Версия 0.4 beta
26 апреля 2022

Введение в высокоточные методы численного решения систем законов сохранения.

Часть 2:
схемы для уравнения переноса

Содержание

5. Схемы высокого порядка для уравнения переноса

В настоящем разделе рассмотрим основные подходы к построению полудискретных схем высокого порядка[1] на примере задачи Коши

$$\frac{\partial u}{\partial t} + \mu \frac{\partial u}{\partial x} = 0, \quad t \in (0, t_{\max}), \quad x \in (0, 2\pi), \tag{5.1}$$

$$u(t,0) = u(t,2\pi), \quad u(0,x) = u_0(x), \tag{5.2}$$

Решение будем предполагать достаточно гладким; конкретные требования к гладкости решения будут понятны из получаемых оценок точности.

Для каждого метода будем уточнять, рассматриваем ли мы равномерные или неравномерные расчётные сетки. Под неравномерной сеткой будем понимать невырожденное разбиение $\{x_j\}$ отрезка $[0, 2\pi]$, а под равномерной сеткой её частный случай, при котором $x_j = jh$. Будем предполагать, что координаты сеточных узлов доопределены формулой $x_{j+kN_x} \equiv x_j + 2\pi k$ для всех $j = 0, \ldots, N_x - 1$ и $k \in \mathbb{Z}$.

5.1. Конечно-разностные схемы высокого порядка. В этом разделе мы будем рассматривать равномерные сетки и случай $\mu > 0$; случай $\mu < 0$ сводится к предыдущему заменой x на $-x$.

Если построение полностью дискретной схемы методом характеристик сводится к вычислению значения интерполяционного полинома в точке, то построение полудискретной конечно-разностной схемы – к вычислению производной от интерполяционного многочлена.

Пространственную производную можно аппроксимировать с порядком p производной от интерполяционного многочлена Лагранжа порядка p, построенного по значениям в $p+1$ точках. Например, если построить интерполяционный многочлен по точкам x_{j-1} и x_j, он будет равен

$$I(x) = v_{j-1} \frac{x_j - x}{h} + v_j \frac{x - x_{j-1}}{h},$$

а его производная

$$I'(x) = \frac{v_j - v_{j-1}}{h},$$

[1] Понятие схем высокого порядка в вычислительной газовой динамике менялось со временем. В 80-х – 90-х годах под ними понимались схемы второго порядка точности. Однако проведённый в 2012 году опрос ведущих специалистов показал, что большинство понимает под ними схемы, обладающие, как минимум, 3-м порядком точности для гладких решений (Wang и др., High-Order CFD Methods: Current Status and Perspective, IJNMF, 2013). В настоящем курсе под схемами высокого порядка будем понимать подходы, позволяющие построить схемы любого наперёд заданного порядка аппроксимации.

то есть производная по x будет аппроксимирована направленной разностью. Если же построить интерполяционный многочлен второго порядка по точкам x_{j-1}, x_j, x_{j+1}, получим

$$I(x) = v_{j-1}\frac{(x-x_j)(x-x_{j+1})}{h^2} + \\ + v_j\frac{(x-x_{j-1})(x-x_{j+1})}{-h^2} + v_{j+1}\frac{(x-x_{j-1})(x-x_j)}{h^2},$$

и его производная при $x = x_j$ равна

$$I'(x_j) = \frac{v_{j+1} - v_{j-1}}{2h},$$

то есть производная по x будет аппроксимирована центральной разностью и схема приобретёт вид центральной разностной схемы 2-го порядка:

$$\frac{dv_j}{dt} + \frac{\mu}{2h}(v_{j+1} - v_{j-1}) = 0, \quad v_j(0) = u(0, jh).$$

Используя узлы $j-2$, $j-1$, j, $j+1$, можно получить схему 3-го порядка аппроксимации

$$\frac{dv_j}{dt} + \frac{\mu}{h}\left(\frac{1}{6}v_{j-2} - v_{j-1} + \frac{1}{2}v_j + \frac{1}{3}v_{j+1}\right) = 0, \quad v_j(0) = u(0, jh). \tag{5.3}$$

Нас будут интересовать два семейства конечно-разностных схем: на симметричном шаблоне $j-m, \ldots, j+m$ и на скошенном шаблоне $j-m-1, \ldots, j+m$. Будем обозначать эти схемы символом FDn, где $n = 2m$ или $n = 2m+1$, соответственно.

Если мы запишем эти схемы в операторном виде

$$\frac{dv}{dt} + \frac{\mu}{h}A_h v = 0, \quad v(0) = \Pi_h u_0, \tag{5.4}$$

то A_h будет циркулянтом. Всюду далее будем предполагать, что число узлов в сетке больше ширины шаблона схемы, то есть $N_x > 2m+1$.

Начнём со схем на симметричном шаблоне.

Лемма 5.1. *Пусть $m \in \mathbb{N}$. Пусть $I(x)$ – интерполяционный многочлен порядка $2m$, построенный по точкам $x_{-m}, \ldots, x_m$. Пусть $b_k \in \mathbb{R}$, $k = -m, \ldots, m$, такие, что*

$$I'(0) = \frac{1}{h}\sum_{k=-m}^{m} b_k v_k.$$

Тогда $b_{-k} = -b_k$ и любое N_x-периодическое решение

$$\frac{dv_j}{dt} + \frac{\mu}{h} \sum_{k=-m}^{m} b_k v_{j+k} = 0$$

удовлетворяет равенству $\|v(t)\|_2 = \|v(0)\|_2$, где

$$\|f\|_2 = \left(\sum_{j=0}^{N_x-1} h|f_j|^2 \right)^{1/2}.$$

Доказательство. Покажем вначале, что $b_{-k} = -b_k$. Действительно, b_k является производной в нуле от многочлена $I_k(x)$ порядка $2m$, равного 1 при $x = k$ и нулю в остальных целых числах в диапазоне от $-m$ до m. Аналогично, b_{-k} является производной в нуле от многочлена $I_{-k}(x)$ порядка $2m$, равного 1 при $x = -k$ и нулю в остальных целых числах в диапазоне от $-m$ до m. Очевидно, что $I_{-k}(x) = I_k(-x)$, откуда $I'_k(0) = -I'_{-k}(0)$, что и требовалось доказать.

Перепишем теперь уравнение в матричном виде:

$$\frac{dv}{dt} + \frac{\mu}{h} A_h v = 0.$$

Матрица A_h является кососимметрическим циркулянтом. В этом случае имем

$$v^T A_h v = (v^T A_h v)^T = v^T A_h^T v = v^T (-A_h) v$$

и, следовательно, $v^T A_h v = 0$. Значит,

$$\frac{d\|v(t)\|_2^2}{dt} = \frac{d}{dt}(hv^T v) = 2hv^T \frac{dv}{dt} = -2h\frac{\mu}{h} v^T A_h v = 0,$$

что и требовалось доказать. □

Сохранение нормы решения говорит о том, что все собственные значения A_h являются чисто мнимыми, то есть ни одна из мод решения не затухает. Такие схемы называются *бездиссипативными*. Это свойство может быть в одних случаях полезным, а в других – вредным.

Упражнение 1. *Для схемы FDn, где n – чётное, найти такое $\sigma = \sigma(n)$, что при использовании 3-стадийного метода Рунге – Кутты 3-го порядка полученная полностью дискретная схема будет устойчивой при $\tau = \sigma h/\mu$ (неулучшаемую оценку на σ получать не нужно). Доказать, что при использовании явного метода Эйлера при любом $\sigma > 0$ такая схема не будет устойчивой.*

Рассмотрим теперь случай схемы на скошенном шаблоне $j - m, \ldots, j + m - 1$, где $m \in \mathbb{N}$. Обратим внимание, что мы рассматриваем не произвольный несимметричный шаблон, а шаблон, имеющий против направления переноса ровно на один узел больше, чем по направлению переноса.

Теорема 5.2. *Пусть $m \in \mathbb{N}$. Пусть $I(x)$ – интерполяционный многочлен порядка $2m - 1$, построенный по точкам $x_{-m}, \ldots, x_{m-1}$. Пусть $a_k \in \mathbb{R}$, $k = -m, \ldots, m - 1$, такие, что*

$$I'(0) = \frac{1}{h} \sum_{k=-m}^{m-1} a_k v_k.$$

Пусть $\mu \geqslant 0$. Тогда любое N_x-периодическое решение

$$\frac{dv_j}{dt} + \frac{\mu}{h} \sum_{k=-m}^{m-1} a_k v_{j+k} = 0$$

удовлетворяет оценке $\|v(t)\|_2 \leqslant \|v(0)\|_2$.

Доказательство. Введём линейное пространство $\mathbb{R}^{\mathbb{Z}}$ числовых последовательностей, бесконечных в обе стороны, т. е. множество отображений из $\mathbb{Z}$ в $\mathbb{R}$. Пусть $\mathcal{L}$ – пространство линейных операторов на $\mathbb{R}^{\mathbb{Z}}$ вида

$$(Lf)_j = \sum_{k=-m}^{m} \alpha_k f_{j+k},$$

где $\alpha_k \in \mathbb{R}$ – некоторые коэффициенты.

Пусть $\mathcal{L}_A \subset \mathcal{L}$ – множество таких операторов из $\mathcal{L}$, что $h^{-1} \sum \alpha_k f(kh)$ аппроксимирует первую производную в $x = 0$ с порядком $2m - 1$ (очевидно, $\mathcal{L}_A$ не является линейным пространством, так как ему не принадлежит нулевой элемент). Условие принадлежности к $\mathcal{L}_A$ можно представить в виде системы линейных уравнений:

$$\sum_{k=-m}^{m} \alpha_k k^p = \begin{cases} 1, & p = 1; \\ 0, & otherwise \end{cases} \tag{5.5}$$

для всех $p = 0, \ldots, 2m - 1$. В этой системе $2m + 1$ неизвестное и $2m$ уравнений. Строки этой системы линейно независимы, поскольку их элементы совпадают с элементами первых $2m$ строк матрицы Вандермонда порядка $2m + 1$. Следовательно, фундаментальная система решений этой системы состоит из одного вектора.

Введём на $\mathbb{R}^{\mathbb{Z}}$ оператор C равенством $(Cf)_j = f_{j+1} - 2f_j + f_{j-1}$. Очевидно, что C^m (т. е. m-я степень этого оператора) действует как

$$(C^m f)_j = \sum_{k=-m}^{m} c_k^{(m)} f_{j+k},$$

то есть $C^m \in \mathcal{L}$. Покажем, что набор коэффициентов $c_k^{(m)}$ является решением однородной системы, полученной из (5.5) занулением правой части. Действительно, пусть $p \in \mathbb{N} \cup \{0\}$. Рассмотрим последовательность чисел $f_j = j^p$, $j \in \mathbb{Z}$. Если $p = 0$ или $p = 1$, то $(Cf)_j = 0$ для всех j; если же $p \geqslant 2$, то

$$(Cf)_j = (j+1)^p + (j-1)^p - 2j^p.$$

Используя бином Ньютона, легко увидеть, что $(Cf)_j$ является многочленом от j порядка не выше $p - 2$. Поэтому для любого $p = 0, \ldots, 2m-1$ последовательность $f = \{f_j = j^p\}$ удовлетворяет $C^m f = 0$. В частности, это равенство выполняется для нулевого элемента последовательности, что и означает, что набор коэффициентов $c_k^{(m)}$ удовлетворяет однородной системе.

Мы знаем два частных решения этой системы (5.5): набор коэффициентов a_k (определённых в условии теоремы), дополненный $a_m = 0$, и набор b_k, определённый в условии леммы 5.1. Также мы знаем, что фундаментальная система решений этой системы состоит из одного набора $c_k^{(m)}$. Следовательно, для всех $k = -m, \ldots, m$ справедливо равенство

$$a_k = b_k + \beta c_k^{(m)},$$

где β – некоторый коэффициент, не зависящий от k. Удобно определить его при $k = m$, поскольку по построению $a_m = 0$, а $c_m^{(m)} = 1$. Таким образом, $\beta = -b_m$.

Чтобы определить знак b_m, вспомним, что эта аппроксимация получается дифференцированием интерполяционного полинома

$$I(x) = u(x_{j+m}) \frac{(x - x_{j-m}) \ldots (x - x_{j+m-1})}{(x_{j+m} - x_{j-m}) \ldots (x_{j+m} - x_{j+m-1})} + \ldots.$$

Нас интересует производная порядка от этого многочлена в точке $x = x_j$, а, точнее, коэффициент при $u(x_{j+m})$ в выражении для этой производной. Непосредственно вычисляя производную, получаем

$$\frac{dI}{dx}(x_j) = \frac{1}{h} b_m u(x_{j+m}) + \ldots, \quad b_m = \frac{\prod\limits_{k=-m,\ldots,m-1;k\neq 0} (-k)}{\prod\limits_{k=-m}^{m-1} (m-k)}.$$

Числитель в выражении для b_m имеет знак $(-1)^{m+1}$, а знаменатель положительный. Таким образом, $(-1)^{m+1}b_m \geqslant 0$, а, следовательно, $\beta(-1)^m \geqslant 0$.

Пусть теперь C_h – ограничение C на пространство N_x-периодических последовательностей. В стандартном базисе в этом пространстве C_h является циркулянтом. Пусть A_h и B_h – циркулянты с коэффициентами a_k и b_k, соответственно. Тогда

$$A_h = B_h + \beta C_h^m.$$

В силу $b_{-k} = -b_k$ имеем $B_h = -B_h^T$, также имеем $C_h = C_h^T$. Матрица C_h является отрицательно полуопределённой, поскольку её собственные значения равны

$$\lambda_k = \lambda\left(2\pi\frac{k}{N_x}\right), \quad \lambda(\phi) = e^{i\phi} - 2 + e^{-i\phi} = -2(1-\cos\phi) \leqslant 0.$$

Домножим (5.4) слева на hv^T и представим $C_h^m = (-1)^m(-C_h)^m$. Получим

$$\frac{d}{dt}\frac{hv^Tv}{2} + \mu v^T B_h v + \mu(-1)^m\beta v^T(-C_h)^m v = 0.$$

Из $v^TB_hv = 0$, $v^T(-C_h)^m v > 0$, $\mu \geqslant 0$ и $(-1)^m\beta \geqslant 0$ следует

$$\frac{d}{dt}\frac{\|v(t)\|^2}{2} = \frac{d}{dt}\frac{hv^Tv}{2} \leqslant 0,$$

откуда искомая оценка очевидна. □

Заметим, что дифференциальное приближение для построенного нами оператора имеет вид

$$\frac{1}{h}\sum_{k=-m}^{m-1} a_k u(x_{j+k}) \approx \frac{du}{dx}(x_j) + \beta h^{2m-1}\frac{d^{2m}u}{dx^{2m}} + O(h^{2m}).$$

Мы показали, что оператор численного дифференцирования, построенный на “скошенном влево” шаблоне, имеет собственные значения с неотрицательной действительной частью. Аналогично, если бы мы выбрали “скошенный вправо” шаблон, мы получили бы оператор численного дифференцирования, все собственные значения которого имели бы неположительную действительную часть. Основанная на нём схема была бы устойчивой при $\mu < 0$ и неустойчивой при $\mu > 0$.

