, будем говорить, что $P = Q = \infty$. Если для всех $P' > P$, $Q' \geqslant P'$ и для всех $P' = P$, $Q' > Q$ не существует оценки вида (2.8)*, то такие параметры P и Q будем называть оптимальными.*

Отметим, что определение 3 вводит понятия формального порядка точности и порядка точности в длительном счёте совместно. Фраза “схема обладает порядком точности в длительном счёте q” не имеет однозначного толкования, поскольку при каком-то значении $p > 0$ схема может обладать формальным порядком точности p и порядком точности в длительном счёте q, тогда как для других (в частности оптимальных) значений P, Q, при которых выполняется (2.8), может быть $Q < q$. Понятие оптимальности оценки введено так, что увеличение значения Q не может достигаться за счёт снижения формального порядка точности.

3. Спектральное представление схемы

Для $\phi \in \mathbb{C}$ введём матрицы

$$Z(\phi) = \sum_{\eta \in \mathcal{S}} Z_\eta \exp(i\phi\eta), \quad L(\phi) = \sum_{\eta \in \mathcal{S}} L_\eta \exp(i\phi\eta),$$

$$A(\phi) = -Z^{-1}(\phi) L(\phi) + i\omega\phi I. \tag{3.1}$$

Для $\phi \in \mathbb{R}$ детерминант $Z(\phi)$ отделён от нуля. Схема (2.3) устойчива с константой K тогда и только тогда, когда при всех $\phi \in \mathbb{R}$ и $\nu > 0$ выполняется

$$\|\exp(A(\phi)\nu)\| \leqslant K. \tag{3.2}$$

Функции аппроксимационной ошибки и ошибки решения в образах Фурье имеют вид

$$\hat{\epsilon}(\phi, \tilde{\Pi}_h) = A(\phi)(\tilde{\Pi}_1 e^{i\phi x})_0, \tag{3.3}$$

$$\hat{\varepsilon}(\phi, \nu, \tilde{\Pi}_h) = \left(e^{\nu A(\phi)} - I\right)(\tilde{\Pi}_1 e^{i\phi x})_0. \tag{3.4}$$

Здесь и ниже индекс 1 означает подстановку $h = 1$, а индекс 0 – взятие блочной компоненты в блоке $\eta = 0$. При этом для α, таких что $\alpha/(2\pi) \in \mathbb{Q}$, выполняется

$$\|\varepsilon_h(t, e^{i\alpha x}, \tilde{\Pi}_h)\| = \|\hat{\varepsilon}(\alpha h, t/h, \tilde{\Pi}_h)\|. \tag{3.5}$$

Следующие два утверждения позволяют переносить оценки функций (3.3) и (3.4) на произвольные достаточно гладкие периодические функции.

Утверждение 3.1. *Пусть $\tilde{\Pi}_h$ – некоторый оператор вида* (2.5) *и $P_A \in \mathbb{N}$. Тогда следующие утверждения эквивалентны:*

- *для некоторого $C > 0$ в некоторой окрестности $\phi = 0$ выполняется $\|\hat{\epsilon}(\phi, \tilde{\Pi}_h)\| \leqslant C|\phi|^{P_A+1}$;*
- *для всех $m = 1, \ldots, P_A$ выполняется $(\epsilon_1(x^m, \tilde{\Pi}_1))_0 = 0$;*
- *для любого $r \geqslant \max\{P_A, q\} + 1$ найдутся $C, \tilde{C} > 0$, такие что для любой $v_0 \in H^r_{per}(\mathbb{R})$ справедлива оценка*

$$\|\epsilon_h(v_0, \tilde{\Pi}_h)\| \leqslant Ch^{P_A}\|\nabla^{P_A+1} v_0\| + \tilde{C}h^{r-1}\|\nabla^r v_0\|. \tag{3.6}$$

Подчеркнём, что порядок малости по ϕ функции $\hat{\epsilon}(\phi, \tilde{\Pi}_h)$ на единицу превосходит порядок аппроксимации схемы.

Утверждение 3.2. *Пусть схема* (2.3) *является устойчивой. Пусть $\tilde{\Pi}_h$ – некоторый оператор вида* (2.5)*. Пусть $P, Q \in \mathbb{N}$, $Q \geqslant P$. Тогда следующие утверждения эквивалентны:*

- *для некоторых $C_1, C_2 \geqslant 0$ в некоторой окрестности $\phi = 0$ справедливо*

$$\|\hat{\varepsilon}(\phi, \nu, \tilde{\Pi}_h)\| \leqslant C_1|\phi|^P + C_2\nu|\phi|^{Q+1}; \tag{3.7}$$

- *для некоторых $C_1, C_2 \geqslant 0$ при любых начальных данных $v_0 \in H^{Q+1}_{per}(\mathbb{R})$ для ошибки решения справедлива оценка* (2.8)*, то есть схема обладает порядком точности P и порядком точности в длительном счёте Q.*

Пусть $\mathbb{R}^{n\times n}$ и $\mathbb{C}^{n\times n}$ – пространства действительно- и комплекснозначных матриц размера n. Через $\mathcal{A}(\ , \mathbb{R}^{n\times n})$ будем обозначать множество функций из $\mathbb{C}$ в $\mathbb{C}^{n\times n}$, аналитических при $\phi = 0$, таких что образ мнимой оси лежит в $\mathbb{R}^{n\times n}$. Проводимый ниже анализ опирается на следующую теорему, доказанную в [9].

Теорема 3.3. *Пусть $A \in \mathcal{A}(\ ,\mathbb{R}^{n\times n})$. Пусть существует $K > 0$, такое что для всех $\phi \in \mathbb{R}$ и всех $\nu > 0$ справедливо $\|\exp(\nu A(\phi))\| \leqslant K$. Тогда в окрестности $\phi = 0$ матрица $A(\phi)$ представима в блочно-диагональном виде*

$$A(\phi) = S(\phi)M(\phi)S^{-1}(\phi), \tag{3.8}$$

$$M(\phi) = \begin{pmatrix} M_0(\phi) & 0 & \dots & 0 & 0 \\ 0 & \phi M_1(\phi) & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \phi^m M_m(\phi) & 0 \\ 0 & 0 & \dots & 0 & M_\infty(\phi) \equiv 0 \end{pmatrix}, \tag{3.9}$$

где $S, M, S^{-1} \in \mathcal{A}(\ ,\mathbb{R}^{n\times n})$, блоки $M_j(\phi)$, $j \in \mathbb{N} \cup \{0, \infty\}$ квадратные, невырожденные при $\phi = 0$ (за исключением $j = \infty$), и некоторые из них могут отсутствовать (иметь нулевой размер).

4. Несколько свойств матричной экспоненты

Утверждение 4.1. *Для любых матриц X и Y верно*

$$\|e^{X+Y} - e^X\| \leqslant e^{\|X\|}\left(e^{\|Y\|} - 1\right). \tag{4.1}$$

Это неравенство было доказано в [10]. Приведём доказательство для полноты изложения. Имеем

$$\|e^{X+Y} - e^X\| = \left\|\sum_{k=0}^{\infty} \frac{1}{k!}((X+Y)^k - X^k)\right\| \leqslant$$

$$\leqslant \sum_{k=0}^{\infty} \frac{1}{k!}\left\|(X+Y)^k - X^k\right\| = \sum_{k=0}^{\infty} \frac{1}{k!}\left\|\sum_{j=1}^{k} \frac{k!}{j!(k-j)!} Y^j X^{k-j}\right\| \leqslant$$

$$\leqslant \sum_{k=0}^{\infty}\sum_{j=1}^{k} \frac{\|Y\|^j}{j!}\frac{\|X\|^{k-j}}{(k-j)!} = \sum_{j=1}^{\infty}\sum_{k=j}^{\infty} \frac{\|Y\|^j}{j!}\frac{\|X\|^{k-j}}{(k-j)!} =$$

$$= \sum_{j=1}^{\infty}\sum_{m=0}^{\infty} \frac{\|Y\|^j}{j!}\frac{\|X\|^m}{m!} = \left(\sum_{j=1}^{\infty} \frac{\|Y\|^j}{j!}\right)\left(\sum_{m=0}^{\infty} \frac{\|X\|^m}{m!}\right) = e^{\|X\|}\left(e^{\|Y\|} - 1\right).$$

Утверждение 4.2. *Для любой матрицы Y справедливо*

$$\|e^Y - I\| \leqslant (\|e^Y\| + e - 1)\min\{1, \|Y\|\}. \tag{4.2}$$

При $\|Y\| \geqslant 1$ в силу неравенства треугольника

$$\|e^Y - I\| \leqslant \|e^Y\| + 1 \leqslant \|e^Y\| + e - 1.$$

При $\|Y\| \leqslant 1$, используя (4.1) при $X = 0$ и выпуклость функции $e^y - 1$, получаем

$$\|e^Y - I\| \leqslant e^{\|Y\|} - 1 \leqslant (e-1)\|Y\| \leqslant (\|e^Y\| + e - 1)\|Y\|.$$

Таким образом, при любой матрице $\|Y\|$ выполняется (4.2).

Утверждение 4.3. *Пусть $A \neq 0$, $\|A\| \leqslant 1$ и*

$$f(A) = \sum_{k=1}^{\infty} \frac{A^{k-1}}{k!}. \tag{4.3}$$

Тогда $\|(f(A))^{-1}\| \leqslant 4$.

Действительно,

$$\|f(A) - I\| = \left\| \sum_{k=2}^{\infty} \frac{A^{k-1}}{k!} \right\| \leqslant \sum_{k=2}^{\infty} \frac{\|A\|^{k-1}}{k!} \leqslant \sum_{k=2}^{\infty} \frac{1}{k!} = \left(\sum_{k=0}^{\infty} \frac{1}{k!} \right) - 2 = e - 2.$$

Отсюда

$$\|(f(A))^{-1}\| = \|(I + (f(A) - I))^{-1}\| \leqslant \frac{1}{1 - \|f(A) - I\|} \leqslant \frac{1}{1 - (e-2)} \leqslant 4.$$

Утверждение 4.4. *Пусть A – невырожденная матрица и $\|A\| \leqslant 1$. Тогда*

$$\|(e^A - I)^{-1}\| \leqslant 4\|A^{-1}\|. \tag{4.4}$$

Действительно, $e^A - I = Af(A)$, где $f(A)$ определено (4.3). Поэтому

$$\|(e^A - I)^{-1}\| \leqslant \|(f(A))^{-1}\|\|A^{-1}\| \leqslant 4\|A^{-1}\|.$$

5. Структура ошибки аппроксимации

Всюду далее под $S(\phi)$ и $M(\phi)$ будем понимать матрицы, доставляющие представление (3.8)–(3.9) для матрицы $A(\phi)$, определённой (3.1). Символом $\bar{\aleph}$ будем обозначать множество $j \in \mathbb{N} \cup \{0, \infty\}$, таких что в матрице $M(\phi)$ присутствует блок $\phi^j M_j(\phi)$.

Пусть $\tilde{\Pi}_h$ – оператор вида (2.5). Для произвольного $\phi \in \mathbb{C}$ через $v(\phi)$ обозначим вектор

$$v(\phi) = S^{-1}(\phi) \left(\Pi_1 e^{i\phi x}\right)_0 . \tag{5.1}$$

Для $j \in \bar{\aleph}$ символом $v_j(\phi)$ обозначим j-й блок $v(\phi)$ (размерность вектора v_j соответствует размерности M_j). Функция $v_j(\phi)$ является аналитической функцией ϕ в некоторой окрестности $\phi = 0$, поскольку таковой является $S^{-1}(\phi)$. Поэтому либо найдётся такое $p_j \in \mathbb{N} \cup \{0\}$, что для достаточно малых ϕ имеет место

$$\frac{1}{2} c_j |\phi|^{p_j} \leqslant \|v_j(\phi)\| \leqslant c_j |\phi|^{p_j}, \tag{5.2}$$

либо в некоторой окрестности нуля выполняется $v_j(\phi) \equiv 0$. В последнем случае положим $p_j = \infty$. Также формально положим $p_j = \infty$ для $j \notin \bar{\aleph}$. Обозначим

$$\aleph = \{j \in \mathbb{N} \cup \{0\} : p_j < \infty\} \subseteq \bar{\aleph}.$$

То есть $j \in \bar{\aleph} \setminus \aleph$, если $j = \infty$ или $p_j = \infty$. Ниже мы покажем, что значениями p_j, $j \in \aleph$, описывается структура численной ошибки. Через них выражаются порядки аппроксимации и точности, и ими описывается специфика поведения ошибки в длительном счёте.

Утверждение 5.1. *Хотя бы одно из значений p_j, $j \in \bar{\aleph}$, равно нулю.*

Действительно, если $p_j \neq 0$, то $v_j(\phi) \to 0$ при $\phi \to 0$. Если предположить, что все p_j, $j \in \bar{\aleph}$, отличны от нуля, то $v(\phi) \to 0$, и поэтому $(\tilde{\Pi}_1 e^{i\phi x})_0 \to 0$. Но это невозможно, поскольку $(\tilde{\Pi}_1 e^{i\phi x})_0 \to (\tilde{\Pi}_1 1)_0 = \mathfrak{e}$, где $\mathfrak{e}$ – вектор из единиц размерности $|M^0|$.

Утверждение 5.2. *Пусть скорость переноса $\omega \neq 0$. Тогда в матрице $M(\phi)$ блок $M_\infty(\phi)$ отсутствует и $\aleph \neq \varnothing$.*

Допустим, что в (3.9) есть блок $M_\infty(\phi) \equiv 0$. Тогда $A(\phi)$ вырождена в некоторой конечной окрестности нуля. Значит, вырожденной является и матрица $-L(\phi) + i\phi Z(\phi)\omega$, то есть в некоторой окрестности нуля выполняется $det(-L(\phi) + i\phi Z(\phi)\omega) \equiv 0$. Все компоненты $L(\phi)$ и $Z(\phi)$ являются линейной комбинацией $e^{\pm 2\pi i k\phi}$, поэтому детерминант является конечной суммой величин вида $c(i\phi\omega)^m e^{2\pi i n\phi}$. Чтобы детерминант был равен нулю, нулю должно быть

равно каждое такое слагаемое, в частности слагаемое при $m = |M_0|$, которое равно $detZ(\phi)(i\phi\omega)^m$. То есть $detZ(\phi)$ тождественно равен нулю в некоторой окрестности $\phi = 0$, что очевидно невозможно.

