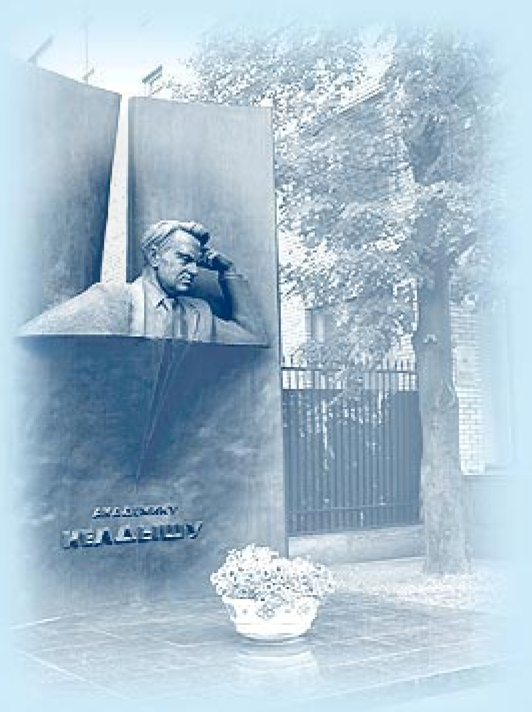
ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 4 за 2024 г.



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

П.А. Бахвалов, М.Д. Сурначёв

Об устойчивости и точности конечно-объёмных схем на неравномерных сетках

Рекомендуемая форма библиографической ссылки: Бахвалов П.А., Сурначёв М.Д. Об устойчивости и точности конечно-объёмных схем на неравномерных сетках // Препринты ИПМ им. М.В.Келдыша. 2024. № 4. 39 с. **https://doi.org/10.20948/prepr-2024-4**
https://library.keldysh.ru/preprint.asp?id=2024-4

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

Об устойчивости и точности конечно-объёмных схем на неравномерных сетках

Москва — 2024

П. А. Бахвалов, М. Д. Сурначёв

Об устойчивости и точности конечно-объёмных схем на неравномерных сетках

В работе изучается поведение конечно-объёмных схем для одномерного уравнения переноса на неравномерных сетках. Рассматриваются схемы с полиномиальной реконструкцией и схемы на основе разделённых разностей. Доказывается достаточное условие устойчивости при малых деформациях расчётной сетки. Также устанавливаются оценки ошибки численного решения.

Ключевые слова: метод конечных объёмов, суперсходимость

Pavel Alexeevich Bakhvalov, Mikhail Dmitrievich Surnachev

On stability and accuracy of finite-volume schemes on non-uniform meshes

The paper studies the behavior of high-order finite-volume schemes for the 1D transport equation on non-uniform meshes. We consider finite-volume schemes with a polynomial reconstruction and schemes based on divided differences. We prove a sufficient stability condition on mildly deformed meshes and establish estimates for the solution error.

Key words: finite volume method, consistency and accuracy, supra-convergence, long-time simulation accuracy

Содержание

1. Введение

В настоящей работе изучаются свойства схем для одномерного уравнения переноса с постоянной скоростью на периодических сетках. Рассматриваются конечно-объёмные схемы с полиномиальной реконструкцией, лежащие в основе конечно-объёмных схем высокого порядка на неструктурированных сетках [1,2]. Также рассматриваются схемы на основе разделённых разностей (далее будем называть их R3 и R5), лежащие в основе схем семейства EBR [3,4]. Целью настоящей работы является объяснение на простой модели эффектов, наблюдаемых в газодинамических расчётах в зоне, где доминирует конвекция и решение при этом является достаточно гладким.

Стандартный и разрывный методы Галёркина применительно к линейному уравнению переноса устойчивы в L_2 на произвольной неструктурированной сетке. Для конечно-объёмных методов высокого порядка это неверно. На равномерных декартовых сетках они вырождаются в конечно-разностные схемы, свойства которых хорошо изучены. На неструктурированных сетках есть только практические рекомендации, как добиться устойчивости (см., например, [5]), но искусственным построением “плохой” сетки устойчивость можно нарушить.

Численные эксперименты для одномерного уравнения переноса, проведённые по схемам с полиномиальной реконструкцией, демонстрируют следующие эффекты:

- схемы остаются устойчивыми, по меньшей мере, при небольшой деформации сетки относительно равномерной;
- если порядок аппроксимации равен p на произвольной сетке и $p+1$ на равномерной сетке, то порядок точности на неравномерной сетке равен $p+1$.

Здесь и далее под порядком аппроксимации понимается порядок дифференциального приближения, а под порядком точности – скорость стремления к нулю нормы ошибки решения при измельчении сетки. Первой задачей настоящей работы является теоретическое обоснование этих фактов.

Вторая задача заключается в объяснении свойств схемы R3. На неравномерной сетке она точна только на линейной функции, а на равномерной вырождается в конечно-разностную схему 3-го порядка. Наш анализ показывает, что решение по этой схеме сходится со 2-м порядком. Однако член 2-го порядка имеет вид $O(h_{\max}(h_{\max} - h_{\min}) + (\Delta h)^2_{\max} t)$, где $h_{\max}$ и $h_{\min}$ – максимальный и минимальный шаги сетки, $(\Delta h)_{\max}$ – максимальная разность между соседними шагами, а t – момент времени. Это означает, что схема обладает так называемой повышенной точностью в длительном счёте. Этот эффект известен для разрывного метода Галёркина [6,7], который в длительном счёте имеет порядок $2p+1$ при формальном порядке $p+1$. В нашем случае порядок точности в длительном счёте такой же, как и формальный порядок. Но с ростом времени счёта ошибка численного решения становится менее чувствительной к неравномерности сетки.

Препринт структурирован следующим образом. Основные результаты формулируются в разделе 2. В разделе 3 мы вводим схему общего вида, для которой в разделе 4 доказывается достаточное условие устойчивости при малой деформации сетки. С помощью этого результата доказываются оценки точности: для схемы с полиномиальной реконструкцией в разделе 5, а для схемы R3 в разделе 6. В разделе 7 полученные результаты иллюстрируются на примере сетки с чередующимся шагом.

2. Постановка задачи и основной результат

В настоящей работе будем рассматривать задачу Коши для одномерного уравнения переноса с единичной скоростью:

$$\begin{gathered}\frac{\partial v}{\partial t} + \frac{\partial v}{\partial x} = 0, \quad 0 < t < t_{\max}, \quad x \in \mathbb{R}; \\ v(0, x) = v_0(x), \quad x \in \mathbb{R}.\end{gathered} \tag{2.1}$$

Начальные данные $v_0(x)$ предполагаются 2π-периодическими и достаточно гладкими, конкретные требования к гладкости будут уточнены позже.

Введём следующие определения. *Сеткой* (или расчётной сеткой) будем называть монотонно возрастающую последовательность чисел

$$X = \{x_j \in \mathbb{R}, j \in \mathbb{Z}\},$$

такую, что для некоторого $N \in \mathbb{N}$ при всех $j \in \mathbb{Z}$ выполняется $x_{j+N} = x_j + 2\pi$. При этом x_j будем называть *узлами* сетки, $N \equiv N(X)$ – *числом узлов* в сетке, а $h_{av} \equiv h_{av}(X) = 2\pi/N(X)$ – *средним шагом* сетки. Для $j \in \mathbb{Z}$ будем использовать стандартные обозначения

$$h_{j+1/2} = x_{j+1} - x_j, \quad x_{j+1/2} = \frac{x_j + x_{j+1}}{2}, \quad \hbar_j = x_{j+1/2} - x_{j-1/2}.$$

Также введём

$$h_{\max} = \max_{j\in\mathbb{Z}} h_{j+1/2}, \quad h_{\min} = \min_{j\in\mathbb{Z}} h_{j+1/2}, \quad (\Delta h)_{\max} = \max_{j\in\mathbb{Z}} |h_{j+1/2} - h_{j-1/2}|. \tag{2.2}$$

Сеточной функцией на сетке X будем называть $N(X)$-периодическую последовательность комплексных чисел $u = \{u_j \in \mathbb{C}, j \in \mathbb{Z}\}$.

Периодом сетки X будем называть минимальное число $m \equiv m(X)$, такое, что для всех j выполняется $h_{j+m+1/2} = h_{j+1/2}$. Очевидно, что такое число существует и не превосходит N. Сетку с периодом $m = 1$ будем называть *равномерной*, у неё $h_{j+1/2} = h_{av}$ при всех j.

В настоящей работе будем рассматривать два класса полудискретных схем для решения (2.1): конечно-объёмные схемы с полиномиальной реконструкцией и схемы на основе разделённых разностей. Схемы с полиномиальной реконструкцией будем записывать “на дуальной сетке”, то есть в качестве контрольных объёмов будем использовать интервалы $(x_{j-1/2}, x_{j+1/2})$, а не (x_j, x_{j+1}), хотя все рассуждения остаются справедливыми и для схем “на исходной сетке”.

Пусть $X = \{x_j\}$ – некоторая сетка, $p = 2s$, $s \in \mathbb{N} \cup \{0\}$. Конечно-объёмная схема с реконструкцией многочленом порядка p имеет вид

$$\frac{du_j}{dt} + \frac{p_j(x_{j+1/2}) - p_{j-1}(x_{j-1/2})}{\hbar_j} = 0, \tag{2.3}$$

$$u_j(0) = \frac{1}{\hbar_j} \int\limits_{x_{j-1/2}}^{x_{j+1/2}} v_0(x)dx, \tag{2.4}$$

где $p_j(x)$ – многочлен порядка p, такой, что для всех $k = -p/2, \ldots, p/2$ выполняется

$$\frac{1}{\hbar_{j+k}} \int\limits_{x_{j+k-1/2}}^{x_{j+k+1/2}} p_j(x)dx = u_{j+k}. \tag{2.5}$$

Эта система содержит $p+1$ уравнений и $p+1$ неизвестных коэффициентов многочлена; её невырожденность хорошо известна (см., например, [8]).

Схема на основе разделёных разностей, которую будем называть R3, имеет вид

$$\frac{du_j}{dt} + \frac{F_{j+1/2} - F_{j-1/2}}{\hbar_j} = 0, \tag{2.6}$$

$$u_j(0) = v_0(x_j), \tag{2.7}$$

где

$$F_{j+1/2} = u_j + \frac{h_{j+1/2}}{2} \left(\frac{2}{3} \frac{u_{j+1} - u_j}{h_{j+1/2}} + \frac{1}{3} \frac{u_j - u_{j-1}}{h_{j-1/2}} \right). \tag{2.8}$$

Также будем рассматривать схему R5, имеющую вид (2.6), (2.7),

$$\begin{aligned} F_{j+1/2} = u_j + \frac{h_{j+1/2}}{2} \left(-\frac{1}{10} \frac{u_{j+2} - u_{j+1}}{h_{j+3/2}} + \frac{4}{5} \frac{u_{j+1} - u_j}{h_{j+1/2}} + \right. \\ \left. + \frac{11}{30} \frac{u_j - u_{j-1}}{h_{j-1/2}} - \frac{1}{15} \frac{u_{j-1} - u_{j-2}}{h_{j-3/2}} \right). \end{aligned} \tag{2.9}$$

Коэффициенты перед разделёнными разностями выбраны таким образом, чтобы обеспечить максимальный порядок точности на равномерных сетках, а именно: 3-й для R3 и 5-й для R5. Ниже будем рассматривать только $N(X)$-периодические решения (2.3) и (2.6).

Для сетки X введём обозначение

$$\mathcal{M}(X) = \frac{(\Delta h)_{\max}}{h_{av}},$$

где $(\Delta h)_{\max}$ определено (2.2), а $h_{av} = 2\pi/N(X)$. Величина $\mathcal{M}(X)$ характеризует степень неравномерности сетки.

Для $\mu : \mathbb{N} \to (0, \infty)$ будем обозначать через $\mathcal{F}_\mu$ множество сеток

$$\mathcal{F}_\mu = \{X \;:\; \mathcal{M}(X) \leqslant \mu(m(X))\}. \tag{2.10}$$

Для $q \in \mathbb{N}$ и $f \in C^q(\mathbb{R})$ будем использовать обозначение

$$|f|_q = \sup_{x\in\mathbb{R}} \left|\frac{d^q f}{dx^q}(x)\right|.$$

Основным результатом настоящей работы являются следующие теоремы.

Теорема 2.1. *Рассмотрим схему* (2.3), (2.4) *с реконструкцией многочленом порядка $p = 2s$, $s \in \mathbb{N}\cup\{0\}$. Для любого $\delta > 0$ существует такое $\mu : \mathbb{N} \to (0, \infty)$, зависящее только от p и δ, что для каждой сетки $X \in \mathcal{F}_\mu$ и для каждого $v_0(x) \in C^{p+2}(\mathbb{R})$ решение $u(t)$ по схеме* (2.3), (2.4) *удовлетворяет оценке*

$$\begin{gathered}\frac{1}{\sqrt{2\pi}}\left(\sum_{j=1}^{N(X)} \hbar_j \left| u_j(t) - \frac{1}{\hbar_j}\int\limits_{x_{j-1/2}}^{x_{j+1/2}} v_0(x-t)dx\right|^2\right)^{1/2} \leqslant \\ \leqslant c|v_0|_{p+1} h_{\max}^p (h_{\max} - h_{\min}) + (c_s+\delta)|v_0|_{p+2} h_{\max}^{p+1} t,\end{gathered} \tag{2.11}$$

причём c зависит только от $m(X)$ и p, а $c_s = s!\,(s+1)!/(2s+2)!$.