Упражнение 2. *Пусть N_x чётное и A_h – кососимметрический циркулянт. Пусть его первая строка имеет вид*

$$(a_0, \ldots, a_m, 0, \ldots, 0, a_{-m}, \ldots, a_{-1}),$$

причём

$$\frac{1}{h}\sum_{j=-m}^{m} a_j f(jh)$$

аппроксимирует $f'(0)$. *Доказать, что* A_h^2 *является циркулянтом, симметрическим и отрицательно полуопределённым, причём размерность его ядра больше 1. Доказать, что если его первая строка имеет вид* $(b_0, \ldots, b_m, 0, \ldots, 0, b_{-m}, \ldots, b_{-1})$, *то*

$$\frac{1}{h}\sum_{j=-m}^{m} b_j f(jh)$$

аппроксимирует $f''(0)$.

Аппроксимируя производную производной от интерполяционного многочлена Лагранжа, можно построить схему любого наперёд заданного порядка порядка аппроксимации и на неравномерных сетках. Но такие схемы не используются.

В завершение этого раздела сделаем замечание относительно использования конечно-разностных схем высокого порядка для решения начально-краевых задач. Рассмотрим, например, задачу

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0, \quad t \in (0, t_{\max}), \quad x \in (0,1),$$

с некоторым начальным условием $u(0,x) = u_0(x)$ и граничным условием $u(t,0) = u_b(t)$. При записи полудискретной схемы в граничном узле можно задать условие $v_0(t) = u_b(t)$. Но помимо постановки граничного условия как такового, требуется изменение аппроксимации в узлах $j = 1, \ldots, m$ (или $j = 1, \ldots, m+1$), поскольку иначе мы бы использовали значения в несуществующих узлах с отрицательными индексами. Универсального способа записи схемы в этой приграничной зоне не существует; это является общей проблемой всех численных методов, использующих широкий шаблон.

5.2. Конечно-объёмные схемы высокого порядка. На неравномерных сетках достаточно хорошо распространены конечно-объёмные схемы высокого порядка. В них значения сеточной функции интерпретируются как интегральные средние по сеточным ячейкам, и интерполяционный многочлен строится по этим интегральным средним.

Рассмотрим неравномерную сетку с узлами x_j, $j = 0, \ldots, N_x$; доопределим $x_{j+kN_x} = x_j + 2\pi k$, $j, k \in \mathbb{Z}$. Мы будем рассматривать дуальную сетку,

то есть под сеточной ячейкой мы будем понимать интервалы $(x_{j-1/2}, x_{j+1/2})$, где $x_{j\pm1/2} = (x_j + x_{j\pm1})/2$. Длину этого интервала обозначим через $\hbar_j = x_{j+1/2} - x_{j-1/2}$. Конечно-объёмные схемы на исходной сетке строятся аналогично, тогда значения сеточных функций будут относиться к ячейкам (x_j, x_{j+1}).

Под сеточной функцией будем понимать набор значений, по одному на ячейку (интервал) $(x_{j-1/2}, x_{j+1/2})$. Интегрируемой функции f будем сопоставлять сеточную функцию

$$\Pi_h f = \{f_j, j = 0, \ldots, N_x - 1\}, \quad f_j = \frac{1}{\hbar_j} \int\limits_{x_{j-1/2}}^{x_{j+1/2}} f(x)dx. \tag{5.6}$$

Проинтегрируем (5.1) по интервалу $(x_{j-1/2}, x_{j+1/2})$ и разделим на его длину:

$$\frac{d}{dt}\frac{1}{\hbar_j} \int\limits_{x_{j-1/2}}^{x_{j+1/2}} u(t,x)dx + \frac{1}{\hbar_j}\left(\mu u(t, x_{j+1/2}) - \mu u(t, x_{j-1/2})\right) = 0. \tag{5.7}$$

Под производной по времени стоит именно то выражение, которое мы сопоставляем функции $u(t, \cdot)$ на интервале $(x_{j-1/2}, x_{j+1/2})$. В то же время, значения в точках $x_{j-1/2}$ и $x_{j+1/2}$ нам не известны. Запишем полудискретную схему

$$\frac{dv_j(t)}{dt} + \frac{1}{\hbar_j}(F_{j+1/2}(v(t)) - F_{j-1/2}(v(t))) = 0, \tag{5.8}$$

$$v_j(0) = \Pi_h u_0, \tag{5.9}$$

где $F_{j+1/2}(v)$ – некоторые функционалы от сеточной функции v. Отметим, что $F_{j+1/2}(v(t))$ входит в два уравнения: для $v_j(t)$ (в качестве $F_{j+1/2}$) и для $v_{j+1}(t)$ (в качестве $F_{(j+1)-1/2}$). Если эти два значения совпадают и зависят только от $v_{j-S}, \ldots, v_{j+S}$ (число $S \in \mathbb{N} \cup \{0\}$ не зависит от сетки и решения), то такая схема называется *локально консервативной* или *дивергентной*. Формулу (5.8) иногда называют *общим видом конечно-объёмной схемы*, а выражения $F_{j+1/2}$ – *численными потоками*.

Нужно уточнить, что локальная консервативность схемы – это возможность её представить в указанном виде. Локальная консервативность – это свойство самой схемы, а не её конкретного представления. Например, схема (5.3) является локально консервативной, поскольку представима в виде (5.8) с потоками $F_{j+1/2}(v) = (2v_{j+1} + 5v_j - v_{j-1})/6$.

Конкретная схема определяется выбором $F_{j+1/2}(v)$. Например, если $\mu \geqslant 0$ и $F_{j+1/2}(v) = v_j$, то при использовании для интегрирования по времени явного

метода Эйлера получается схема

$$\frac{v_j^{n+1} - v_j^n}{\tau} + \mu \frac{v_j^n - v_{j-1}^n}{\hbar_j} = 0, \quad v_j^0 = u(0, x_j), \tag{5.10}$$

для которой мы доказали сходимость с первым порядком при $\tau \leqslant h_{\min}/\mu$ в смысле точечных значений.

Упражнение 3. *Доказать сходимость схемы* (5.8)–(5.9) *с численными потоками* $F_{j+1/2}(v) = v_j$ *в смысле интегральных средних, то есть получить оценку на* $\varepsilon(t) = v(t) - \Pi_h u(t, \cdot)$, *где* Π_h *определён* (5.6).

Для дальнейшего изложения нам понадобится понятие точности на многочленах. Поскольку многочлены не являются периодическими функциями, мы не можем говорить о решении по схеме с начальными данными в виде многочлена. Однако ничто не мешает подставить многочлены и проверить, удовлетворяют ли они каждому из определяющих схему уравнений в отдельности. Поэтому введём следующее определение. Будем говорить, что система равенств

$$\frac{dv_j}{dt} + \sum_{k=-S}^{S} a_{jk} v_{j+k} = 0, \quad j \in \mathbb{Z},$$

точна на многочленах порядка p в смысле точечного отображения, если для любого многочлена $P(x)$ порядка p подстановка в неё $v_j(t) = P(x_j - \mu t)$ даёт точное равенство. Аналогично, будем говорить, что система точна на многочленах порядка p в смысле интегрального отображения, если для любого многочлена $P(x)$ порядка p подстановка в неё

$$v_j(t) = \frac{1}{\hbar_j} \int\limits_{x_{j-1/2}}^{x_{j+1/2}} P(x - \mu t) dx \tag{5.11}$$

даёт точное равенство. Другими словами, говоря про точность на многочленах, мы “забываем”, что сеточная функция должна быть N_x–периодической и рассматриваем её просто как бесконечную в обе стороны числовую последовательность.

Покажем теперь, как построить схему наперёд заданного порядка аппроксимации. Зададимся числом $m \in \mathbb{N} \cup \{0\}$. Рассмотрим следующую задачу: найти многочлен $p_j(x)$ порядка $2m$, такой что для всех $k = -m, \ldots, m$ выполняется

$$\frac{1}{\hbar_{j+k}} \int\limits_{x_{j+k-1/2}}^{x_{j+k+1/2}} p_j(x) dx = v_{j+k}. \tag{5.12}$$

Напомним, что мы допускаем равенство нулю любого из коэффициентов многочлена, то есть многочленом порядка $2m$ мы называем любой многочлен порядка не выше $2m$.

Система (5.12), $k = -m, \ldots, m$, очевидно, является линейной относительно коэффициентов многочлена; число уравнений в ней равно числу неизвестных.

Лемма 5.3. *Система* (5.12)*, $k = -m, \ldots, m$, как система линейных алгебраических уравнений относительно коэффициентов многочлена имеет единственное решение.*

Доказательство. Поскольку число уравнений в системе совпадает с числом неизвестных, достаточно показать разрешимость системы для любой правой части v_{j+k}.

Пусть $P_j(x)$ – многочлен порядка $2m+1$, равный нулю в $x_{j-m-1/2}$ и удовлетворяющий равенствам

$$P_j(x_{j+k+1/2}) - P_j(x_{j+k-1/2}) = v_{j+k}\hbar_{j+k}$$

для всех $k = -m, \ldots, m$. Эти равенства разрешаются последовательно относительно значений P_j в точках с полуцелыми индексами; $P_j(x)$ строится однозначно как интерполяционный многочлен Лагранжа с этими значениями. Остаётся заметить, что $p_j(x) = P_j'(x)$ удовлетворяет системе (5.12). □

Рассмотрим схему (5.8)–(5.9) с потоками

$$F_{j+1/2}(v) = \begin{cases} \mu p_j(x_{j+1/2}), & \mu \geqslant 0; \\ \mu p_{j+1}(x_{j+1/2}), & \mu < 0, \end{cases} \tag{5.13}$$

где p_j и p_{j+1} – найденные многочлены.

Лемма 5.4. *В смысле интегрального отображения система равенств* (5.8) *с потоками* (5.13) *точна на многочленах порядка $2m$ на произвольной сетке и точна на многочленах порядка $2m+1$ на равномерной сетке.*

Доказательство. Положим для простоты $t = 0$. Пусть $P(x)$ – многочлен порядка $2m$. Пусть $v_j(0)$, $j \in \mathbb{Z}$, определены (5.11). При всех j многочлен $p_j(x) = P(x)$ будет удовлетворять всем равенствам системы (5.12) с правой частью $v_{j+k}(0)$. Таким образом, численный поток $F_{j+1/2}(v)$, определённый (5.13), равен $F_{j+1/2}(v) = \mu P(x_{j+1/2})$. Поэтому выражение (5.8) с подстановкой (5.11) превращается в точное равенство (5.7).

Рассмотрим теперь случай равномерной сетки. В силу точности на многочленах порядка $2m$, значение $p_j(x_{j+1/2})$ имеет дифференциальное приближение вида

$$p_j(x_{j+1/2}) = u(t, x_{j+1/2}) + ch^{2m+1}\frac{d^{2m+1}}{dx^{2m+1}}(u(t, x_{j+1/2})) + O(h^{2m+2})$$

Тогда

$$p_{j-1}(x_{j-1/2}) = u(t, x_{j-1/2}) + ch^{2m+1}\frac{d^{2m+1}}{dx^{2m+1}}(u(t, x_{j-1/2})) + O(h^{2m+2})$$

с той же константой c. Если решение локально совпадает с многочленом порядка $2m+1$, то разность этих величин в точности равна $u(t, x_{j+1/2}) - u(t, x_{j-1/2})$, и выражение (5.8) с подстановкой (5.11) превращается в точное равенство (5.7). □

На равномерной сетке точность на многочленах порядка $2m+1$ автоматически влечёт порядок аппроксимации $2m+1$. На неравномерной сетке, если выполняется $1/M \leqslant h_{j+1/2}/h_{j-1/2} \leqslant M$, где M не меняется при измельчении сетки, то схема обладает порядком аппроксимации $2m$.

На равномерной сетке построенную схему можно записать в виде

$$\frac{dv_j}{dt} + \frac{\mu}{h}\sum_{k=-m-1}^{m} a_k v_{j+k} = 0, \quad j = 0, \ldots, N_x - 1; \tag{5.14}$$

$$v_j(0) = \frac{1}{h}\int_{(j-1/2)h}^{(j+1/2)h} u_0(x)dx, \quad j = 0, \ldots, N_x - 1. \tag{5.15}$$

Нам понадобится вспомогательный результат.

Лемма 5.5. *Схема* (5.14)*,* $v_j(0) = u_0(jh)$ *точна на многочленах порядка* p *в смысле точечного отображения тогда и только тогда, когда она точна на многочленах порядка* p *в смысле интегрального отображения.*

Доказательство. По определению, схема (5.14), $v_j(0) = u_0(jh)$ точна на многочленах порядка p в смысле точечного отображения, если для любого многочлена $p(x)$ порядка p для функции $u(t, x) = p(0, x - \mu t)$ выполняется

$$\frac{\partial u}{\partial t}(t, jh) + \frac{\mu}{h}\sum_{k=-m-1}^{m} a_k u(t, x_{j+k}) = 0. \tag{5.16}$$

Аналогично, эта схема точна на многочленах порядка p в смысле интегрального отображения, если для любого многочлена $\tilde{p}(x)$ порядка p для функции $\tilde{u}(t,x) = \tilde{p}(0, x-\mu t)$ выполняется

$$\frac{1}{h}\int\limits_{-h/2}^{h/2} \frac{\partial \tilde{u}}{\partial t}(t, x_j + \xi)d\xi + \frac{\mu}{h}\sum_{k=-m-1}^{m} a_k \frac{1}{h}\int\limits_{-h/2}^{h/2} \tilde{u}(t, x_{j+k} + \xi)dx = 0. \qquad (5.17)$$

Заметим, что формула

$$p(x) = \frac{1}{h}\int\limits_{-h/2}^{h/2} \tilde{p}(x+\xi)d\xi$$

задаёт невырожденное преобразование на пространстве многочленов порядка p. При этом схема точна на многочлене p в смысле точечного отображения тогда и только тогда, когда она точна на многочлене $\tilde{p}$ в смысле интегрального отображения. □

Следствие 5.6. *Коэффициенты a_k в* (5.14) *совпадают с коэффициентами конечно-разностной схемы порядка $2m+1$ на скошенном шаблоне*[2].

Следствие 5.7. *Пусть сетка равномерная. Тогда решение по схеме* (5.8)–(5.9)*,* (5.13) *удовлетворяет неравенству $\|v(t)\|_2 \leqslant \|v(0)\|_2$.*

Аналогичным образом можно доказать устойчивость схемы на основе многочлена порядка $2m+1$, построенного по ячейкам $j-m, \ldots, j+m+1$; такая схема будет бездиссипативной, а $p_j(x_{j+1/2})$ будет равно $p_{j+1}(x_{j+1/2})$.

Про устойчивость конечно-объёмной схемы с полиномиальной реконструкцией переменных на неравномерной сетке нет теоретических результатов (кроме случая полинома нулевого порядка). Численные эксперименты показывают, что использование “центрального” шаблона даёт неустойчивую схему, а при использовании “смещённого” шаблона схема остаётся устойчивой даже при сильной неравномерности сетки. Отсутствие теоретического доказательства устойчивости не означает, что такими схемами нельзя пользоваться. Тем не менее, в некоторых задачах этот недостаток может оказаться существенным.