Далее, поскольку в силу утверждения 5.1 хотя бы одно значение p_j равно нулю, $\aleph = \varnothing$ возможно только в том случае, когда $p_\infty = 0$. Но это невозможно, поскольку блок $M_\infty(\phi)$ в матрице $M(\phi)$ отсутствует.

Утверждение 5.3. *Если $\aleph = \varnothing$, то схема является точной, то есть для любой $v_0 \in H^1_{per}(\mathbb{R})$ выполняется $\epsilon_h(v_0, \Pi_h) = 0$.*

Действительно, пусть $j \in \bar{\aleph}$. Поскольку $\aleph = \varnothing$, выполняется $j \in \bar{\aleph} \setminus \aleph$. Значит, либо $j = \infty$, либо $p_j = \infty$. В первом случае блок матрицы $M(\phi)$ в некоторой окрестности $\phi = 0$ является нулевым, во втором случае нулевым является вектор $v(\phi)$. Значит, для всех $j \in \bar{\aleph}$ выполняется $M_j(\phi)v_j(\phi) \equiv 0$ и, следовательно, $A(\phi)\left(\Pi_1 e^{i\phi x}\right)_0 \equiv 0$. Домножим это равенство на $Z(\phi)$. В силу утверждения 5.2 из условия $\aleph = \varnothing$ следует, что $\omega = 0$, поэтому получаем $L(\phi)(\Pi_1 e^{i\phi x})_0 = 0$. Поскольку $L(\phi)$ и $\left(\Pi_1 e^{i\phi x}\right)_0$ являются целыми аналитическими функциями ϕ, это равенство справедливо при всех $\phi \in \mathbb{R}$. Поскольку любая функция $v_0 \in H^1_{per}(\mathbb{R})$ представима в виде суммы ряда по волнам вида $e^{i\phi x}$, получаем искомое утверждение.

Утверждение 5.4. *Если $\aleph \neq \varnothing$, то схема обладает порядком аппроксимации*

$$P_A = \min_{j \in \aleph}\{p_j + j - 1\} \tag{5.3}$$

на $H^{P_A+1}_{per}(\mathbb{R})$. Схема не обладает порядком аппроксимации $p > P_A$ на $H^q_{per}(\mathbb{R})$ ни при каком q.

Поскольку матрицы $Z(\phi)$, $S(\phi)$ и $M_j(\phi)$ при всех j аналитические в окрестности $\phi = 0$, при достаточно малом ϕ можно записать

$$\|\hat{\epsilon}(\phi, \Pi_h)\| \leqslant C \sum_{j \in \aleph} |\phi|^j \|v_j(\phi)\| \leqslant \tilde{C} \max_{j \in \aleph} |\phi|^{j+p_j}.$$

В силу утверждения 3.1 для $v_0 \in H^{P_A+1}_{per}(\mathbb{R})$ отсюда следует порядок аппроксимации P_A.

Обратно, пусть на $H^q_{per}(\mathbb{R})$ имеет место p-й порядок аппроксимации, тогда в силу утверждения 3.1 имеем $\|\hat{\epsilon}(\phi, \Pi_h)\| \leqslant C|\phi|^{p+1}$. Учитывая невырожденность $S(\phi)$, при достаточно малых ϕ при всех $j \in \aleph$ выполняется

$$\|\phi^j M_j(\phi)v_j(\phi)\| \leqslant \tilde{c}|\phi|^{p+1}.$$

Поскольку матрица $M_j(\phi)$ невырождена при $\phi = 0$, обратная к ней ограничена в некоторой окрестности. Значит, $|\phi|^j\|v_j(\phi)\| \leqslant \hat{c}|\phi|^{p+1}$, и в силу (5.2) получаем $p_j \geqslant p + 1 - j$. Ввиду произвольности $j \in \aleph$ отсюда получаем, что $p \leqslant P_A$.

6. Структура ошибки решения

Теперь подставим даваемое утверждением 3.3 представление матрицы в формулу (3.4) для функции ошибки решения. С использованием обозначения (5.1) это равенство можно переписать в виде

$$\hat{\varepsilon}(\phi,\nu,\tilde{\Pi}_h) = S(\phi)\left(e^{\nu M(\phi)} - I\right) v(\phi) = S(\phi)\begin{pmatrix} \vdots \\ \left(e^{\nu\phi^j M_j(\phi)} - I\right) v_j(\phi) \\ \vdots \end{pmatrix}.$$

Введём вектор $E_j(\phi,\nu)$ размерности m_j, равной размерности блока M_j матрицы M, данной (3.9), равенством

$$E_j(\phi,\nu) = \left(e^{\nu\phi^j M_j(\phi)} - I\right) v_j(\phi). \tag{6.1}$$

Будем называть его *компонентой ошибки*. При $j \in \bar{\aleph} \setminus \aleph$ в окрестности $\phi = 0$ имеем $E_j(\phi,\nu) \equiv 0$. При $j \in \aleph$ воспользуемся равенством (4.2) с подстановкой $Y = \nu\phi^j M_j(\phi)$. Из условия устойчивости (3.2) имеем $\|e^Y\| \leqslant K$. В результате получаем

$$\|E_j(\phi,\nu)\| \leqslant (K+e)\min\left\{1, \nu|\phi|^j \,\|M_j(\phi)\|\right\} c_{j,2}|\phi|^{p_j}. \tag{6.2}$$

Покажем, что для каждой компоненты справедлива одна из двух оценок: либо $O(|\phi|^P)$, либо $O(\nu|\phi|^{Q+1})$.

Утверждение 6.1. *Пусть схема обладает порядком точности P и порядком точности в длительном счёте $Q \geqslant P$. Тогда для всех $j \in \aleph$ справедливо*

$$p_j \geqslant \min\{P, Q+1-j\}. \tag{6.3}$$

Представим

$$M_j(\phi) = M_j(0) + \phi\varkappa(\phi),$$

где $\varkappa(\phi)$ – матричная функция, аналитическая в окрестности $\phi = 0$, а матрица $M_j(0)$ невырожденная (см. теорему 3.3). Тогда (6.1) переписывается в виде

$$E_j(\phi,\nu) = \left[\exp\left(\phi^j\nu M_j(0) + \phi^{j+1}\nu\varkappa(\phi)\right) - I\right] v_j(\phi).$$

Обозначим

$$\bar{E}_j(\phi,\nu) = \left[\exp(\phi^j\nu M_j(0)) - I\right] v_j(\phi). \tag{6.4}$$

Оценим разность между E_j и $\bar{E}_j$ по формуле (4.1):

$$\|E_j(\phi,\nu) - \bar{E}_j(\phi,\nu)\| \leqslant \exp\left(\nu|\phi|^j\|M_j(0)\|\right)\left[\exp(\nu|\phi|^{j+1}\|\varkappa(\phi)\|) - 1\right]\|v_j(\phi)\|.$$

В силу утверждения 3.2 справедливо (3.7), откуда ввиду невырожденности $S(0)$ следует оценка на каждую компоненту:

$$\|E_j(\phi,\nu)\| \leqslant C_1'|\phi|^P + C_2'(\nu+1)|\phi|^{Q+1}, \tag{6.5}$$

Тогда, пользуясь неравенством $e^y - 1 \leqslant ye^y$ при $y \geqslant 0$, получаем

$$\begin{gathered}\|\bar{E}_j(\phi,\nu)\| \leqslant \left(C_1'|\phi|^P + C_2'|\phi|^{Q+1}(\nu+1)\right) + \\ + \exp\left(\nu\|M_j(0)\||\phi|^j\right)\nu|\phi|^{j+1}\|\varkappa(\phi)\|\exp(\nu|\phi|^{j+1}\|\varkappa(\phi)\|)\|v_j(\phi)\|.\end{gathered} \tag{6.6}$$

Получим теперь неравенство в обратную сторону. Из равенства (6.4) имеем

$$\|\bar{E}_j(\phi,\nu)\| \geqslant \left\|\left[\exp(\nu\phi^j M_j(0)) - I\right]^{-1}\right\|^{-1}\|v_j(\phi)\|.$$

Обозначим $\gamma = |\phi|^j\|M_j(0)\|$. При $\gamma\nu \leqslant 1$ мы можем воспользоваться оценкой (4.4). Получаем

$$\|\bar{E}_j(\phi,\nu)\| \geqslant \frac{1}{4}\nu|\phi|^j\|(M_j(0))^{-1}\|^{-1}\|v_j(\phi)\|. \tag{6.7}$$

Неравенство (6.6) должно выполняться при любом ν, а (6.7) – при $\gamma\nu \leqslant 1$. Поэтому положим $\nu = 1/\gamma$. Сопоставим при этом (6.6) с (6.7) и учтём оценку (5.2):

$$\begin{gathered}\frac{1}{8}\frac{1}{\|M_j(0)\|\,\|(M_j(0))^{-1}\|}c_j|\phi|^{p_j} \leqslant \\ \leqslant \left(C_1'|\phi|^P + C_2'|\phi|^{Q+1-j}\|M_j(0)\|^{-1} + C_2'|\phi|^{Q+1}\right) + \\ + e|\phi|\|M_j(0)\|^{-1}\|\varkappa(\phi)\|\exp(|\phi|\|M_j(0)\|^{-1}\|\varkappa(\phi)\|)c_j|\phi|^{p_j}.\end{gathered}$$

Сравнивая степени $|\phi|$, получаем (6.3).

Следствие 6.2. *В условиях утверждения 6.1 для каждого $j \in \aleph$ справедлива одна из двух оценок: $\|E_j(\phi,\nu)\| \leqslant \tilde{C}|\phi|^P$ или $\|E_j(\phi,\nu)\| \leqslant \tilde{C}\nu|\phi|^{Q+1}$.*

В силу утверждения 6.1 для каждого $j \in \aleph$ справедливо либо $p_j \geqslant P$, либо $p_j \geqslant Q + 1 - j$. В первом случае, заменив в (6.2) минимум из двух величин на первое значение, получаем $\|E_j(\phi,\nu)\| \leqslant \tilde{C}|\phi|^P$. Во втором случае, заменив минимум в (6.2) на второе значение, с учётом $j + p_j \geqslant Q + 1$ получаем $\|E_j(\phi,\nu)\| \leqslant \tilde{C}\nu|\phi|^{Q+1}$.

Утверждение 6.3. *Если* $p_0 = 0$ *или* $p_1 = 0$*, то схема не обладает никаким порядком точности.*

Действительно, допустим, что схема обладает порядком точности $P > 0$ и порядком точности в длительном счёте $Q \geqslant P$. Тогда по утверждению 6.1 получаем $p_0 \geqslant \min\{P, Q+1\}$, $p_1 \geqslant \min\{P, Q\}$. Но поскольку $Q \geqslant P$, при выполнении любого из условий $p_0 = 0$ или $p_1 = 0$ немедленно получаем $P \leqslant 0$.

Утверждение 6.4. *Пусть* $p_0 \neq 0$, $p_1 \neq 0$. *Тогда схема* (2.3), (2.4) *обладает формальным порядком точности* P *и порядком точности в длительном счёте* Q*, где*

$$P = \min_{j \in \aleph} (\max\{p_j + j - 1, p_j\}) = \min\{p_0, \min_{j \geqslant 1}(p_j + j - 1)\}, \tag{6.8}$$

$$Q = \min_{j \in \aleph:\ p_j < P} (p_j + j - 1). \tag{6.9}$$

Если $\omega \neq 0$*, то множества, по которым берутся минимумы, непусто. Если* $\omega = 0$ *и какой-то из минимумов берётся по пустому множеству, то он полагается равным бесконечности. Значения* P *и* Q*, даваемые* (6.8) *и* (6.9)*, являются оптимальными в смысле определения 3.*

Из условия $p_0 \neq 0$, $p_1 \neq 0$ следует, что $P \neq 0$. Теперь покажем, что $Q \geqslant P$. Так как в силу (6.8) выполняется $p_0 \geqslant P$, то

$$Q = \min_{j : p_j < P} (p_j + j - 1) = \min_{j \geqslant 1 : p_j < P} (p_j + j - 1) \geqslant \min_{j \geqslant 1} (p_j + j - 1) \geqslant P.$$

Покажем наличие формального порядка P и порядка точности в длительном счёте Q. Действительно, из (6.2) в некоторой окрестности $\phi = 0$ имеем

$$\|\hat{\varepsilon}(\phi, \nu, \Pi_h)\| \leqslant \hat{C} \sum_{j \in \aleph} \|E_j(\phi, \nu, \Pi_h)\| \leqslant C \sum_{j \in \aleph} \min\{|\phi|^{p_j}, \nu|\phi|^{p_j + j}\}.$$

Отсюда

$$\begin{aligned} \|\hat{\varepsilon}(\phi, \nu, \Pi_h)\| &\leqslant C \sum_{j : p_j < P} \nu|\phi|^{p_j + j} + C \sum_{j : p_j \geqslant P} |\phi|^{p_j} \leqslant \\ &\leqslant C' \nu |\phi|^{Q+1} + C' |\phi|^P. \end{aligned} \tag{6.10}$$

В силу утверждения 3.2 получаем искомую оценку (2.8).

Предположим теперь, что $\omega \neq 0$. Тогда по утверждению 5.2 множество $\aleph$ непусто, поэтому P конечно. Поскольку по утверждению 5.1 хотя бы одно из p_j, $j \in \bar{\aleph}$, равно нулю, Q может быть бесконечным только в том случае,

когда $p_\infty = 0$. Но по утверждению 5.2 блок M_∞ в матрице отсутствует, то есть $p_\infty = \infty$. Следовательно, Q также конечно.

Покажем теперь, что значения P и Q оптимальны. Действительно, пусть для некоторых натуральных чисел Q' и P', таких что $Q' \geqslant P' > 0$, выполняется оценка (2.8) с подстановками $P = P'$ и $Q = Q'$. Тогда для каждой компоненты ошибки, определённой (6.1), выполняется (6.5) с той же подстановкой. По утверждению 6.1, для всех j выполняется хотя бы одно из неравенств

$$\left[\begin{array}{l} p_j \geqslant P', \\ p_j + j - 1 \geqslant Q'. \end{array}\right. \tag{6.11}$$

Поскольку $Q' \geqslant P'$, из (6.11) для всех j следует $P' \leqslant \max\{p_j, p_j + j - 1\}$. Взяв минимум по j, в силу определения (6.8) получаем $P' \leqslant P$.

Предположим, что $P' = P$. Тогда из (6.11) следует

$$Q' \leqslant \min_{p_j < P'}(p_j + j - 1) = \min_{p_j < P}(p_j + j - 1) = Q.$$

Таким образом, из оценки (2.8) с подстановками $P = P'$ и $Q = Q'$ следует, что выполняется одно из двух условий:

$$\left[\begin{array}{ll} P' < P, & \\ P' = P, & Q' \leqslant Q. \end{array}\right.$$

Это и означает, что значения P и Q являются оптимальными в смысле определения 3.