Теорема 2.2. *Для любого $\delta > 0$ существует такое $\mu : \mathbb{N} \to (0, \infty)$, что для каждой сетки $X \in \mathcal{F}_\mu$ и для каждого $v_0(x) \in C^4(\mathbb{R})$ решение $u(t)$ по схеме* (2.6), (2.7), (2.8) *удовлетворяет оценке*

$$\begin{gathered}\frac{1}{\sqrt{2\pi}}\left(\sum_{j=1}^{N(X)} \hbar_j \left| u_j(t) - v_0(x_j - t)\right|^2\right)^{1/2} \leqslant \\ \leqslant 2|v_0|_2\, h_{\max}(h_{\max} - h_{\min}) + 14|v_0|_3\, h_{\max}^2 (h_{\max} - h_{\min}) + \\ +2|v_0|_3\,(\Delta h)_{\max}^2 t + (1/12+\delta)|v_0|_4\, h_{\max}^3 t.\end{gathered} \tag{2.12}$$

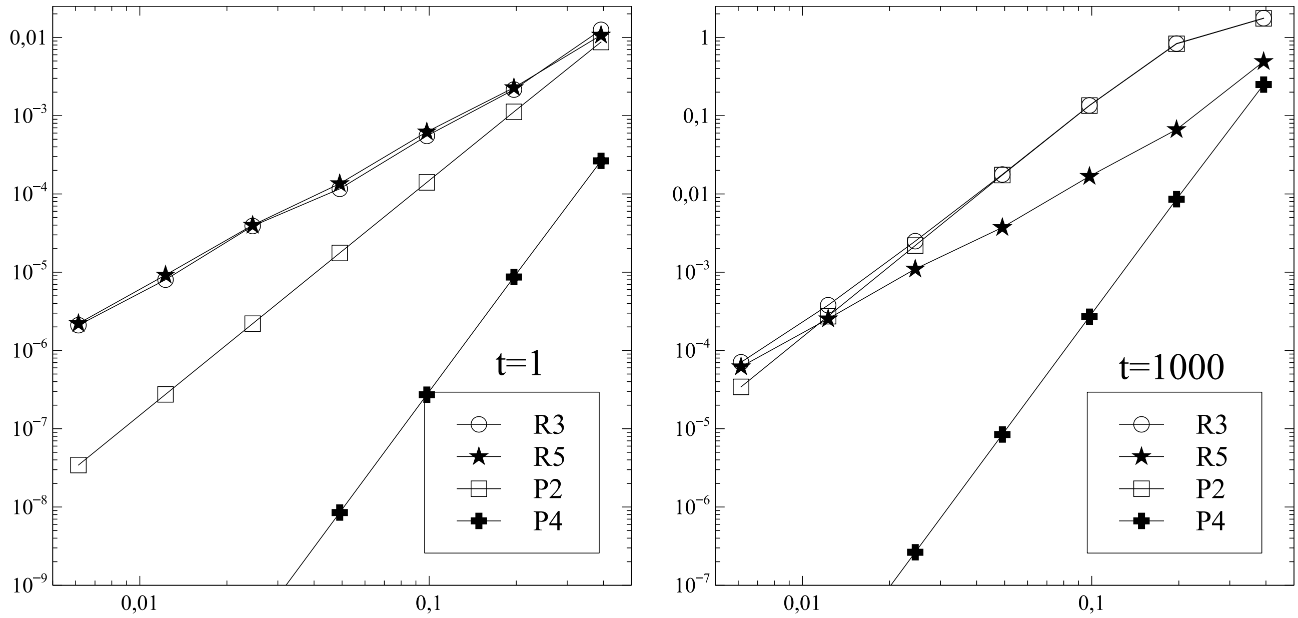


Рис. 1. Сходимость решений по схемам с полиномиальной реконструкцией (P2 и P4) и схемам на основе разделённых разностей (R3 и R5). Слева: $t = 1$, справа: $t = 1000$.

На равномерной сетке можно положить $\delta = 0$, тогда правая часть (2.11) сведётся к $c_s|v_0|_{p+2}h_{\max}^{p+1}t$, а правая часть (2.12) – к этому же выражению с подстановкой $s = 1$. Эти оценки совпадают со стандартными оценками ошибки конечно-разностных схем порядка $2s + 1$.

Значения множителей перед первыми тремя слагаемыми в правой части (2.12) не являются оптимальными. Также полученные оценки можно уточнить, заменив $|v_0|_q$, $q \in \mathbb{N}$, на квадратичную норму от q-й производной и соответствующим образом ослабив требование на гладкость функции $v_0(x)$. Но мы не будем этим заниматься, поскольку это потребовало бы технических рассуждений, загромождающих изложение основного результата.

Для схемы R5 нет оценок вида (2.12), поскольку, как мы покажем в разделе 7, эта схема неустойчива на любом семействе сеток вида (2.10).

Прежде чем начать доказательство основных результатов, приведём результаты численных экспериментов. Возьмём начальные данные $v_0(x) = \cos x$ и N-периодическую сетку с узлами $x_j = (j + d_j)h_{av}$, где $d_0 = 0$, а d_j, $j = 1, \dots, N-1$, – независимые случайные числа, равномерно распределённые на $[-0.1, 0.1]$. Сравним схемы с полиномиальной реконструкцией 2-го и 4-го порядка (обозначим их через P2 и P4) со схемами R3 и R5. Результаты на время $t = 1$ и $t = 1000$ приведены на рис. 1. По горизонтали отложено h_{av}, а по вертикали – норма ошибки решения, использованная в формулировках теорем.

Результаты показывают, что в длительном счёте R3 даёт почти ту же точность, что и P2, в широком диапазоне h_{av}. Это значит, что член второго порядка в ошибке решения меньше, чем член 3-го порядка. Переход от R3 к R5 уничтожает ошибку 3-го порядка, что и объясняет поведение R5, показанное на рис. 1. Заметим, что если решение имеет финитный спектр, то отсутствие устойчивости, вообще говоря, не является препятствием для сходимости.

3. Абстрактная запись схемы

Будем рассматривать системы уравнений, зависящие от сетки X, вида

$$\frac{du_j(t)}{dt} + \sum_{k=-S}^{S} a_k(h_{j-S+1/2}, \dots, h_{j+S-1/2})\, u_{j+k}(t) = 0, \quad j \in \mathbb{Z}. \tag{3.1}$$

Под решением этой системы будем понимать $N(X)$-периодическую последовательность $u(t) = \{u_j(t) \in \mathbb{C}, j \in \mathbb{Z}\}$. Ниже по тексту $N \equiv N(X)$.

Будем предполагать, что $a_k(r_{-S}, \dots, r_{S-1})$ определена в окрестности точки $r_{-S} = \dots = r_S = 1$, является аналитической в этой точке и в области определения удовлетворяет свойству

$$a_k(\alpha r_{-S}, \dots, \alpha r_{S-1}) = \frac{1}{\alpha} a_k(r_{-S}, \dots, r_{S-1}). \tag{3.2}$$

Тогда (3.1) можно переписать в виде

$$\frac{du_j(t)}{dt} + \frac{1}{h_{av}} \sum_{k=-S}^{S} a_k\left(\frac{h_{j-S+1/2}}{h_{av}}, \dots, \frac{h_{j+S-1/2}}{h_{av}}\right) u_{j+k}(t) = 0. \tag{3.3}$$

На равномерной сетке (3.1) приобретает вид

$$\frac{du_j(t)}{dt} + \frac{1}{h} \sum_{k=-S}^{S} \mathring{a}_k u_{j+k}(t) = 0, \tag{3.4}$$

где

$$\mathring{a}_k = a_k(1, \dots, 1), \quad h \equiv h_{av}.$$

На пространстве последовательностей с комплексными компонентами seq $(\mathbb{Z}, \mathbb{C})$ определим оператор $\mathring{\mathcal{L}}$ через

$$(\mathring{\mathcal{L}} u)_j = \sum_{k=-S}^{S} \mathring{a}_k u_{j+k}.$$

Оператор $\mathring{\mathcal{L}}$ отображает N-периодические последовательности в N-периодические для любого $N \in \mathbb{N}$. Для $\phi \in \mathbb{C}$ через $w(\phi)$ обозначим последовательность с компонентами

$$(w(\phi))_j = e^{i\phi j}, \quad j \in \mathbb{Z}.$$

Тогда

$$\mathring{\mathcal{L}}w(\phi) = \mathring{\lambda}(\phi)w(\phi), \tag{3.5}$$

где

$$\mathring{\lambda}(\phi) = \sum_{k=-S}^{S} \mathring{a}_k e^{i\phi k}. \tag{3.6}$$

Ограничение оператора $\mathring{\mathcal{L}}$ на пространство N-периодических последовательностей имеет базис из собственных векторов $w(2\pi l/N)$, $l = 0, \ldots, N-1$.

Отождествляя u как N-периодическую последовательность с её фрагментом $u = \{u_j, j = 0, \ldots, N-1\}$, можно переписать (3.4) в виде

$$\frac{du}{dt} + \frac{1}{h}Lu = 0, \tag{3.7}$$

где L – циркулянт размера N. Циркулянт унитарным преобразованием (переходом к базису $N^{-1/2}w(2\pi l/N)$, $l = 0, \ldots, N-1$) преобразуется к диагональной матрице, а его собственные значения равны $\lambda_l = \mathring{\lambda}(2\pi l/N)$.

На схему на равномерной сетке наложим два дополнительных условия. Во-первых, схема *точна на линейной функции*, то есть для некоторых $c \in \mathbb{C}\backslash\{0\}$ и $P_0 \in \mathbb{N}$ при $\phi \to 0$ выполняется

$$\mathring{\lambda}(\phi) = i\phi + c\phi^{P_0+1} + O(|\phi|^{P_0+2}). \tag{3.8}$$

Величина P_0 является порядком аппроксимации на равномерной сетке. Во-вторых, выполняется условие

$$\mathrm{Re}\mathring{\lambda}(\phi) > 0, \quad \phi \in \mathbb{R}, \quad \phi/(2\pi) \notin \mathbb{Z}. \tag{3.9}$$

Из этого условия следует, что для любой ненулевой сеточной функции u с нулевым средним выполняется $u^*Lu > 0$. Схемы на равномерной сетке, удовлетворяющие (3.9), будем называть *строго диссипативными*. Заметим, что при нечётных P_0 это влечёт $c > 0$ в (3.8).

Введём отображение $\hat{\Pi}$, сопоставляющее сетке X и функции $f \in C(\mathbb{R})$ числовую последовательность $\hat{\Pi}_X f$ с компонентами

$$(\hat{\Pi}_X f)_j = \frac{1}{\hbar_j} \int\limits_{x_{j-1/2}}^{x_{j+1/2}} f(x)dx, \quad j \in \mathbb{Z}. \tag{3.10}$$

Аналогично введём отображение $\mathring{\Pi}$, сопоставляющее сетке X и функции $f \in C(\mathbb{R})$ числовую последовательность $\mathring{\Pi}_X f$ с компонентами

$$(\mathring{\Pi}_X f)_j = f(x_j), \quad j \in \mathbb{Z}. \tag{3.11}$$

Эти отображения использовались для задания начальных данных (см. (2.4) и (2.7)).

Для формулировки ряда лемм нам понадобится более общее определение. *Локальным отображением* будем называть отображение Π, для некоторого $q \in \mathbb{N} \cup \{0\}$ сопоставляющее сетке X и функции $f \in C^q(\mathbb{R})$ последовательность значений $\Pi_X f = \{(\Pi_X f)_j \in \mathbb{C}, j \in \mathbb{Z}\}$ вида

$$(\Pi_X f)_j = \langle \mu_j^{(X)}, f((\,\cdot\, + j)h_{av})\rangle,$$

где $\mu_j^{(X)} \in (C^q(G))^*$ для некоторой ограниченной области G, периодична по j с периодом $m(X)$, не меняется при сжатии сетки (то есть при преобразовании $\{x_j\} \to \{x_j/k\}$, $k \in \mathbb{N}$) и удовлетворяет условию $\langle \mu_j^{(X)}, 1\rangle \neq 0$, $j \in \mathbb{Z}$.

Легко показать, что для 2π-периодической функции f последовательность $\Pi_X f$ является $N(X)$-периодической, то есть сеточной функцией на X. Примерами локальных отображений являются $\hat{\Pi}$ и $\mathring{\Pi}$. В первом случае

$$\mu_j^{(X)}(x) = \begin{cases} h_{av}/\hbar_j, & x \in (x_{j-1/2}/h_{av} - j, x_{j+1/2}/h_{av} - j); \\ 0, & otherwise; \end{cases}$$

во втором случае $\mu_j^{(X)}(x) = \delta(x - (x_j/h_{av} - j))$.

Ошибкой аппроксимации на функции f в смысле локального отображения Π будем называть последовательность $\epsilon_X(f, \Pi)$ с компонентами

$$(\epsilon_X(f,\Pi))_j = -(\Pi_X f')_j + \frac{1}{h_{av}} \sum_{k=-S}^{S} a_k\left(\frac{h_{j-S+1/2}}{h_{av}}, \ldots, \frac{h_{j+S-1/2}}{h_{av}}\right) (\Pi_X f)_{j+k}.$$

Здесь $f' \equiv df/dx$. Будем говорить, что система (3.1) точна на многочленах порядка q в смысле Π, если для любого многочлена $f \equiv f(x)$ порядка не выше q выполняется $\epsilon_X(f, \Pi) = 0$.

Ошибкой решения с начальными данными v_0 в смысле локального отображения Π на момент $t \geqslant 0$ будем называть сеточную функцию

$$\varepsilon_X(t, v_0, \Pi) = u(t) - \Pi_X v(t,\,\cdot\,),$$

где $v(t,x) = v_0(x-t)$, а $u(t)$ – $N(X)$-периодическое решение (3.1) с начальным условием $u(0) = \Pi_X v_0$.

4. Спектральный анализ

В этом разделе мы докажем достаточное условие устойчивости на слабо неравномерной сетке. Всюду в настоящем разделе будем рассматривать сетку X с периодом $m \equiv m(X) > 1$.

Введём обозначение

$$\gamma_j = \frac{h_{j+1/2}}{h_{av}} - 1, \quad j \in \mathbb{Z}, \tag{4.1}$$

и $\gamma = \{\gamma_j, j = 0, \dots, m-1\}$. Набор γ будем называть *структурой* сетки. Обозначим $|\gamma| = \max_j |\gamma_j|$. Очевидно, что

$$\sum_{j=0}^{m-1} \gamma_j = 0. \tag{4.2}$$

Доопределим $a_k(\dots)$ нулём при $|k| > S$.

Пусть $\mathcal{L}(\gamma)$ обозначает оператор на пространстве $\mathrm{seq}\,(\mathbb{Z}, \mathbb{C})$, определённый как

$$(\mathcal{L}(\gamma)u)_j = \sum_{k=-S}^{S} a_k(\gamma_{j-S}+1, \dots, \gamma_{j+S-1}+1)\ u_{j+k}.$$

Ясно, что оператор $\mathcal{L}(\gamma)$ отображает m-периодические последовательности в m-периодические и $\mathcal{L}(0) = \mathring{\mathcal{L}}$. Тогда на пространстве комплекснозначных последовательностей схема (3.3) запишется как

$$\frac{du}{dt} + \frac{1}{h_{av}} \mathcal{L}(\gamma) u = 0.$$

Запишем схему (3.3) в блочном виде. Для $m \in \mathbb{N}$ определим оператор $\mathcal{B}_m :$ $\mathrm{seq}\,(\mathbb{Z}, \mathbb{C}) \to \mathrm{seq}\,(\mathbb{Z}, \mathbb{C}^m)$ как

$$(\mathcal{B}_m u)_\eta = (u_{\eta m}, \dots, u_{\eta m + m - 1})^T.$$

На пространстве $\mathrm{seq}\,(\mathbb{Z}, \mathbb{C}^m)$ определим оператор $\mathcal{L}_m(\gamma)$ соотношением

$$\mathcal{L}_m(\gamma)\mathcal{B}_m = \mathcal{B}_m \mathcal{L}(\gamma). \tag{4.3}$$

Этот оператор выражается в виде

$$(\mathcal{L}_m(\gamma)U)_\eta = \sum_{\zeta=-\lceil S/m \rceil}^{\lceil S/m \rceil} L_\zeta(\gamma) U_{\eta+\zeta}, \tag{4.4}$$

где $L_\zeta(\gamma)$ – действительнозначная матрица размера $m \times m$ с элементами

$$(L_\zeta(\gamma))_{jk} = a_{\zeta m + k - j}(\gamma_{j-S}+1, \dots, \gamma_{j+S-1}+1), \quad j, k = 0, \dots, m-1.$$

Таким образом, если ввести обозначение $U_\eta = (\mathcal{B}_m u)_\eta$, то (3.3) перепишется в виде

$$\frac{dU_\eta}{dt} + \frac{1}{h_{av}} \sum_{\zeta=-\lceil S/m \rceil}^{\lceil S/m \rceil} L_\zeta(\gamma)\, U_{\eta+\zeta} = 0, \quad \eta \in \mathbb{Z}. \tag{4.5}$$

Системы уравнений вида (4.5) были подробно рассмотрены в [9]. Обозначения в этой работе соответствуют нашим следующим образом: характерный шаг $h \equiv h_{av}$, смещение между сеточными блоками в расчёте на $h = 1$ равно $T = m$, множество степеней свободы на одном сеточном блоке есть $M^0 = \{0, \ldots, m-1\}$.