В заключение отметим, что при программной реализации полиномиальной реконструкции, если представлять многочлен в виде $p_j(x) = \sum_l c_{jl}x^l$ и записать систему уравнений на нахождение c_{jl}, то при $h \ll 1$ или $|x_j| \gg h$ такая

[2]Из этой формулировки видно, что в этом и предыдущем разделах обозначение m использовалось для разных величин, отличающихся друг от друга на единицу.

система будет плохо обусловленной. Удобно представлять многочлен в виде

$$p_j(x) = \sum_{l=0}^{2m} c_{jl} \left(\frac{x - x_j}{h} \right)^l .$$

Тогда число обусловленности системы не будет зависеть от h и x_j, и при умеренных значениях m (скажем, при использовании многочленов до 10-го порядка) такой подход можно использовать.

При решении начально-краевых задач конечно-объёмные схемы высокого порядка страдают от той же проблемы, которая была описана выше для конечно-разностных схем.

5.3. Стандартный метод Галёркина. Рассмотрим два подхода к построению схем высокого порядка, решение которых на произвольных неравномерных сетках в *некоторой* L_2-норме удовлетворяет неравенству $\|u(t)\| \leqslant \|u(0)\|$. Начнём со стандартного (непрерывного) метода Галёркина. Стандартный метод Галёркина наиболее часто применяется к решению эллиптических задач на неравномерных и неструктурированных сетках; ему посвящено большое количество учебной и научной литературы. Мы же рассмотрим его применение к уравнению переноса; при этом мы не будем рассматривать метод Галёркина в общем виде, а ограничимся одним частным случаем.

Пусть $p \in \mathbb{N}$ – параметр метода. Рассмотрим неравномерную сетку $\{x_j, j \in \mathbb{Z}\}$, $x_{j+N_x} = x_j + 2\pi$. Обозначим через S_h пространство 2π-периодических непрерывных действительнозначных функций, являющихся многочленами порядка p на каждом отрезке $[x_j, x_{j+1}]$. Это пространство имеет размерность $N_x(p-1)$.

Первой нашей задачей является задание базиса в пространстве S_h, чтобы иметь возможность представления функции из S_h набором коэффициентов разложения по этому базису. Поступим следующим образом. Пусть $\{\xi_m, m = 0, \ldots, p\}$ – некоторое разбиение отрезка [0,1], например, $\xi_m = m/p$. Определим следующее разбиение отрезка $[0,2\pi]$:

$$0, \xi_1 h_{1/2}, \ldots, \xi_{m-1} h_{1/2}, x_1, x_1 + \xi_1 h_{3/2}, \ldots, x_{N_x-1} + h_{N_x-1/2} \xi_{m-1}, 2\pi.$$

Пример такого разбиения для $N_x = 3$ и $p = 2$ изображён на рис. 1. Закрашенными кругами отмечены узлы сетки, незакрашенными – дополнительные точки.

Ввиду периодических условий условимся считать $x = 0$ и $x = 2\pi$ одной точкой разбиения. Тогда число точек в заданном разбиении равно размерности пространства S_h. Базис на S_h мы будем строить из следующих соображений.

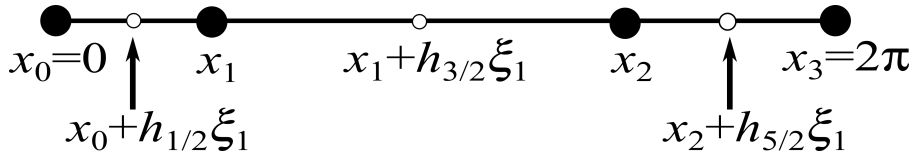

Рис. 1. Пример разбиения отрезка $[0,2\pi]$

Во-первых, каждая базисная функция должна быть равна единице в одной точке разбиения и нулю в остальных точках. Во-вторых, пересечение носителя[3] базисной функции с $[0,2\pi]$ должно состоять из не более чем двух сеточных отрезков. Реализацией этих условий является базисные функции следующего вида.

Пусть $p_k(x)$, $j = 0, \ldots, p$, – многочлены порядка p, такие что

$$p_k(\xi_m) = \delta_{km},$$

где δ_{km} – символ Кронекера ($\delta_{kk} = 1$; $\delta_{km} = 0$ при $k \neq m$). Теперь определим функции $\tilde{\phi}_{k,m}$, $k \in \mathbb{Z}$, $m = 0, \ldots, p-1$, следующим образом. При $m = 0$

$$\tilde{\phi}_{k,0}(x) = \begin{cases} p_0((x - x_k)/h_{k+1/2}), & x_k \leqslant x < x_{k+1}; \\ p_p((x - x_{k-1})/h_{k-1/2}), & x_{k-1} \leqslant x < x_k; \\ 0, & otherwise. \end{cases} \tag{5.18}$$

При $m = 1, \ldots, p-1$

$$\tilde{\phi}_{k,m}(x) = \begin{cases} p_m((x - x_k)/h_{k+1/2}), & x_k \leqslant x < x_{k+1}; \\ 0, & otherwise. \end{cases} \tag{5.19}$$

Наконец, продолжим[4] построенные функции периодическим образом:

$$\phi_{k,m}(x) = \sum_{l \in \mathbb{Z}} \tilde{\phi}_{k+lN_x,m}(x), \quad k = 0, \ldots, N_x - 1, \quad m = 0, \ldots, p-1. \tag{5.20}$$

Базисные функции для $p = 1$ и $p = 2$ изображены на рис. 2.

Лемма 5.8. *Функции $\phi_{k,m}(x)$, определённые* (5.20)*, образуют базис в пространстве S_h.*

Доказательство. Принадлежность $\phi_{k,m}(x)$ к S_h очевидна. Число функций совпадает с размерностью пространства, а их линейная независимость следует из того, что каждая из них равна единице ровно в одной точке использованного разбиения. □

[3] Напомним, что носителем функции называется замыкание множества точек, на которых она не равна нулю.

[4] 2π-периодическое продолжение функции лишено смысла в случае $N_x = 1$, поскольку носитель функции $\tilde{\phi}_{k,0}$ имеет длину 4π. Определение базисных функций через сумму в этом случае остаётся корректной.

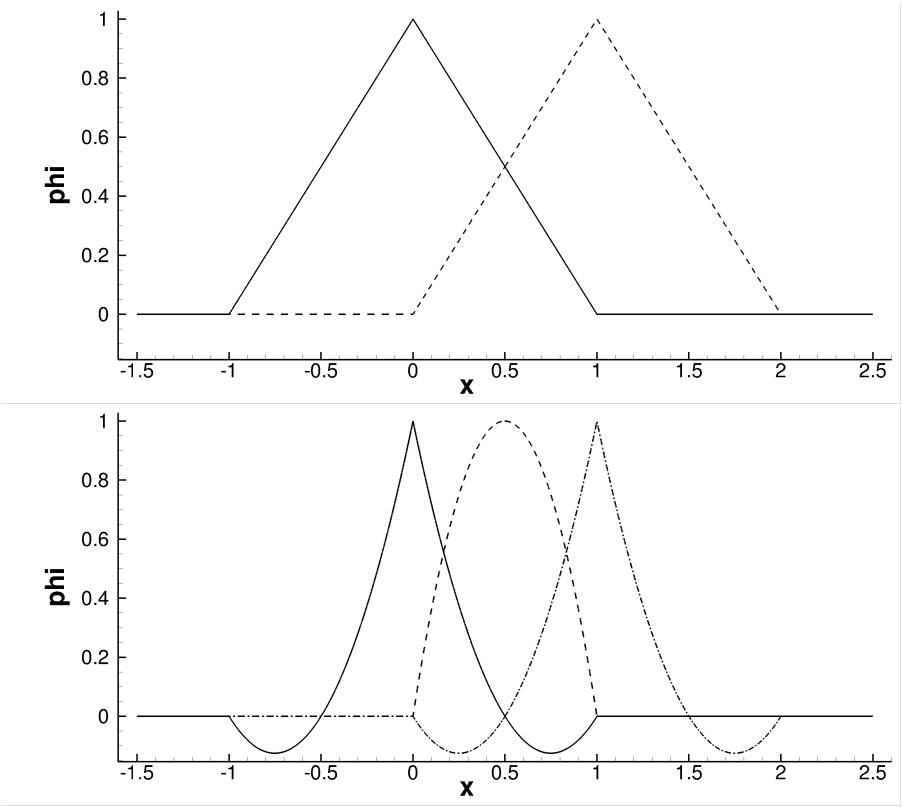

Рис. 2. Примеры базисных функций для стандартного метода Галёркина. Верхний рисунок: $p = 1$. Нижний рисунок: $p = 2$

Введём отображение Π_h непрерывных функций на S_h равенством[5]

$$\Pi_h f = \sum_{k=0}^{N_x-1} \sum_{m=0}^{p-1} f(x_k + h_{k+1/2}\xi_m)\phi_{k,m}(x).$$

Лемма 5.9. *Для любой $p+1$ раз непрерывно дифференцируемой функции $f(x)$ выполняется*

$$\sup_{x\in\mathbb{R}} |f(x) - (\Pi_h f)(x)| \leqslant C \sup_{x\in\mathbb{R}} \left|\frac{d^{p+1} f}{dx^{p+1}}\right| h_{\max}^{p+1}, \tag{5.21}$$

$$\sup_{x\in\mathbb{R}, x\neq x_k} \left|\frac{df(x)}{dx} - \frac{d(\Pi_h f)(x)}{dx}\right| \leqslant C \sup_{x\in\mathbb{R}} \left|\frac{d^{p+1} f}{dx^{p+1}}\right| h_{\max}^{p}, \tag{5.22}$$

где C зависит только от выбора точек ξ_m.

Доказательство. Поскольку на каждом сеточном отрезке функция $(\Pi_h f)(x)$ является лагранжевым интерполянтом, утверждение леммы следует из соответствующих свойств лагранжевых интерполянтов. □

[5] В методе Галёркина чаще используется отображение $\Pi_h f = \arg\min_{g\in S_h} \|f - g\|_2$. Но для него непонятно, выполняется ли (5.22), и требуются дополнительные технические шаги, чтобы получить оценку ошибки.

Здесь мы для удобства нумеровали базисные функции двумя индексами k и m; легко их перенумеровать единым индексом, меняющимся от 1 до $N = pN_x$.

Будем искать приближённое решение задачи Коши в виде линейной комбинации базисных функций с коэффициентами, зависящими от времени:

$$v(t,x) = \sum_{k=1}^{N} v_k(t)\phi_k(x). \tag{5.23}$$

Определим ошибку этого решения как

$$\varepsilon(t,x) = v(t,x) - u(t,x). \tag{5.24}$$

Формально подставим в уравнение

$$\frac{\partial u}{\partial t} + \mu\frac{\partial u}{\partial x} = 0$$

функцию $u(t,x) = v(t,x) - \varepsilon(t,x)$, тогда

$$\sum_{k=1}^{N} \frac{dv_k(t)}{dt}\phi_k(x) + \mu\sum_{k=1}^{N} v_k(t)\frac{d\phi_k(x)}{dx} = \epsilon(t,x), \tag{5.25}$$

где

$$\epsilon(t,x) = \frac{\partial\varepsilon}{\partial t} + \mu\frac{\partial\varepsilon}{\partial x}. \tag{5.26}$$

Можно выбрать некоторый набор точек y_j, в которых $\epsilon(t,y_j)$ определено, и записать систему уравнений $\epsilon(t,y_j) = 0$. Такой метод, называемый *методом коллокаций*, не пользуется популярностью. Значительно шире распространён *метод Галёркина*:

$$\int_0^{2\pi} \epsilon(t,x)\phi_j(x)dx = 0, \quad j = 1,\ldots,N. \tag{5.27}$$

Говоря словами, метод Галёркина заключается в условии ортогональнальности невязки базисным функциям в смысле скалярного произведения в L_2. Подставляя выражение (5.25) для невязки в (5.27), получаем

$$\sum_{k=1}^{N} \frac{dv_k(t)}{dt}\int_0^{2\pi} \phi_j(x)\phi_k(x)dx + \mu\sum_{k=1}^{N} v_k(t)\int_0^{2\pi} \frac{d\phi_k(x)}{dx}\phi_j(x)dx = 0. \tag{5.28}$$

Дополним систему уравнений (5.28) начальными условиями

$$v(0) = \Pi_h u(0, \cdot). \tag{5.29}$$

Полудискретная схема (5.28)–(5.29), как и подход к её построению, называется *стандартным методом Галёркина*.

Систему (5.28) можно переписать в матричной форме

$$M_h \frac{dv}{dt} + A_h v = 0, \tag{5.30}$$

где $M_h = \{m_{jk}\}$, $A_h = \{a_{jk}\}$,

$$m_{jk} = \int_0^{2\pi} \phi_j(x)\phi_k(x)dx, \quad a_{jk} = \int_0^{2\pi} \frac{d\phi_k(x)}{dx}\phi_j(x)dx.$$

Очевидно, что матрица M_h симметрическая. Легко также увидеть, что она положительно определённая. Действительно, для любой сеточной функции v, имеем

$$\begin{gathered} v^T M v = \sum_{j=1}^{N}\sum_{k=1}^{N} v_j v_k \int_0^{2\pi} \phi_j(x)\phi_k(x)dx = \\ = \int_0^{2\pi} \left(\sum_{j=1}^{N} v_j\phi_j(x)\right)\left(\sum_{k=1}^{N} v_k\phi_k(x)\right)dx = \int_0^{2\pi} \left(\sum_{j=1}^{N} v_j\phi_j(x)\right)^2 dx. \end{gathered} \tag{5.31}$$

Поскольку базисные функции линейно независимы, подынтегральное выражение равно нулю тогда и только тогда, когда $v_j = 0$ для всех j. Отсюда видно, что $v^T M v > 0$ при $v \neq 0$, то есть матрица M является положительно определённой.

Матрица A_h является кососимметрической. Действительно, интегрируя по частям и пользуясь 2π-периодичностью базисных функций, получаем

$$a_{kj} = \int_0^{2\pi} \frac{d\phi_j(x)}{dx}\phi_k(x)dx = \phi_j(x)\phi_k(x)\Big|_0^{2\pi} - \int_0^{2\pi} \phi_j(x)\frac{d\phi_k(x)}{dx}dx = -a_{jk}.$$

Можно показать, что число обусловленности матрицы M зависит только от порядка многочленов p и отношения $h_{\max}/h_{\min}$. Это вытекает из следующей леммы.