7. Основная теорема

Предположим, что при некоторых $C_1, C_2 \geqslant 0$ и некоторых P и $Q \geqslant P$ в некоторой окрестности $\phi = 0$ справедлива оценка

$$\|\hat{\varepsilon}(\phi, \nu, \Pi_h)\| = \left\| \left(e^{\nu A(\phi)} - I \right) (\Pi_1 e^{i\phi x})_0 \right\| \leqslant C_1 |\phi|^P + C_2 \nu |\phi|^{Q+1}. \tag{7.1}$$

Определим в окрестности $\phi = 0$ аналитическую функцию

$$V(\phi) = S(\phi) \begin{pmatrix} \delta_0 I & 0 & \dots & 0 & 0 \\ 0 & \delta_1 I & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \delta_m I & 0 \\ 0 & 0 & \dots & 0 & \delta_\infty I \end{pmatrix} S^{-1}(\phi) (\Pi_1 e^{i\phi x})_0, \tag{7.2}$$

где $S(\phi)$ дана теоремой 3.3, размеры блоков соответствуют представлению (3.9) матрицы M, а $\delta_j = 0$, если $p_j \geqslant P$, и $\delta_j = 1$, если $p_j < P$. Вектор $V(\phi)$ является

действительнозначным при $i\phi \in \mathbb{R}$, поскольку таковыми являются матрицы $S(\phi)$ и $S^{-1}(\phi)$. По построению функция $V(\phi)$ удовлетворяет условию

$$\|V(\phi) - (\Pi_1 e^{i\phi x})_0\| \leqslant C|\phi|^P. \tag{7.3}$$

Матрица $A(\phi)V(\phi)$ отличается от $V(\phi)$ заменой δ_j на $\delta_j\phi^j M_j(\phi)$, и поэтому в силу утверждения 6.1 в некоторой окрестности $\phi = 0$ справедливо

$$\|A(\phi)V(\phi)\| \leqslant \tilde{C} \max_{j:p_j<P} |\phi|^j \|M_j(\phi)\|\|v_j(\phi)\| \leqslant \tilde{\tilde{C}}|\phi|^{Q+1}. \tag{7.4}$$

Теорема 7.1. *Пусть отображения* Π_h *и* $\mathcal{P}_h$ *определены* (2.4). *Пусть схема* (2.3) *в смысле* Π_h *обладает формальным порядком точности* P *и порядком точности в длительном счёте* Q, $P \leqslant Q < \infty$ *(эти значения необязательно оптимальные). Тогда существуют такие действительнозначные диагональные матрицы* $\mathfrak{C}^{(m)} \in \mathbb{R}^{M^0}$, *что схема обладает порядком аппроксимации* Q *в смысле модифицированного отображения* $\tilde{\Pi}_h : H^{Q+1}_{per}(\mathbb{R}) \to V_{per}$, *определяемого* (2.5).

В силу утверждения 3.2 имеем оценку (7.1). Это позволяет определить $V(\phi)$ равенством (7.2). Определим диагональную матрицу $\mathfrak{C}(\phi)$ равенством

$$V(\phi) - (\Pi_1 e^{i\phi x})_0 = \mathfrak{C}(\phi)(\mathcal{P}_1 e^{i\phi x})_0. \tag{7.5}$$

Она определена однозначно, поскольку $(\mathcal{P}_1 e^{i\phi x})_0 \to \mathfrak{e}$ при $\phi \to 0$. Далее, положим

$$\mathfrak{C}^{(m)} = \frac{(-i)^m}{m!} \left.\frac{d^m \mathfrak{C}(\phi)}{d\phi^m}\right|_{\phi=0}.$$

Поскольку $\mathfrak{C}(\phi)$ по построению действительнозначная при $i\phi \in \mathbb{R}$, то коэффициенты $\mathfrak{C}^{(m)}$ действительнозначные.

Подставим значения $\mathfrak{C}^{(m)}$ в выражение (2.5). Имеем

$$\left(\tilde{\Pi}_1^{(P,Q)} e^{i\phi x}\right)_0 = \left(\Pi_1 e^{i\phi x}\right)_0 - \sum_{m=P}^{Q} \frac{\phi^m}{m!} \left.\frac{d^m \mathfrak{C}(\phi)}{d\phi^m}\right|_{\phi=0} \left(\mathcal{P}_1 e^{i\phi x}\right)_0. \tag{7.6}$$

Из (7.5) и (7.3) видно, что $\mathfrak{C}(\phi) = O(\phi^P)$. Поэтому сумма в (7.6) представляет собой сумму первых $Q+1$ членов разложения $\mathfrak{C}(\phi)$ в ряд Тейлора. Таким образом,

$$\left(\tilde{\Pi}_1^{(P,Q)} e^{i\phi x}\right)_0 = \left(\Pi_1 e^{i\phi x}\right)_0 + \mathfrak{C}(\phi)\left(\mathcal{P}_1 e^{i\phi x}\right)_0 + O(\phi^{Q+1}) = V(\phi) + O(\phi^{Q+1}),$$

и в силу (3.3) и (7.4) имеем

$$\left\|\hat{\epsilon}(\phi, \tilde{\Pi}_h^{(P,Q)})\right\| = \left\|A(\phi)\left(\tilde{\Pi}_1^{(P,Q)} e^{i\phi x}\right)_0\right\| = O(\phi^{Q+1}).$$

В силу утверждения 3.1 отсюда следует, что в смысле $\tilde{\Pi}_h^{(P,Q)}$ схема обладает Q-м порядком аппроксимации на $H_{per}^{Q+1}(\mathbb{R})$.

Если $P = Q = \infty$, то $\varepsilon_h(t, v_0, \Pi_h) \equiv 0$, и введение вспомогательного отображения не имеет смысла. Если $P < Q = \infty$, то в силу теоремы 7.1 для любого конечного q можно ввести оператор $\tilde{\Pi}_h^{(P,q)}$, в смысле которого будет иметь место q-й порядок аппроксимации. Также можно ввести (нелокальный) оператор $\tilde{\Pi}_h^{(P,\infty)} : L_{2,per}(\mathbb{R}) \to V_{per}$, действие которого на функцию $v \in L_{2,per}(\mathbb{R})$ с периодом N_0 определяется как

$$\left(\tilde{\Pi}_h^{(P,\infty)} v\right)_\eta = \sum_{k \in I_{\beta N}} v_k V\left(\frac{2\pi k h}{N_0}\right) e^{2\pi i k h \eta / N_0},$$

где $N = N_0/h$, v_k – коэффициенты ряда Фурье $v = \sum_{k\in\mathbb{Z}} v_k e^{2\pi i k x/N_0}$, а β выбрано так, чтобы все $\phi = 2\pi k/N$, $k \in I_{\beta N}$, лежали в области определения $V(\phi)$, заданной (7.2). Покажем, что в смысле $\tilde{\Pi}_h^{(P,\infty)}$ схема будет точной на $H_{per}^{P+1}(\mathbb{R})$. Действительно, для $\phi = 2\pi k/N$, $k \in I_{\beta N}$, имеем

$$A(\phi)\left(\tilde{\Pi}_1^{(P,\infty)} e^{i\phi x}\right)_0 = A(\phi)V(\phi).$$

По построению в $S^{-1}(\phi)V(\phi)$ остаются только компоненты, соответствующие $p_j < P$. Но поскольку $Q = \infty$, такая компонента единственна и соответствует $j = \infty$, и соответствующий блок $M_\infty(\phi) \equiv 0$. Таким образом, выполняется $A(\phi)(\tilde{\Pi}_1^{(P,\infty)} e^{i\phi x})_0 = A(\phi)V(\phi) = 0$. Дальнейшее доказательство точности на $H_{per}^{P+1}(\mathbb{R})$ повторяет доказательство утверждения 5.3.

Подводя итог, можно сделать следующие выводы о структуре и свойствах ошибки численного решения. При использовании оператора Π_h схема обладала ошибкой аппроксимации

$$\|\epsilon_h(v_0, \Pi_h)\| \leqslant C \sum_{j\in\aleph} \|\nabla^{p_j+j} v_0\| h^{p_j+j-1}$$

и ошибкой решения (точностью)

$$\|\varepsilon_h(t, v_0, \Pi_h)\| \leqslant C \sum_{j\in\aleph} \min\left\{\|\nabla^{p_j} v_0\| h^{p_j}, \|\nabla^{p_j+j} v_0\| h^{p_j+j-1}(t+h)\right\}. \qquad (7.7)$$

Последнее следует из (6.2) и утверждения 3.2. Теорема 7.1 показывает, что можно построить оператор $\tilde{\Pi}_h^{(P,Q)}$, где P и Q заданы (6.8)–(6.9), такой что при его использовании схема обладает ошибкой аппроксимации

$$\|\epsilon_h(v_0, \tilde{\Pi}_h)\| \leqslant C\|\nabla^{Q+1} v_0\| h^Q.$$

В силу устойчивости отсюда следует оценка на ошибку решения

$$\|\varepsilon_h(t, v_0, \tilde{\Pi}_h)\| \leqslant C\|\nabla^{Q+1} v_0\| h^Q t. \tag{7.8}$$

И вновь возвращаясь к оценке ошибке в смысле Π_h, используя неравенство треугольника, получаем

$$\|\varepsilon_h(t, v_0, \Pi_h)\| \leqslant C_1\|\nabla^P v_0\| h^P + C_2\|\nabla^{Q+1} v_0\| h^Q t, \tag{7.9}$$

Сравнивая (7.9) с (2.8), можно заметить отсутствие члена при h^{Q+1}. Проведённые рассуждения показывают, что эти оценки эквивалентны. Также отметим, что если вдруг для всех $j \in \aleph$ выполняется $p_j < P$, то $P = Q = P_A$. При этом $V(\phi) = (\Pi_1 e^{i\phi x})_0 = 0$, следовательно, $\mathfrak{C}(\phi) \equiv 0$ и $\tilde{\Pi}_h \equiv \Pi_h$. В этом случае (7.9) формально ухудшает оценку (7.8).

8. Точечное отображение

Утверждение о существовании модифицированного отображения легко перенести на случай нормы максимума. Определим операторы $\mathring{\Pi}_h, \mathring{\mathcal{P}}_h : C(\mathbb{R}) \to \mathbb{C}^M$, определяемые равенством

$$(\mathring{\Pi}_h f)_{\eta,\xi} = f\left(h\left(\rho_\xi + \eta\right)\right), \tag{8.1}$$

$$(\mathring{\mathcal{P}}_h f)_{\eta,\xi} = f\left(h\left(\hat{\rho}_\xi + \eta\right)\right), \tag{8.2}$$

где ρ_ξ и $\hat{\rho}_\xi$ – координаты точек, в которых определяются значения при $h = 1$. Напомним, что h такое, что $1/h \in \mathbb{N}$.

Также будем использовать отображение $\mathring{\tilde{\Pi}}_h^{(P,Q)} : C^Q_{per}(\mathbb{R}) \to V_{per}$, определяемое как

$$\left(\mathring{\tilde{\Pi}}_h^{(P,Q)} f\right)_\eta = \left(\mathring{\Pi}_h f\right)_\eta + \sum_{m=P}^{Q} h^m \mathfrak{C}^{(m)} \left(\mathring{\mathcal{P}}_h \left(\frac{d^m f}{df^m}\right)\right)_\eta. \tag{8.3}$$

Нам понадобится следующее утверждение, доказанное в [8].

Утверждение 8.1. *Для любого $q \in \mathbb{N}$ можно построить оператор Π_h вида* (2.4)*, такой что $\|\Pi_h f - \mathring{\Pi}_h f\| \leqslant Ch^{q+1}\|\nabla^{q+1} f\|_\infty$. Можно построить оператор Π_h вида* (2.4)*, такой что $\|\Pi_h f - \mathring{\tilde{\Pi}}_h^{(P,Q)} f\| \leqslant Ch^{Q+1}\|\nabla^{Q+1} f\|_\infty$.*

Докажем обобщение теоремы 7.1 на случай точечного отображения функции на сеточное пространство.

Утверждение 8.2. *Пусть схема* (2.3) *в смысле отображения* $\mathring{\Pi}_h$*, определённого* (8.1)*, обладает оценкой ошибки*

$$\|\varepsilon_h(t, v_0, \mathring{\Pi}_h)\| \leqslant C_1 h^P \|\nabla^P v_0\|_\infty + C_2(t+h)h^Q \|\nabla^{Q+1} v_0\|_\infty, \tag{8.4}$$

где $Q \geqslant P > 0$*. Тогда существуют такие диагональные матрицы* $\mathfrak{C}^{(m)} \in \mathbb{R}^{M^0}$*, что схема обладает порядком аппроксимации* Q *в смысле модифицированного отображения* $\mathring{\tilde{\Pi}}_h$*, определённого* (8.3)*.*

Действительно, если оценка (8.4) выполняется для произвольной $v_0 \in C^Q_{per}(\mathbb{R})$, то она выполняется, в частности, на всех функциях $v_0(x) = e^{i\phi x/h}$, $\phi/(2\pi) \in \mathbb{Q}$. Напомним, что $1/h \in \mathbb{N}$. В силу утверждения 8.1 существуют такие Π_h и $\mathcal{P}_h$, что

$$\|\Pi_h f - \mathring{\Pi}_h f\| \leqslant C h^{Q+1} \|\nabla^{Q+1} f\|_\infty \tag{8.5}$$

и

$$\|\mathcal{P}_h f - \mathring{\mathcal{P}}_h f\| \leqslant C h^\sigma \|\nabla^\sigma f\|_\infty, \quad \forall \sigma = 0, \ldots, Q+1. \tag{8.6}$$

Отсюда для операторов $\tilde{\Pi}_h^{(P,Q)}$ и $\mathring{\tilde{\Pi}}_h^{(P,Q)}$ с одинаковыми коэффициентами $\mathfrak{C}^{(m)}$ получаем

$$\|\tilde{\Pi}_h^{(P,Q)} f - \mathring{\tilde{\Pi}}_h^{(P,Q)} f\| \leqslant C' h^{Q+1} \|\nabla^{Q+1} f\|_\infty. \tag{8.7}$$

Из (8.5) с учётом устойчивости схемы получаем оценку

$$\|\hat{\varepsilon}(\phi, \nu, \Pi_h)\| = \|\varepsilon_h(\nu h, e^{i\phi x/h}, \Pi_h)\| \leqslant$$

$$\leqslant \|\varepsilon_h(\nu h, e^{i\phi x/h}, \mathring{\Pi}_h)\| + (K+1)\|\Pi_h e^{i\phi x/h} - \mathring{\Pi}_h e^{i\phi x/h}\| \leqslant$$

$$\leqslant C_1 |\phi|^P + C_2(\nu+1)|\phi|^{Q+1} + C(K+1)|\phi|^{Q+1}.$$

В силу утверждения 3.2 получаем, что в смысле Π_h схема обладает порядком точности P и порядком точности в длительном счёте Q. По теореме 7.1 получаем, что существует отображение $\tilde{\Pi}_h^{(P,Q)}$ вида (2.5), в смысле которого схема обладает порядком аппроксимации Q, то есть для всех $v_0 \in H^{Q+1}_{per}(\mathbb{R})$ выполняется

$$\|\epsilon_h(v_0, \tilde{\Pi}_h^{(P,Q)})\| \leqslant C' \|\nabla^{Q+1} v_0\| h^Q \leqslant C' \|\nabla^{Q+1} v_0\|_\infty h^Q.$$

Последнее неравенство написано в силу того, что для любой функции $f \in L_{2,per}(\mathbb{R})$ выполняется $\|f\| \leqslant \|f\|_\infty$. Остаётся воспользоваться (8.7), чтобы получить аналогичную оценку на аппроксимационную ошибку в смысле $\mathring{\tilde{\Pi}}_h$.