На пространстве сеточных функций на сетке X будем использовать норму

$$\|f\|_{av} = \left(\frac{1}{N(X)} \sum_{j=0}^{N(X)-1} |f_j|^2 \right)^{1/2} = \frac{1}{\sqrt{2\pi}} \left(\sum_{j=0}^{N(X)-1} h_{av} |f_j|^2 \right)^{1/2}. \tag{4.6}$$

На $\mathbb{C}^m$ будем использовать евклидову норму. Её, как и порождённую ей матричную норму, будем обозначать через $\|\cdot\|$.

Локальному отображению Π сопоставим функцию, отображающую структуру γ и $\phi \in \mathbb{C}$ в

$$v(\gamma, \phi, \Pi) = \left\{ \left[\Pi_X \exp\left(i\phi \frac{x}{h_{av}} \right) \right]_j, j = 0, \ldots, m-1 \right\} \in \mathbb{C}^m,$$

где X – сетка со средним шагом h_{av}, структурой γ и смещением $x_0 = 0$. Из определения локального отображения следует, что $v(\gamma, \phi, \Pi)$ не зависит от h_{av}.

Пусть $\mathcal{S}_\phi$ – пространство последовательностей $U \in \mathrm{seq}\,(\mathbb{Z}, \mathbb{C}^m)$, таких, что $U_\eta = \exp(i\phi m \eta) U_0$, $\eta \in \mathbb{Z}$. Пусть e_k – векторы стандартного базиса в $\mathbb{C}^m$. Тогда векторы $e_k(\phi)$, $k = 0, \ldots, m-1$, с компонентами

$$(e_k(\phi))_\eta = \exp(i\phi m \eta) e_k \tag{4.7}$$

образуют базис в $\mathcal{S}_\phi$. Введём матрицу ограничения $\mathcal{L}_m(\gamma)$ на $\mathcal{S}_\phi$ в этом базисе:

$$L(\gamma, \phi) = \sum_{\zeta=-\lceil S/m \rceil}^{\lceil S/m \rceil} L_\zeta(\gamma) \exp(i\phi m \zeta). \tag{4.8}$$

Образ Фурье (N/m)-периодической последовательности V с элементами $V_\eta \in \mathbb{C}^m$ может быть определён как

$$\hat{V}(\phi) = \frac{m}{N} \sum_{\eta=0}^{N/m-1} \exp\left(-im\phi\eta\right) V_\eta, \quad \phi = \frac{2\pi k}{N}, \quad k \in \mathbb{Z},$$

тогда обратное преобразование имеет вид

$$V_\eta = \sum_\phi \exp\left(im\phi\eta\right)\hat{V}(\phi),$$

где ϕ пробегает значения $0, 2\pi/N, \ldots, 2\pi/m - 1/N$. Справедливо равенство Парсеваля

$$\frac{m}{N}\sum_{j=0}^{N/m-1}\|V_j\|^2 = \sum_\phi \|\hat{V}(\phi)\|^2. \tag{4.9}$$

В частности, определим

$$\hat{U}(t,\phi) = \frac{m}{N}\sum_{\eta=0}^{N/m-1}\exp\left(-im\phi\eta\right)U_\eta(t), \quad \phi = \frac{2\pi k}{N}, \quad k \in \mathbb{Z}.$$

В образах Фурье система (4.5) принимает вид

$$\frac{d}{dt}\hat{U}(t,\phi) + \frac{1}{h_{av}}L(\gamma,\phi)\hat{U}(t,\phi) = 0, \quad \phi = \frac{2\pi k}{N}, \quad k \in \mathbb{Z}. \tag{4.10}$$

Вектор-функция $\hat{U}(t,\phi)$ и матричная функция $L(\gamma,\phi)$ имеют период $2\pi/m$ по ϕ, поэтому в (4.10) имеем N/m независимых уравнений.

Функции $v(\gamma,\phi,\Pi)$ и $L(\gamma,\phi)$ определены для γ, таких что $|\gamma|$ достаточно мало и $\sum\gamma_j = 0$; доопределим их на некоторую окрестность нуля равенствами

$$v(\gamma,\phi,\Pi) = v\left(\gamma - \frac{\mathbf{1}}{m}\sum_j \gamma_j, \phi, \Pi\right), \quad L(\gamma,\phi) = L\left(\gamma - \frac{\mathbf{1}}{m}\sum_j \gamma_j, \phi\right),$$

где $\mathbf{1} = (1,\ldots,1)^T$. Также введём

$$A(\gamma,\phi) = i\phi I - L(\gamma,\phi),$$

$$\hat{\epsilon}(\gamma,\phi,\Pi) = A(\gamma,\phi)v(\gamma,\phi,\Pi),$$

$$\hat{\varepsilon}(\gamma,\nu,\phi,\Pi) = \left(\exp(\nu\, A(\gamma,\phi)) - I\right)v(\gamma,\phi,\Pi).$$

Лемма 4.1. *Пусть Π – локальное отображение, X – сетка, h_{av} и γ – её средний шаг и структура, $\alpha \in \mathbb{Z}$, $t \geqslant 0$. Тогда справедливы равенства*

$$\|\epsilon_X(\exp(i\alpha x),\Pi)\|_{av} = \frac{1}{\sqrt{m}\,h_{av}}\|\hat{\epsilon}(\gamma,\alpha h_{av},\Pi)\|, \tag{4.11}$$

$$\|\varepsilon_X(t,\exp(i\alpha x),\Pi)\|_{av} = \frac{1}{\sqrt{m}}\|\hat{\varepsilon}(\gamma,t/h_{av},\alpha h_{av},\Pi)\|. \tag{4.12}$$

Доказательство. Обозначим $\phi = \alpha h_{av}$. Поскольку

$$\Pi_X e^{i\alpha x} \in S_\phi, \quad (\mathcal{B}_m \Pi_X e^{i\alpha x})_0 = v(\gamma, \alpha h_{av}, \Pi),$$

имеем $\epsilon_X(\exp(i\alpha x, \Pi)) \in S_\phi$ и

$$(\mathcal{B}_m \epsilon_X(e^{i\alpha x}, \Pi))_0 = \frac{1}{h_{av}} A(\gamma, \alpha h_{av}) v(\gamma, \alpha h_{av}, \Pi) = \frac{1}{h_{av}} \hat{\epsilon}(\gamma, \alpha h_{av}, \Pi).$$

Поэтому (4.11) следует из (4.9). Равенство (4.12) доказывается аналогично. □

Лемма 4.2. *Пусть $S_m(\phi)$ – матрица с элементами*

$$(S_m(\phi))_{jk} = m^{-1/2} \exp(2\pi i jk/m + i\phi j), \quad j, k = 0, \ldots, m-1. \tag{4.13}$$

Тогда $S_m(\phi)$ унитарная и

$$L(0, \phi) = S_m(\phi) \operatorname{diag}\{\mathring{\lambda}(\phi + 2\pi l/m),\ l = 0, \ldots, m-1\}\, S_m^{-1}(\phi). \tag{4.14}$$

Доказательство. Унитарность $S_m(\phi)$ проверяется непосредственно.

Обозначим через $Y(\phi)$ правую часть (4.14). Используя выражение (3.6) для $\mathring{\lambda}(\phi)$, получаем

$$Y_{jk}(\phi) = \sum_{p=-S}^{S} \mathring{a}_p \exp(i\phi(j+p-k)) \left[\frac{1}{m}\sum_{l=0}^{m-1} \exp\left(2\pi i l \frac{j+p-k}{m}\right)\right] =$$

$$= \sum_{\zeta \in \mathbb{Z}} \mathring{a}_{k-j+\zeta m} \exp(i\phi m \zeta).$$

Теперь легко видеть, что $Y_{jk}(\phi)$ совпадает с $(L(0,\phi))_{jk}$, определённым в (4.8).

Есть и другой способ понять (4.14) без явных вычислений. Матрица $L(0, \phi)$ является матрицей ограничения оператора $\mathcal{L}(0)$ на пространство последовательностей $\tilde{S}_\phi \subset \operatorname{seq}(\mathbb{Z}, \mathbb{C})$, таких что $u_{k+m\eta} = e^{im\eta\phi} u_k$, в базисе из последовательностей $E_k(\phi)$, $k = 0, \ldots, m-1$, таких что $(E_k(\phi))_j = \delta_{kj}$, $j = 0, \ldots, m-1$.

Последовательности $m^{-1/2} w(\phi + 2\pi k/m)$, $k = 0, \ldots, m-1$, принадлежат $\tilde{S}_\phi$ и в силу (3.5) являются собственными векторами этого оператора, соответствующими собственным значениям $\mathring{\lambda}(\phi + 2\pi k/m)$. В базисе $E_k(\phi)$ эти векторы имеют координаты, записанные в столбцах матрицы $S_m(\phi)$. □

Обозначим через $\mathbb{C}^{n\times n}$ пространство комплекснозначных матриц размера $n \times n$, а через $\mathcal{A}(\mathbb{C}^d, \mathbb{C}^{n\times n})$ – множество функций из $\mathbb{C}^d$ в $\mathbb{C}^{n\times n}$, аналитических в нуле. Через $\sigma(A)$ будем обозначать спектр матрицы A.

Нам понадобится следующий результат теории возмущений.

Лемма 4.3. *Пусть* $A \in \mathcal{A}(\mathbb{C}^d, \mathbb{C}^{n\times n})$. *Пусть* $\sigma(A(0)) = G \cup H$, $G \cap H = \varnothing$. *Тогда в некоторой окрестности* $\boldsymbol{x} = 0$ *справедливо представление* $A(\boldsymbol{x}) = S(\boldsymbol{x})M(\boldsymbol{x})S^{-1}(\boldsymbol{x})$, *где* $S, M, S^{-1} \in \mathcal{A}(\mathbb{C}^d, \mathbb{C}^{n\times n})$,

$$M(\boldsymbol{x}) = \begin{pmatrix} M^{(G)}(\boldsymbol{x}) & 0 \\ 0 & M^{(H)}(\boldsymbol{x}) \end{pmatrix}, \tag{4.15}$$

причём $\sigma(M^{(G)}(0)) = G$, $\sigma(M^{(H)}(0)) = H$, *а размер* n_G *матрицы* $M^{(G)}(\boldsymbol{x})$ *равен суммарной кратности собственных значений, лежащих в* G.

Функция $L(\gamma, \phi)$ является аналитической функцией $m + 1$ переменных ($\gamma_0, \dots, \gamma_{m-1}$, ϕ) в каждой точке $(0, \phi)$, $\phi \in \mathbb{R}$. В частности, она аналитична в точке $(0{,}0)$. Если схема на равномерной сетке строго диссипативна, то по лемме 4.2 матрица $L(0{,}0)$ имеет нулевое собственное значение кратности 1. В силу леммы 4.3 в некоторой окрестности нуля справедливо представление

$$L(\gamma, \phi) = \tilde{S}(\gamma, \phi) \begin{pmatrix} \lambda_0(\gamma, \phi) & 0 \\ 0 & \tilde{M}^{(*)}(\gamma, \phi) \end{pmatrix} \tilde{S}^{-1}(\gamma, \phi), \tag{4.16}$$

где матрицы $\tilde{S}$, $\tilde{M}^{(*)}$, $\tilde{S}^{-1}$ и скалярная функция λ_0 аналитические в нуле, а $\tilde{M}^{(*)}(0{,}0)$ невырожденная. Очевидно, что в некоторой окрестности $\phi = 0$ выполняется $\lambda_0(0, \phi) = \mathring{\lambda}(\phi)$. Пусть $S_m(\phi)$ – унитарная матрица, определённая (4.13). Поскольку каждая из матриц $\tilde{S}(0, \phi)$ и $S_m(\phi)$ приводит матрицу $L(0, \phi)$ к блочно-диагональному виду, справедливо равенство

$$\tilde{S}(0, \phi) = S_m(\phi) \begin{pmatrix} a(\phi) & 0 \\ 0 & T(\phi) \end{pmatrix}$$

с некоторым $a(\phi) \in \mathbb{C} \setminus \{0\}$ и некоторой невырожденной матрицей $T(\phi)$. Введя $S(\gamma, \phi) = \tilde{S}(\gamma, \phi)\tilde{S}^{-1}(0, \phi)S_m(\phi)$, мы приходим к разложению

$$L(\gamma, \phi) = S(\gamma, \phi) \begin{pmatrix} \lambda_0(\gamma, \phi) & 0 \\ 0 & M^{(*)}(\gamma, \phi) \end{pmatrix} S^{-1}(\gamma, \phi). \tag{4.17}$$

При этом $S(0, \phi) = S_m(\phi)$ и, следовательно, для любого $\tilde{K} > 1$ найдётся такая окрестность нуля, что в ней выполняется

$$\|S(\gamma, \phi)\| \leqslant \tilde{K}, \quad \|S^{-1}(\gamma, \phi)\| \leqslant \tilde{K}. \tag{4.18}$$

Кроме того, $M^{(*)}(0, \phi)$ является диагональной.