Лемма 5.10. *Пусть* $\{\phi_j\}$ *– базисные функции, введённые* (5.18)–(5.20). *Тогда существует такая константа* $c > 0$, *что для любой сетки выполняется* $M - ch_{\min} I \geqslant 0$ *и* $ch_{\max} I - M \geqslant 0$, *где* I *– единичная матрица.*

Доказательство. Докажем первое неравенство, второе доказывается аналогично. Рассмотрим пространство многочленов порядка p (это конечномерное пространство размерности $p+1$). Введём на нём две нормы:

$$\|f\|_a = \left(\sum_{m=0}^{p} |f(\xi_m)|^2 \right)^{1/2},$$

$$\|f\|_2 = \left(\int_0^1 |f(x)|^2 dx \right)^{1/2}.$$

Поскольку любые две нормы на конечномерном пространстве являются эквивалентными, существует такая константа $c > 0$, что $\|f\|_2 \geqslant \sqrt{c}\|f\|_a$. Продолжим цепочку равенств (5.31).

$$v^T M v = \sum_{k=0}^{N_x-1} \int_{x_k}^{x_{k+1}} \left(\sum_{j=1}^{N} v_j \phi_j(x) \right)^2 dx =$$

$$= \sum_{k=0}^{N_x-1} h_{k+1/2} \int_0^1 \left(\sum_{j=1}^{N} v_j \phi_j(x_k + \xi h_{k+1/2}) \right)^2 d\xi.$$

Подынтегральная функция является многочленом порядка p, поэтому

$$v^T M v \geqslant c \sum_{k=0}^{N_x-1} h_{k+1/2} \sum_{m=0}^{p} \left(\sum_{j=1}^{N} v_j \phi_j(x_k + \xi_m h_{k+1/2}) \right)^2 dx \geqslant$$

$$\geqslant ch_{\min} \sum_{k=0}^{N_x-1} \sum_{m=0}^{p-1} \left(\sum_{j=1}^{N} v_j \phi_j(x_k + \xi_m h_{k+1/2}) \right)^2 dx$$

(заметим, что мы заменили p на $p-1$). Выражение $\phi_j(x_k + \xi_m h_{k+1/2})$ равно единице, если номер базисной функции j соответствует числам k и m; в противном случае оно равно нулю. Поэтому во всю тройную сумму каждое значение v_j^2 входит ровно по одному разу, и других слагаемых нет. Таким образом, получаем

$$v^T M v \geqslant ch_{\min} \sum_{j=1}^{N} v_j^2 = ch_{\min} v^T v,$$

то есть $M - ch_{\min} I \geqslant 0$, что и требовалось доказать. □

Покажем, что стандартный метод Галёркина устойчив по начальным данным. Введём норму $\| \cdot \|_2$ равенством

$$\|f\|_2^2 = f^T M f = \int_0^{2\pi} \left(\sum_{j=1}^{N} f_j \phi_j(x) \right)^2 dx.$$

Если сеточную функцию понимать не как совокупность коэффициентов v_j, а как непрерывную кусочно-полиномиальную функцию, то норма $\| \cdot \|_2$ – это обычная L_2-норма.

Лемма 5.11. *Любое решение* (5.30) *удовлетворяет равенству*

$$\|v(t)\|_2 = \|v(0)\|_2. \tag{5.32}$$

Доказательство. Домножим уравнение (5.30) слева на v^T. Получаем

$$v^T M \frac{dv}{dt} = 0.$$

Отсюда следует

$$\frac{d}{dt}(v^T M v) = \frac{dv^T}{dt} M v + v^T M \frac{dv}{dt} = \left(\frac{dv^T}{dt} M v\right)^T + v^T M \frac{dv}{dt} = 0,$$

и остаётся воспользоваться определением нормы $\| \cdot \|_2$. □

Построенный метод является устойчивым и точным на многочленах порядка p. Поэтому можно ожидать, что решение будет сходиться к точному с порядком p. Можно было бы доказать этот факт при помощи подхода Лакса – Рябенького (и использовать леммы 5.10 и 5.11). Но мы докажем его при помощи конечно-элементной техники, позволяющей получить более изящное доказательство.

Теорема 5.12. *Пусть $u(t,x)$ – решение задачи Коши* (5.1)–(5.2). *Пусть $\{\phi_j\}$ – базисные функции, введённые* (5.18)–(5.20). *Пусть $v_j(t)$ – решение по полудискретной схеме* (5.28)–(5.29). *Тогда справедливы оценки*

$$\|v(t) - \Pi_h u(t, \cdot)\|_2 \leqslant C t h_{\max}^p, \tag{5.33}$$

$$\left(\int_0^{2\pi} \left(u(t,x) - \sum_{j=1}^{N} v_j(t)\phi_j(x) \right)^2 dx \right)^{1/2} \leqslant C(t + h_{\max}) h_{\max}^p, \tag{5.34}$$

где C не зависит от расчётной сетки.

Доказательство. Введём на пространстве непрерывных 2π-периодических функций скалярное произведение

$$(f,g) = \int_0^{2\pi} f(x) g(x) dx.$$

В том числе, оно определено для сеточных функций; если сеточную функцию понимать как совокупность значений, то $(f, g) = f^T M g$.

Будем пользоваться обозначениями $v(t, x)$ и $\varepsilon(t, x)$, определёнными (5.23) и (5.24). Из (5.26) и (5.27) для всех j и t получаем

$$\left(\frac{\partial \varepsilon}{\partial t}, \phi_j\right) + \mu \left(\frac{\partial \varepsilon}{\partial x}, \phi_j\right) = 0. \tag{5.35}$$

Введём $\tilde{\varepsilon}(t) = v(t) - \Pi_h u(t, \,\cdot\,)$. Домножим (5.35) на $(\tilde{\varepsilon}(t))_j$ и просуммируем по j. Получаем

$$\left(\frac{\partial \varepsilon}{\partial t}, \tilde{\varepsilon}\right) + \mu \left(\frac{\partial \varepsilon}{\partial x}, \tilde{\varepsilon}\right) = 0.$$

Теперь представим $\varepsilon = \tilde{\varepsilon} - (u - \Pi_h u)$. Тогда

$$\left(\frac{\partial \tilde{\varepsilon}}{\partial t}, \tilde{\varepsilon}\right) + \mu \left(\frac{\partial \tilde{\varepsilon}}{\partial x}, \tilde{\varepsilon}\right) = \left(\frac{\partial u}{\partial t} - \Pi_h \frac{\partial u}{\partial t}, \tilde{\varepsilon}\right) + \mu \left(\frac{\partial u}{\partial x} - \frac{\partial (\Pi_h u)}{\partial x}, \tilde{\varepsilon}\right). \tag{5.36}$$

Первое слагаемое в левой части даёт

$$\left(\frac{\partial \tilde{\varepsilon}}{\partial t}, \tilde{\varepsilon}\right) = \int\limits_0^{2\pi} \tilde{\varepsilon} \frac{\partial \tilde{\varepsilon}}{\partial t} dx = \frac{1}{2} \frac{d}{dt} \int\limits_0^{2\pi} \tilde{\varepsilon}^2 dx = \frac{1}{2} \frac{d}{dt} \|\tilde{\varepsilon}\|_2^2 = \|\tilde{\varepsilon}\|_2 \frac{d\|\tilde{\varepsilon}\|_2}{dt}.$$

Второе слагаемое в левой части (5.36) равно нулю в силу периодичности $\tilde{\varepsilon}$. Пользуясь неравенством Коши – Буняковского и сокращая одну степень $\|\tilde{\varepsilon}\|_2$, получаем

$$\frac{d\|\tilde{\varepsilon}\|_2}{dt} \leqslant \left\| \frac{\partial u}{\partial t} - \Pi_h \frac{\partial u}{\partial t} \right\|_2 + |\mu| \left\| \frac{\partial u}{\partial x} - \frac{\partial \Pi_h u}{\partial x} \right\|_2.$$

Мы получили, что оценка точности стандартного метода Галёркина сводится к оценке точности представления решения. Этот результат является характерным для конечно-элементных методов.

Применяя оценки (5.21) и (5.22), получаем

$$\frac{d\|\tilde{\varepsilon}\|_2}{dt} \leqslant c h_{\max}^p,$$

откуда с учётом $\tilde{\varepsilon}(0) = 0$ следует $\|\tilde{\varepsilon}(t)\|_2 \leqslant c t h_{\max}^p$. Остаётся заметить, что разность между $\varepsilon(t)$ и $\tilde{\varepsilon}(t)$ имеет величину порядка $h_{\max}^{p+1}$, откуда получаем итоговые оценки

$$\|\tilde{\varepsilon}(t)\|_2 \leqslant c t h_{\max}^p, \quad \|\varepsilon(t)\|_2 \leqslant c(h_{\max} + t) h_{\max}^p,$$

расшифровывающиеся как (5.33) и (5.34). □

При $p = 1$ на равномерной сетке стандартный метод Галёркина вырождается в следующую схему:

$$\frac{1}{6}\frac{dv_{j-1}}{dt} + \frac{2}{3}\frac{dv_j}{dt} + \frac{1}{6}\frac{dv_{j+1}}{dt} + \mu\frac{v_{j+1} - v_{i-1}}{2h} = 0, \tag{5.37}$$

$$v_j(0) = u(0, jh). \tag{5.38}$$

Напомним, что $v_{N_x} \equiv v_0$, $v_{-1} \equiv v_{N_x-1}$.

Исследуя эту схему обычным образом, мы можем обнаружить, что она точна на многочленах четвёртого порядка. С учётом устойчивости отсюда следует, что эта схема обладает оценкой $\|\tilde{\varepsilon}(t)\| \leqslant cth^4$. Формула (5.37) носит название *формулы компактного дифференцирования* или *компактной схемы* 4-го порядка. Этот пример показывает, что оценка ошибки решения, которую мы доказали в теореме 5.12, вообще говоря, не является оптимальной.

Достоинством стандартного метода Галёркина является теоретически доказанная устойчивость на неравномерных сетках. Но у него есть ряд недостатков, один из которых заключается в наличии матрицы масс перед производной по времени. Даже при решении системы ОДУ явными методами Рунге – Кутты полностью дискретная схема всё равно получается неявной, то есть требует решения системы алгебраических уравнений. Например, если применить для решения (5.37)–(5.38) явный метод Эйлера, получаем полностью дискретную схему

$$\frac{1}{6}v_{j-1}^{n+1} + \frac{2}{3}v_j^{n+1} + \frac{1}{6}v_{j+1}^{n+1} = \frac{1}{6}v_{j-1}^{n} + \frac{2}{3}v_j^{n} + \frac{1}{6}v_{j+1}^{n} - \tau\mu\frac{v_{j+1} - v_{i-1}}{2h} = 0, \tag{5.39}$$

$$v_j^0 = u(0, jh), \quad j = 0, \ldots, N_x - 1. \tag{5.40}$$

Чтобы по известным значениям v_j^n, $j = 0, \ldots, N_x - 1$, найти значения v_j^{n+1}, нужно решить систему линейных алгебраических уравнений с трёхдиагональной матрицей. В одномерном случае такая система легко решается методом прогонки, а в многомерном случае требует применения итерационного процесса. Хотя на квазиравномерной сетке эта система хорошо обусловлена (лемма 5.10 обобщается на многомерный случай) и, следовательно, итерационный процесс будет достаточно быстро сходиться, это всё равно увеличивает затраты машинного времени.

Упражнение 4. *Существует ли такое $\sigma > 0$, что схема* (5.39)–(5.40) *устойчива при $\tau = \sigma h/|\mu|$?*

5.4. Метод Галёркина с разрывными базисными функциями. Метод Галёркина с разрывными базисными функциями (DG – discontinuous Galerkin method), он же разрывный метод Галёркина, сочетает в себе свойства конечно-объёмных и конечно-элементных методов.

Под сеточной функцией понимается функция, на каждом интервале (x_j, x_{j+1}) являющаяся многочленом порядка $p \in \mathbb{N} \cup \{0\}$. В отличие от стандартного метода Галёркина, условие непрерывности не накладывается. На каждом интервале мы можем выбрать любой базис в пространстве многочленов порядка p и продолжить эти функции нулём вне интервала; объединение таких функций образует базис в пространстве сеточных функций. Будем обозначать базисную функцию символом $\phi_j^m(x)$, где $j = 0, \ldots, N_x - 1$ и $m = 0, \ldots, p$. Как минимум, часть этих функций являются разрывными, что и дало название численному методу.

Как и в случае стандартного метода Галёркина, мы будем искать решение в виде линейной комбинации базисных функций с зависящими от времени коэффициентами:

$$v(t,x) = \sum_{k=0}^{N_x-1} \sum_{m=0}^{p} v_{k,m}(t)\phi_k^m(x). \tag{5.41}$$

Домножим (5.1) на $\phi_j^m(x)$ и проинтегрируем по x по периоду. Поскольку базисная функция равна нулю вне одного отрезка, получаем

$$\int_{x_j}^{x_{j+1}} \phi_j^m(x)\frac{\partial u}{\partial t}(t,x)dx + \mu \int_{x_j}^{x_{j+1}} \phi_j^m(x)\frac{\partial u}{\partial x}(t,x)dx = 0. \tag{5.42}$$

В случае стандартного метода Галёркина мы подставляли в это равенство $v(t,x)$ вместо $u(t,x)$ и рассматривали полученное равенство в качестве полудискретной схемы. Здесь же аналогичный подход напрямую неприменим, поскольку $v(t,x)$ не дифференцируемо.

Проинтегрируем второе слагаемое в (5.42) по частям:

$$\begin{gathered} \frac{d}{dt}\int_{x_j}^{x_{j+1}} \phi_j^m(x)u(t,x)dx + \\ +\mu\phi_j^m(x_{j+1})u(t,x_{j+1}) - \mu\phi_j^m(x_j)u(t,x_j) - \\ -\mu\int_{x_j}^{x_{j+1}} \frac{d\phi_j^m(x)}{dx}u(t,x)dx = 0. \end{gathered} \tag{5.43}$$

Мы по-прежнему не можем определить полудискретную схему подстановкой $v(t,x)$ вместо $u(t,x)$ в (5.43), так как $v(t,x)$ терпит разрывы в сеточных узлах и, следовательно, выражения $v(t,x_{j+1})$ и $v(t,x_j)$ не определены.

Обратим внимание, что при $p=0$ каждая из базисных функций является ненулевой константой внутри своей ячейки и после деления на эту константу выражение (5.43) сводится к

$$\frac{d}{dt}\int\limits_{x_j}^{x_{j+1}} u(t,x)dx + \mu u(t,x_{j+1}) - \mu u(t,x_j) = 0.$$

Это равенство мы уже встречали при рассмотрении конечно-объёмного подхода к построению схем (с тем отличием, что мы интегрировали по интервалам вида $(x_{j-1/2}, x_{j+1/2})$). Конечно-объёмный подход предполагал замену $\mu u(t,x_{j+1})$ на некоторые функции, называемые численными потоками. В разрывном методе Галёркина поступают таким же образом.

Запишем

$$\begin{aligned}&\int\limits_{x_j}^{x_{j+1}} \phi_j^m(x)\frac{\partial v}{\partial t}(t,x)dx + \\ &+\phi_j^m(x_{j+1})F_{j+1}(t) - \phi_j^m(x_j)F_j(t) - \\ &-\mu\int\limits_{x_j}^{x_{j+1}} \frac{d\phi_j^m(x)}{dx}v(t,x)dx = 0,\end{aligned} \tag{5.44}$$

где

$$F_j(t) = v(t,x_j-0)\frac{1+\sigma\,\mathrm{sign}\mu}{2} + v(t,x_j+0)\frac{1-\sigma\,\mathrm{sign}\mu}{2}, \tag{5.45}$$

а $0 \leqslant \sigma \leqslant 1$. Символами $v(t,x_j \pm 0)$ обозначены предельные значения слева и справа от точки x_j. Часто полагают $\sigma \equiv 1$, но мы оставим возможность варьировать этот параметр. Выражение для $F_{j+1}(t)$ получается заменой j на $j+1$. В принципе, σ можно выбирать независимо для каждого j. Но при фиксированном j выражение $F_j(t)$ входит в схему $2(p+1)$ раз, и во всех вхождениях σ должно быть одинаковым. Чтобы завершить описание схемы, запишем

$$v(0,\,\cdot\,) = \Pi_h u(0,\,\cdot\,),$$

где Π_h – тот же оператор, который мы использовали при рассмотрении стандартного метода Галёркина.