9. Алгоритм нахождения оптимальной оценки ошибки

В этом разделе будем считать, что дана схема вида (2.3), устойчивая на V_{per}, удовлетворяющая условию точности на константе: $L(0)\mathfrak{e} = 0$, и оператор Π_h вида (2.4). Предположим, что оптимальные значения формального порядка точности P и порядка точности в длительном счёте Q конечны (достаточным условием для этого является $\omega \neq 0$). Покажем, что задача нахождения оптимальных значений P и Q сводится к проверке совместности системы линейных алгебраических уравнений.

Зададимся некоторым оператором $\mathcal{P}_h$ вида (2.4), например $\mathcal{P}_h = \Pi_h$. Получаемые ниже результаты (оптимальные значения P и Q) не зависят от его выбора. С учётом $L(0) = \sum_\eta L_\eta$ при $q \geqslant m$ можно записать

$$\left(\epsilon_1\left(\frac{x^m}{m!}, \tilde{\Pi}_1^{(p,q)}\right)\right)_0 = \left(\epsilon_1\left(\frac{x^m}{m!}, \tilde{\Pi}_1^{(p,m)}\right)\right)_0 = L(0)\mathfrak{C}^{(m)}\mathfrak{e} + \left(\epsilon_1\left(\frac{x^m}{m!}, \tilde{\Pi}_1^{(p,m-1)}\right)\right)_0.$$

Здесь предполагается, что коэффициенты $\mathfrak{C}$ во всех операторах одни и те же. Отсюда с учётом утверждения 3.1 получаем следующее.

Утверждение 9.1. *Пусть* $\tilde{\Pi}_h^{(p,q)}$ *– оператор вида* (2.5)*. Схема обладает порядком аппроксимации* $P_A \geqslant 1$ *в смысле оператора* $\tilde{\Pi}_h^{(p,q)}$ *тогда и только тогда, когда для всех* $m = 1, \ldots, P_A$ *верна система равенств*

$$L(0)\mathfrak{C}^{(m)}\mathfrak{e} = -\left(\epsilon_1\left(\frac{x^m}{m!}, \tilde{\Pi}_1^{(p,\min\{m-1,q\})}\right)\right)_0. \tag{9.1}$$

Уравнение (9.1) можно записать в развёрнутом виде

$$L(0)\mathfrak{C}^{(m)}\mathfrak{e} = -\omega\sum_{\eta\in\mathcal{S}} Z_\eta\left[\left(\Pi_1\frac{x^{m-1}}{(m-1)!}\right)_\eta + \sum_{n=p}^{\min\{m-1,q\}}\mathfrak{C}^{(n)}\left(\mathcal{P}_1\frac{x^{m-n-1}}{(m-n-1)!}\right)_\eta\right] +$$
$$+\sum_{\eta\in\mathcal{S}} L_\eta\left[\left(\Pi_1\frac{x^m}{m!}\right)_\eta + \sum_{n=p}^{\min\{m-1,q\}}\mathfrak{C}^{(n)}\left(\mathcal{P}_1\frac{x^{m-n}}{(m-n)!}\right)_\eta\right].$$

Введём обозначение

$$\mathcal{F}_m(\mathfrak{C}^{(1)}, \ldots, \mathfrak{C}^{(Q)}) = \begin{pmatrix} (\epsilon_1(x, \tilde{\Pi}_1^{(1,Q)}))_0 \\ \vdots \\ (\epsilon_1(x^m/m!, \tilde{\Pi}_1^{(1,Q)}))_0 \end{pmatrix}, \tag{9.2}$$

где $\tilde{\Pi}_1^{(1,Q)}$ дано (2.5). В силу теоремы 7.1 и утверждения 3.1 схема обладает порядком аппроксимации Q в смысле $\tilde{\Pi}_h^{(1,Q)}$ тогда и только тогда, когда $\mathcal{F}_m(\mathfrak{C}^{(1)}, \ldots, \mathfrak{C}^{(Q)}) = 0$, где $\mathfrak{C}^{(1)}, \ldots, \mathfrak{C}^{(Q)}$ – коэффициенты $\tilde{\Pi}_h^{(1,Q)}$.

Приведём алгоритм определения оптимальных значений формального порядка точности и порядка точности в длительном счёте.

Алгоритм 1.

1. *Определяем* $P_A = \max\{m \in \mathbb{N} \cup \{0\} \,:\, \mathcal{F}_m(0,\dots,0) = 0\}$.
2. *Если* $(\epsilon_1(x^{P_A+1}, \tilde{\Pi}_1^{(P_A+1,P_A+1)}))_0 = 0$ *совместна как система относительно* $\mathfrak{C}^{(P_A+1)}$*, полагаем* $P' = P_A + 1$*, иначе* $P' = P_A$.
3. *Полагаем* $Q' = P'$.
4. *Если* $\mathcal{F}_{Q'+1}(0,\dots,\mathfrak{C}^{(P')},\dots,\mathfrak{C}^{(Q'+1)}) = 0$ *как система относительно* $\{\mathfrak{C}^{(m)}, m = P',\dots,Q'+1\}$ *совместна, увеличиваем* Q' *на единицу и повторяем п. 4.*

Теорема 9.2. *Значения* P' *и* Q'*, найденные алгоритмом 1, совпадают с оптимальными значениями* P *и* Q *в смысле определения 3.*

Пусть в ходе работы алгоритма были найдены значения P' и Q'. Покажем наличие оценки (2.8). Если ни одна система уравнений, проверенная алгоритмом, не оказалась совместной, он выдаёт значение $P' = Q' = P_A$, и оценкой (2.8) при $P = Q = P_A$ схема обладает в силу устойчивости. Предположим теперь, что хотя бы одна система оказалась совместной. Последняя система уравнений, которая в ходе работы алгоритма была помечена как совместная, имеет некоторое решение $\{\mathfrak{C}^{(m)}, m = P',\dots,Q'\}$. Обозначим через $\tilde{\Pi}_h^{(P',Q')}$ оператор вида (2.5), в который подставлены эти значения. По утверждению 3.1 отсюда следует оценка $\|\epsilon_h(v_0, \tilde{\Pi}_h^{(P',Q')})\| \leqslant ch^{Q'}\|\nabla^{Q'+1}v_0\|$. При этом по построению

$$\|\tilde{\Pi}_h^{(P',Q')}f - \Pi_h f\| \leqslant \sum_{m=P'}^{Q'} C_m h^m \|\nabla^m f\| \leqslant (h^{P'}\|\nabla^{P'}f\| + h^{Q'+1}\|\nabla^{Q'+1}f\|)\sum_{m=P}^{Q'} C_m.$$

Отсюда в силу устойчивости и неравенства треугольника получаем (2.8) при $P = P'$ и $Q = Q'$.

Обратно, пусть выполнено (2.8). Тогда по теореме 7.1 найдутся такие $\tilde{\mathfrak{C}}^{(m)}$, $m = P,\dots,Q$, что $\mathcal{F}_Q(0,\dots,0,\tilde{\mathfrak{C}}^{(P)},\dots,\tilde{\mathfrak{C}}^{(Q)}) = 0$. Из (5.3) и (6.8) очевидно, что либо $P = P_A$, либо $P = P_A + 1$. Поскольку $P' \geqslant P_A$, в первом случае алгоритм установит $P' \geqslant P = P_A$. Во втором случае $\mathcal{F}_P(0,\dots,0,\tilde{\mathfrak{C}}^{(P)}) = 0$, поэтому $\tilde{\mathfrak{C}}^{(P)}$ будет решением системы, проверяемой в пункте 2, и $P' = P_A + 1$ будет установлено. Далее, в обоих случаях $\{\tilde{\mathfrak{C}}^{(m)}\}$, $m = P,\dots,Q$, будет решением системы, проверяемой в пункте 4 на шаге $Q' - 1$, поэтому алгоритм выдаст $Q' \geqslant Q$.

Отметим, что условие, проверяемое в п. 2, равносильно

$$(\epsilon_1(x^{P_A+1}, \Pi_1))_0 \in \operatorname{Im} L(0).$$

Точность на константе влечёт за собой то, что матрица $L(0)$ является вырожденной, поскольку $L(0)\mathfrak{e} = 0$. Если её собственное значение $\lambda = 0$ является простым, то с точностью до $c_m I$ все элементы диагональной матрицы $\mathfrak{C}^{(m)}$ находятся однозначно из условия (9.1), записанного для значения m. Будем называть такой случай *простым*. Условие простого случая кратко записывается в виде $\mathrm{Ker}L(0) = \mathrm{span}(\mathfrak{e})$. В этом случае алгоритм может быть упрощён.

Докажем два вспомогательных утверждения.

Утверждение 9.3. *Пусть* $\tilde{\Pi}_h^{(p,P_A)}$ *– оператор вида* (2.5)*, в смысле которого схема обладает порядком аппроксимации* P_A*. Пусть* $n \leqslant P_A$ *и* $c_n \in \mathbb{R}$*. Тогда найдётся оператор* $\hat{\tilde{\Pi}}_h^{(p,P_A)}$ *с коэффициентами* $\hat{\mathfrak{C}}^{(m)}$*, такими что* $\hat{\mathfrak{C}}^{(m)} = \mathfrak{C}^{(m)}$ *при* $m < n$ *и* $\hat{\mathfrak{C}}^{(n)} = \mathfrak{C}^{(n)} + c_n I$*, что схема обладает порядком аппроксимации* P_A *в смысле* $\hat{\tilde{\Pi}}_h^{(p,P_A)}$*.*

Действительно, порядок аппроксимации P_A в смысле $\tilde{\Pi}_h^{(p,P_A)}$ равносилен условию $\hat{\varepsilon}(\phi, \tilde{\Pi}_h^{(p,P_A)}) = O(|\phi|^{P_A+1})$, то есть

$$A(\phi)\left[(\Pi_1 e^{i\phi x})_0 + \left(\sum_{m=p}^{P_A} \phi^m \mathfrak{C}^{(m)}\right)(\mathcal{P}_1 e^{i\phi x})_0\right] = O(|\phi|^{P_A+1}).$$

Обозначим вектор в квадратных скобках через $V(\phi)$. Тогда имеем $A(\phi)V(\phi) = O(|\phi|^{P_A+1})$. Введём вектор

$$\hat{V}(\phi) = (1 + c_n\phi^n)V(\phi).$$

Очевидно, $A(\phi)\hat{V}(\phi) = O(|\phi|^{P_A+1})$. Введём диагональную матрицу $\hat{\mathfrak{C}}(\phi)$ с компонентами

$$\hat{\mathfrak{C}}_{\xi,\xi}(\phi) = \left[(\hat{V}(\phi))_\xi - (\Pi_1 e^{i\phi x})_{0,\xi}\right]\left[(\mathcal{P}_1 e^{i\phi x})_{0,\xi}\right]^{-1}$$

и определим $\hat{\mathfrak{C}}^{(m)}$, $m = p, \ldots, P_A$ как коэффициенты её разложения в ряд по степеням ϕ вблизи $\phi = 0$. Очевидно, что $\hat{\mathfrak{C}}^{(m)} = \mathfrak{C}^{(m)}$ при $m < n$ и $\hat{\mathfrak{C}}^{(n)} = \mathfrak{C}^{(n)} + c_n I$. Тогда имеем

$$A(\phi)(\Pi_1 e^{i\phi x})_0 + A(\phi)\left(\sum_{m=p}^{P_A} \phi^m \hat{\mathfrak{C}}^{(m)}\right)(\mathcal{P}_1 e^{i\phi x})_0 =$$

$$= A(\phi)(\Pi_1 e^{i\phi x})_0 + A(\phi)\hat{\mathfrak{C}}(\phi)(\mathcal{P}_1 e^{i\phi x})_0 + O(|\phi|^{P_A+1}) =$$

$$= A(\phi)\hat{V}(\phi) + O(|\phi|^{P_A+1}) = O(|\phi|^{P_A+1}),$$

что и требовалось доказать.

Этот результат можно интерпретировать следующим образом. Введение аддитивной добавки $c_n I$ соответствует тому, что перед отображением функции f на сетку вначале от неё берётся дифференциальный оператор $1 + c_n h^n (d/dx)^n$. Поскольку такое отображение является отображением на сетку функции $(1 + c_n h^n (d/dx)^n) f$, которая также является решением уравнения (2.1), все свойства схемы (аппроксимация, формальный порядок точности и порядок точности при длительном счёте) при этом не меняются.

Утверждение 9.4. *Пусть собственное значение $\lambda = 0$ матрицы $A(0)$ простое, и $|M^0| > 1$. Пусть P и Q определяются* (6.8) *и* (6.9) *соответственно, и $P > 0$. Тогда имеет место альтернатива: либо $P_A = P = Q$, либо $Q \geqslant P = P_A + 1$. Кроме того, если $P < Q$, то $\bar{\aleph} = \{0, Q + 1\}$, и*

$$p_0 = P, \quad p_{Q+1} = 0. \tag{9.3}$$

Доказательство. Поскольку собственное значение $\lambda = 0$ матрицы $A(0)$ простое, в матрице $M(\phi)$ вида (3.9), даваемой утверждением 3.3, размерность матрицы $M_0(\phi)$ равна $|M^0| - 1$. Следовательно, в матрице $M(\phi)$ есть только один блок $M_m(\phi)$, для которого $m \neq 0$, и его размерность равна 1. Величина m может быть как конечной, так и бесконечной.