Лемма 4.4. *Пусть* B *и* D *– квадратные матрицы. Пусть* D *– нормальная матрица, собственные значения которой удовлетворяют условию* $\operatorname{Re} d_j \leqslant \mu$. *Тогда*

$$\| \exp(B + D)\| \leqslant \exp(\mu + \|B\|).$$

Доказательство. Воспользуемся формулой Ли:

$$\exp(B+D) = \lim_{n\to\infty} \left(\exp(B/n)\exp(D/n)\right)^n .$$

Отсюда

$$\|\exp(B+D)\| \leqslant \lim_{n\to\infty} \|\exp(B/n)\|^n \lim_{n\to\infty} \|\exp(D/n)\|^n .$$

Поскольку $\|\exp(B/n)\| \leqslant \exp(\|B\|/n)$, первый множитель не превосходит $\exp(\|B\|)$. А поскольку матрица D/n нормальная, она унитарным преобразованием преобразуется к диагональной, и $\|\exp(D/n)\|$ равно экспоненте от максимальной действительной части её собственных значений. Отсюда следует искомое неравенство. □

Из точности на линейной функции следует $\mathring{\lambda}(0) = 0$ и $\mathring{\lambda}'(0) = i$, поэтому в некотором интервале $\phi \in (0, \phi_{\max})$ выполняется $\mathrm{Im}\mathring{\lambda}(\phi) > 0$ и $d\mathring{\lambda}(\phi)/d\phi \neq 0$. А из (3.9) следует, что функция $\mathrm{Re}\mathring{\lambda}(\phi)$ задаёт взаимнооднозначное отображение

$$\mathrm{Re}\mathring{\lambda}(\phi) \ : \ [0, \phi_{\max}] \to [0, x_{\max}] \tag{4.19}$$

для некоторых $\phi_{\max} > 0$ и $x_{\max} > 0$, причём

$$\mathrm{Re}\mathring{\lambda}(\phi) \geqslant x_{\max}, \quad \phi_{\max} \leqslant \phi \leqslant 2\pi - \phi_{\max}. \tag{4.20}$$

Лемма 4.5. *Пусть в некоторой окрестности* (0,0) *справедливо* $\mathrm{Re}\,\lambda_0(\gamma,\phi) \geqslant 0$. *Тогда для любого* $K > 1$ *найдётся такое* $\gamma_{\max} > 0$, *что при всех* $\phi \in \mathbb{R}$, $|\gamma| \leqslant \gamma_{\max}$ *и* $\nu \geqslant 0$ *выполняется*

$$\|\exp(-\nu L(\gamma,\phi))\| \leqslant K. \tag{4.21}$$

Если $\lambda_0(0,\phi) \equiv \mathring{\lambda}(\phi)$ образует контур без самопересечений, это утверждение очевидно. Приведём доказательство для общего случая.

Доказательство. Пусть $\phi_{\max}$ и $x_{\max}$ определены (4.19)–(4.20). Разложение (4.17), оценка (4.18) с $\tilde{K} = K^{1/2}$ и условие $\mathrm{Re}\,\lambda_0(\gamma,\phi) \geqslant 0$ справедливы в некотором параллелепипеде вида

$$\left\{(\gamma,\phi) \ : \ \gamma \in [-\tilde{\gamma}, \tilde{\gamma}]^m, \quad \phi \in [-\tilde{\phi}, \tilde{\phi}]\right\}.$$

Без ограничения общности будем считать, что $\tilde{\phi} \leqslant \phi_{\max}$, и определим $\tilde{x} = \mathrm{Re}\mathring{\lambda}(\tilde{\phi})$.

Вначале рассмотрим случай $|\phi| \leqslant \tilde{\phi}$. Из представления (4.17) имеем

$$\|\exp(-\nu L(\gamma,\phi))\| \leqslant \tilde{K}^2 \max\{|\exp(-\nu\lambda_0(\gamma,\phi))|, \|\exp(-\nu M^*(\gamma,\phi))\|\}.$$

Первый аргумент максимума не превосходит единицы по условию леммы. Для оценки второго аргумента воспользуемся леммой 4.4 с $D = -\nu M^{(*)}(0,\phi)$ и $B = -\nu(M^{(*)}(\gamma,\phi) - M^{(*)}(0,\phi))$. Все собственные значения $M^{(*)}(0,\phi)$ имеют действительную часть, большую или равную $x_{\max} > 0$. При некотором, возможно, более строгом ограничении на $|\gamma|$ выполняется $\|B\| \leqslant \nu x_{\max}$, поэтому второй аргумент также не превосходит единицы. Отсюда имеем искомое неравенство (4.21).

Теперь рассмотрим случай $|\phi| > \tilde{\phi}$, $|\phi| \leqslant \pi/m$. Действительная часть всех собственных значений матрицы $L(0,\phi)$ больше или равна $\tilde{x}$. Определим $D = -\nu L(0,\phi)$ (D – нормальная матрица) и $B = -\nu(L(\gamma,\phi) - L(0,\phi))$. При некотором, возможно, более строгом ограничении на $|\gamma|$ выполняется $\|B\| \leqslant \nu\tilde{x}$. По лемме 4.4 получаем (4.21).

Мы доказали (4.21) при $|\phi| \leqslant \pi/m$. Поскольку $L(\gamma,\phi)$ периодична по ϕ с периодом $2\pi/m$, оно переносится на $\phi \in \mathbb{R}$. □

Условием леммы 4.5 является $\operatorname{Re}\lambda_0(\gamma,\phi) \geqslant 0$ в некоторой окрестности $(0{,}0)$. Получим достаточное условие, при котором оно выполняется.

Напомним, что через $P_0 + 1$ мы обозначали порядок малости функции $\lambda_0(0,\phi) - i\phi$ при $\phi \to 0$. Обозначим через $q + 1$ минимальную степень ϕ в разложении $\lambda_0(\gamma,\phi) - i\phi$ в окрестности точки $(0{,}0)$.

Лемма 4.6. *Пусть число P_0 нечётное, а $q \geqslant \max\{1, P_0 - 1\}$. Тогда в некоторой окрестности $(0{,}0)$ выполняется $\operatorname{Re}\lambda_0(\gamma,\phi) \geqslant 0$.*

Доказательство. Очевидно, что $q \leqslant P_0$, поэтому либо $q = P_0$, либо $q = P_0 - 1$. Напомним, что $\lambda_0(0,\phi) = \mathring{\lambda}(\phi)$, где $\mathring{\lambda}(\phi)$ определено (3.6).

Пусть $q = P_0$. Тогда

$$\lambda_0(\gamma,\phi) = i\phi + \phi^{P_0+1}(c + O(|\gamma| + |\phi|)),$$

где $c > 0$ в силу (3.8)–(3.9), откуда утверждение леммы очевидно.

Пусть $q = P_0 - 1$. Тогда

$$\lambda_0(\gamma,\phi) = i\phi + c\phi^{P_0+1} + (i\phi)^{P_0}(c_1(\gamma) + c_2(\gamma)\phi) + O(|\phi|^{P_0+2}),$$

где $c > 0$, $c_1 = O(|\gamma|)$ и $c_2 = O(|\gamma|)$. Покажем, что $c_1(\gamma) \in \mathbb{R}$ в окрестности $\gamma = 0$, тогда утверждение леммы будет очевидно.

Матрица $L(\gamma, i\phi)$ по построению действительнозначная. При достаточно малых $|\gamma|$ и $|\phi|$ в силу (4.17) единственным собственным значением $L(\gamma, i\phi)$ в некотором шаре $B_\epsilon(0)$ является $\lambda_0(\gamma, i\phi)$. Если допустить, что $\lambda_0(\gamma, i\phi) \notin \mathbb{R}$, то у $L(\gamma, i\phi)$ будет ещё одно собственное значение, $\bar{\lambda}_0(\gamma, i\phi)$, также лежащее в этом шаре. Таким образом, $\lambda_0(\gamma, i\phi) \in \mathbb{R}$, а значит все производные по ϕ от $\lambda_0(\gamma, i\phi)$ также действительнозначные. □

Остаётся указать, каким образом можно определить q, не прибегая к исследованию матрицы $L(\gamma,\phi)$.

Лемма 4.7. *Пусть* Π *– локальное отображение. Пусть ошибка аппроксимации в смысле* Π *для некоторых q и $C(\gamma)$ удовлетворяет оценке*

$$\|\epsilon_X(f,\Pi)\|_{av} \leqslant C(\gamma)\ (h_{av})^q |f|_{q+1}, \tag{4.22}$$

где h_{av} и γ – средний шаг и структура сетки X. Тогда в окрестности (0,0) *справедлива оценка*

$$|\lambda_0(\gamma,\phi) - i\phi| \leqslant 2C(\gamma)\ |\phi|^{q+1} \tag{4.23}$$

с той же $C(\gamma)$.

Доказательство. Из (4.11) в окрестности $\phi = 0$ имеем

$$\|\hat{\epsilon}(\gamma,\phi,\Pi)\| \leqslant \sqrt{m}C(\gamma)\ |\phi|^{q+1}.$$

Пользуясь разложением (4.17), запишем

$$S^{-1}(\gamma,\phi)\hat{\epsilon}(\gamma,\phi,\Pi) =$$
$$= \begin{pmatrix} i\phi - \lambda_0(\gamma,\phi) & 0 \\ 0 & i\phi I - M^{(*)}(\gamma,\phi) \end{pmatrix} S^{-1}(\gamma,\phi)v(\gamma,\phi,\Pi).$$

Поскольку $v(\gamma,\phi,\Pi)$ имеет предел $(1,\dots,1)^T$ при $\phi \to 0$, а $S^{-1}(\gamma,\phi) \to S_m$ при $(\gamma,\phi) \to$ (0,0), то первая компонента $S^{-1}(\gamma,\phi)v(\gamma,\phi,\Pi)$ стремится к $\sqrt{m}$. С другой стороны, норма левой части в некоторой окрестности нуля не превосходит $\sqrt{2m}C(\gamma)|\phi|^{q+1}$. Отсюда имеем (4.23). □

Теперь мы можем сформулировать общий результат, устанавливающий устойчивость и, таким образом, объясняющий первый эффект, упомянутый во введении. Ниже мы применим этот результат к схемам, описанным в разделе 2.

Теорема 4.8. *Рассмотрим схему* (3.1)*, где a_k аналитические в точке* $(1,\dots,1)$ *и удовлетворяют* (3.2)*. Пусть*

- *на равномерной сетке эта схема имеет порядок $P_0 = 2k-1$, $k \in \mathbb{N}$;*
- *на равномерной сетке эта схема является строго диссипативной;*
- *на некотором семействе сеток вида* (2.10) *она точна на многочленах порядка $q \geqslant \max\{1, P_0 - 1\}$ в смысле некоторого локального отображения.*

Тогда для любого $K > 1$ существует такое $\mu : \mathbb{N} \to (0,\infty)$, что любое решение (3.1) *на любой сетке $X \in \mathcal{F}_\mu$, где $\mathcal{F}_\mu$ определено* (2.10)*, удовлетворяет оценке*

$$\|u(t)\|_{av} \leqslant K\|u(0)\|_{av}. \tag{4.24}$$

Доказательство. Если $m = 1$, сетка равномерная. Для любого K положим $\mu_1 = 1$. Из строгой диссипативности следует $\|u(t)\|_{av} \leqslant \|u(0)\|_{av}$.

Пусть $m \in \mathbb{N} \setminus \{1\}$. По лемме 4.7 выполняется (4.23). Значит, предположения леммы 4.6 выполнены и в окрестности $(0,0)$ выполняется $\operatorname{Re} \lambda_0(\gamma, \phi) \geqslant 0$. По лемме 4.5 для любого $K > 1$ найдётся $\gamma_{\max}$ (зависящее от m и от схемы), такое, что для всех $|\gamma| \leqslant \gamma_{\max}$ любое решение удовлетворяет (4.24). Поскольку для сетки X со структурой γ выполняется $|\gamma| \leqslant m\mathcal{M}(X)$, можно положить $\mu_m = m^{-1}\gamma_{\max}$. □

Заметим, что в терминах [9] в условиях теоремы 4.8 при фиксированной структуре сетки схема (3.1) является “простой” и обладает порядком точности в длительном счёте q.

5. Схема с полиномиальной реконструкцией

Пусть $p = 2s$, $s \in \mathbb{N} \cup \{0\}$, – порядок многочлена, используемого при реконструкции. Через $\hat{\Pi}$ будем обозначать локальное отображение (3.10). На равномерной сетке $X = \{jh, j \in \mathbb{Z}\}$ схема (2.3), (2.4), (2.5) вырождается в конечно-разностную схему порядка $P_0 = p + 1$ вида

$$\frac{du_j(t)}{dt} + \frac{1}{h} \sum_{k=-s-1}^{s} \mathring{a}_k u_{j+k}(t) = 0 \tag{5.1}$$

с начальными данными $u(0) = \hat{\Pi}_X v_0$. Свойства этой схемы хорошо изучены (см., например, [10]). В частности, её аппроксимационная ошибка имеет представление

$$(\epsilon_X(f, \hat{\Pi}))_j = c_s h^{2s+1} \frac{d^{2s+2} f}{dx^{2s+2}}((j + \theta)h), \quad -s - 1 \leqslant \theta \leqslant s, \tag{5.2}$$

где

$$c_s = \frac{s!(s+1)!}{(2s+2)!},$$

а функция $\mathring{\lambda}(\phi)$, определённая (3.6), удовлетворяет равенству

$$\operatorname{Re}\mathring{\lambda}(\phi) = (2^s c_s) \sin^{2s+2}(\phi/2). \tag{5.3}$$

Отсюда видно, что $\operatorname{Re}\mathring{\lambda}(\phi) > 0$ при $\phi/(2\pi) \notin \mathbb{Z}$, то есть эта схема является строго диссипативной.

Лемма 5.1. *Справедливо равенство*

$$\sum_{j=0}^{m-1} \hbar_j (\epsilon_X(x^{p+1}, \hat{\Pi}))_j = 0. \tag{5.4}$$

Доказательство. Рассмотрим многочлены $p_j(x)$ порядка p, определённые системой (2.5) при $u_k = (\hat{\Pi}_X x^{p+1})_k$. По построению имеем

$$(\epsilon_X(f, \hat{\Pi}))_j = \frac{1}{\hbar_j}(\Delta_{j+1/2} - \Delta_{j-1/2}),$$

где $\Delta_{j+1/2} = p_j(x_{j+1/2}) - x^{p+1}_{j+1/2}$. Таким образом, нужно проверить, что

$$p_m(x_{m+1/2}) - p_0(x_{1/2}) = x^{p+1}_{m+1/2} - x^{p+1}_{1/2}. \tag{5.5}$$

Система (2.5) для определения p_0 имеет вид

$$\int\limits_{x_{k-1/2}}^{x_{k+1/2}} p_0(x)dx = \int\limits_{x_{k-1/2}}^{x_{k+1/2}} x^{p+1}dx, \quad k = -p/2, \ldots, p/2. \tag{5.6}$$

Поскольку $x_{k+m} = x_k + mh_{av}$, система (2.5) для определения p_m может быть переписана в виде

$$\int\limits_{x_{k-1/2}}^{x_{k+1/2}} p_m(x + mh_{av})dx = \int\limits_{x_{k-1/2}}^{x_{k+1/2}} (x + mh_{av})^{p+1}dx, \quad k = -p/2, \ldots, p/2.$$

Решением этой системы является многочлен p_m порядка p, определённый равенством

$$p_m(x + mh_{av}) - p_0(x) \equiv (x + mh_{av})^{p+1} - x^{p+1}.$$

Подставив в последнее равенство $x = x_{1/2}$, мы получаем (5.5), что завершает доказательство леммы. □

Лемма 5.2. *Существует такое* $\mu : \mathbb{N} \to (0, \infty)$ *и локальное отображение* $\tilde{\Pi}$ *вида*

$$(\tilde{\Pi}_X f)_j = \frac{1}{\hbar_j} \int\limits_{x_{j-1/2}}^{x_{j+1/2}} f(x)dx + \mathfrak{C}^{(X)}_j (h_{av}(X))^{p+1} \frac{d^{p+1}f}{dx^{p+1}}(x_j), \tag{5.7}$$

где $\mathfrak{C}^{(X)}_j$ *–* $m(X)$*-периодическая последовательность действительных чисел, что система* (2.3) *точна на многочленах порядка* $p+1$ *на любой сетке* X *из семейства* $\mathcal{F}_\mu$*, заданного* (2.10). *При этом* $|\mathfrak{C}^{(X)}_j| \leqslant c_1(h_{\max} - h_{\min})/h_{av}$ *и* c_1 *зависит только от* p *и* $m(X)$.