Систему (5.44) можно переписать в терминах коэффициентов $v_j(x)$ разложения (5.41):

$$\sum_{n=0}^{p} \int_{x_j}^{x_{j+1}} \phi_j^m(x)\phi_j^n(x)dx \frac{v_j^n(t)}{dt} + \\ +\phi_j^m(x_{j+1})F_{j+1}(t) - \phi_j^m(x_j)F_j(t,x) - \\ -\mu \sum_{n=0}^{p} \int_{x_j}^{x_{j+1}} \frac{d\phi_j^m(x)}{dx}\phi_j^n(x)dx\ v_j^n(t) = 0, \tag{5.46}$$

$$F_j(t) = \mu v(t, x_j - 0)\frac{1 + \sigma\,\mathrm{sign}\mu}{2} + \mu v(t, x_j + 0)\frac{1 - \sigma\,\mathrm{sign}\mu}{2},$$

$$v(t, x_j - 0) = \sum_{m=0}^{p} v_{j-1,m}\phi_{j-1}^m(x_j),$$

$$v(t, x_j + 0) = \sum_{m=0}^{p} v_{j,m}\phi_j^m(x_j).$$

Также её можно представить в виде

$$M_h \frac{dv}{dt} + A_h v = 0,$$

где матрица M_h симметрическая и блочно-диагональная (на диагонали стоят блоки размера $p+1$), причём, если во всех ячейках базисные функции выбраны одинаковым образом, то диагональные блоки отличаются друг от друга только множителем $h_{j+1/2}$. Если обратить один такой блок на инициализации, то при явном интегрировании по времени для вычисления значения на $(n+1)$-м слое по времени по значениям на предыдущем (предыдущих) слоях не требуется решать систему уравнений. Последнее свойство выгодно отличает разрывный метод Галёркина от стандартного.

При $p = 0$ метод Галёркина вырождается в схему

$$\frac{dv_j}{dt} + \frac{1}{h_{j+1/2}}\left(F_{j+1}(t) - F_j(t)\right) = 0, \tag{5.47}$$

$$v(0) = \Pi_h u(0, \cdot),$$

где

$$F_j(t) = \mu v_{j-1}\frac{1 + \sigma\,\mathrm{sign}\mu}{2} + \mu v_j \frac{1 - \sigma\,\mathrm{sign}\mu}{2}.$$

Если дополнительно положить $\sigma = 1$, то мы получаем схему с направленными разностями с точностью до переобозначения и выбора оператора Π_h.

Введём норму

$$\|f\|_2 = \left(\int_0^{2\pi} |f(x)|^2 dx \right)^{1/2} \tag{5.48}$$

на пространстве 2π-периодических функций, для которых этот функционал определён. Сеточные функции, очевидно, принадлежат этому пространству. Если под сеточной функцией f понимать набор коэффициентов её разложения по базису, но норма может быть представлена в виде

$$\|f\|_2 = \left(\sum_{k=0}^{N_x-1} \int_{x_k}^{x_{k+1}} \left| \sum_{m=0}^{p} f_{k,m} \phi_k^m(x) \right|^2 dx \right)^{1/2} = (f^T M_h f)^{1/2}.$$

Докажем устойчивость по начальным данным для разрывного метода Галёркина.

Лемма 5.13. *Если $\sigma \geqslant 0$ то для любого решения* (5.46), (5.45) *верно*

$$\|u(t)\|_2 \leqslant \|u(0)\|_2;$$

если $\sigma = 0$, то выполняется $\|u(t)\|_2 = \|u(0)\|_2$.

Доказательство. Домножая (5.44) на $u_{j,m}$ и суммируя по j и m, получаем

$$\begin{gathered} \int_0^{2\pi} v \frac{\partial v}{\partial t} dx + \mu \sum_{j=0}^{N_x-1} \left(v(t, x_{j+1} - 0) F_{j+1}(t) - v(t, x_j + 0) F_j(t) \right) - \\ -\mu \sum_{j=0}^{N_x-1} \int_{x_j}^{x_{j+1}} \frac{\partial v}{\partial x} v dx = 0. \end{gathered} \tag{5.49}$$

Теперь заметим, что

$$\int_{x_j}^{x_{j+1}} \frac{\partial v}{\partial t} v dx = \frac{1}{2} \frac{d}{dt} \int_{x_j}^{x_{j+1}} v^2 dx = \frac{1}{2} \frac{d}{dt} \left(\|v(t)\|_2^2 \right) \tag{5.50}$$

и

$$\int_{x_j}^{x_{j+1}} \frac{\partial v}{\partial x} v dx = \frac{1}{2} \int_{x_j}^{x_{j+1}} \frac{\partial (v^2)}{\partial x} dx = \frac{\left(v(t, x_{j+1} - 0) \right)^2 - \left(v(t, x_j + 0) \right)^2}{2}. \tag{5.51}$$

Подставляя (5.50) и (5.51) в (5.49), получим

$$\begin{gathered}\frac{1}{2}\frac{d}{dt}\left(\|v(t)\|_2^2\right)+\\+\sum_{j=0}^{N_x-1}\left(v(t,x_{j+1}-0)F_{j+1}(t)-v(t,x_j+0)F_j(t)\right)-\\-\mu\sum_{j=0}^{N_x-1}\frac{\left(v(t,x_{j+1}-0)\right)^2-\left(v(t,x_j+0)\right)^2}{2}=0.\end{gathered}\tag{5.52}$$

Теперь разобьём каждую из сумм на две и сдвинем индекс суммирования в одной из них, пользуясь периодичностью. В результате получим

$$\begin{gathered}\frac{1}{2}\frac{d}{dt}\left(\|v(t)\|_2^2\right)+\\+\sum_{j=0}^{N_x-1}\left(v(t,x_j-0)F_j(t)-v(t,x_j+0)F_j(t)\right)-\\-\mu\sum_{j=0}^{N_x-1}\frac{\left(v(t,x_j-0)\right)^2-\left(v(t,x_j+0)\right)^2}{2}=0\end{gathered}\tag{5.53}$$

или, группируя слагаемые,

$$\begin{gathered}\frac{1}{2}\frac{d}{dt}\left(\|v(t)\|_2^2\right)+\\+\sum_{j=0}^{N_x-1}\left(v(t,x_j-0)-v(t,x_j+0)\right)\left(F(t,x_j)-\mu\frac{v(t,x_j-0)+v(t,x_j+0)}{2}\right)=0.\end{gathered}\tag{5.54}$$

Остаётся подставить выражение (5.45) для численного потока $F_j(t)$. Получаем

$$\frac{1}{2}\frac{d}{dt}\left(\|v(t)\|_2^2\right)+\frac{|\mu|\sigma}{2}\sum_{j=0}^{N_x-1}\left(v(t,x_j-0)-v(t,x_j+0)\right)^2=0.\tag{5.55}$$

По условию σ неотрицательное, поэтому второе слагаемое неотрицательное и $d(\|v(t)\|_2^2)/dt \leqslant 0$, значит, $\|v(t)\|_2$ не возрастает. При $\sigma = 0$ имеем $d(\|v(t)\|_2^2)/dt = 0$, поэтому $\|v(t)\|_2$ сохраняется во времени. □

Напомним следующий результат.

Лемма 5.14. *Пусть схема устойчива с константой $K(t)$, где $K(t)$ не убывает с ростом t. Тогда любое решение неоднородной системы*

$$\frac{dv(t)}{dt}+A_hv(t)=y(t)\tag{5.56}$$

удовлетворяет оценке

$$\|v(t)\| \leqslant K(t)\|v(0)\| + K(t)t \sup_{0\leqslant t'\leqslant t} \|y(t)\|.$$

Для метода Галёркина с разрывными базисными функциями справедлива та же оценка ошибки, что и для стандартного метода Галёркина.

Теорема 5.15. *Пусть $u(t,x) - p+1$ раз непрерывно дифференцируемое решение задачи Коши* (5.1)–(5.2)*. Пусть $p \in \mathbb{N}$. Пусть $v(t) = \{v_j(t)\}$ – решение по полудискретной схеме* (5.44), (5.45)*, $v(0) = \Pi_h u(t,\,\cdot\,)$. Тогда справедливы оценки*

$$\|v(t) - \Pi_h u(t,\,\cdot\,)\|_2 \leqslant C t h_{\max}^p, \tag{5.57}$$

$$\left(\int_0^{2\pi} \left(u(t,x) - \sum_{j=1}^{N} v_j(t)\phi_j(x) \right)^2 dx \right)^{1/2} \leqslant C(t + h_{\max}) h_{\max}^p, \tag{5.58}$$

где C не зависит от расчётной сетки.

Доказательство. Заметим, что оператор Π_h выбран таким образом, что для любой непрерывной функции f сеточная функция $\Pi_h f$ является непрерывной, причём $(\Pi_h f)(x_j) = f(x_j)$ для всех j. Продолжим оператор Π_h на пространство ограниченных непрерывных на $\mathbb{R}$ функций естественным образом, тогда образ функции под действием этого оператора будет кусочно-полиномиальной, но, вообще говоря, не периодической функцией.

Представим численное решение в виде $v(t) = \Pi_h u(t,\,\cdot\,) + \tilde{\varepsilon}(t)$. Тогда имеем

$$M_h \frac{d\tilde{\varepsilon}}{dt} + A_h \tilde{\varepsilon} = -\tilde{\epsilon},$$

где

$$\tilde{\epsilon}(t) = M_h \frac{d\Pi_h u(t,\,\cdot\,)}{dt} + A_h \Pi_h u(t,\,\cdot\,).$$

Рассмотрим некоторую пару значений $j = 0, \ldots, N_x - 1$, $m = 0, \ldots, p$ и оценим величину $\tilde{\epsilon}_{j,m}(t)$. Представим $u(t,x)$ в виде

$$u(t,x) = g_{j,m}(t,x) + w_{j,m}(t,x),$$

где $g(t,x)$ при каждом t – многочлен порядка p, полученный разложением $u(t,x)$ в ряд Тейлора около точки x_j. Легко заметить, что $g_{j,m}(t,x) = g_{j,m}(0, x - \mu t)$.

Тогда запишем

$$\tilde{\epsilon}_{j,m}(t) = \left(M_h \frac{d\Pi_h g_{j,m}(t,\,\cdot\,)}{dt} + A_h \Pi_h g_{j,m}(t,\,\cdot\,) \right)_{j,m} + \\ + \left(M_h \frac{d\Pi_h w_{j,m}(t,\,\cdot\,)}{dt} + A_h \Pi_h w_{j,m}(t,\,\cdot\,) \right)_{j,m}. \tag{5.59}$$

Поскольку $g_{j,m}(t,x)$ является многочленом, в силу (5.21) оператор Π_h оставляет её без изменения. Тогда первое слагаемое в правой части (5.59) вырождается в левую часть (5.43) с подстановкой $g_{j,m}$ вместо u и поэтому равно нулю. (Другими словами, оно равно нулю в силу точности Π_h и схемы на многочленах порядка p).

Остаётся оценить второе слагаемое в правой части (5.59). Функция $w_{j,m}(t,x)$ является $p+1$ раз непрерывно дифференцируемой и, по свойству остаточного члена ряда Тейлора,

$$|g_{j,m}(t,x)| \leqslant c(x-x_j)^{p+1} \sup \left| \frac{\partial^{p+1} g_{j,m}}{\partial x^{p+1}} \right|,$$

$$\left| \frac{\partial g_{j,m}(t,x)}{\partial t} \right| = |\mu| \left| \frac{\partial g_{j,m}(t,x)}{dx} \right| \leqslant c(x-x_j)^{p} \sup \left| \frac{\partial^{p+1} g_{j,m}}{\partial x^{p+1}} \right|.$$

По свойствам (5.21) и (5.22) эти же оценки имеют место и для $\Pi_h g_{j,m}$. Тогда, раскрывая выражения для M_h и A_h, получаем оценку

$$|\tilde{\epsilon}_{j,m}(t)| \leqslant c h_{j+1/2}^{p+1} \left| \frac{\partial^{p+1} g_{j,m}}{\partial x^{p+1}} \right| = c h_{j+1/2}^{p+1} \sup \left| \frac{\partial^{p+1} u}{\partial x^{p+1}} \right|.$$

Нам нужна оценка на $\left(M_h^{-1}\tilde{\epsilon}\right)_{j,m}$, но мы знаем, что матрица M_h блочно-диагональная, причём j-й диагональный блок равен некоторой невырожденной матрице M_1 размера $(p+1)\times(p+1)$, умноженной на $h_{j+1/2}$. Поэтому

$$\left| \left(M_h^{-1}\tilde{\epsilon}\right)_{j,m} \right| \leqslant c' h_{j+1/2}^{p} \sup \left| \frac{\partial^{p+1} u}{\partial x^{p+1}} \right|.$$

Складывая теперь оценки ошибки по всем интервалам, получаем оценку $O(h_{\max}^p)$. Для доказательства (5.57) остаётся применить лемму 5.14, а (5.58) следует из (5.57) по неравенству треугольника. □

Оценки, полученные в теореме 5.15, не являются оптимальными. Неравенство (5.57) может быть уточнено до

$$\|v(t) - \Pi_h u(t,\,\cdot\,)\|_2 \leqslant C_1 t h^{2p+1} + C_2 h^{p+1}, \tag{5.60}$$

и правая часть в (5.57) также может быть заменена на $C_1 t h^{2p+1} + C_2 h^{p+1}$. Идея одного из доказательств этого факта заключается в следующем. Существует[6] такое отображение Π_h, удовлетворяющее условиям (5.21), (5.22), что ошибка аппроксимации полудискретной схемы (5.44), (5.45), $v(0) = \Pi_h u(t, \cdot)$ имеет величину $O(h^{2p+1})$ (тогда как для рассмотренного нами оператора Π_h оценка $O(h^p)$ при $p \neq 0$ неулучшаема). Остаётся воспользоваться устойчивостью схемы и неравенством треугольника.

Сравним разрывный метод Галёркина (DG) с конечно-объёмными схемами высокого порядка, основанными на полиномиальной реконструкции (кратко обозначая их как FV). В схемах FV на каждой сеточной ячейке задано одно значение, интерпретируемое как интегральное среднее от искомой функции по ячейке. В схемах DG на каждой ячейке задано полиномиальное распределение. В обеих схемах фигурирует численный поток F_j (или, если записывать схемы на дуальной сетке, то $F_{j+1/2}$), который определяется с учётом направления скорости переноса. В DG для его вычисления используется предельное значение распределения в одной из соседних ячеек. В FV, поскольку внутри ячейки нет полиномиального представления сеточной функции, оно восстанавливается по имеющимся интегральным средним на расширенном шаблоне.