Величина p_0 не может быть равна 0, поскольку отсюда следовало бы $P = 0$. С другой стороны, в силу утверждения 5.1 найдётся $j \in \bar{\aleph}$, такое что $p_j = 0$. Таким образом, имеем $p_m = 0$. Если m конечно, то по формуле (6.9) с учётом $p_0 \geqslant P$ получаем $Q = p_m + m - 1$, откуда $m = Q + 1$, и $P = \min\{p_0, Q\}$. Если же $m = \infty$, то $Q = \infty$ и $P = p_0$.

Из определений (5.3) и (6.8) очевидно $P_A \leqslant P$. Поэтому для доказательства первой части утверждения фактически нужно доказать, что из $P_A = P$ следует $P = Q$. Действительно,

$$P_A = \min_j \{p_j + j - 1\} = \min\{p_0 - 1, Q\}, \quad P = \min\{p_0, Q\}.$$

Если $P_A = p_0 - 1 < Q$, то равенство $P_A = P$ сводится к $p_0 - 1 = \min\{p_0, Q\}$ и потому невозможно. Значит, имеем $P_A = Q$. Но это означает, что $P = Q$.

Если $Q > P$, получаем $p_0 = P$ и $P_A = p_0 - 1$. Таким образом, (9.3) доказано.

В простом случае оптимальные значения формального порядка точности и порядка точности в длительном счёте для схемы (2.3) в смысле Π_h могут быть найдены следующим алгоритмом. Как и в предыдущем алгоритме, зададимся некоторым оператором $\mathcal{P}_h$ вида (2.4).

Алгоритм 2 (определения формального порядка точности и порядка точности в длительном счёте для простого случая)**.**

1. *Определяем* $P_A = \max\{m \in \mathbb{N} \cup \{0\} \ : \ \mathcal{F}_m(0, \ldots, 0) = 0\}$.
2. *Полагаем* $m = P_A + 1$.
3. *Вычисляем* $f^m = -(\epsilon_1(x^m/m!, \tilde{\Pi}_1^{(P_A+1,m-1)}))_0$*, подставляя в* $\tilde{\Pi}_1^{(P_A+1,m-1)}$ *ранее найденные коэффициенты* $\mathfrak{C}^{(n)}$, $n = P_A + 1, \ldots, m-1$.
4. *Если* $f^m \in \mathrm{Im}L(0)$*, то:*
   - *находим диагональную матрицу* $\mathfrak{C}^{(m)}$ *из системы* $L(0)\mathfrak{C}^{(m)}\mathfrak{e} = f^m$ *(с точностью до* $c_m I$, $c_m \in \mathbb{R}$*);*
   - *увеличиваем значение* m *на единицу;*
   - *возвращаемся к п. 3.*
5. *Полагаем* $Q' = m - 1$.
6. *Если* $Q' = P_A$*, полагаем* $P' = P_A$*, иначе полагаем* $P' = P_A + 1$.

Теорема 9.5. *Пусть* $\mathrm{Ker}L(0) = \mathrm{span}(\mathfrak{e})$. *Значения* P' *и* Q'*, являющиеся результатом работы алгоритма 2, совпадают с оптимальными значениями* P *и* Q *в смысле определения 3.*

Прежде всего, отметим, что значение P_A, выдаваемое алгоритмом, совпадает с оптимальным значением порядка аппроксимации (см. утверждение (3.1).

Пусть в ходе работы алгоритма были найдены значения P' и Q'. Покажем наличие оценки (2.8). Возможны две альтернативы: либо условие, проверяемое в п. 4, ни разу не было выполнено, и тогда $P' = Q' = P_A$, либо выполняется $Q' \geqslant P' = P_A + 1$. В первом случае оценка (2.8) напрямую следует из условия устойчивости. Во втором случае в ходе выполнения алгоритма были определены такие диагональные матрицы $\mathfrak{C}^{(m)}$, $m = P', \ldots, Q'$, что $\mathcal{F}_{Q'}(\mathfrak{C}^{(P_A+1)}, \ldots, \mathfrak{C}^{(Q')}) = 0$. Доказательство оценки (2.8) для $P = P'$ и $Q = Q'$ повторяет утверждение 9.2.

Обратно, пусть имеет место оценка (2.8). Покажем, что алгоритм выдаст значения не меньшие, чем P и Q. По утверждению 9.4 имеет место альтернатива: либо $P = Q = P_A$, либо $Q \geqslant P = P_A + 1$. Значения P и Q, меньшие P_A, алгоритм, очевидно, выдать не может, поэтому в первом случае алгоритм даёт оптимальные значения. Пусть теперь $Q \geqslant P = P_A + 1$. По теореме 7.1 существует отображение $\tilde{\Pi}_h^{(P,Q)}$ вида (2.5), в смысле которого имеется Q-й порядок аппроксимации. Следовательно, в смысле $\tilde{\Pi}_h^{(P,P)}$ имеет место P-й порядок аппроксимации, и поэтому коэффициент $\mathfrak{C}^{(P)}$ в этой формуле удовлетворяет равенству (9.1) для $m = P$. Значит, условие, проверяемое в п. 4, как минимум один раз выполнится, поэтому $Q' > P_A$ и $P' = P_A + 1$ алгоритмом будет установлено.

Предположим теперь, что $Q > P$, но алгоритм выдаст значение Q', такое что $Q > Q' \geqslant P$. Значит, алгоритмом были найдены некоторые значения $\tilde{\mathfrak{C}}^{(m)}$,

$m \leqslant Q'$, но на шаге $m = Q' + 1$ система на нахождение коэффициентов $\mathfrak{C}^{(m)}$ оказалась несовместной. То есть несовместной является система

$$\mathcal{F}_{Q'+1}(0, \ldots, \tilde{\mathfrak{C}}^{(P)}, \ldots, \tilde{\mathfrak{C}}^{(Q')}, \mathfrak{C}^{(Q'+1)}) = 0$$

на нахождение $\mathfrak{C}^{(Q'+1)}$. Поскольку схема обладает формальным порядком точности P и порядком точности в длительном счёте Q, она также обладает порядком точности в длительном счёте $Q' + 1 \geqslant P$ при том же формальном порядке P. Согласно теореме 7.1, существует такой набор диагональных матриц $\hat{\mathfrak{C}}^{(n)}$, $n = P, \ldots, Q' + 1$, которые обеспечивают порядок аппроксимации $Q' + 1$ в смысле $\tilde{\Pi}_h^{(P,Q'+1)}$, то есть система

$$\mathcal{F}_{Q'+1}(0, \ldots, \mathfrak{C}^{(P)}, \ldots, \mathfrak{C}^{(Q')}, \mathfrak{C}^{(Q'+1)}) = 0$$

совместна. Следовательно, существует $m \in \{P, \ldots, Q'\}$, такое что система

$$\mathcal{F}_{Q'+1}(0, \ldots, \tilde{\mathfrak{C}}^{(P)}, \ldots, \tilde{\mathfrak{C}}^{(m)}, \mathfrak{C}^{(m+1)}, \ldots, \mathfrak{C}^{(Q')}, \mathfrak{C}^{(Q'+1)}) = 0 \qquad (9.4)$$

совместна, но система

$$\mathcal{F}_{Q'+1}(0, \ldots, \tilde{\mathfrak{C}}^{(P)}, \ldots, \tilde{\mathfrak{C}}^{(m)}, \tilde{\mathfrak{C}}^{(m+1)}, \mathfrak{C}^{(m+2)}, \ldots, \mathfrak{C}^{(Q'+1)}) = 0$$

несовместна. Обозначим некоторое решение (9.4) через $\{\hat{\mathfrak{C}}^{(n)}\}$, $n = m + 1, \ldots, Q' + 1$. Тогда имеем

$$\mathcal{F}_{m+1}(0, ..., \tilde{\mathfrak{C}}^{(P)}, ..., \tilde{\mathfrak{C}}^{(m)}, \hat{\mathfrak{C}}^{(m+1)}) = \mathcal{F}_{m+1}(0, ..., \tilde{\mathfrak{C}}^{(P)}, ..., \tilde{\mathfrak{C}}^{(m)}, \hat{\mathfrak{C}}^{(m+1)}, ..., \mathfrak{C}^{(Q'+1)}) = 0.$$

С другой стороны, по построению $\mathcal{F}_{m+1}(0, ..., \tilde{\mathfrak{C}}^{(P)}, ..., \tilde{\mathfrak{C}}^{(m)}, \tilde{\mathfrak{C}}^{(m+1)}) = 0$. Таким образом, $\tilde{\mathfrak{C}}^{(m+1)}$ и $\hat{\mathfrak{C}}^{(m+1)}$ – два решения (9.1) с подстановкой $m + 1$ вместо m. Но по условию $\mathrm{Ker} L(0) = \mathrm{span}(\mathfrak{e})$, следовательно, $\tilde{\mathfrak{C}}^{(m+1)}$ может отличаться от $\hat{\mathfrak{C}}^{(m+1)}$ только на величину вида $c_{m+1} I$, $c_{m+1} \in \mathbb{R}$. Таким образом, мы приходим к противоречию с утверждением 9.3. Следовательно, значение Q, найденное алгоритмом, совпадает с оптимальным.

Если порядок точности бесконечен, то процедуры 1 и 2 зацикливаются. Поэтому в случае $\omega = 0$ их формально нельзя рассматривать как алгоритмы. Тем не менее, если в процедурах 1 и 2 ограничить итерационный процесс некоторым значением $Q_{\max}$, то процедуры становятся конечными и позволяют получить один из трёх результатов относительно оптимальных значений формального порядка точности P и порядка точности в длительном счёте Q:

(a) $P \leqslant Q \leqslant Q_{\max}$, значения P и Q определены алгоритмом;
(b) $P \leqslant Q_{\max} < Q$, значение P определено алгоритмом, Q может быть конечным или бесконечным;
(c) $Q_{\max} < P \leqslant Q$, значения P и Q могут быть конечными или бесконечными.

10. Примеры схем и замечания

10.1. Однородные конечно-разностные схемы. В этом случае имеем $|M^0| = 1$. Если такая схема обладает в точности k-м порядком (то есть обладает k-м и не обладает $(k+1)$-м порядком), то $\aleph = \{k+1\}$, и $p_{k+1} = 0$. Отсюда следует оценка ошибки $\|\varepsilon_h(t, v_0, \Pi_h)\| \leqslant Ch^k t \|\nabla^{k+1} v_0\|$.

10.2. Метод Галёркина с разрывными базисными функциями (DG). Пусть k – порядок используемого полинома. Размером блока $|M^0|$ является число степеней свободы на одной ячейке, равное $k+1$. Хорошо известно, что никакое решение, кроме константного, не является стационарным решением по разрывному методу Галёркина. Значит, собственное значение $\lambda = 0$ матрицы $A(0)$ является простым. Поэтому, зная оценку (1.1) (доказанную в [2]) и её оптимальность, из утверждения 9.4 при $k > 0$ следует, что $\aleph = \{0, 2k+2\}$, $p_0 = k+1$, $p_{2k+2} = 0$. Размер блока $M_0(\phi)$ равен k, а размер блока $M_{2k+2}(\phi)$ равен 1. При $k = 0$ схема совпадает с конечно-разностной схемой 1-го порядка, поэтому имеем $\aleph = \{2\}$, $p_2 = 0$. Таким образом объясняется структура матрицы $A(\phi)$, полученная при помощи системы символьных вычислений в [4].

Зная значения p_j, при $k \neq 0$ для метода Галёркина с разрывными базисными функциями можно записать оценку точности (7.7):

$$\|\varepsilon_h(t, v_0, \Pi_h)\| \leqslant C_1 \|\nabla^{k+1} v_0\| \min\{h^{k+1}, h^k t\} + C_2 \min\{\|v_0\|, h^{2k+1}(t+h)\|\nabla^{2k+2} v_0\|\}.$$

На разных интервалах времени доминирующее слагаемое в ошибке разное:

- при $t \lesssim h$ имеем $\|\varepsilon_h(t, v_0, \Pi_h)\| \sim h^k t$;
- при $h \lesssim t \lesssim h^{-k}$ имеем $\|\varepsilon_h(t, v_0, \Pi_h)\| \sim h^{k+1}$;
- при $h^{-k} \lesssim t \lesssim h^{-(2k+1)}$ имеем $\|\varepsilon_h(t, v_0, \Pi_h)\| \sim t h^{2k+1}$;
- при $t \gtrsim h^{-(2k+1)}$ имеем $\|\varepsilon_h(t, v_0, \Pi_h)\| \sim 1$.

Зависимость численной ошибки от времени на примере решения уравнения $\partial v/\partial t + \partial v/\partial x = 0$ с начальными данными $v_0(x) = \sin(2\pi x)$ приведена на рис. 1 (конечные разности 5-го порядка) и рис. 2 (метод Галёркина с разрывными базисными функциями на основе многочленов 4-го порядка). Каждый тип маркера соответствует расчёту на одной сетке. Поскольку у DG на одну ячейку приходится $|M^0| = k+1 = 5$ степеней свободы, расчёт по DG на сетке с шагом h можно сравнивать с расчётом по конечно-разностной схеме с шагом $h/5$.