Доказательство. Поскольку система (2.3) по построению точна на многочленах порядка p в смысле $\hat{\Pi}$, она точна на многочленах порядка p в смысле любого $\tilde{\Pi}$ вида (5.7). В силу линейности условие точности на всех многочленах порядка $p+1$ равносильно условию точности на любом выбранном многочлене порядка $p+1$ с ненулевым старшим коэффициентом, например, $f(x) = x^{p+1}/(p+1)!$.

Запишем выражение для ошибки аппроксимации на функции $f(x) = x^{p+1}/(p+1)!$ в смысле $\tilde{\Pi}$. Заметим, что $(\tilde{\Pi}_X f')_j = (\hat{\Pi}_X f')_j$ и $(\tilde{\Pi}_X f)_j = (\hat{\Pi}_X f)_j + h_{av}^{p+1}\mathfrak{C}_j^{(X)}$. Тогда

$$(\epsilon_X(f,\tilde{\Pi}))_j = -(\hat{\Pi}_X f')_j +$$
$$+\frac{1}{h_{av}}\sum_{k=-s-1}^{s} a_k\left(\frac{h_{j-S+1/2}}{h_{av}},\ldots,\frac{h_{j+S-1/2}}{h_{av}}\right)\left[(\hat{\Pi}_X f)_{j+k} + h_{av}^{p+1}\mathfrak{C}_{j+k}^{(X)}\right] =$$
$$= (\epsilon_X(f,\hat{\Pi}))_j + \sum_{k=-s-1}^{s} a_k\left(\frac{h_{j-S+1/2}}{h_{av}},\ldots,\frac{h_{j+S-1/2}}{h_{av}}\right) h_{av}^{p}\mathfrak{C}_{j+k}^{(X)}.$$

Приравнивая правую часть к нулю, получаем систему уравнений

$$\mathcal{L}(\gamma)\mathfrak{C}^{(X)} = -h_{av}^{-p}\epsilon_X(f,\hat{\Pi})$$

относительно $\mathfrak{C}^{(X)} = \{\mathfrak{C}_j^{(X)}, j \in \mathbb{Z}\}$. В силу $m(X)$-периодичности $\mathfrak{C}^{(X)}$ эта система сводится к

$$L(\gamma,0)\begin{pmatrix}\mathfrak{C}_0^{(X)}\\ \vdots \\ \mathfrak{C}_{m-1}^{(X)}\end{pmatrix} = -h_{av}^{-p}\begin{pmatrix}(\epsilon_X(f,\hat{\Pi}))_0\\ \vdots \\ (\epsilon_X(f,\hat{\Pi}))_{m-1}\end{pmatrix}. \tag{5.8}$$

По построению схема (2.3) является консервативной: для любой N-периодической последовательности u выполняется

$$\sum_{j=0}^{N-1}\hbar_j(\mathcal{L}(\gamma)u)_j = 0. \tag{5.9}$$

Для $m(X)$-периодических последовательностей свойство (5.9) принимает вид

$$\sum_{j=0}^{m-1}\hbar_j(\mathcal{L}(\gamma)u)_j = 0.$$

Отсюда, вспоминая (4.3), (4.4), (4.8), получаем

$$\sum_{j=0}^{m-1}\hbar_j(L(\gamma,0)U)_j = 0 \quad \text{для всех} \quad U \in \mathbb{C}^m,$$

то есть $(\hbar_0, \dots, \hbar_{m-1})$ является левым собственным вектором матрицы $L(\gamma, 0)$, соответствующим нулевому собственному значению. Спектр матрицы $L(0{,}0)$ дан леммой (4.2); в силу (5.3) нулевое собственное значение $L(0{,}0)$ является простым. По непрерывности оно остаётся простым при достаточно малой $|\gamma|$. Следовательно, условие совместности системы (5.8) имеет вид (5.4); по лемме 5.1 оно выполняется.

Поскольку $L(\gamma{,}0)$ непрерывна по γ, а $L(0{,}0)$ является циркулянтом, при достаточно малых $|\gamma|$ система (5.8) допускает решение, такое, что

$$\left(\sum_j |\mathfrak{C}_j^{(X)}|^2\right)^{1/2} \leqslant \frac{2}{|\lambda|_{\min}} \left(\sum_j |h_{av}^{-p}(\epsilon_X(f, \hat{\Pi}))_j|^2\right)^{1/2},$$

где $|\lambda|_{\min}$ – минимальный модуль ненулевого собственного значения $L(\gamma, 0)$. Величины $h_{av}^{-p}(\epsilon_X(f, \hat{\Pi}))_j$ не зависят от h_{av} при фиксированном γ, равны нулю при $\gamma = 0$ и являются гладкими функциями γ, поэтому при малых $|\gamma|$ они по модулю не превосходят $C|\gamma|$.

Пусть $\mathcal{F}_\mu$ – множество сеток с достаточно малыми $|\gamma|$, чтобы выполнялись оговоренные выше условия. Для $X \in \mathcal{F}_\mu$ определим $\mathfrak{C}_j^{(X)}$ условием (5.8), а для остальных сеток положим $\mathfrak{C}_j^{(X)} = 0$. Неравенство $|\mathfrak{C}_j^{(X)}| \leqslant c_1(h_{\max} - h_{\min})/h_{av}$ следует из оценки $|\gamma| \leqslant (h_{\max} - h_{\min})/h_{av}$.

Локальность отображения (5.7) следует из того, что при фиксированном γ значения $\mathfrak{C}_j^{(X)}$ не зависят от h_{av}. □

Доказательство теоремы 2.1. Поскольку на равномерной сетке порядок аппроксимации равен P_0, а на неравномерной сетке в смысле $\tilde{\Pi}$ схема точна на многочленах порядка $p = P_0$, то выполняются условия теоремы 4.8 (в терминах предыдущего раздела $q = P_0 = p + 1$). Значит, для любой $K > 1$ найдётся такое $\mu : \mathbb{N} \to (0, \infty)$, что на любой сетке $X \in \mathcal{F}_\mu$ схема устойчива с константой K.

Пусть отображение (5.7) вместе с его коэффициентами $\mathfrak{C}_j(X)$ дано леммой 5.2. Пусть $\tilde{u}(t)$ – решение с начальными данными $\tilde{\Pi}_X v_0$. Тогда

$$\begin{gathered}\|u(t) - \hat{\Pi}_X v(t,\,\cdot\,)\|_{av} \leqslant \\ \leqslant \|u(t) - \tilde{u}(t)\|_{av} + \|\tilde{u}(t) - \tilde{\Pi}_X v(t,\,\cdot\,)\|_{av} + \|\hat{\Pi}_X v(t,\,\cdot\,) - \tilde{\Pi}_X v(t,\,\cdot\,)\|_{av}.\end{gathered} \tag{5.10}$$

Последнее слагаемое в правой части (5.10) легко оценивается через $c_1|v|_{p+1}h_{\max}^p(h_{\max} - h_{\min})$. В силу устойчивости для первого слагаемого получаем

$$\begin{gathered}\|u(t) - \tilde{u}(t)\|_{av} \leqslant K\|u(0) - \tilde{u}(0)\|_{av} = \\ = K\|\hat{\Pi}_X v(0,\,\cdot\,) - \tilde{\Pi}_X v(0,\,\cdot\,)\|_{av} \leqslant Kc_1|v|_{p+1}h_{\max}^p(h_{\max} - h_{\min}).\end{gathered}$$

Второе слагаемое в правой части (5.10) является ошибкой решения по устойчивой схеме, поэтому

$$\|\tilde{u}(t) - \tilde{\Pi}_X v(t,\,\cdot\,)\|_{av} \leqslant Kt \max_{0<t'<t} \|\epsilon_X(v(t',\,\cdot\,), \tilde{\Pi})\|_{av}.$$

Для получения оценки на $(\epsilon_X(v(t',\,\cdot\,), \tilde{\Pi}))_j$ при фиксированном t' воспользуемся стандартным представлением

$$v(t', x) = p(x) + q(x),$$

где $p(x)$ – многочлен Тейлора порядка $p+1$ функции $v(t', x)$ в окрестности $x = x_j$. Отсюда $|q(x)| \leqslant (x - x_j)^{p+2}|v|_{p+2}/(p+2)!$. Поскольку схема точна на многочленах порядка $p+1$ по построению,

$$(\epsilon_X(v(t',\,\cdot\,), \tilde{\Pi}))_j = (\epsilon_X(q, \tilde{\Pi}))_j.$$

Величина в правой части последнего равенства оценивается напрямую из определения. На равномерной сетке она имеет представление (5.2). А поскольку коэффициенты $\mathfrak{C}_j^{(X)}$ стремятся к нулю при $\gamma \to 0$, для любого $\delta' > 0$ можно подобрать такое $\mathcal{F}_\mu$, что

$$\|\epsilon_X(v(t',\,\cdot\,), \tilde{\Pi})\|_{av} \leqslant (c_s + \delta')h_{\max}^{p+1}|v|_{p+2}.$$

Выбирая $\delta' = K^{-1}(c_s + \delta) - c_s$ (при K, достаточно близких к единице, эта величина положительная), получаем

$$\|\tilde{u}(t) - \tilde{\Pi}_X v(t,\,\cdot\,)\|_{av} \leqslant (c_s + \delta)th_{\max}^{p+1}|v|_{p+2}.$$

Складывая оценки на каждое из слагаемых в правой части (5.10), получаем

$$\|u(t) - \hat{\Pi}_X v(t,\,\cdot\,)\|_{av} \leqslant c|v|_{p+1}h_{\max}^{p}(h_{\max} - h_{\min}) + (c_s + \delta)|v|_{p+2}h_{\max}^{p+1}t. \quad (5.11)$$

Поскольку за счёт выбора μ отношение норм, стоящих в левых частях (2.11) и (5.11), можно сделать сколь угодно малым, отсюда следует искомая оценка (2.11). □

6. Схема R3

На равномерной сетке схема R3 вырождается в конечно-разностную схему 3-го порядка вида (5.1) с $s = 1$ и начальными данными $u(0) = \mathring{\Pi}_X v_0$. Коэффициенты схемы равны $a_{-2} = 1/6$, $a_{-1} = -1$, $a_0 = 1/2$, $a_1 = 1/3$. Как частный случай (5.3) справедливо $\mathrm{Re}\mathring{\lambda}(\phi) = \sin^4(\phi/2)/6 > 0$ при $\phi/(2\pi) \notin \mathbb{Z}$, поэтому эта схема является строго диссипативной.

Мы рассмотрим два метода анализа схемы R3. Первый метод заключается в предъявлении вспомогательного отображения и является очень громоздким. Второй метод – с опорой на спектральный анализ – не использует явного вида схемы. Но оценка, которая им будет получена, несколько слабее, чем даваемая теоремой 2.2. В частности, она допускает рост констант при $m(X) \to \infty$ и содержит дополнительный член $C_1|v|_4\, h^4_{\max}$.

Лемма 6.1. *Существует локальное отображение вида*

$$(\tilde{\Pi}_X f)_j = f(x_j) + \mathfrak{C}^{(X)}_j \frac{d^2 f}{dx^2}(x_j) + \mathfrak{D}^{(X)}_j \frac{d^3 f}{dx^3}(x_j), \tag{6.1}$$

где $\mathfrak{C}^{(X)}_j$ и $\mathfrak{D}^{(X)}_j$ – $m(X)$–периодические последовательности, такие, что

$$|\mathfrak{C}^{(X)}_j| \leqslant c_0 h_{\max}(h_{\max} - h_{\min}), \quad |\mathfrak{D}^{(X)}_j| \leqslant c_1 h^2_{\max}(h_{\max} - h_{\min}), \tag{6.2}$$

в смысле которого на любой сетке X, удовлетворяющей $\Lambda_{\max} := h_{\max}/h_{\min} < 3$, ошибка аппроксимации имеет оценку

$$\|\epsilon_X(f, \tilde{\Pi})\| \leqslant c_2|f|_3(\Delta h)^2_{\max} + c_3|f|_4 h^3_{\max}. \tag{6.3}$$

Константы c_0, c_1, c_2 и c_3 зависят только от $\Lambda_{\max}$, непрерывны при $\Lambda_{\max} \in (1{,}3)$ и имеют предельные значения $c_0 = 15/16$, $c_1 = 165/12$, $c_2 = 16/9$, $c_3 = 1/12$ при $\Lambda_{\max} \to 1$.

Значения констант c_0, c_1 и c_2, даваемые леммой 6.1, предположительно, не являются оптимальными. Предельное значение c_3, равное $1/12$, сохраняется на равномерной сетке и поэтому неулучшаемо.

Все необходимые технические выкладки для доказательства леммы 6.1 были сделаны в [11], однако общий ход рассуждений привёл к появлению в итоговой оценке дополнительных членов. Исправленное доказательство приведено в приложении.

Доказательство теоремы 2.2. Поскольку на равномерной сетке порядок аппроксимации равен 3, а на неравномерной сетке в смысле $\tilde{\Pi}$ схема точна на многочленах порядка 2, выполняются условия теоремы 4.8. Значит, для любого $K > 1$ найдётся такое семейство сеток $\mathcal{F}_\mu$, на котором схема устойчива с константой K.

Пусть отображение (5.7) вместе с его коэффициентами $\mathfrak{C}^{(X)}_j$ и $\mathfrak{D}^{(X)}_j$ даны леммой 6.1. Пусть $\tilde{u}(t)$ – решение с начальными данными $\tilde{\Pi}_X v_0$. Повторяя доказательство теоремы 1, получаем

$$\begin{gathered} \|u(t) - \hat{\Pi}_X v(t,\,\cdot\,)\|_{av} \leqslant \\ \leqslant K\|\hat{\Pi}_X v(0,\,\cdot\,) - \tilde{\Pi}_X v(0,\,\cdot\,)\|_{av} + \|\hat{\Pi}_X v(t,\,\cdot\,) - \tilde{\Pi}_X v(t,\,\cdot\,)\|_{av} + \\ + Kt \max_{0<t'<t} \|\epsilon_X(v(t',\,\cdot\,), \tilde{\Pi})\|_{av}. \end{gathered} \tag{6.4}$$

Первые два члена в правой части выражаются через коэффициенты $\mathfrak{C}_j^{(X)}$ и $\mathfrak{D}_j^{(X)}$, оценка для которых даётся (6.2). Оценка последнего слагаемого даётся (6.3).