Можно ли сочетать эти два подхода: использовать на ячейке полином некоторого порядка, а для определения численного потока по широкому шаблону строить приближение искомой функции многочленом более высокого порядка? Ответ на этот вопрос положительный. За схемами, сочетюающими эти подходы, закрепилось название PnPm-схем. Однако их популярность уступает как конечно-объёмным схемам, так и разрывному методу Галёркина.

[6] Явным образом вид этого отображения приведён работе Cao W., Zhang Z., Zou Q., Superconvergence of discontinuous Galerkin methods for linear hyperbolic equations. SIAM J. Numer. Anal. 2014. Vol. 52. P. 2555–2573. Этот результат сохраняется и для неравномерной сетки.

6. Как сравнивать численные методы

6.1. Общие соображения. Предположим, что мы имеем дело с корректно поставленной[7] математической задачей, например, для системы дифференциальных уравнений. Нашей целью является нахождение приближённого решения этой задачи при помощи компьютера. У любого способа достижения этой цели есть две характеристики: точность результата и затраты ресурсов. Основным ресурсом является машинное время, т. е. время, требуемое компьютеру для вычисления решения. Более эффективным мы можем считать тот подход, который позволяет добиться лучшей точности при меньших затратах.

Для уравнения переноса с постоянным коэффициентом задача поиска наиболее эффективного способа численного решения бессмысленна, так как его точное решение известно. “Самая лучшая численная схема” для её решения может выглядеть, например, так:

$$u_j^n = u_0(x_j - \mu t_n).$$

Тем не менее, сравнение схем на уравнении переноса не лишено смысла. Во многих сложных физических процессах (например, в газовой динамике) одной из составляющих является конвективной перенос. Качество воспроизведения численным методом конвективного переноса даёт частичную информацию о их точности на сложной задаче.

Затраты машинного времени на решение конкретной задачи зависят от многих факторов. Можно разделить их на четыре группы:

- используемый численный метод;
- используемая расчётная сетка;
- быстродействие компьютера (производительность вычислительных устройств, пропускная способность памяти и т. д);
- качество программного кода (квалификация программиста, выбор компилятора и т. д.).

Последние две группы факторов лежат заведомо вне области, изучаемой теорией численных методов. Поэтому иногда затраты машинного времени заменяют числом арифметических операций с плавающей запятой (FLOP – floating point operation).

Часто поступают ещё проще: вместо числа FLOP рассматривают размерность пространства сеточных функций. Эту размерность называют *числом степеней свободы* на расчётной сетке. При этом отбрасывается два фактора. Первый – число арифметических операций на одну степень свободы. Второй – число шагов по времени. При расчётах по явным схемам шаг по времени обычно

[7] Формальное определение корректности см. в книге: А. Н. Тихонов, В. Я. Арсенин, Методы решения некорректных задач. Говоря кратко, задача называется корректно поставленной, если на пространстве допустимых значений параметров её решение единственно и является непрерывной по Липшицу функцией этих параметров.

ограничивается условием устойчивости, поэтому число шагов по времени обратно пропорционально ограничению на шаг. Для грубой оценки этими двумя факторами можно пренебречь, но для более точного сравнения их полезно учитывать.

Почему берётся именно число степеней свободы, а не количество узлов в расчётной сетке? Рассмотрим в качестве примера схему с направленной разностью на равномерной сетке:

$$\frac{dv_j}{dt} + \frac{\mu}{h}(v_j - v_{j-1}) = 0, \quad j = 0, \ldots, N_x - 1;$$

$$v_j(0) = u(0, jh).$$

Наряду с этой схемой рассмотрим схему, в которой в одной ячейке определено два значения v_j и w_j:

$$\frac{dv_j}{dt} + \frac{2\mu}{h}(v_j - w_{j-1}) = 0, \quad j = 0, \ldots, N_x - 1;$$

$$\frac{dw_j}{dt} + \frac{2\mu}{h}(w_j - v_j) = 0, \quad j = 0, \ldots, N_x - 1;$$

$$v_j(0) = u(0, jh); \quad w_j(0) = u(0, (j + 1/2)h).$$

Применяя этот метод, на одной и той же сетке мы получим примерно вдвое меньшую ошибку, чем по схеме с направленной разностью.

Но нетрудно заметить, что это та же самая схема с направленной разностью, просто под сеточной ячейкой мы стали понимать интервал вдвое большей длины. Число операций на одну ячейку сетки при этом увеличилось вдвое (а если учитывать необходимость вдвое меньшего шага по времени для обеспечения устойчивости – то в четыре раза). Если мы будем сравнивать схемы на фиксированной сетке и игнорировать число операций на одну ячейку, то этот нехитрый приём нам позволит строить всё более и более точные схемы первого порядка аппроксимации. Если же в качестве меры вычислительной стоимости мы будем брать число степеней свободы на ячейке, то этот трюк ничего не даст: при одном и том же числе степеней свободы точность у первой и второй схем будут совпадать.

В разделе 5 мы доказали оценки сходимости для ряда численных методов. У этих оценок есть, как минимум, два недостатка: 1) они могут не быть оптимальными (см. пример в разделе 5.3) и 2) они получены с точностью до константного множителя. Поэтому, наряду с теоретическим анализом численных методов, проводится их экспериментальный анализ, то есть анализ их применения к тем или иным задачам.

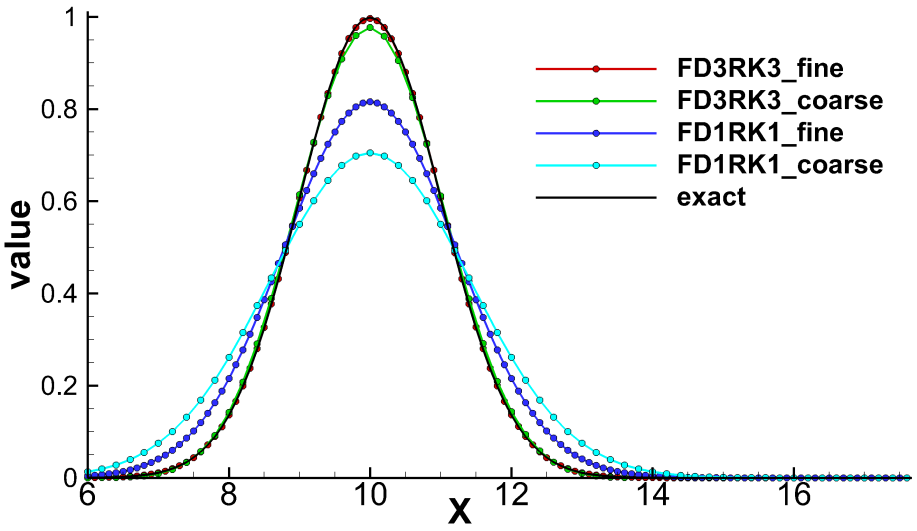

Рис. 3. Численное решение уравнения переноса по схемам 1-го и 3-го порядка

6.2. Визуальное сравнение результатов. Проведём следующий тест. Рассмотрим задачу Коши

$$\frac{\partial u}{\partial t} + \mu \frac{\partial u}{\partial x} = 0, \quad 0 < t < t_{\max}, \quad -50 < x < 50, \tag{6.1}$$

с периодическими граничными условиями и начальными данными

$$u(0, x) = u_0(x) = \exp(-x^2/2), \quad -50 < x < 50. \tag{6.2}$$

Скорость переноса положим равной $\mu = 1$. Периодическое продолжение начальных данных не является гладкой функцией, но скачки производных настолько малы ($\exp(-50^2/2) \approx 10^{-542}$), что не могут сказаться на точности. Рассмотрим сетки с шагом $h = 1/5$ и $h = 1/10$ и проведём расчёт по двум схемам:

- “левый уголок”;
- конечно-разностной схеме 3-го порядка (5.3) с использованием метода Рунге – Кутта 3-го порядка для решения системы ОДУ.

Расчёт будем проводить до достижения времени $t_{\max} = 10$. Результаты расчётов сведены на рис. 3. В расчётах наблюдается сходимость по сетке по обеим схемам; также видно, что решение по схеме 3-го порядка значительно ближе к точному, чем по схеме 1-го порядка.

6.3. Экспериментальное исследование сходимости. Визуальное сопоставление решений, полученных по разным схемам на одной сетке, может быть наглядным. Однако результат расчёта на одной сетке даёт недостаточно информации о точности схемы. Часто приводят величину численной ошибки на последовательности из нескольких (обычно от трёх до пяти) сеток. На рис. 3 мы

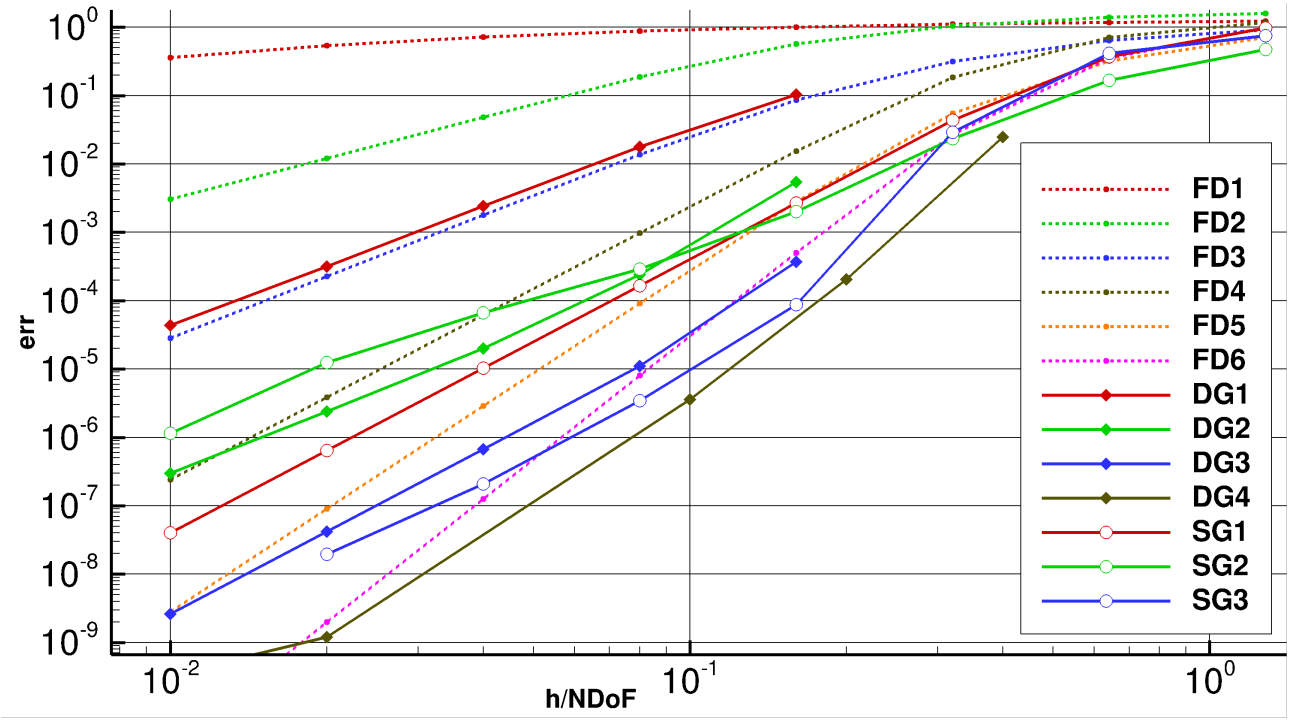

Рис. 4. Ошибка численного решения уравнения переноса для конечно-разностных схем, стандартного и разрывного методов Галёркина

привели результаты расчётов на двух сетках; наложение на него результатов расчётов на ещё более мелких сетках было бы неинформативным, так как все линии, посчитанные по схеме 3-го порядка, накладывались бы друг на друга.

Удобно представлять результаты в виде графика зависимости нормы ошибки решения от шага сетки. Поскольку мы будем сравнивать схемы с разным числом степеней свободы на ячейку, по горизонтальной оси будем откладывать шаг сетки, делённый на число степеней свободы, приходящихся на одну ячейку. Поскольку эта зависимость обычно имеет степенной вид ($\|\varepsilon\| \sim h^k$), этот график рисуют в двойном логарифмическом масштабе: по горизонтальной оси откладывают шаг, а по вертикальной – норму ошибки, тогда тангенс угла наклона будет характеризовать порядок малости величины ошибки при $h \to 0$. Это порядок, посчитанный по паре сеток, часто называют *численным порядком сходимости*. Численный порядок легко оценить графически, подсчитав, на сколько делений график ошибки смещается вниз при $h = 10^{-2}$ в сравнении с $h = 10^{-1}$, см. рис. 4.

Итак, рассмотрим задачу Коши (6.1)–(6.2) с $\mu = 1$. Будем проводить расчёты по конечно-разностной схеме (FD), стандартному методу Галёркина (SG) и методу Галёркина с разрывными базисными функциями (DG). Для DG положим коэффициент диссипации равным $\sigma = 1$, а в качестве Π_h выберем оператор, на каждом интервале (x_j, x_{j+1}) сопоставляющий функции её лагранжев

интерполянт по узлам $x_j + h_{j+1/2}m/p$, $m = 0, \ldots, p$. Расчёт будем проводить до момента времени $t_{\max} = 100$. Результаты сведены на рис. 4.

Для конечно-разностных схем 2-го порядка и выше численный порядок близок к теоретически ожидаемому. В правой части графика все кривые загибаются: поскольку схемы являются устойчивыми с константой устойчивости $K = 1$, норма ошибки не может превзойти удвоенную норму самого решения. Измерять численный порядок по расчётам на таких сетках не имеет большого смысла. Для схемы 1-го порядка весь диапазон использованных сеток является "правой частью графика", поэтому численный порядок на них оказывается существенно меньше единицы.

Поведение метода Галёркина с разрывными базисными функциями может показаться довольно странным. В правой части графика численный порядок предсказуемо близок к нулю, в середине достигает своего максимального значения, а затем уменьшается. Причина такого поведения заключается в следующем. Существует[8] такое отображение Π_h, удовлетворяющее условиям (5.21), (5.22), что ошибка аппроксимации полудискретной схемы (5.44), (5.45), $v(0) = \Pi_h u(t, \cdot)$ имеет величину $O(h^{2p+1})$ (тогда как для рассмотренного нами оператора Π_h оценка $O(h^p)$ при $p \neq 0$ неулучшаема). Пользуясь устойчивостью схемы и неравенством треугольника (мы уже рассматривали это подробно в предыдущей части курса), получаем, что решение удовлетворяет оценке (5.60):

$$\|v(t) - \Pi_h u(t, \cdot)\|_2 \leqslant C_1 t h^{2p+1} + C_2 h^{p+1}.$$

При достаточно большом значении $t = t_{\max}$ и не слишком малых значений h первое слагаемое доминирует, и в этом случае мы можем наблюдать численный порядок точности, близкий к $2p+1$. При дальнейшем измельчении сетки (или при уменьшении $t_{\max}$) начинает доминировать второе слагаемое, и начинает наблюдаться численный порядок $p + 1$. Таким образом, численный порядок может существенно зависеть от выбора $t_{\max}$ и в целом постановки задачи, на которой проводится тестирование схемы. Также отметим, что, например, сравнивая FD4 и DG2, мы видим диапазон h, при котором точность DG2 выше (т. е. численая ошибка меньше), но при этом численный порядок сходимости у DG2 ниже.