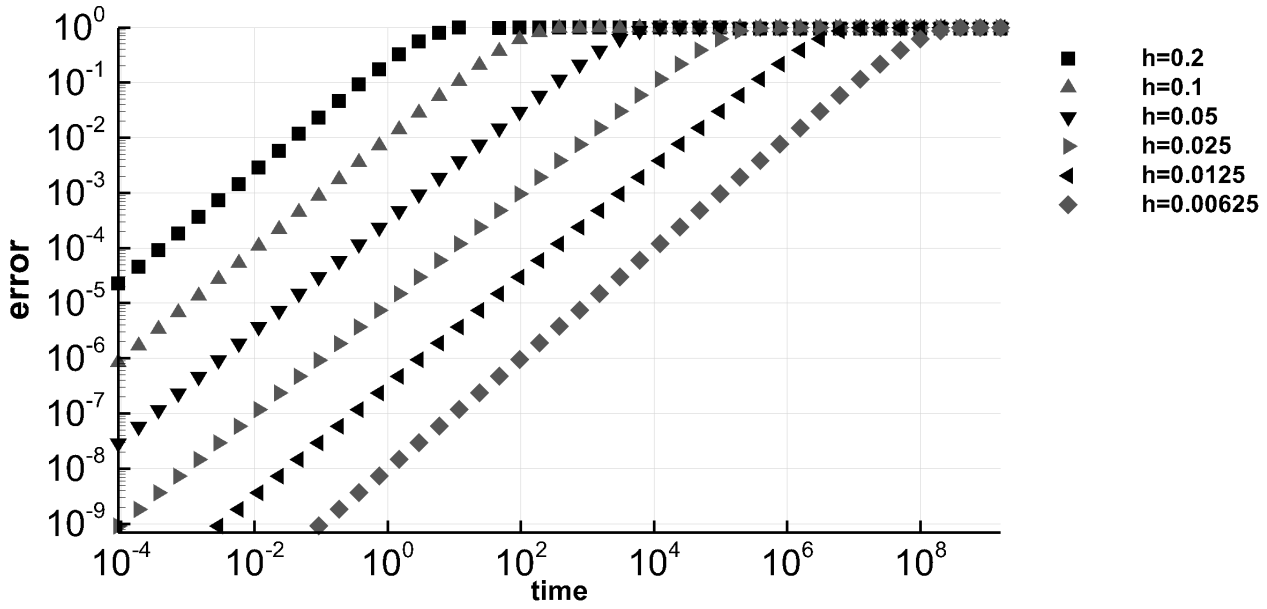

Рис. 1. Ошибка решения $\partial v/\partial t + \partial v/\partial x = 0$, $u_0 = \sin(2\pi x)$, по конечно-разностной схеме порядка $k = 5$

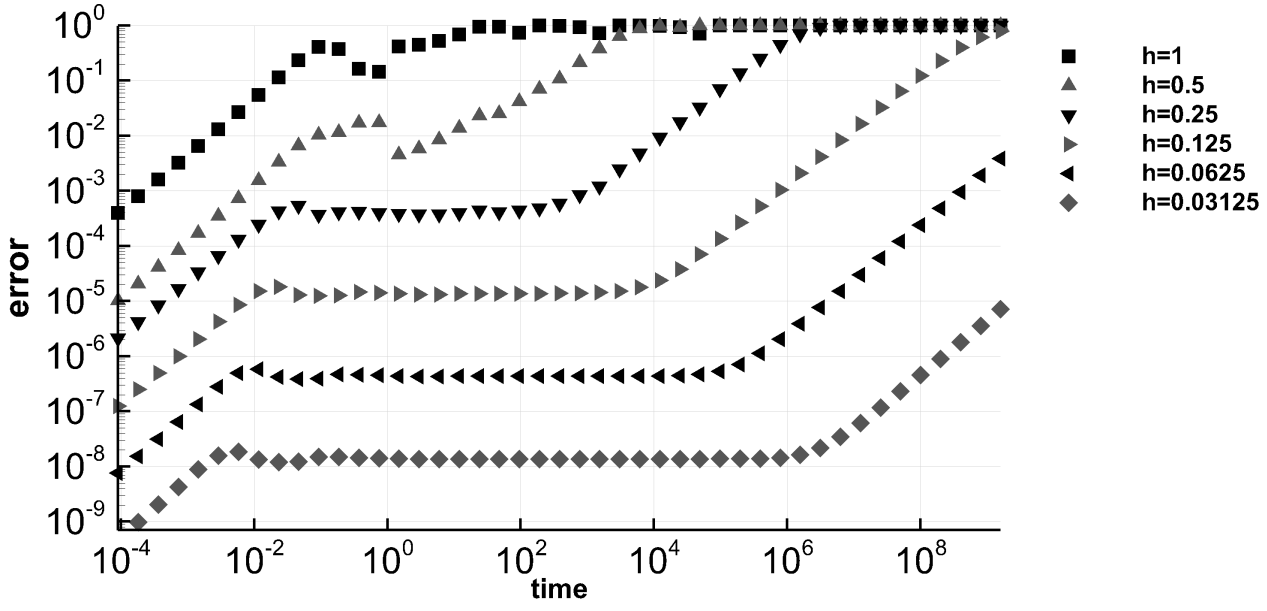

Рис. 2. Ошибка решения $\partial v/\partial t + \partial v/\partial x = 0$, $u_0 = \sin(2\pi x)$, DG, $k = 4$

10.3. Построение схемы с заданными свойствами. Рассмотрим уравнение переноса $\partial v/\partial t + \omega \partial v/\partial x = 0$, $\omega \neq 0$. Пусть $\bar{\aleph}$ – некоторое конечное подмножество $\mathbb{N} \cup \{0\}$. Пусть $p_j \in \mathbb{N} \cup \{0, \infty\}$, $j \in \bar{\aleph}$, – некоторая последовательность, содержащая хотя бы один нулевой элемент. Построим схему вида (2.3), которая будет обладать значениями p_j.

Положим множество переменных на одной ячейке $M^0 = \{1, \ldots, |\bar{\aleph}|\}$. Упорядочим некоторым образом все значения $j \in \bar{\aleph}$ и тем самым обозначим их через $j_1, \ldots, j_M$. Обозначим через $\hat{j}$ любое из значений j, при котором $p_j = 0$, и пусть $\hat{j} = j_{\hat{\xi}}$.

Пусть $c_{m,\eta}$, $m \in \mathbb{N} \cup \{0\}$ – коэффициенты некоторой конечно-разностной аппроксимации производной на равномерной сетке с единичным шагом с порядком, в точности равным m, то есть для всех $m' \leqslant m$ выполняется

$$\sum_{\eta \in \mathcal{S}_m} c_{m,\eta} \eta^{m'} = \delta_{m',1}, \tag{10.1}$$

где $\mathcal{S}_m \subset \mathbb{Z}$ – некоторое конечное множество, а при $m' = m + 1$ это равенство не выполняется. Для $m = -1$ формально положим $\mathcal{S}_{-1} = \{0\}$, $c_{-1,0} = 1$. Для $\eta \notin \mathcal{S}_m$ положим $c_{m,\eta} = 0$.

Пусть S – матрица, элементы $S_{\xi,\xi'}$, $\xi, \xi' \in M^0$, которой равны

$$S_{\xi,\xi} = 1, \quad S_{\xi,\hat{\xi}} = \begin{cases} 1, & p_{j_\xi} > 0, \\ 0, & p_{j_\xi} = 0,\ \xi \neq \hat{\xi}, \end{cases}$$

а остальные элементы равны 0. Очевидно, эта матрица невырождена. Определим схему вида (2.3) следующим образом. Положим

$$\begin{gathered} \mathcal{S} = \bigcup_{m \in \bar{\aleph}} \mathcal{S}_m; \quad Z_0 = I; \quad Z_\eta = 0, \eta \neq 0; \\ L_\eta = \omega S \begin{pmatrix} c_{j_1-1,\eta} & \ldots & 0 \\ \vdots & \ldots & \vdots \\ 0 & \ldots & c_{j_M-1,\eta} \end{pmatrix} S^{-1}. \end{gathered} \tag{10.2}$$

Тогда

$$A(\phi) = i\omega\phi - \sum_\eta L_\eta e^{i\phi\eta} =$$

$$= S \begin{pmatrix} i\omega\phi - \omega \sum\limits_\eta c_{j_1-1,\eta} e^{i\phi\eta} & \ldots & 0 \\ \vdots & \ldots & \vdots \\ 0 & \ldots & i\omega\phi - \omega \sum\limits_\eta c_{j_M-1,\eta} e^{i\phi\eta} \end{pmatrix} S^{-1}.$$

Из (10.1) получаем

$$i\omega\phi - \omega\sum_\eta c_{j_l-1,\eta}e^{i\phi\eta} = i\omega\phi - \omega\sum_{r=0}^{\infty}\frac{(i\phi)^r}{r!}\sum_\eta c_{j_l-1,\eta}\eta^r = \gamma_l\phi^{j_l} + O(\phi^{j_l+1}),$$

где $\gamma_l \neq 0$. Таким образом, имеем представление (3.8)–(3.9), причём присутствуют блоки с $j = j_l$, $l = 1, \ldots, M$, размера 1 и только они.

Чтобы схема обладала заданными значениями p_j, нужно, чтобы

$$v(\phi) = S^{-1}(\phi)\left(\Pi_1 e^{i\phi x}\right)_0 = \begin{pmatrix} \phi^{p_{j_1}}(c_1 + O(\phi)) \\ \ldots \\ \phi^{p_{j_M}}(c_M + O(\phi)) \end{pmatrix}, \tag{10.3}$$

где $c_m \neq 0$. Положим, что элементы этого вектора, стоящие в позиции m, где $p_{j_m} = 0$, в точности равны 1. Тогда требуется, чтобы у вектора $(\Pi_1 \exp(i\phi x))_0 = Sv(\phi)$ компоненты были равны

$$(\Pi_1 \exp(i\phi x))_{0,\xi} = (Sv(\phi))_\xi = \begin{cases} 1, & p_{j_\xi} = 0, \\ 1 + \phi^{p_{j_\xi}}(c_{j_\xi} + O(\phi)), & p_{j_\xi} > 0. \end{cases}$$

Оператор с таким свойством легко построить:

$$(\mathring{\tilde{\Pi}}_h f)_{\eta,\xi} = \begin{cases} f(\eta h), & p_{j_\xi} = 0, \\ f(\eta h) + h^{p_{j_\xi}}\frac{d^{p_{j_\xi}} f}{dx^{p_{j_\xi}}}(\eta h), & p_{j_\xi} > 0. \end{cases}$$

Он имеет вид (8.3), но в силу утверждения 8.1 можно построить и Π_h вида (2.4), также обладающий свойством (10.3). Полученное отображение вместе с коэффициентами (10.2) составляют искомую схему. Вопрос устойчивости этой схемы сводится к устойчивости конечно-разностных схем с коэффициентами $c_{m,\eta}$: если эти схемы устойчивы, то и построенная схема также будет устойчивой.

10.4. Схема с быстроосциллирующим решением. Приведём пример, в котором $A^*(\phi) = -A(\phi)$, $P_A = 0$, $P = 1$. При этом в решении присутствуют гармоники, осциллирующие с периодом $T \sim h$.

Рассмотрим уравнение переноса (2.1) при скорости переноса $\omega = 1$. Положим $M^0 = \{1, 2, 3, 4\}$. Введём вспомогательные матрицы

$$X = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

и определим коэффициенты схемы следующим образом:

$$Z_0 = I, \quad L_{\pm1} = \pm I/2 - X/4, \quad L_0 = -X/2 - Y.$$

Остальные коэффициенты равны нулю. Введём отображение $\overline{\Pi}_h$ равенствами $(\overline{\Pi}_h f)_{\eta,1} = f((\eta+1)h)$, $(\overline{\Pi}_h f)_{\eta,2} = (\overline{\Pi}_h f)_{\eta,3} = (\overline{\Pi}_h f)_{\eta,4} = f(\eta h)$.

Матрица $A(\phi)$ имеет вид

$$A(\phi) = i\phi I - \big((I/2 - X/4)e^{i\phi} + (-X/2 - Y) + (-I/2 - X/4)e^{-i\phi}\big) =$$

$$= i(\phi - \sin\phi)I + Y + X\cos^2(\phi/2).$$

Обозначим $v = \phi - \sin\phi$, $w = \sin^2(\phi/2)$. Тогда матрица $A(\phi)$ представима в виде

$$A(\phi) = ivI + Y + X(1-w) = S\Lambda(\phi)S^{-1},$$

где

$$\Lambda(\phi) = \begin{pmatrix} i(v+w-2) & 0 & 0 & 0 \\ 0 & i(v-w+2) & 0 & 0 \\ 0 & 0 & i(v+w) & 0 \\ 0 & 0 & 0 & i(v-w) \end{pmatrix},$$

$$S^{-1} = \frac{1}{2}\begin{pmatrix} i & -1 & -i & 1 \\ -i & -1 & i & 1 \\ -i & 1 & -i & 1 \\ i & 1 & i & 1 \end{pmatrix}, \quad S = (S^{-1})^*.$$

Первые два собственных значения стремятся к $\pm 2i$ при $\phi \to 0$ и соответствуют паразитным быстро осциллирующим волнам, последние два собственных значения соответствуют физическим компонентам ($j = 2$). Рассмотрим величину $v(\phi)$, определённую (5.1):

$$v(\phi) = S^{-1}\begin{pmatrix} e^{i\phi} \\ 1 \\ 1 \\ 1 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} i(e^{i\phi}-1) \\ -i(e^{i\phi}-1) \\ -i(e^{i\phi}+1)+2 \\ i(e^{i\phi}+1)+2 \end{pmatrix}.$$

Первые две компоненты имеют первый порядок малости по ϕ, поэтому $p_0 = 1$. Последние две компоненты имеют конечный предел при $\phi = 0$, поэтому $p_2 = 0$. Вычисляя значения порядка аппроксимации, порядка точности и порядка точности в длительном счёте, получаем $P_A = 0$, $P = Q = 1$.

10.5. Об оптимальности оценки. В определении оптимальной оценки порядка точности при длительном счёте мы поставили условие, что эта оптимальность не может достигаться за счёт ухудшения формального порядка точности. Поясним это на примере. Положим $\aleph = \{3{,}5\}$, $p_3 = 1$, $p_5 = 0$. Схема с такими параметрами обладает 3-м порядком аппроксимации и оценкой точности (7.7):

$$\|\varepsilon_h(t, v_0, \Pi_h)\| \leqslant C_1 \min\{h\|\nabla v_0\|, h^3 t\|\nabla^4 v_0\|\} + C_2 \min\{\|v_0\|, h^4 t\|\nabla^5 v_0\|\}.$$

Значит, она обладает формальным порядком точности $P = 3$ и порядком точности в длительном счёте $Q = 3$, причём эти значения оптимальны в смысле данного определения. При этом мы можем записать оценку

$$\|\varepsilon_h(t, v_0, \Pi_h)\| \leqslant C_1 h\|\nabla v_0\| + C_2 h^4 t\|\nabla^5 v_0\|,$$

имеющую вид (2.8). Таким образом, можно сказать, что схема обладает 1-м формальным порядком точности и 4-м порядком точности в длительном счёте. Однако в смысле данного определения эта оценка не будет оптимальной. Отметим, что по теореме 7.1 существует отображение, отличающееся от Π_h на величину $O(h)$, в смысле которого имеется 4-й порядок точности, как формальный, так и в длинном счёте.