Поскольку за счёт выбора μ отношение норм, стоящих в левых частях (2.12) и (6.4), можно сделать сколь угодно малым, отсюда следует искомая оценка (2.12). $\square$

Перейдём к другому способу анализа точности схемы R3 (2.6), (2.7), (2.8).

Лемма 6.2. *Справедливо равенство*

$$\sum_{j=0}^{m-1} \hbar_j (\epsilon_X(x^2, \mathring{\Pi}))_j = 0. \tag{6.5}$$

Доказательство. Пусть $F_{j+1/2}[f]$ определено (2.8) с подстановкой $u_k = f(x_k)$, $k \in \mathbb{Z}$. Легко убедиться, что численный поток $F_{j+1/2}[f]$ удовлетворяет следующим свойствам:

1) линейность: $F_{j+1/2}[f]$ линейна по f;
2) точность на линейной функции: $F_{j+1/2}[x+c] = x_{j+1/2} + c$;
3) если $g(x) = f(x + mh_{av})$, то $F_{j+m+1/2}[f] = F_{j+1/2}[g]$.

Рассмотрим ещё один численный поток, а именно

$$\hat{F}_{j+1/2}[f] = f(x_j) + \frac{h_{j+1/2}}{2}\frac{df}{dx}(x_j).$$

Он также удовлетворяет свойствам 1), 2) и 3).

По определению

$$(\epsilon_X(x^2, \mathring{\Pi}))_j = -2x_j + \frac{1}{\hbar_j}(F_{j+1/2}[x^2] - F_{j-1/2}[x^2]).$$

Поскольку

$$2x_j = \frac{1}{\hbar_j}(\hat{F}_{j+1/2}[x^2] - \hat{F}_{j-1/2}[x^2]),$$

получаем

$$(\epsilon_X(x^2, \mathring{\Pi}))_j = \frac{1}{\hbar_j}(\tilde{F}_{j+1/2}[x^2] - \tilde{F}_{j-1/2}[x^2]), \quad \tilde{F}_{j+1/2}[x^2] = F_{j+1/2}[x^2] - \hat{F}_{j+1/2}[x^2].$$

Остаётся показать, что $\tilde{F}_{j+m+1/2}[x^2] = \tilde{F}_{j+1/2}[x^2]$. Действительно,

$$\tilde{F}_{j+m+1/2}[x^2] = \tilde{F}_{j+1/2}[(x + mh_{av})^2] =$$

$$= \tilde{F}_{j+1/2}[x^2] + \tilde{F}_{j+1/2}[2xmh_{av} + (mh_{av})^2] = \tilde{F}_{j+1/2}[x^2].$$

Первое равенство написано в силу свойства 3) для F и $\hat{F}$, второе – в силу их линейности. Поскольку F и $\hat{F}$ точны на линейной функции, $\tilde{F}$ на ней даёт ноль, что влечёт третье равенство. □

Отметим, что поток $\hat{F}_{j+1/2}[f]$, использованный в качестве "точного", соответствует методу коррекции потоков [12, 13].

Лемма 6.3. *Существует локальное отображение* $\tilde{\Pi}$ *вида*

$$(\tilde{\Pi}_X f)_j = f(x_j) + \mathfrak{C}_j^{(X)}(h_{av}(X))^2 \frac{d^2 f}{dx^2}(x_j), \tag{6.6}$$

где $\mathfrak{C}_j^{(X)}$ *–* $m(X)$*-периодическая последовательность действительных чисел, такая, что система* (2.6)*,* (2.8) *точна на многочленах порядка* 2 *на любой сетке* X *из некоторого семейства* $\mathcal{F}_\mu$ *вида* (2.10)*. При этом* $|\mathfrak{C}_j^{(X)}| \leqslant c(\Delta h)_{\max}/h_{\max}$ *и* c *зависит только от* $m(X)$*.*

Доказательство этой леммы повторяет доказательство леммы 5.2, только вместо леммы 5.1 используется лемма 6.2.

Следствие 6.4. *Для любого* $K > 1$ *существует такое семейство* $\mathcal{F}_\mu$ *вида* (2.10)*, что любое решение* (2.6)*,* (2.8) *на любой сетке* $X \in \mathcal{F}_\mu$ *удовлетворяет оценке* $\|u(t)\|_{av} \leqslant K\|u(0)\|_{av}$*.*

Доказательство. Схема R3 на равномерной сетке вырождается в конечно-разностную схему порядка $P_0 = 3$ вида (5.1) при $p = 2$, поэтому условие строгой диссипативности (3.9) следует из (5.3). На неравномерной сетке в смысле $\tilde{\Pi}$ схема точна на полиномах порядка $q = P_0 - 1$ на некотором семействе сеток вида (2.10). Поэтому доказываемое утверждение следует из теоремы 4.8. □

Из леммы 6.3 с учётом леммы 4.7 вытекает следующий результат.

Следствие 6.5. *В окрестности нуля справедливо*

$$|\lambda_0(\gamma, \phi) - i\phi| \leqslant C(\gamma)|\phi|^3. \tag{6.7}$$

Лемма 6.6. *Пусть* $f(\boldsymbol{x}) \equiv f(x_1, \ldots, x_n)$ *– функция, гладкая в нуле, инвариантная относительно циклических перестановок аргументов. Тогда справедливо*

$$f(\boldsymbol{x}) = f(0) + O\left(\left|\sum x_j\right| + \sum |x_j|^2\right).$$

Доказательство. Любой многочлен первого порядка, инвариантный относительно циклических перестановок аргументов, имеет вид $c\sum x_j$. Отсюда утверждение леммы очевидно. □

Лемма 6.7. *Справедливо представление*

$$\lambda_0(\gamma,\phi) = \lambda_0(0,\phi) + \sum_{j,k=0}^{m-1} c_{jk}(\gamma,\phi)\gamma_j\gamma_k\phi^3, \tag{6.8}$$

где c_{jk} аналитические при $\phi \in \mathbb{R}$.

Доказательство. Пусть $\tilde{\gamma}$ – циклическая перестановка γ, то есть для некоторого натурального s выполняется $\tilde{\gamma}_j = \gamma_{j+s}$, $j \in \mathbb{Z}$. По определению

$$(L_\zeta(\tilde{\gamma}))_{j,k} = a_{\zeta m+k-j}(\gamma_{j+s-S}+1,\ldots,\gamma_{j+s+S-1}+1) = (L_\zeta(\gamma))_{j+s,k+s},$$

поэтому $(L(\tilde{\gamma},\phi))_{j,k} = (L(\gamma,\phi))_{j+s,k+s}$ и $(A(\tilde{\gamma},\phi))_{j,k} = (A(\gamma,\phi))_{j+s,k+s}$ (в индексах матриц прибавление s понимается по модулю m). Значит, $A(\tilde{\gamma},\phi)$ и $A(\gamma,\phi)$ обладают одинаковым набором собственных значений.

Контур $\mathring{\lambda}(\phi)$ не имеет самопересечений. Следовательно, при каждом ϕ в окрестности $\gamma = 0$ все собственные значения матрицы $A(\gamma,\phi)$ различны и $\lambda_0(\tilde{\gamma},\phi) = \lambda_0(\gamma,\phi)$. Таким образом, $\lambda_0(\gamma,\phi)$ как функция γ инвариантна относительно её циклических перестановок. Из предыдущей леммы с учётом $\sum_j \gamma_j = 0$ следует, что при каждом ϕ в окрестности $\gamma = 0$ выполняется

$$|\lambda_0(\gamma,\phi) - \lambda_0(0,\phi)| \leqslant C(\phi)\sum_{j=0}^{m-1}|\gamma_j|^2.$$

Сопоставляя полученный результат с (6.7), получаем (6.8). □

Лемма 6.8. *Пусть M – матрица размера $n \times n$, все собственные значения которой удовлетворяют условию $\mathrm{Re}\lambda \geqslant \mu$, где $\mu > 0$. Тогда для всех $\nu > 0$ выполняется $\|\exp(-\nu M)\| \leqslant c(n, \mu^{-1}\|M\|)$.*

Это утверждение является частью теоремы Крайса о матрицах [14]. Его доказательство с явной оценкой константы приведено в приложении.

Теперь мы можем установить основной результат. Идея заключается в использовании представления (4.17), чтобы разложить ошибку численного решения на физическую компоненту (возникающую из-за $\lambda_0(\gamma,\phi) \neq i\phi$) и паразитные компоненты (соответствующие последним $m(X)-1$ компонентам $S^{-1}(\gamma,\phi)v(\gamma,\phi,\mathring{\Pi})$). Оценку на паразитные компоненты мы получим исходя из ошибки аппроксимации, а на физическую – исходя из (6.8).

Лемма 6.9. *В окрестности $(\phi,\gamma) = (0{,}0)$ выполняется*

$$\|\hat{\varepsilon}(\gamma,\nu,\phi,\mathring{\Pi})\| \leqslant c\left[(|\gamma|\,|\phi|^2 + |\phi|^4) + \nu(|\gamma|^2|\phi|^3 + |\phi|^4)\right]$$

где c не зависит от ϕ,γ,ν.

Доказательство. Используя разложение в ряд Тейлора, для $f \in C^4(\mathbb{R})$ легко получить оценку на ошибку аппроксимации:

$$\|\epsilon_X(f,\Pi)\|_{av} \leqslant c_1(\Delta h)_{\max}|f|_2 + c_2(\Delta h)_{\max}h_{\max}|f|_3 + c_3h^3_{\max}|f|_4,$$

где c_1, c_2, c_3 – некоторые константы, не зависящие от f и сетки. Пользуясь равенством (4.11), с учётом $(\Delta h)_{\max} \leqslant \sqrt{2}h_{av}|\gamma|$ получаем

$$\|\hat{\epsilon}(\gamma,\phi,\Pi)\| \leqslant \hat{c}_1|\gamma|\,|\phi|^2 + \hat{c}_2|\gamma|\,|\phi|^3 + \hat{c}_3|\phi|^4 \leqslant c(|\gamma|\,|\phi|^2 + |\phi|^4).$$

Воспользуемся разложением (4.17). Поскольку $M^{(*)}(0{,}0)$ невырождена, в окрестности $(0{,}0)$ получаем

$$|(S^{-1}(\gamma,\phi)v(\gamma,\phi,\mathring{\Pi}))_j| \leqslant \tilde{c}(|\gamma|\,|\phi|^2 + |\phi|^4), \quad j = 1,\ldots,m(X)-1, \tag{6.9}$$

где $\tilde{c}$ зависит только от $m(X)$.

Если взять оценку на $\lambda_0(\gamma,\phi)$, даваемую леммой 4.7, мы получим ту же оценку ошибки решения, которая получается непосредственно из аппроксимации и устойчивости. Вместо неё возьмём более точную оценку (6.8). Пользуясь разложением (4.17), запишем

$$\hat{\varepsilon}(\gamma,\nu,\phi,\Pi) = S(\gamma,\phi)\mathcal{A}(\gamma,\nu,\phi)S^{-1}(\gamma,\phi)v(\gamma,\phi,\Pi),$$
$$\mathcal{A}(\gamma,\nu,\phi) = \begin{pmatrix} \exp(-\nu(\lambda_0(\gamma,\phi)-i\phi))-1 & 0 \\ 0 & \exp(i\phi\nu I - \nu M^{(*)}(\gamma,\phi)) - I \end{pmatrix}.$$

Начнём с паразитных компонент. По лемме 6.8 имеем

$$\|\exp(i\phi\nu I - \nu M^{(*)}(\gamma,\phi))\| = \|\exp(-\nu M^{(*)}(\gamma,\phi))\| \leqslant \tilde{C}_m,$$

где $\tilde{C}_m$ зависит от m, $\|M^{(*)}(\gamma,\phi)\|$ и минимальной действительной части собственных значений $M^{(*)}(\gamma,\phi)$. Все эти величины при некотором ограничении вида $|\gamma| \leqslant \tilde{\gamma}$, где $\tilde{\gamma}$ зависит от m и исходной системы (3.1), также зависят только от m и исходной системы. Соответствующие компоненты вектора $S^{-1}(\gamma,\phi)v(\gamma,\phi,\Pi)$ оцениваются по (6.9).

Теперь рассмотрим физическую компоненту. По лемме 4.6 в некоторой окрестности нуля имеем $\mathrm{Re}\lambda_0(\gamma,\phi) \geqslant 0$. Поэтому, используя неравенство

$$\mathrm{Re}z \leqslant 0 \quad \Rightarrow \quad |e^z - 1| \leqslant |z|,$$

получаем

$$|\exp(i\phi\nu - \nu\lambda_0(\gamma,\phi)) - 1| \leqslant \nu|i\phi - \lambda_0(\gamma,\phi)| \leqslant c\,\nu(|\phi|^4 + |\gamma|^2|\phi|^3).$$

Последнее неравенство написано в силу (6.8) и $\lambda_0(\phi) - i\phi = O(\phi^4)$. Соответствующая компонента $S^{-1}(\gamma,\phi)v(\phi,\Pi_X)$ стремится к $\sqrt{m}$ при $(\gamma,\phi) \to (0{,}0)$.

Собирая полученные оценки и пользуясь унитарностью $S^{-1}(0{,}0)$, получаем утверждение леммы. $\square$

Утверждение 6.10. *Существует такое семейство $\mathcal{F}_\mu$ вида* (2.10)*, что для каждой сетки $X \in \mathcal{F}_\mu$ и для каждого $v_0(x) \in C^4(\mathbb{R})$ решение $u(t)$ по схеме* (2.6)*,* (2.7)*,* (2.8) *удовлетворяет оценке*

$$\left(\sum_{j=1}^{N(X)} \hbar_j \left| u_j(t) - v(t, x_j) \right|^2 \right)^{1/2} \leqslant$$
$$\leqslant C_0 |v|_2 \, h_{\max} (h_{\max} - h_{\min}) + C_1 |v|_4 \, h_{\max}^4 + C_2 |v|_3 \, (\Delta h)_{\max}^2 t + C_3 |v|_4 \, h_{\max}^3 t, \tag{6.10}$$

причём константы C_0, C_1, C_2 и C_3 зависят только от $m(X)$.