Для стандартного метода Галёркина при $p = 1$ наблюдается 4-й порядок сходимости; это объясняется тем, что метод вырождается в компактную схему (5.37)–(5.38). При $p > 1$ полученные результаты дают основание полагать, что оценка, даваемая теоремой 5.12, не является оптимальной (по меньшей мере, на раномерной сетке), однако теоретические результаты, устанавливающие более высокий порядок сходимости, автору не известны.

[8] Явным образом вид этого отображения приведён работе Cao W., Zhang Z., Zou Q., Superconvergence of discontinuous Galerkin methods for linear hyperbolic equations. SIAM J. Numer. Anal. 2014. Vol. 52. P. 2555–2573. Этот результат сохраняется и для неравномерной сетки.

6.4. Сопоставление спектров. Следующий метод предназначен для анализа и сравнения численных схем для линейного уравнения переноса на равномерных сетках. Мы будем обсуждать его только для схем вида

$$\frac{dv_j}{dt} + \frac{\mu}{h}\sum_{k=-S}^{S} a_k v_{j+k} = 0, \quad j = 0, \dots, N_x - 1, \tag{6.3}$$

$$v_j(0) = u_0(jh), \quad j \in \mathbb{Z}, \tag{6.4}$$

где $v(t) = \{v_j(t)\}$ – N_x-периодическая последовательность комплексных чисел. Описываемый метод также может быть применён для схем, имеющих большое одной степени свободы на ячейку. Обобщение этого метода на неравномерные сетки также возможно, но не входит в число общепринятых методов анализа.

Конкретный вид оператора $\Pi_h f = \{f(jh), j \in \mathbb{Z}\}$ выбран для удобства изложения; как видно из леммы 5.5, выбор отображения Π_h для таких схем не играет существенной роли. Также будем использовать для системы (6.3) представление

$$\frac{dv(t)}{dt} + \frac{\mu}{h} A_h v(t) = 0. \tag{6.5}$$

Каждой схеме вида (7.3)–(7.3) можно поставить в соответствие функцию $\lambda(\phi)$, определённую формулой (7.5):

$$\lambda(\phi) = \sum_{q=-S}^{S} a_q \exp(iq\phi). \tag{6.6}$$

Очевидно, что по функции $\lambda(\phi)$ восстанавливается равенство (6.3). Таким образом, можно считать, что схема определяется только функцией $\lambda(\phi)$. Диагональная матрица с элементами $\lambda(2\pi k/N_x)$, $k = 0, \dots, N_x$, является представлением оператора A_h в базисе, составленном из столбцов матрицы S_{N_x}. Рассмотрим подробнее эти базисные функции и решения уравнения переноса (5.1), сеточными образами которых эти базисные функции являются.

Поскольку мы предполагаем, что $u_0(x)$ – 2π-периодическая непрерывно дифференцируемая функция $x \in \mathbb{R}$, она представима в виде ряда Фурье:

$$u_0(x) = \sum_{k\in\mathbb{Z}} c_k \exp(ikx),$$

причём $\sum |c_k| < \infty$. Если $u_0(x)$ – действительнозначная, то $c_{-k} = \overline{c_k}$. Функции вида $\exp(ikx)$, $k \in \mathbb{Z}$ будем называть *гармониками*, число k – *волновым числом*; период гармоники с волновым числом k равен $2\pi/|k|$.

Рассмотрим теперь сеточный образ одной гармоники:

$$\Pi_h \exp(ikx) = \{\exp(ijkh), j = 0, \ldots, N_x - 1\} =$$
$$= \left\{\exp\left(2\pi i \frac{jk}{N_x}\right), j = 0, \ldots, N_x - 1\right\}.$$

Заметим, что при любом $m \in \mathbb{Z}$

$$\Pi_h \exp(i(k + mN_x)x) = \{\exp(ij(k + mN_x)h)\} = \{\exp(ijkh + ijmN_xh)\} =$$
$$= \{\exp(ijkh + 2\pi imj)\} = \{\exp(ijkh)\} = \Pi_h \exp(ijx),$$

то есть гармоники с волновыми числами, отличающимися на mN_x, $m \in \mathbb{Z}$, имеют один и тот же образ под действием Π_h. Имея сеточную функцию с компонентами $f_j = \exp(2\pi ijk/N_x) = \exp(2\pi ij(k + mN_x)/N_x)$, логично интерпретировать её как образ той гармоники, у которой волновое число $k + mN_x$ наименьшее по модулю. Поскольку любая сеточная функция представляется в виде суммы сеточных образов гармоник, мы можем интерпретировать её как образ некоторой функции вида

$$f(x) = \sum_{k=-\lfloor N_x/2 \rfloor}^{\lfloor (N_x-1)/2 \rfloor} c_k \exp(ikx).$$

Здесь $\lfloor \cdot \rfloor$ – округление вниз. Гармоники, входящие в эту сумму, называются *разрешаемыми сеткой*. Максимальный модуль волнового числа, соответствующего разрешаемой сеткой гармонике, равен $\lfloor N_x/2 \rfloor$. Соответствующий ему пространственный период равен

$$L = \frac{2\pi}{\lfloor N_x/2 \rfloor} \approx \frac{2\pi}{N_x/2} = \frac{4\pi}{2\pi/h} = 2h.$$

Говорят, что гармоники, период которых меньше $2h$, *не разрешаются расчётной сеткой*. Их сеточный образ неотличим от образа возмущений, соответствующих меньшим по модулю волновым числам, поэтому вряд ли можно рассчитывать на корректное моделирование численной схемой динамики этих гармоник. Поэтому дальше будем рассматривать только *разрешаемые сеткой гармоники*, то есть волны, период которых больше или равен $2h$.

Пусть $u_0(x) = \exp(ikx)$, $k = \phi/h \in \mathbb{Z}$. Тогда сеточная функция в начальный момент времени равна $v(0) = \Pi_h u_0 = \{\exp(ij\phi), j \in \mathbb{Z}\}$. Величина ϕ характеризует сеточное разрешение: $\phi = \pi$ соответствует гармонике с периодом $2h$, $\phi = \pi/2$ – гармонике с периодом $4h$ и т. д. Тогда, поскольку $v(0)$ является собственным вектором оператора A_h, из (6.5) имеем

$$v(t) = \exp\left(-\mu\lambda_k \frac{t}{h}\right) v(0) = \exp\left(-\mu\lambda(\phi)\frac{t}{h}\right) v(0).$$

где $\lambda_k = \lambda(\phi)$ – соответствующее собственное значение. С другой стороны,

$$\Pi_h u(t,\ \cdot\ ) = \Pi_h \exp(ik(x-\mu t)) = \exp(ik\mu t)\Pi_h \exp(ikx) = \exp\left(i\mu\phi\frac{t}{h}\right) v(0).$$

Значит, ошибка решения равна

$$\varepsilon(t) = v(t) - \Pi_h u(t,\ \cdot\ ) = \left[\exp\left(-\mu\lambda(\phi)\frac{t}{h}\right) - \exp\left(i\mu\phi\frac{t}{h}\right)\right] v(0) =$$
$$= \left[\exp\left(-\mu(\lambda(\phi) - i\phi)\frac{t}{h}\right) - 1\right] \Pi_h u(t,\ \cdot\ ).$$

Напомним, что полудискретная схема (7.3)–(7.4) устойчива тогда и только тогда, когда $\mu \mathrm{Re}\,\lambda(\phi) \geqslant 0$ для всех $\phi \in \mathbb{R}$. Поэтому, если схема устойчива, в показателе экспоненты стоит величина с неположительной действительной частью. Заметим, что при $\mathrm{Re}\, z \leqslant 0$ выполняется

$$|e^z - 1| = \left| z \int_0^1 e^{\alpha z} d\alpha \right| \leqslant |z| \int_0^1 |e^{\alpha z}|\, d\alpha \leqslant |z|,$$

поэтому

$$\|\varepsilon(t)\| \leqslant |\lambda(\phi) - i\phi|\ |\mu| h^{-1} t\ \|\Pi_h u(t,\ \cdot\ )\|$$

с подстановкой $\phi = kh$. Видно, что функцию $\lambda(\phi) - i\phi$ мы можем рассматривать в качестве меры ошибки численной схемы. Если бы $\lambda(\phi)$ в точности равнялось $i\phi$, то все разрешаемые сеткой гармоники переносились бы схемой точно. Однако такая схема имела бы бесконечный по пространству шаблон.

Величина $\lambda(\phi) - i\phi$ является комплекснозначной функцией. Принято отдельно изображать её действительную и мнимую компоненты. Действительная компонента соответствует амплитудной ошибке, мнимая – фазовой.

Графики $\mathrm{Re}\,\lambda(\phi)$ для схем FDn, где $n = 1, 3, 5, 7$ приведены на рис. 5. Для схем с кососимметрической матрицей A_h выполняется $\mathrm{Re}\,\lambda(\phi) = 0$, поэтому изображать бессмысленно.

Рисунок 5, однако, не имеет большого смысла. Дело в том, что величина $\mathrm{Re}\,\lambda(\phi)$ характеризует скорость затухания волны. Если положить для простоты, что $\mu > 0$ и шаг по времени равен $\tau = h/\mu$, то за шаг по времени амплитуда волны убывает в $\exp(\lambda(\phi))$ раз. Если, допустим, мы допускаем затухание волны в e раз за 10000 шагов по времени, то нам требуется, чтобы $0 \leqslant \mathrm{Re}\,\lambda(\phi) \leqslant 10^{-4}$. Поэтому удобно представить результат в логарифмическом масштабе, см. рис. 6. Для схемы FD1 это условие выполняется при $\phi \lesssim 0.016$, что соответствует сеточному разрешению 393 узла на период. Для схем FD3, FD5 и FD7 аналогично получаем 35, 15 и 10 точек на период.

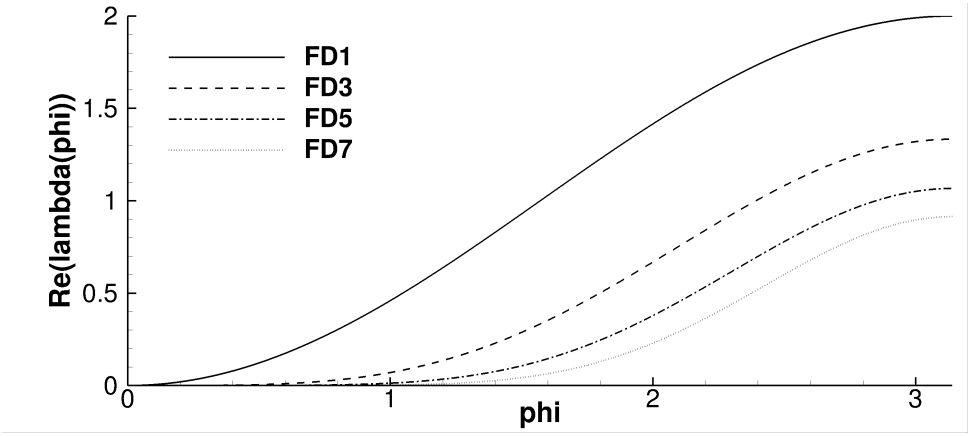

Рис. 5. Графики $\mathrm{Re}\,\lambda(\phi)$ для схем FDn

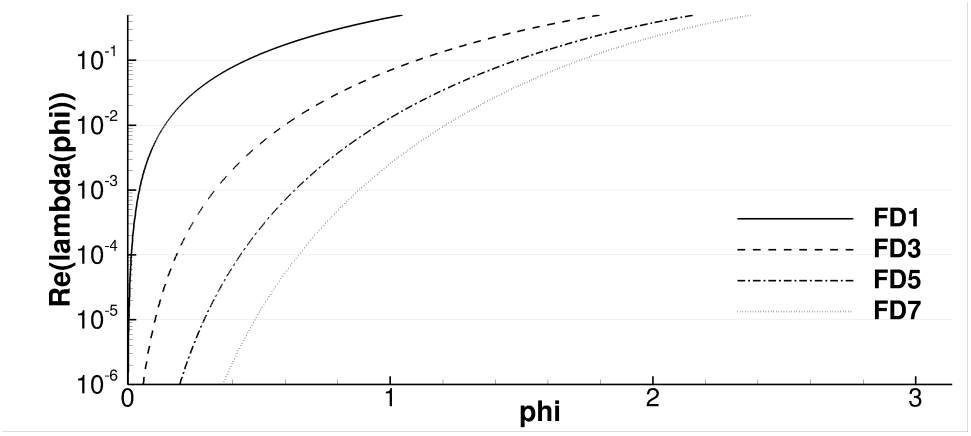

Рис. 6. Графики $\mathrm{Re}\,\lambda(\phi)$ для схем FDn

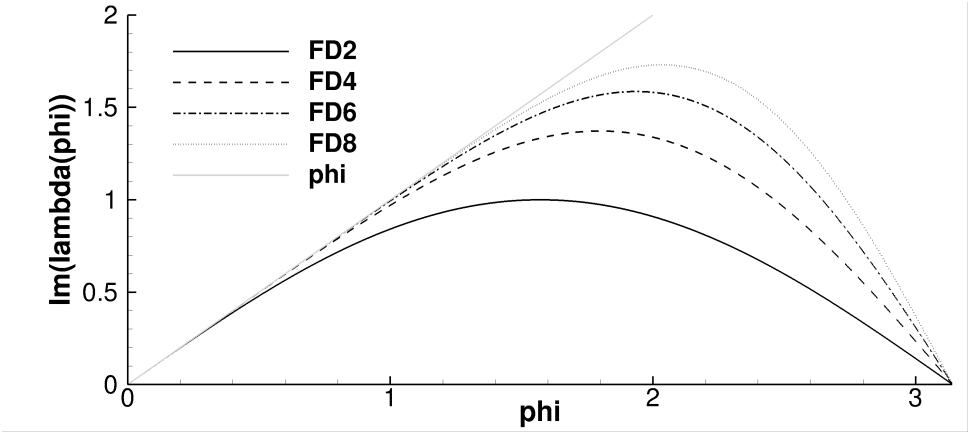

Рис. 7. Графики $\mathrm{Im}\,\lambda(\phi)$ для схем FDn

Графики $\mathrm{Im}\,\lambda(\phi)$, наоборот, приведём на рис. 5 только для схем FDn, где $n = 2, 4, 6, 8$. Для схем FDn с нечётным n волны с большими волновыми числами быстро затухают, поэтому соответствующая им фазовая ошибка не представляет интереса.

Спектральное сравнение схем, имеющих больше одной степени свободы на ячейку, или схем на неравномерных сетках требует дополнительного уточнения.

7. Случай разрывного решения

Назовём 2π-периодическую функцию f кусочно непрерывно дифференцируемой, если существуют такие $0 = y_1 < \ldots < y_n = 2\pi$, что:

- f непрерывно дифференцируема на каждом интервале (y_j, y_{j+1});
- в каждой точке y_j существуют две односторонние производные, равные предельному значению производных изнутри соответствующих интервалов;
- в каждой точке y_j величина $f(y_j)$ равна либо $f(y_j - 0)$, либо $f(y_j + 0)$.