10.6. О нелинейном росте ошибки со временем. Линейная скорость роста второго слагаемого со временем в оценке (2.8) является существенной. Рассмотрим конечно-разностную схему 3-го порядка на равномерной сетке. Она обладает ошибкой $\|\varepsilon_h(t, e^{i\alpha x}, \Pi_h)\| \leqslant C|\alpha|^4 h^3 t$. Отсюда напрямую следует, что $\|\varepsilon_h(t, e^{i\alpha x})\| \leqslant C(|\alpha|^2 h^2 + |\alpha|^6 h^4 t^2)/2$. Следовательно, $\|\varepsilon_h(t, v_0, \Pi_h)\| \leqslant \tilde{C}(\|\nabla^2 v_0\|^2 h^2 + \|\nabla^6 v_0\| h^4 (t^2 + h^2))$. Физически нет никаких двух компонент ошибки, но с формальной точки зрения мы можем сказать, что *оценка* нормы ошибки складывается из двух компонент. Выбором отображения решения на пространство сеточных функций мы на первое слагаемое повлиять по существу не можем.

10.7. Достаточное условие отсутствия сверхсходимости.

Утверждение 10.1. *Пусть матрица $A(\phi)$, заданная* (3.1)*, удовлетворяет условию $A(0) = 0$. Тогда $P_A = P$.*

Действительно, если $A(0) = 0$, то в матрице $M(\phi)$ блока, соответствующего $j = 0$, нет, поэтому $0 \notin \aleph$ и выражение (6.8), определяющее P, совпадает с выражением (5.3), определяющим значение P_A.

Отметим, что условие $A(0) = 0$ равносильно $\sum L_\eta = 0$. Например, такое имеет место, если $L_\eta = -L_{-\eta}$.

11. Примеры применения алгоритма

В настоящем разделе рассмотрим две схемы для решения уравнения переноса $\partial v/\partial t + \partial v/\partial x = 0$, на равномерной сетке с узлами $x_j = jh$.

11.1. P1-DG. Рассмотрим метод Галёркина с разрывными базисными функциями на основе многочленов 1-го порядка. На каждой сеточной ячейке определим две базисные функции:

$$\phi_j^L(x) = \frac{x_{j+1} - x}{h}, \quad \phi_j^R(x) = \frac{x - x_j}{h}.$$

При этом схема запишется в виде

$$\begin{pmatrix} \frac{1}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{3} \end{pmatrix} \frac{du_j}{dt} + \frac{1}{h}\left[\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} u_j + \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix} u_{j-1} \right] = 0.$$

Здесь $u_j = (u_j^L, u_j^R)^T$ – коэффициенты разложения решения по базисным функциям в пределах ячейки (x_j, x_{j+1}). Отсюда имеем

$$A(\phi) = i\phi I - \begin{pmatrix} \frac{1}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{3} \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} - e^{-i\phi} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} i\phi - 3 & -1 + 4e^{-i\phi} \\ 3 & i\phi - 1 - 2e^{-i\phi} \end{pmatrix}.$$

Собственные значения этой матрицы равны $\lambda_\pm(\phi) = i\phi - (2 + e^{-i\phi}) \pm \varkappa(\phi)$, где $\varkappa(\phi) = (e^{-2i\phi} + 10e^{-i\phi} - 2)^{1/2}$. Матрица правых собственных векторов

$$S(\phi) = \frac{1}{2\varkappa(\phi)(4e^{-i\phi} - 1)} \begin{pmatrix} 4e^{-i\phi} - 1 & -4e^{-i\phi} + 1 \\ 1 - e^{-i\phi} + \varkappa(\phi) & -1 + e^{-i\phi} + \varkappa(\phi) \end{pmatrix},$$

матрица левых собственных векторов

$$S^{-1}(\phi) = \begin{pmatrix} -1 + e^{-i\phi} + \varkappa(\phi) & 4e^{-i\phi} - 1 \\ -1 + e^{-i\phi} - \varkappa(\phi) & 4e^{-i\phi} - 1 \end{pmatrix}.$$

Первый столбец $S(\phi)$ и, соответственно, первая строка $S^{-1}(\phi)$ соответствует $\lambda_+(\phi)$, второй столбец и вторая строка – $\lambda_-(\phi)$. Дифференциальное приближение для $\lambda_\pm$ имеет вид $\lambda_+(\phi) \approx -\phi^4/72 + O(\phi^5)$, $\lambda_-(\phi) = -6 + O(\phi)$. Собственное значение $\lambda_+(\phi)$ соответствует физической волне, а собственное значение $\lambda_-(\phi)$ – нефизической волне.

Чтобы говорить об ошибке аппроксимации и о точности схемы, нужно выбрать оператор Π_h. Для простоты будем использовать точечное отображение: $(\Pi_h f)_j^L = f(x_j)$, $(\Pi_h f)_j^R = f(x_{j+1})$. Тогда

$$v(\phi) = S^{-1}(\phi)(\Pi_1 e^{i\phi x})_0 = \begin{pmatrix} -1 + e^{-i\phi} + \varkappa(\phi) & 4e^{-i\phi} - 1 \\ -1 + e^{-i\phi} - \varkappa(\phi) & 4e^{-i\phi} - 1 \end{pmatrix} \begin{pmatrix} 1 \\ e^{i\phi} \end{pmatrix} =$$

$$= \begin{pmatrix} 3 + e^{-i\phi} - e^{i\phi} + \varkappa(\phi) \\ 3 + e^{-i\phi} - e^{i\phi} - \varkappa(\phi) \end{pmatrix} = \begin{pmatrix} 6 + O(\phi) \\ \frac{1}{2}\phi^2 + O(\phi^3) \end{pmatrix}.$$

Таким образом, в ранее введённых терминах имеем $p_0 = 2$, $p_4 = 0$, остальные $p_j = \infty$. Поэтому схема обладает 1-м порядком аппроксимации, а оптимальная оценка ошибки решения имеет вид $O(h^2 + h^3 t)$.

Заменив отображение Π_h на $\tilde{\Pi}_h$, можно добиться увеличения значения p_0 и, следовательно, порядка аппроксимации и формального порядка. Условием на это отображение является то, чтобы вторая компонента вектора $S^{-1}(\phi)(\tilde{\Pi}_1 e^{i\phi x})_0$ имела более высокий порядок малости. Формально мы можем её занулить, положив, например

$$(\tilde{\Pi}_1 e^{i\phi x})_0 = \begin{pmatrix} 1 \\ \dfrac{1 + \varkappa(\phi) - e^{-i\phi}}{4e^{-i\phi} - 1} \end{pmatrix}. \tag{11.1}$$

Введённое отображение является нелокальным; чтобы определить оператор вида (2.5) или (8.3), нужно взять аппроксимацию (11.1). Например, аппроксимации

$$(\tilde{\Pi}_1 e^{i\phi x})_0 = \begin{pmatrix} 1 \\ e^{i\phi} - \frac{1}{6}\phi^2 - \frac{5}{18} i\phi^3 \end{pmatrix} \tag{11.2}$$

соответствует отображение

$$\begin{gathered} (\tilde{\Pi}_h f)_{\eta,L} = f(\eta h), \\ (\tilde{\Pi}_h f)_{\eta,R} = f((\eta+1)h) + \frac{1}{6}h^2 f''(\eta h) + \frac{5}{18}h^3 f'''(\eta h). \end{gathered} \tag{11.3}$$

Оно имеет вид (8.3) при $\mathcal{P}_h$, определённом равенством $(\mathcal{P}_h f)_{\eta,L} = (\mathcal{P}_h f)_{\eta,R} = f(h\eta)$. При этом получим $p_0 = 4$, $p_4 = 0$, и схема будет обладать порядком аппроксимации 3 и оценкой ошибки $O(h^3(h+t))$. Если же при определении отображения (11.3) опустить слагаемое порядка h^3, получим $p_0 = 3$, $p_4 = 0$, порядок аппроксимации 2 и оценку ошибки $O(h^3(1+t))$.

Существует множество других локальных отображений, доставляющих третий порядок аппроксимации. Например, равенство

$$(\tilde{\Pi}_1 e^{i\phi x})_0 = \begin{pmatrix} 1 \\ e^{i\phi}\left(1 - \frac{1}{6}\phi^2 - \frac{1}{9} i\phi^3\right) \end{pmatrix},$$

также аппроксимирующее (11.1), соответствовало бы выбору $\mathcal{P}_h = \Pi_h$ и отображению

$$\begin{gathered} (\tilde{\Pi}_h f)_{\eta,L} = f(\eta h), \\ (\tilde{\Pi}_h f)_{\eta,R} = f((\eta+1)h) + \frac{1}{6}h^2 f''((\eta+1)h) + \frac{1}{9}h^3 f'''((\eta+1)h). \end{gathered}$$

А равенство

$$(\tilde{\Pi}_1 e^{i\phi x})_0 = \frac{1}{3}\begin{pmatrix} 4e^{-i\phi}-1 \\ 1-e^{-i\phi}+(3+2e^{-i\phi}-\frac{1}{2}e^{-2i\phi}+\frac{1}{3}e^{-3i\phi}) \end{pmatrix},$$

с 3-м порядком аппроксимирующее правую часть (11.1), умноженную на $(4e^{i\phi}-1)/3$, получается для отображения $\tilde{\Pi}_h$, заданного равенствами

$$(\tilde{\Pi}_h f)_{\eta,L} = \frac{4}{3}f((\eta-1)h) - \frac{1}{3}f(\eta h),$$
$$(\tilde{\Pi}_h f)_{\eta,R} = \frac{1}{9}f((\eta-3)h) - \frac{1}{6}f((\eta-2)h) + \frac{1}{3}f((\eta-1)h) + \frac{4}{3}f(\eta h).$$

Это отображение не имеет представления (8.3) ни для какого $\mathcal{P}_h$.

Покажем теперь, как находятся коэффициенты $\mathfrak{C}^{(m)}$ построенным алгоритмом, не прибегая к спектральному анализу. Поскольку собственное значение $\lambda = 0$ матрицы $A(0)$ простое, будем использовать алгоритм 2. Подставив в схему функцию $f = x$, установим, что схема обладает первым порядком аппроксимации, а подставив функцию $f = x^2/2$, – что вторым порядком аппроксимации она не обладает.

Вычислим $f^2 = -(\epsilon_1(x^2/2, \Pi_1))_0$. Имеем

$$f^2 = \begin{pmatrix} \frac{1}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{3} \end{pmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix} - \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}\begin{bmatrix} 0 \\ 1/2 \end{bmatrix} - \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix}\begin{bmatrix} 1/2 \\ 0 \end{bmatrix} = \begin{bmatrix} -1/12 \\ 1/12 \end{bmatrix}.$$

Система уравнений

$$\begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}\mathfrak{C}^{(2)}\mathfrak{e} = \begin{pmatrix} -1/12 \\ 1/12 \end{pmatrix},$$

где $\mathfrak{e} = (1,1)^T$, имеет решение

$$\mathfrak{C}^{(2)} = \begin{pmatrix} 0 & 0 \\ 0 & 1/6 \end{pmatrix} + c_2 I.$$

Положим $c_2 = 0$. Коэффициенты $\mathfrak{C}^{(3)}$ уже будут существенно зависеть от выбора дополнительного отображения $\mathcal{P}_h$. Определим его, например, следующим образом: $(\mathcal{P}_h f)_{\eta,L} = (\mathcal{P}_h f)_{\eta,R} = f(h\eta)$. Вычислим теперь $f^3 = -(\epsilon_1(x^3/6, \tilde{\Pi}_1^{(2,2)}))_0$, где $\tilde{\Pi}_h^{(2,2)}$ имеет вид (2.5) с подстановкой найденного значения $\mathfrak{C}^{(2)}$. Имеем

$$f^3 = \begin{pmatrix} \frac{1}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{3} \end{pmatrix} \left[\begin{pmatrix} 0 \\ 1/2 \end{pmatrix} + \mathfrak{C}^{(2)} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right] - \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \left[\begin{pmatrix} 0 \\ 1/6 \end{pmatrix} + \mathfrak{C}^{(2)} \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right] -$$
$$- \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix} \left[\begin{pmatrix} -1/6 \\ 0 \end{pmatrix} + \mathfrak{C}^{(2)} \begin{pmatrix} -1 \\ -1 \end{pmatrix} \right] = \begin{pmatrix} -5/36 \\ 5/36 \end{pmatrix}.$$

Система уравнений

$$\begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \mathfrak{C}^{(3)} \mathfrak{e} = \begin{pmatrix} -5/36 \\ 5/36 \end{pmatrix}$$

имеет решение

$$\mathfrak{C}^{(3)} = \begin{pmatrix} 0 & 0 \\ 0 & 5/18 \end{pmatrix} + c_3 I.$$

Оператор $\tilde{\Pi}_h^{(2,3)}$, порождённый выбранными Π_h, $\mathcal{P}_h$ и матрицами $\mathfrak{C}^{(2)}$ и $\mathfrak{C}^{(3)}$ при $c_2 = c_3 = 0$, имеет вид 11.3.

Система $L(0)\mathfrak{C}^{(4)}\mathfrak{e} = f^4$ будет несовместной. Поэтому $P = 2$ и $Q = 3$ являются оптимальными значениями формального порядка точности и порядка точности в длительной счёте.

11.2. Комбинация центральных разностей. Пусть множество степеней свободы на один блок (шаг) сетки $M^0 = \{\text{“L”}, \text{“R”}\}$, $|M^0| = 2$. Будем интерпретировать первую степень свободы как значение в целом узле, а вторую – как значение в полуцелом узле. Соответственно введём точечный проектор $\overline{\Pi}_h$ равенствами $(\overline{\Pi}_h f)_{j,L} = f(jh)$, $(\overline{\Pi}_h f)_{j,R} = f(jh + h/2)$, $j \in \mathbb{Z}$. Для краткости будем обозначать $(u_j)_L = v_j$, $(u_j)_R = v_{j+1/2}$.

Рассмотрим следующую полудискретную схему для уравнения переноса (2.1) при скорости переноса $\omega = 1$. В целых узлах для определения производной будем использовать двухточечную центральная разность 2-го порядка аппроксимации, а полуцелых узлах – 4-точечную центральная разность 4-го порядка аппроксимации. Эта схема имеет вид

$$\begin{gathered} \frac{dv_j}{dt} + \frac{v_{j+1/2} - v_{j-1/2}}{h} = 0, \quad j \in \mathbb{Z}; \\ \frac{dv_{j+1/2}}{dt} + \frac{4}{3}\frac{v_{j+1} - v_j}{h} - \frac{1}{3}\frac{v_{j+3/2} - v_{j-1/2}}{2h} = 0, \quad j \in \mathbb{Z}. \end{gathered} \tag{11.4}$$

В смысле $\overline{\Pi}_h$ схема (11.4) обладает вторым порядком аппроксимации.