Доказательство. В силу $|\gamma| h_{\max} \leqslant c(\Delta h)_{\max}$ и (4.12) для функций вида $f = \exp(i\alpha x)$, $\alpha \in \mathbb{Z}$, таких, что αh_{av} лежит в достаточно малой окрестности нуля, выполняется

$$\|\varepsilon_X(t, f, \mathring{\Pi})\|_{av} \leqslant \breve{c} \left[(\Delta h)_{\max} h_{\max} |f|_2 + h_{\max}^4 |f|_4 + t h_{\max}^3 |f|_4 + t (\Delta h)_{\max}^2 |f|_3 \right].$$

По теореме 6.17 из [9] (с использованием $r = 4$) эта оценка переносится с изменением константы на все 2π-периодические функции $f \in C^4(\mathbb{R})$.

Остаётся заметить, что $\hbar_j \leqslant m h_{av}$, и поэтому

$$\frac{1}{\sqrt{2\pi}} \left(\sum_{j=1}^{N(X)} \hbar_j \left| u_j(t) - v(t, x_j) \right|^2 \right)^{1/2} \leqslant \sqrt{m} \|\varepsilon_X(t, v_0, \mathring{\Pi})\|_{av}.$$

□

Подчеркнём, что в проведённых рассуждениях мы не использовали явный вид схемы R3. Поэтому утверждение леммы 6.7 также верно для схемы R5, имеющей вид (2.6), (2.7), (2.9). Однако для этой схемы $P_0 = 5$ и $q = 2$, поэтому её устойчивость не вытекает из теоремы 4.8. В разделе 7 мы покажем, что эта схема действительно неустойчива на любом семействе сеток вида (2.10).

7. Сетки с чередущимся шагом

Чтобы проиллюстрировать результаты настоящей работы, рассмотрим сетки с чередующимся шагом, т. е. с периодом $m = 2$. Будем считать, что N целое. Пусть $\xi \in [0,1)$ и $\gamma = (\xi, -\xi)$ – структура сетки. В терминах узлов сетки $x_k = k h_{av}$ для чётных k и $x_k = (k + \xi) h_{av}$ для нечётных k. Тогда $\hbar_j = h_{av}$, $h_{\max} = (1 + \xi) h_{av}$, $h_{\min} = (1 - \xi) h_{av}$, $(\Delta h)_j = 2\xi h_{av} (-1)^j$.

Начнём с анализа устойчивости для фиксированного ξ. Для этой цели удобно использовать форму (4.10), (4.8). Схема является устойчивой тогда и только тогда, когда

$$\sup_{\phi\in\mathbb{R}}\sup_{\nu\geqslant 0}\|\exp(-\nu L(\gamma,\phi))\|<\infty.$$

Для схем с полиномиальной реконструкцией такая неравномерность сетки снимается построением дуальных ячеек. Поэтому в качестве контрольных объёмов будем рассматривать интервалы (x_j, x_{j+1}) и положим

$$(\hat{\Pi}f)_j = \frac{1}{h_{j+1/2}}\int\limits_{x_j}^{x_{j+1}} f(x)dx.$$

Начнём со схем с реконструкцией многочленом 2-го порядка. Матрицы $L_\zeta(\gamma)$ можно получить напрямую из формулировки схемы. Они имеют вид

$$L_{-1} = \frac{1}{2(9-\xi^2)}\begin{pmatrix} 3-4\xi+\xi^2 & -18+10\xi \\ 0 & 3+4\xi+\xi^2 \end{pmatrix},$$

$$L_0 = \frac{1}{2(9-\xi^2)}\begin{pmatrix} 9-8\xi-\xi^2 & 6+2\xi \\ -18-10\xi & 9+8\xi-\xi^2 \end{pmatrix},$$

$$L_1 = \frac{1}{2(9-\xi^2)}\begin{pmatrix} 0 & 0 \\ 6-2\xi & 0 \end{pmatrix}.$$

Матрица $L(\gamma,\phi)$ имеет два семейства собственных значений:

$$\lambda_*(\phi) = \frac{12}{9-\xi^2} + O(\phi), \quad \lambda_0(\phi) = i\phi + \frac{1-\xi^2}{12}\phi^4 + O(\phi^5)$$

при $\phi \to 0$. Численное вычисление показывает, что условие $\mathrm{Re}\lambda(\phi) > 0$ выполняется для обоих собственных значений, всех $\phi \neq k\pi$ ($k \in \mathbb{Z}$) и всех $\xi \in [0{,}1)$. Таким образом, схема на основе реконструкции многочленом 2-го порядка устойчива на любой сетке с периодом $m = 2$.

Теперь рассмотрим схему R3. Обозначим $\mathcal{H} = (1+\xi)/(1-\xi)$, тогда коэффициенты блочного представления схемы имеют вид

$$L_{-1} = \frac{1}{6}\begin{pmatrix} \mathcal{H}^{-1} & -4-\mathcal{H}-\mathcal{H}^{-1} \\ 0 & \mathcal{H} \end{pmatrix},$$

$$L_0 = \frac{1}{6}\begin{pmatrix} 2+\mathcal{H} & 2 \\ -4-\mathcal{H}-\mathcal{H}^{-1} & 2+\mathcal{H}^{-1} \end{pmatrix}, \; L_1 = \frac{1}{6}\begin{pmatrix} 0 & 0 \\ 2 & 0 \end{pmatrix}.$$

Матрица $L(\gamma,\phi)$ имеет два семейства собственных значений:

$$\lambda_*(\gamma,\phi) = \frac{4}{3(1-\xi^2)} + O(\phi),$$

$$\lambda_0(\gamma,\phi) = i\phi + \frac{1}{3} i\xi^2\phi^3 + \frac{1}{12}(1-\xi^2)(1-4\xi^2)\phi^4 + O(\phi^5)$$

при $\phi \to 0$. Если $\xi < 1/2$, оба собственных значения имеют неотрицателюную действительную часть при $\phi = 0$. Более точный анализ показывает, что при $\xi \leqslant 1/2$ условие $\mathrm{Re}\lambda(\phi) > 0$ выполняется при всех $\phi \neq \pi k$. Однако, если $\xi > 1/2$, в проколотой окрестности $\phi = 0$ выполняется $\mathrm{Re}\lambda_0(\phi) < 0$. Это означает, что при $\xi > 1/2$ схема R3 неустойчива и неустойчивость развивается на низкочастотных модах.

Пусть $\xi \in (0{,}1/2)$, $v_0(x) = \exp(i\alpha x)$, $\alpha \in \mathbb{N}$. Получим для этого случая выражение для старшего члена ошибки в предположениях $t/h_{av} \gg 1$, $\alpha h_{av} \ll 1$. Напомним, что в силу (4.12) выполняется

$$\|\varepsilon_X(t, \exp(i\alpha x), \Pi)\|_{av} = \frac{1}{\sqrt{2}}\|\hat{\varepsilon}(\gamma, t/h_{av}, \alpha h_{av}, \Pi)\|,$$

где

$$\hat{\varepsilon}(\gamma, t/h_{av}, \alpha h_{av}, \Pi) =$$
$$= \left(\exp\left(\frac{t}{h_{av}}(i\alpha h_{av} - L(\gamma, \alpha h_{av}))\right) - I\right)\begin{pmatrix} 1 \\ \exp(i\alpha h_{av}(1+\xi)) \end{pmatrix}.$$

Пусть $r_*(\gamma,\phi)$ и $r_0(\gamma,\phi)$ – правые собственные векторы, соответствующие λ_* и λ_0, а l_* и l_0 – соответствующие левые собственные векторы. Нормируем их так, чтобы при $\xi \to 0$ они стремились к ортонормированному базису в $\mathbb{C}^2$. Поскольку $\mathrm{Re}\lambda_*$ ограничено снизу, при сделанных предположениях

$$\|\varepsilon_X(t, \exp(i\alpha x), \Pi)\|_{av} = (1 + O(\xi))\,(E_1 + E_2) + O(\exp(-ct/h_{av})),$$

где

$$E_1 = \left|\exp\left(\frac{t}{h_{av}}\left(i\alpha h_{av} - \lambda_0(\gamma, \alpha h_{av})\right)\right) - 1\right|$$

является главным членом ошибки в длительном счёте, а

$$E_2 = \frac{1}{\sqrt{2}}\left|l_*(\gamma, \alpha h_{av})\begin{pmatrix} 1 \\ \exp(i\alpha h_{av}(1+\xi)) \end{pmatrix}\right|$$

овечает за неоптимальную интерпретацию численного решения. Если $E_1 \ll 1$, то, используя выражение для λ_0, разложение экспоненты в ряд Тейлора и равенство $\xi = (\Delta h)_{\max}/2$, получаем

$$E_1 = \left|\frac{i}{12}(\Delta h)^2_{\max}\alpha^2 + \frac{1}{12}h^3_{\max}\alpha^3 + O((\alpha h_{\max})^4)\right|\,\alpha t.$$

Разложение в ряд Тейлора λ_* и l_* даёт

$$E_2 = \frac{1}{2}\xi\alpha^2 h_{av}^2 + O(\xi(\alpha h_{av})^3) = \frac{1}{4}(\Delta h)_{\max} h_{av}\alpha^2 + O(\xi(\alpha h_{av})^3).$$

Это показывает, что в теореме 2.2 в случае $m = 2$ множитель перед $|v|_2$ можно положить равным $1/4$, а множитель перед $|v|_3(\Delta h)_{\max}^2 t$ – равным $1/12$.

Наконец, рассмотрим схему R5. Коэффициенты в блочном представлении схемы равны

$$L_{-2} = \frac{1}{60}\begin{pmatrix} 0 & -2 \\ 0 & 0 \end{pmatrix}, \quad L_{-1} = \frac{1}{60}\begin{pmatrix} 4 + 11\mathcal{H}^{-1} & -38 - 11\mathcal{H} - 11\mathcal{H}^{-1} \\ -2 & 4 + 11\mathcal{H} \end{pmatrix},$$

$$L_0 = \frac{1}{60}\begin{pmatrix} 12 + 11\mathcal{H} - 3\mathcal{H}^{-1} & 24 + 3\mathcal{H} + 3\mathcal{H}^{-1} \\ -38 - 11\mathcal{H} - 11\mathcal{H}^{-1} & 12 + 11\mathcal{H}^{-1} - 3\mathcal{H} \end{pmatrix},$$

$$L_1 = \frac{1}{60}\begin{pmatrix} -3\mathcal{H} & 0 \\ 24 + 3\mathcal{H} + 3\mathcal{H}^{-1} & -3\mathcal{H}^{-1} \end{pmatrix}.$$

Матрица $L(\gamma, \phi)$ имеет два семейства собственных значений:

$$\lambda_*(\gamma, \phi) = \frac{16}{15(1 - \xi^2)} + O(\phi),$$

$$\lambda_0(\gamma, \phi) = i\phi + \frac{1}{3}i\xi^2\phi^3 - \frac{5}{12}\xi^2(1 - \xi^2)\phi^4 + O(\phi^5)$$

при $\phi \to 0$. Очевидно, что при любом $\xi \neq 0$ в окрестности $\phi = 0$ выполняется $\mathrm{Re}\lambda_0(\gamma, \phi) < 0$. Это значит, что схема R5 неустойчива на любой неравномерной сетке с $m = 2$.

Разложения в ряд Тейлора, приведённые в этом разделе, получены с использованием математического пакета Sage.

Приложение

A. Доказательство леммы 6.1

Все необходимые технические выкладки для доказательства этой леммы были сделаны в [11], однако общий ход рассуждений привёл к появлению в итоговой оценке дополнительных членов. Приведём правильный ход рассуждений, заимствуя технические результаты из [11].

Доказательство. Пусть X – некоторая сетка, такая, что $h_{\max}/h_{\min} < 3$. В тексте доказательства будем опускать аргумент X у N и у коэффициентов $\mathfrak{C}_j$ и $\mathfrak{D}_j$. Через $\mathfrak{C}$ будем обозначать набор $\mathfrak{C}_0, \ldots, \mathfrak{C}_{N-1}$. Обозначим $\Lambda_j = h_{j+1/2}/h_{j-1/2}$.

Шаг 1. Представим схему (2.6), (2.8) в виде

$$\frac{du}{dt} + Lu = 0,$$

где $u \in \mathbb{C}^N$. Тогда L задаётся равенством

$$L = H^{-1}DF, \tag{A.1}$$

где $H = \mathrm{diag}\{\hbar_j, j = 0, \ldots, N-1\}$,

$$D = \left\| \begin{array}{lllllll} 1 & 0 & 0 & 0 & 0 & 0 & -1 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{array} \right\|,$$

$$F = \left\| \begin{array}{lllllll} \frac{2}{3} + \frac{\Lambda_0}{6} & \frac{1}{3} & 0 & 0 & 0 & 0 & -\frac{\Lambda_0}{6} \\ -\frac{\Lambda_1}{6} & \frac{2}{3} + \frac{\Lambda_1}{6} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ 0 & -\frac{\Lambda_2}{6} & \frac{2}{3} + \frac{\Lambda_2}{6} & \frac{1}{3} & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \frac{1}{3} & 0 & 0 & 0 & 0 & -\frac{\Lambda_{N-1}}{6} & \frac{2}{3} + \frac{\Lambda_{N-1}}{6} \end{array} \right\|.$$

Схему R3 можно также переписать через разделённые разности:

$$\hbar_j \frac{du_j}{dt} = \left(\frac{2}{3}h_{j-1/2} + \frac{1}{6}h_{j+1/2}\right)\frac{u_j - u_{j-1}}{h_{j-1/2}} -$$

$$-\frac{1}{6}h_{j-1/2}\frac{u_{j-1} - u_{j-2}}{h_{j-3/2}} + \frac{1}{3}h_{j+1/2}\frac{u_{j+1} - u_j}{h_{j+1/2}} = 0.$$

Отсюда, вводя $\tilde{H} = \mathrm{diag}\{h_{j-1/2}, j = 0, \ldots, N-1\}$, получаем представление

$$L = G\tilde{H}^{-1}D.$$

Через $\|\cdot\|_\infty$ будем обозначать векторную норму $\|f\|_\infty = \max_j\{|f|_j\}$ и индуцированную ей матричную норму. Тогда

$$\|F^{-1}\|_\infty \leqslant C_F = 3 + \frac{3}{4}\Lambda_{\max}, \quad \|G^{-1}\|_\infty \leqslant C_G = 3\frac{4 + \Lambda_{\max}}{3 - \Lambda_{\max}}. \tag{A.2}$$

Шаг 2. Изучим ошибку аппроксимации схемы в смысле отображения $\mathring{\Pi}$ на многочленах второго порядка. Справедливо

$$(\epsilon_X(x^2/2, \mathring{\Pi}))_j = \frac{1}{\hbar_j}(\tilde{F}_{j+1/2} - \tilde{F}_{j-1/2}),$$

$$\tilde F_{j+1/2} = \frac{h_{j+1/2}}{4}\left(\frac{2}{3}h_{j+1/2} - \frac{1}{3}h_{j-1/2}\right).$$

Или, в векторной записи,

$$\epsilon_X(x^2/2, \mathring{\Pi}) = H^{-1}D\tilde f, \quad \tilde f = \{\tilde F_{1/2}, \ldots, \tilde F_{N-1/2}\}.$$

Шаг 3. Определим $\mathfrak{C}_j$ из условия точности на многочленах 2-го порядка в смысле $\tilde\Pi$ (очевидно, что коэффициенты $\mathfrak{D}_j$ в этом условии не участвуют). Опять же, достаточно рассмотреть только многочлен $x^2/2$.

$$\left(\epsilon_X\left(x^2/2, \tilde\Pi\right)\right)_j = \left(\epsilon_X\left(x^2/2, \Pi\right)\right)_j + (L\mathfrak{C})_j.$$

Приравнивая это равенство к нулю и пользуясь представлением (A.1), получаем

$$H^{-1}D(F\mathfrak{C} + \tilde f) = 0.$$

Очевидно, эта система совместна, и её общим решением является

$$\mathfrak{C} = -F^{-1}\tilde f + \alpha e = -F^{-1}(\tilde f + \alpha e),$$

где $e = (1, \ldots, 1)^T$. Последнее равенство следует из $Fe = e$. Выберем $\alpha = -h_{av}^2/12$, тогда

$$\|\tilde f + \alpha e\|_\infty = \max_j \left|\frac{h_{j+1/2}}{4}\left(\frac{2}{3}h_{j+1/2} - \frac{1}{3}h_{j-1/2}\right) - \frac{1}{12}h_{av}^2\right| \leqslant$$

$$\leqslant \frac{1}{12}\max_j \left|h_{j+1/2}(h_{j-1/2} - h_{j+1/2}) + (h_{av} - h_{j+1/2})(h_{av} + h_{j+1/2})\right| \leqslant$$

$$\leqslant \frac{1}{4}h_{\max}(h_{\max} - h_{\min}).$$

Пользуясь оценкой (A.2) для $\|F^{-1}\|$, получаем

$$\|\mathfrak{C}\|_\infty \leqslant C_F\|\tilde f + \alpha e\|_\infty \leqslant \left(3 + \frac{3}{4}\Lambda_{\max}\right)\frac{1}{4}h_{\max}(h_{\max} - h_{\min}).$$

Это даёт первое неравенство в (6.2).