В блочном виде она переписывается как

$$\frac{du_j}{dt} + L_{-1}u_{j-1} + L_0 u_j + L_1 u_{j+1} = 0,$$

$$L_{-1} = \begin{pmatrix} 0 & -1 \\ 0 & 1/6 \end{pmatrix}, \quad L_0 = \begin{pmatrix} 0 & 1 \\ -4/3 & 0 \end{pmatrix}, \quad L_1 = \begin{pmatrix} 0 & 0 \\ 4/3 & -1/6 \end{pmatrix}.$$

Рассмотрим матрицу $L(\phi) = e^{-i\phi}L_{-1} + L_0 + e^{i\phi}L_1$:

$$L(\phi) = \begin{pmatrix} 0 & 1 - e^{-i\phi} \\ \frac{4}{3}(-1 + e^{i\phi}) & -\frac{1}{6}(e^{i\phi} - e^{-i\phi}) \end{pmatrix} = \begin{pmatrix} 0 & 2ie^{-i\phi/2}\sin(\phi/2) \\ \frac{8}{3}ie^{i\phi/2}\sin(\phi/2) & -\frac{1}{3}i\sin(\phi) \end{pmatrix}.$$

Тогда матрица $A(\phi)$, определённая (3.1), имеет вид

$$A(\phi) = \begin{pmatrix} i\phi & -2ie^{-i\phi/2}\sin(\phi/2) \\ -\frac{8}{3}ie^{i\phi/2}\sin(\phi/2) & i\phi + \frac{1}{3}i\sin(\phi) \end{pmatrix}.$$

Её собственные значения находятся из решения квадратного уравнения

$$(\lambda(\phi) - i\phi)^2 - \frac{1}{3}i\sin(\phi)(\lambda(\phi) - i\phi) + \frac{16}{3}\sin^2(\phi/2) = 0$$

и равны

$$\lambda_\pm(\phi) = i\phi + \frac{1}{6}i\sin(\phi) \pm \frac{1}{6}i\sin(\phi/2)\varkappa(\phi),$$

где $\varkappa(\phi) = \sqrt{194 + 2\cos\phi} > 0$.

Матрица правых собственных векторов

$$S(\phi) = \frac{1}{4\varkappa(\phi)} \begin{pmatrix} 2\exp(-i\phi/2) & 2\exp(-i\phi/2) \\ i\frac{\lambda_+}{\sin(\phi/2)} & i\frac{\lambda_-}{\sin(\phi/2)} \end{pmatrix}.$$

Матрица левых собственных векторов

$$S^{-1}(\phi) = \begin{pmatrix} \exp(i\phi/2)(\varkappa(\phi) - 2\cos(\phi/2)) & -12 \\ \exp(i\phi/2)(\varkappa(\phi) + 2\cos(\phi/2)) & 12 \end{pmatrix}.$$

Первый столбец $S(\phi)$ и, соответственно, первая строка $S^{-1}(\phi)$ соответствуют $\lambda_+(\phi)$, второй столбец и вторая строка – $\lambda_-(\phi)$. Легко убедиться, что при $\phi \to 0$ матрицы $S(\phi)$ и $S^{-1}(\phi)$ имеют конечные пределы.

Собственные значения $\lambda_\pm(\phi)$ являются чисто мнимыми. Кроме того, $\lambda_+(\phi) = \lambda_-(\phi)$ только при $\phi = 0$, где $A(\phi) = 0$. Таким образом, для всех ϕ матрица $A(\phi)$ имеет два линейно независимых собственных вектора. Поскольку при $\phi \to 0$ матрицы $S(\phi)$ и $S^{-1}(\phi)$ имеют конечный предел, отсюда следует (3.2). Таким образом, схема (11.4) устойчива.

Разлагая собственные значения в ряд Тейлора, получаем

$$\lambda_+(\phi) = \frac{7}{3}i\phi + O(\phi^3), \quad \lambda_-(\phi) = \frac{1}{42}i\phi^3 + O(\phi^5).$$

Видно, что собственное значение $\lambda_-(\phi)$ соответствует физической моде, а $\lambda_+(\phi)$ – нефизической. Матрица S^{-1} имеет ряд Тейлора

$$S^{-1}(\phi) = \begin{pmatrix} 12 + 6i\phi - \frac{9}{7}\phi^2 + \dots & -12 \\ 16 + 8i\phi - \frac{16}{7}\phi^2 + \dots & 12 \end{pmatrix}.$$

Рассмотрим "блочный" проектор Π_h, определённый равенством

$$(\Pi_h f)_{j,L} = (\Pi_h f)_{j,R} = f(jh).$$

Легко убедиться, что в смысле Π_h ошибка аппроксимации имеет первый порядок малости и не обладает вторым порядком малости. Действуя в рамках спектрального анализа, запишем

$$v(\phi) = S^{-1}(\phi)(\Pi_1 e^{i\phi x})_0 = \begin{pmatrix} 6i\phi + O(\phi^2) \\ 28 + O(\phi) \end{pmatrix}.$$

Таким образом, имеем $p_1 = 1$, $p_3 = 0$, остальные $p_j = \infty$. Отсюда следует оценка

$$\varepsilon(t, v_0) \leqslant C_1 h \|\nabla v_0\|_\infty + C_2 t h^2 \|\nabla^3 v_0\|_\infty. \tag{11.5}$$

Эту же оценку можно получить, используя проектор $\overline{\Pi}_h$: если обозначить через $\overline{u}$ решение по схеме (11.4) с начальными данными $\overline{\Pi}_h v_0$ и ввести $v_t(x) = v_0(x - t)$, получим

$$\varepsilon_h(t, v_0) = u - \Pi_h v_t = (u - \overline{u}) + (\overline{u} - \overline{\Pi}_h v_t) + (\overline{\Pi}_h v_t - \Pi_h v_t).$$

Поскольку для любой функции по построению отображений $\overline{\Pi}_h$ и Π_h выполняется верно $\|\overline{\Pi}_h f - \Pi_h f\| \leqslant Ch\|\nabla v_0\|_\infty$, третье слагаемое по норме не превосходит $Ch\|\nabla v_0\|_\infty$. Первое слагаемое есть разница между решениями задач по одной схеме с разными начальными данными, поэтому

$$\|u - \overline{u}\| \leqslant K\|\Pi_h v_0 - \overline{\Pi}_h v_0\| \leqslant CKh\|\nabla v_0\|_\infty.$$

Наконец, второе слагаемое есть ошибка решения $\varepsilon_h(t, v_0)$ при использовании отображения $\overline{\Pi}_h$. Поэтому

$$\|\overline{u} - \overline{\Pi}_h v_t\| \leqslant K \int_0^t \|\epsilon_h(v_\tau, \overline{\Pi}_h)\| d\tau \leqslant CKth^2 \|\nabla^3 v_0\|.$$

Складывая эти три оценки, получаем искомый результат (11.5).

Продемонстрируем теперь, как анализ точности схемы (11.4) с отображением Π_h проводится при помощи алгоритма 1.

Легко проверить, что схема обладает порядком аппроксимации $P_A = 1$ и не обладает 2-м порядком аппроксимации. Поскольку $A(0) = 0$, в силу утверждения 10.1 выполняется $P = P_A = 1$, и шаг 2 можно пропустить. Рассмотрим систему $\mathcal{F}_2(\mathfrak{C}^{(1)}, \mathfrak{C}^{(2)}) = 0$. Она состоит из двух равенств. Поскольку $(\epsilon_1(x, \Pi_1))_0 = 0$ и $L(0) = 0$, для любых $\mathfrak{C}^{(1)}$ и $\mathfrak{C}^{(2)}$ имеем $(\epsilon_1(x, \tilde{\Pi}_1^{(1,2)}))_0 = (\epsilon_1(x, \Pi_1))_0 + L(0)\mathfrak{C}^{(1)}\mathfrak{e} = 0$. Равенство $(\epsilon_1(x^2/2, \tilde{\Pi}_1^{(1,2)}))_0 = 0$ переписывается в виде

$$-\mathfrak{C}^{(1)}\begin{pmatrix}1\\1\end{pmatrix} + \begin{pmatrix}0 & -1\\0 & 1/6\end{pmatrix}\left[\begin{pmatrix}1/2\\1/2\end{pmatrix} + \mathfrak{C}^{(1)}\begin{pmatrix}-1\\-1\end{pmatrix} + \mathfrak{C}^{(2)}\mathfrak{e}\right] +$$
$$+\begin{pmatrix}0 & 1\\-4/3 & 0\end{pmatrix}\mathfrak{C}^{(2)}\mathfrak{e} + \begin{pmatrix}0 & 0\\4/3 & -1/6\end{pmatrix}\left[\begin{pmatrix}1/2\\1/2\end{pmatrix} + \mathfrak{C}^{(1)}\begin{pmatrix}1\\1\end{pmatrix} + \mathfrak{C}^{(2)}\mathfrak{e}\right] = 0.$$

Поскольку $L(0) = 0$, слагаемые, содержащие $\mathfrak{C}^{(2)}$, взаимно уничтожаются. Упрощая, получаем систему уравнений на нахождение элементов диагональной матрицы $\mathfrak{C}^{(1)}$:

$$\begin{cases} -\mathfrak{C}_L^{(1)} + \mathfrak{C}_R^{(1)} - \frac{1}{2} = 0; \\ \frac{4}{3}\mathfrak{C}_L^{(1)} - \frac{4}{3}\mathfrak{C}_R^{(1)} + \frac{2}{3} = 0. \end{cases}$$

Решением этой системы является матрица

$$\mathfrak{C}^{(1)} = \begin{pmatrix}0 & 0\\0 & 1/2\end{pmatrix} + \alpha_1 I. \tag{11.6}$$

Матрица $\mathfrak{C}^{(2)}$ может быть любой. Полученный оператор $\tilde{\Pi}_h$, если положить $\alpha_1 = 0$ и $\mathfrak{C}^{(2)} = 0$, выражается формулами $(\tilde{\Pi}_h f)_{j,L} = f(jh)$, $(\tilde{\Pi}_h f)_{j,L} = f(jh) + (h/2)f'(jh)$. Отметим, что с точностью до $O(h^2)$ оператор $\tilde{\Pi}_h$ совпадает с $\overline{\Pi}_h$.

Запишем теперь систему $\mathcal{F}_3(\mathfrak{C}^{(1)}, \mathfrak{C}^{(2)}, \mathfrak{C}^{(3)}) = 0$. Она состоит из трёх равенств. Первое выполняется при любых $\mathfrak{C}^{(m)}$, второе равносильно (11.6). Третье равенство даёт

$$-\mathfrak{C}^{(2)}\begin{pmatrix}1\\1\end{pmatrix} + \begin{pmatrix}0 & -1\\0 & 1/6\end{pmatrix}\left[\begin{pmatrix}-1/6\\-1/6\end{pmatrix} + \mathfrak{C}^{(1)}\begin{pmatrix}1/2\\1/2\end{pmatrix} + \mathfrak{C}^{(2)}\begin{pmatrix}-1\\-1\end{pmatrix}\right] +$$
$$+\begin{pmatrix}0 & 0\\4/3 & -1/6\end{pmatrix}\left[\begin{pmatrix}1/6\\1/6\end{pmatrix} + \mathfrak{C}^{(1)}\begin{pmatrix}1/2\\1/2\end{pmatrix} + \mathfrak{C}^{(2)}\begin{pmatrix}1\\1\end{pmatrix}\right] = 0$$

(поскольку $A(0) = 0$, слагаемые, содержащие $\mathfrak{C}^{(3)}$, сократились). Легко проверить, что мы получили несовместную систему уравнений. Это доказывает, что значения $P = 1$ и $Q = 2$ являются оптимальными.

12. Заключение

Для разностной схемы общего вида установление порядка точности, а тем более специфики поведения ошибки в длительном счёте, является непростой задачей. Поэтому на практике при анализе схемы в первую очередь устанавливают её порядок аппроксимации P_A . Если схема является устойчивой, то разность между решением разностной задачи и сеточным образом точного решения является величиной порядка h^{P_A}. Но такая оценка часто является грубой.

В настоящей работе предложен новый алгоритм исследования схем вида (2.3). В предположении L_2-устойчивости алгоритм 1 позволяет за конечное число операций в точной арифметике установить оценку вида (2.8) с оптимальными значениями порядка точности P и порядка точности в длительном счёте Q, если это значение конечно. В "простом" случае ($\mathrm{Ker} L(0) = \mathrm{span}(\mathfrak{e})$) можно пользоваться более быстрым алгоритмом 2.

Список литературы

1. Lowrie R. Compact higher-order numerical methods for hyperbolic conservation laws. Ph.D. thesis: The University of Michigan. 1996.
2. 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.
3. Zhang M., Shu C.-W. An analysis of and a comparison between the discontinuous Galerkin and the spectral finite volume methods // Computers and Fluids. 2005. Vol. 34. P. 581–592.
4. Guo W., Zhong X., Qui C.-M. Superconvergence of discontinuous Galerkin and local discontinuous Galerkin methods: Eigen-structure analysis based on Fourier approach // J. Comput. Phys. 2013. Vol. 235. P. 458–485.
5. Cheng Y., Shu C.-W. Superconvergence and time evolution of discontinuous Galerkin finite element solutions // J. Comput. Phys. 2008. Vol. 227, no. 22. P. 9612–9627.
6. Yang Y., Shu C.-W. Analysis of optimal superconvergence of discontinuous Galerkin method for linear hyperbolic equations // SIAM Journal on Numerical Analysis. 2012. Vol. 50, no. 6. P. 3110–3133.
7. Zhang Q., Shu C.-W. Stability analysis and a priori error estimates of the third order explicit Runge–Kutta discontinuous Galerkin method for scalar conservation laws // SIAM Journal of Numerical Analysis. 2010. Vol. 48, no. 3. P. 1038–1063.
8. Бахвалов П. А., Сурначёв М. Д. О спектральном анализе схем для линейного уравнения переноса // Препринты ИПМ им. М.В.Келдыша. 2019. № 70. с. 28.
9. Бахвалов П. А., Сурначёв М. Д. Об аналитических семействах матриц, порождающих ограниченные полугруппы // Сибирский журнал вычислительной математики. (представлено в редакцию).
10. Levis A. H. Some computational aspects of the matrix exponential // IEEE Trans. Automatic Control, AC-14. 1969. P. 410–411.