Также нам понадобится оценка на $|\mathfrak{C}_j - \mathfrak{C}_{j-1}|/h_{j-1/2}$. Имеем

$$\tilde H^{-1}D\mathfrak{C} = -G^{-1}\epsilon_X(x^2/2, \mathring{\Pi}),$$

$$|(\epsilon_X(x^2/2, \mathring{\Pi}))_j| = \frac{1}{\hbar_j}|\tilde F_{j+1/2} - \tilde F_{j-1/2}| =$$

$$= \frac{1}{\hbar_j}\left|-\frac{1}{6}(h^2_{j+1/2}-h^2_{j-1/2})+\frac{1}{12}h_{j-1/2}(h_{j+1/2}-h_{j-3/2})\right| \leqslant \frac{2}{3}(\Delta h)_{\max},$$

поэтому

$$\|\tilde{H}^{-1}D\mathfrak{C}\|_\infty = \max_j \frac{|\mathfrak{C}_j-\mathfrak{C}_{j-1}|}{h_{j-1/2}} \leqslant \frac{2}{3}C_G(\Delta h)_{\max}.$$

Шаг 4. Запишем теперь ошибку аппроксимации на функции $f(x)=(x-x_j)^3/6$ в смысле $\tilde{\Pi}$ с зафиксированными коэффициентами $\mathfrak{C}_j$ и нулевыми $\mathfrak{D}_j$ (обозначим этот оператор через $\tilde{\Pi}^{(2)}$). Имеем

$$(\epsilon_X(f,\tilde{\Pi}^{(2)}))_k = \frac{\eta_{k+1/2}-\eta_{k-1/2}}{\hbar_k}+g_k,$$

где $\eta_{k+1/2}=\eta^{(q)}_{k+1/2}+\eta^{(p)}_{k+1/2}$,

$$\eta^{(q)}_{k+1/2} = \frac{h^2_{k-1/2}h_{k+1/2}+3h_{k-1/2}h^2_{k+1/2}}{36},$$

$$\eta^{(p)}_{k+1/2} = h_{k+1/2}\left[\frac{1}{3}\mathfrak{C}_{k-1}-\frac{1}{2}\mathfrak{C}_k+\frac{1}{6}\mathfrak{C}_{k+1}\right]+\frac{1}{6}h^2_{k+1/2}\frac{\mathfrak{C}_k-\mathfrak{C}_{k-1}}{h_{k-1/2}},$$

$$g_k = \frac{1}{3}(h_{k+1/2}-h_{k-1/2})\frac{\mathfrak{C}_k-\mathfrak{C}_{k-1}}{h_{k-1/2}}+\frac{1}{36\hbar_j}(2h_{j+1/2}+h_{j-1/2})(h_{j+1/2}-h_{j-1/2})^2. \tag{A.3}$$

Теперь выберем $\mathfrak{D}_j$ таким образом, чтобы $\epsilon_k(t,\tilde{\Pi})=g_k$. Система для их нахождения аналогична системе на нахождение коэффициентов $\mathfrak{C}$:

$$H^{-1}D(F\mathfrak{D}+\eta)=0.$$

Выбирая $\mathfrak{D}=h^3_{av}/9-F^{-1}\eta$, получаем

$$|\mathfrak{D}_j| \leqslant C_F\max_k\left|\eta_{k+1/2}-\frac{h^3_{av}}{9}\right|.$$

Легко показать, что для любых трёх шагов $h_{j+1/2}$, $h_{k+1/2}$, $h_{l+1/2}$ справедливо

$$|h_{j+1/2}h_{k+1/2}h_{l+1/2}-h^3_{av}| \leqslant 3h^2_{\max}(h_{\max}-h_{\min}),$$

поэтому

$$\left|\eta^{(q)}_{k+1/2}-\frac{h^2_{av}}{9}\right| \leqslant \frac{1}{3}h^2_{\max}(h_{\max}-h_{\min}).$$

Далее

$$|\eta^{(p)}_{k+1/2}| \leqslant \frac{2}{3}h_{\max}\|D\mathfrak{C}\|_\infty+\frac{1}{6}\Lambda_{\max}(\Delta h)_{\max}\|\mathfrak{C}\|_\infty \leqslant$$

$$\leqslant \frac{4}{3}h_{\max}^2\frac{4+\Lambda_{\max}}{3-\Lambda_{\max}}(\Delta h)_{\max}+\frac{1}{6}\Lambda_{\max}(\Delta h)_{\max}\left(3+\frac{3}{4}\Lambda_{\max}\right)\frac{1}{4}h_{\max}(h_{\max}-h_{\min}).$$

Опуская $h_{\min}$ во втором слагаемом и пользуясь $(\Delta h)_{\max} \leqslant h_{\max}-h_{\min}$, получаем

$$|\eta_{k+1/2}^{(p)}| \leqslant \tilde{c}h_{\max}^2(h_{\max}-h_{\min}), \quad \tilde{c}=\frac{4}{3}\frac{4+\Lambda_{\max}}{3-\Lambda_{\max}}+\frac{\Lambda_{\max}}{24}\left(3+\frac{3}{4}\Lambda_{\max}\right).$$

Отсюда

$$|\mathfrak{D}_j| \leqslant C_F\left(\tilde{c}+\frac{1}{3}\right)h_{\max}^2(h_{\max}-h_{\min}).$$

Это даёт оценку (6.2).

Шаг 5. Рассмотрим ошибку аппроксимации в узле j на произвольной функции $f(x)$ в смысле $\tilde{\Pi}$. Представим $f(x)=p(x)+q(x)$, где p – многочлен Тейлора степени 3 в точке x_j. Тогда

$$(\epsilon_X(f,\tilde{\Pi}))_j=(\epsilon_X(p,\tilde{\Pi}))_j+(\epsilon_X(q,\tilde{\Pi}))_j=g_jp'''(x_j)+(\epsilon_X(q,\tilde{\Pi}))_j.$$

Напрямую из (A.3) имеем

$$|g_j| \leqslant \frac{1}{3}(\Delta h)_{\max}\|\tilde{H}^{-1}D\mathfrak{C}\|_\infty+\frac{1}{9}(\Delta h)_{\max}^2 \leqslant \left(\frac{2}{3}\frac{4+\Lambda_{\max}}{3-\Lambda_{\max}}+\frac{1}{9}\right)(\Delta h)_{\max}^2.$$

Остаётся оценить $(\epsilon_X(q,\tilde{\Pi}))_j$. Поскольку $|q(x)| \leqslant |f|_4|x-x_j|^3/24$, в силу непрерывной зависимости коэффициентов схемы и коэффициентов $\mathfrak{C}_k$ и $\mathfrak{D}_k$ от положений узлов, имеет место оценка

$$|(\epsilon_X(q,\tilde{\Pi}))_j| \leqslant \left(\frac{1}{12}+\delta\right)|f|_4h_{\max}^3,$$

где $\delta \to 0$ при $\Lambda_{\max} \to 1$. □

B. Доказательство леммы 6.8

Лемма B.1. *Пусть* $A=\{a_{ij}\}_{i,j=1}^n$ *– верхнетреугольная матрица размера* $n\times n$*, причём* $\mu_j := \operatorname{Re} a_{jj} \leqslant 0$*,* $j=1,\ldots,n$*, и* $|a_{jk}| \leqslant C\min(|\mu_j|,|\mu_k|)$ *для* $k>j$*. Тогда для всех* $t \geqslant 0$ *выполняется* $\|\exp(At)\| \leqslant (1+C)^{n-1}$.

Доказательство. Рассмотрим форму

$$\Phi(x)=(Sx)^*Sx, \quad x=(x_1,\ldots,x_n)^T \in \mathbb{C}^n.$$

Для любого решения системы $\dot{x}=Ax$ выполняется

$$\frac{d}{dt}\Phi(x)=\Psi(Sx), \quad \text{где} \quad \Psi(z)=\overline{z}^T((SAS^{-1})^*+SAS^{-1})z.$$

Возьмём $S = \mathrm{diag}\,\{1, q, q^2, \ldots, q^{n-1}\}$, $q > 1$. Тогда $(SAS^{-1})_{jk} = a_{jk}q^{j-k}$ и

$$\Psi(z) = 2\sum_{j=1}^{n} \mu_j |z_j|^2 + 2\mathrm{Re}\sum_{k>j} a_{jk}q^{j-k}\overline{z}_j z_k.$$

По условию леммы выполняется

$$2|a_{jk}\overline{z}_j z_k| \leq -C\mu_j|z_j|^2 - C\mu_k|z_k|^2,$$

следовательно,

$$2\left|\sum_{k>j} a_{jk}q^{j-k}\overline{z}_j z_k\right| \leqslant -C\sum_{j=1}^{n}\mu_j|z_j|^2\left(\sum_{l=1}^{j-1} q^{l-j} + \sum_{m=j+1}^{n} q^{j-m}\right) \leqslant$$
$$\leqslant -\frac{2C}{q-1}\sum_{j=1}^{n}\mu_j|z_j|^2.$$

Возьмём $q = 1 + C$, тогда $\Psi(z) \leqslant 0$ для всех $z \in \mathbb{C}^n$. Следовательно, функция $\Phi(x(t))$ не возрастает по t, откуда

$$\|x(t)\|^2 \leqslant \Phi(x(t)) \leqslant \Phi(x(0)) \leqslant q^{2(n-1)}\|x(0)\|^2.$$

Неравенство $\|x(t)\| \leqslant q^{n-1}\|x(0)\|$ эквивалентно доказываемой оценке. □

Лемма B.2. *Пусть M – матрица размера $n \times n$, все собственные значения которой удовлетворяют условию* $\mathrm{Re}\lambda \geqslant \mu$*, где* $\mu > 0$*. Тогда для всех* $\nu > 0$ *выполняется* $\|\exp(-\nu M)\| \leqslant (1 + \|M\|\mu^{-1})^{n-1}$.

Доказательство. По теореме Шура существуют унитарная матрица Q и верхнетреугольная матрица B, такие, что $-M = QBQ^*$. На диагонали в матрице B стоят величины, действительная часть которых не больше $-\mu$; внедиагональные элементы по модулю не превосходят $\|B\| = \|M\|$. Отсюда в силу утверждения B.1 следует оценка $\|\exp(-\nu B)\| \leqslant (1 + \|M\|\mu^{-1})^{n-1}$, а в силу унитарности матрицы Q имеем искомую оценку на $\exp(-\nu M) = Q\exp(-\nu B)Q^*$. □

Список литературы

1. Tsoutsanis P., Titarev V. A., Drikakis D. WENO schemes on arbitrary mixed-element unstructured meshes in three space dimensions // Journal of Computational Physics. 2011. Vol. 230. P. 1585–1601.

2. Antoniadis A. F., Tsoutsanis P., Drikakis D. Assessment of High-Order Finite Volume Methods on Unstructured Meshes for RANS Solutions of Aeronautical Configurations // Journal of Computational Physics. 2017. Vol. 256. P. 254–276.

3. Bakhvalov P. A., Kozubskaya T. K. EBR-WENO scheme for solving gas dynamics problems with discontinuities on unstructured meshes // Computers and Fluids. 2017. Vol. 157. P. 312–324.

4. Bakhvalov P. A., Kozubskaya T. K., Rodionov P. V. EBR schemes with curvilinear reconstructions for hybrid meshes // Computers and Fluids. 2022. Vol. 239, no. 105352.

5. Zangeneh R., Oliver-Gooch C. F. Stability Analysis and Improvement of the Solution Reconstruction for Cell Centered Finite Volume Methods on Unstructured Meshes // Journal of Computational Physics. 2019. Vol. 393. P. 375–405.

6. Enhanced accuracy by post-processing for finite element methods for hyperbolic equations / Cockburn B., Luskin M., Shu C.-W. et al. // Mathematics of Computation. 2003. Vol. 72. P. 577–606.

7. 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.

8. Shu C.-W. Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws: Tech. Rep.: ICASE Report 97-65, 1997.

9. Bakhvalov P. A., Surnachev M. D. Linear schemes with several degrees of freedom for the transport equation and the long-time simulation accuracy // IMA Journal of Numerical Analysis. 2023. URL: http://doi.org/10.1093/imanum/drad006.

10. Iserles A. Order stars and a saturation theorem for first order hyperbolics // IMA Journal of Numerical Analysis. 1982. Vol. 2. P. 49–61.

11. Бахвалов П. А. Метод нестационарного корректора для анализа точности линейных полудискретных схем // Препринты ИПМ им. М.В.Келдыша. 2018. № 123. С. 1–38.

12. Katz A., Sankaran V. An Efficient Correction Method to Obtain a Formally Third-Order Accurate Flow Solver for Node-Centered Unstructured Grids // J. Sci. Comput. 2012. T. 51, № 2. C. 375–393.

13. Nishikawa H. Accuracy-Preserving Source Term Quadrature for Third-Order Edge-Based Discretization // J. Comput. Phys. 2017. T. 344. C. 595–622.

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