В настоящем разделе будем рассматривать задачу Коши

$$\frac{\partial u}{\partial t} + \mu \frac{\partial u}{\partial x} = 0, \quad t \in (0, t_{\max}), \quad x \in (-\pi, \pi), \tag{7.1}$$

$$u(t, -\pi) = u(t, \pi), \quad u(0, x) = u_0(x), \tag{7.2}$$

с 2π-периодическими кусочно непрерывно дифференцируемыми начальными данными $u_0(x)$. Под решением этой задачи понимается функция $u(t, x) = u_0(x - \mu t)$.

Все оценки точности численных схем, полученные в первой части нашего курса, выражаются через норму той или иной производной от решения. Если решение содержит разрывы, эти оценки лишены смысла. В этом разделе мы изучим поведение конечно-разностных схем на равномерных сетках в этой ситуации.

7.1. Численный эксперимент. Прежде чем формулировать результаты, проведём численный эксперимент. Положим скорость переноса равной $\mu = 1$ и выберем начальные данные в виде

$$u_0(x) = \begin{cases} 1, & x \in [-\pi/2, \pi/2), \\ -1, & otherwise. \end{cases}$$

Проведём расчёт задачи с этими начальными данными до времени $t_{\max} = 0.2\pi$ по схеме “левый уголок”, по конечно-разностной схеме FD3, используя для интегрирования по времени 3-стадийный метод Рунге – Кутты 3-го порядка, и по конечно-разностной схеме FD4, используя для интегрирования по времени 4-стадийный метод Рунге – Кутты 4-го порядка. Начальные данные будем задавать с использованием поточечного отображения $\Pi_h f = \{f(x_j)\}$. Будем использовать сетки с $h = 0.002$ и $h = 0.001$; шаг по времени выберем равным $\tau = 0.5h$. Результаты расчётов приведены на рис. 8.

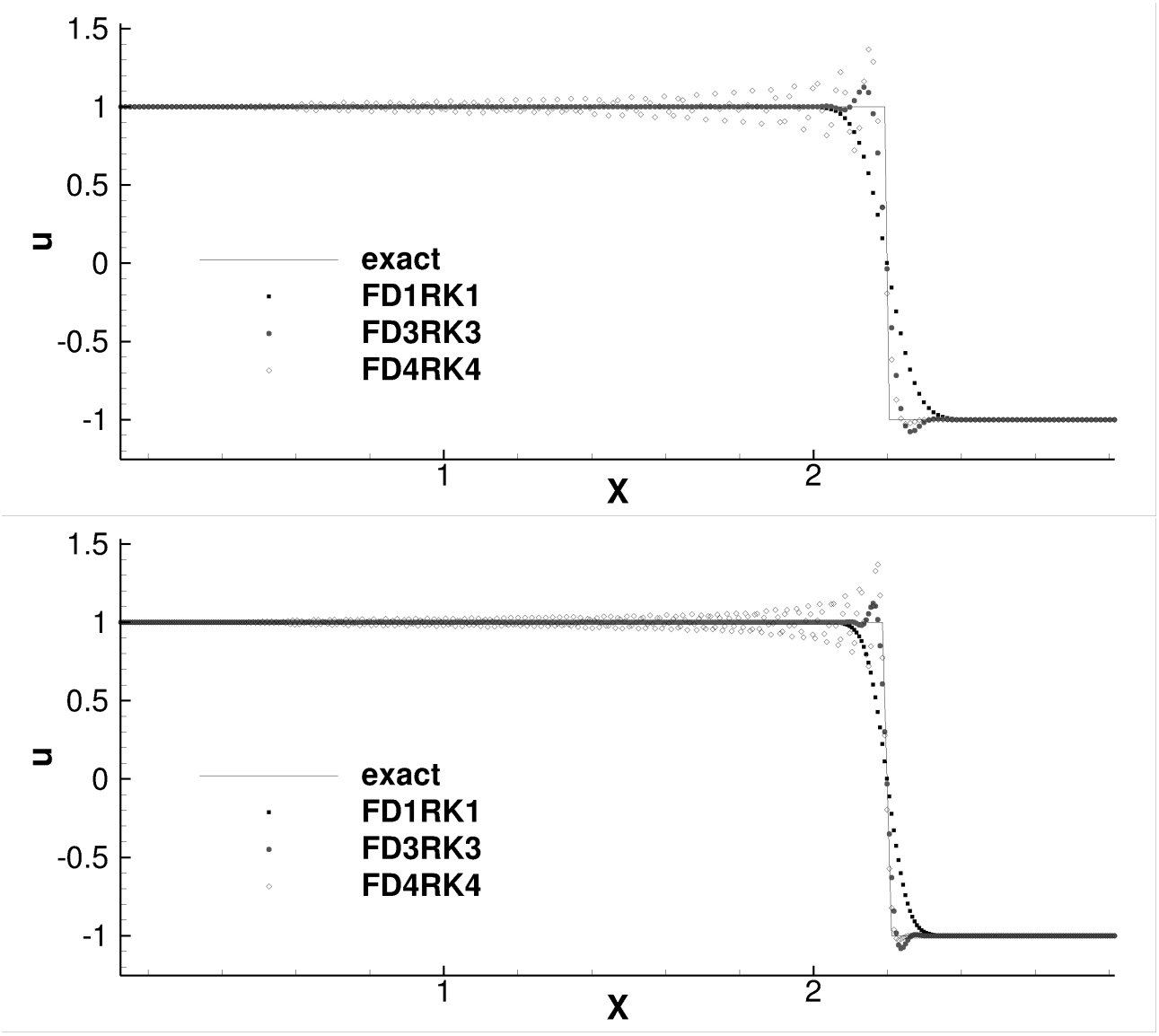

Рис. 8. Численное решение уравнения переноса по схемам “левый уголок”, FD3+RK3, FD4+RK4. Сверху: на грубой сетке $(h = 0.002)$, снизу: на подробной сетке $(h = 0.001)$.

Отметим два результата, общих для всех рассмотренных схем. Во-первых, величина $\|v(t) - \Pi_h u(t, \cdot)\|_\infty = \max_j\{v_j(t) - u(t, x_j)\}$ является величиной порядка амплитуды скачка, поэтому сходимость в $\| \cdot \|_\infty$ не имеет места. Поскольку для любой сеточной функции f выполняется $\|f\|_p \geqslant h^{1/p}\|f\|_\infty$ и, как следствие,

$$\|v(t) - \Pi_h u(t, \cdot)\|_p \geqslant h^{1/p}\|v(t) - \Pi_h u(t, \cdot)\|_\infty,$$

порядок сходимости в $\| \cdot \|_p$ не может быть больше $1/p$. То есть наличие ошибки порядка единицы в одном сеточном узле даёт оценку сверху на порядок сходимости.

Во-вторых, для схем FD1 и FD3 ширина зоны, в которой численное реше-

ние визуально отличается от "точного" (то есть от $\Pi_h u(t,\,\cdot\,)$), стремится к нулю при $h \to 0$ и фиксированном $t_{\max}$, поэтому сходимость в $\|\,\cdot\,\|_2$ наблюдается. Для схемы FD4 сходимость "на глаз" оценить сложно. Приведём численные значения нормы ошибки:

h	0.002	0.001	0.0005	0.00025
$\|\varepsilon\|_2/(2\pi)$	0.1065	0.0808	0.0609	0.0456

Численный порядок точности, посчитанный по этим значениям, примерно равен 0.4.

Рассмотрим полудискретную схему вида

$$\frac{dv_j}{dt} + \frac{\mu}{h}\sum_{k=-S}^{S} a_k v_{j+k} = 0, \quad j = 0, \ldots, N_x - 1; \tag{7.3}$$

$$v(0) = \Pi_h u_0, \tag{7.4}$$

где v_j продолжено периодическим образом на $j \in \mathbb{Z}$, а $a_k \in \mathbb{R}$ – некоторые константы.

Теорема 7.1. *Пусть $u(t,x)$ – численное решение задачи Коши* (5.1)–(5.2) *с кусочно-непрерывно-дифференцируемыми начальными данными $u_0(x)$, Π_h определён формулой $\Pi_h f = \{f(x_j)\}$. Пусть $v(t)$ – решение по устойчивой полудискретной схеме вида* (7.3), (7.4)*, обладающей порядком аппроксимации n на гладких решениях. Тогда для любых $\sigma > 0$ и $t > 0$ имеет место оценка*

$$\|v(t) - \Pi_h u(t,\,\cdot\,)\|_2 \leqslant C(\sigma, t) h^{\frac{n}{2(n+1)} - \sigma}.$$

Для полностью дискретных схем можно получить аналогичный результат.

Заметим, что при $n = 4$ теорема 7.1 утверждает сходимость с порядком $n/(2(n+1)) - \sigma = 0.4 - \sigma$ при сколь угодно малом $\sigma > 0$, что согласуется с полученными выше экспериментальными результатами.

Доказательство теоремы 7.1. Поскольку в каждый момент времени решение периодическое по пространству и её квадрат является интегрируемой функцией, она представима в виде ряда Фурье, сходящегося как ряд элементов L_2:

$$u(t,x) = \sum_{k=-\infty}^{\infty} c_k e^{ik(x-t)}.$$

Коэффициенты ряда удовлетворяют равенству Парсеваля:

$$\sum_{k=-\infty}^{\infty} |c_k|^2 = \|u_0\|_2^2 < \infty.$$

Поскольку начальные данные кусочно непрерывно дифференцируемы, выполняется $|c_k| \leqslant m/k$, где m зависит только от u_0.

Численное решение подчиняется равенству

$$S_{N_x} v(t) = \text{diag}\{\exp(-t\lambda_k)\} S_{N_x} v(0),$$

где $\lambda_k = \lambda(kh)/h$, а

$$\lambda(\phi) = \sum_{q=-S}^{S} a_q \exp(iq\phi) \tag{7.5}$$

(мы это подробно рассматривали в первой части курса). Ошибка решения $\varepsilon(t) = v(t) - \Pi_h u(t, \cdot)$ подчиняется равенству

$$S_{N_x} \varepsilon(t) = \text{diag}\{\exp(-t\lambda_k) - \exp(-ikt)\} S_{N_x} v(0).$$

Пользуясь унитарностью матрицы S_{N_x}, получаем

$$\|\varepsilon\|_2^2 = \sum_{k=-\infty}^{\infty} |c_k|^2 \left(\exp(-t\lambda_k) - \exp(-ikt)\right)^2 =$$
$$= 4 \sum_{k=-\infty}^{\infty} |c_k|^2 \left(\frac{\exp(-t\lambda_k) - \exp(-ikt)}{2}\right)^2.$$

В силу устойчивости $\operatorname{Re} \lambda_k \geqslant 0$, поэтому $|\exp(-t\lambda_k)| \leqslant 1$ и множитель при $|c_k|^2$ не превосходит 1. Поэтому для любого $\delta < 2$ справедливо

$$\|\varepsilon\|_2^2 \leqslant 4 \sum_{k=-\infty}^{\infty} |c_k|^2 \left(\frac{\exp(-t\lambda_k) - \exp(-ikt)}{2}\right)^{\delta}.$$

Поскольку для любого $\alpha \in \mathbb{C}$, такого что $\operatorname{Re} \alpha \leqslant 0$, выполняется $|\exp(\alpha) - 1| \leqslant |\alpha|$ (для доказательства этого факта достаточно заметить, что $\exp(\alpha) - 1$ равно интегралу от $\exp(z)$ по отрезку $[0, \alpha]$),

$$|\exp(-t\lambda_k) - \exp(-ikt)| = |\exp(-ikt)|\,|\exp(-t(\lambda_k - ik)) - 1| \leqslant$$
$$\leqslant t|\lambda_k - ik| = \frac{t}{h}\,|\lambda(kh) - ikh|\,.$$

Поскольку $|\lambda(\phi) - i\phi| \leqslant C|\phi|^{n+1}$, где n – порядок схемы на гладком решении, отсюда получаем

$$\|\varepsilon\|_2^2 \leqslant 4 \sum_{k=-\infty}^{\infty} |c_k|^2 \left(C\frac{t(kh)^{n+1}}{2h}\right)^{\delta}.$$

Поскольку $|c_k| \leqslant 2/k$, при любом $\delta < 1/(n+1)$ выражение под знаком суммы не будет превосходить $\tilde{C}k^{\alpha}$, где $\alpha = (n+1)\delta - 2 > -1$ и, следовательно,

ряд будет сходиться. При этом сумма ряда будет пропорциональна $h^{\delta n}$. Поскольку в левой части стоит квадрат ошибки, от этой суммы нам нужно будет взять корень. Таким образом мы доказали, что при любом $p < n/(2(n+1))$ будет иметь место сходимость с порядком p в норме $\|\cdot\|_2$. □

7.2. Эффект Гиббса. Между результатами расчётов по разным схемам наблюдаются и различия. У решения по схеме "левый уголок" при условии $\tau\mu/h \in [0,1]$ локальные максимумы решения не возрастают (и минимумы не убывают), а новые экстремумы возникать не могут. Поэтому численное решение имеет только один локальный максимум и один локальный минимум (при достаточно малом $t_{\max}$ равные 1 и -1 соответственно). У решений по схемам, обладающим порядком выше 1 на задачах с гладкими решениями, имеются паразитные "забросы" выше максимума и ниже минимума. Наличие этих осцилляций называется *эффектом Гиббса*.

Объясним эффект Гиббса на примере полудискретных схем вида (7.3)–(7.4), таких что $a_k = a_{-k}$ (и, в частности, $a_0 = 0$). Будем предполагать, что $\mu \neq 0$, хотя бы один коэффициент a_k отличен от нуля, а N_x чётное и больше $2S+1$. Численное решение по такой схеме удовлетворяет равенству

$$\frac{d}{dt}\|v(t)\|_2 = 0.$$

Напомним, что

$$\|f\|_2 = \left(\sum_{j=0}^{N_x-1} h|f_j|^2\right)^{1/2}.$$

Начальные данные и сетка нами выбраны таким образом, что $v_j(0) = \pm 1$, поэтому $\|v(0)\|_2 = 1$ и $\|v(t)\|_2 = 1$ при всех $t \geqslant 0$.

Покажем, что в некотором интервале $t \in (0, \tilde{t})$ выполняется $\|v(t)\|_\infty > 1$. Допустим противное. Тогда в некоторой окрестности $t = 0$ выполняется $|v_j(t)| \leqslant 1$ и $\sum |v_j(t)|^2 = N_x$. Это возможно только тогда, когда $v_j(t) = \pm 1$. Но поскольку каждое значение $v_j(t)$ непрерывно по времени, это означает, что оно не меняется во времени. При сделанных предположениях такое невозможно.

7.3. Является ли эффект Гиббса злом? Снова о критериях качества. Выше мы показали, что несмотря на наличие осцилляций на разрыве, численное

решение по устойчивой конечно-разностной схеме сходится в $\| \cdot \|_2$. Но является ли сходимость единственным условием, которое представляет интерес? Чтобы ответить на этот вопрос, нужно вернуться к общим принципам разработки численных методов.

Если бы мы рассматривали готовые программные продукты, предназначенные для промышленного применения, мы бы могли соотнести их по точности получаемых результатов и затратам машинных ресурсов на проведение расчётов. Но промышленные задачи слишком сложны, чтобы мы могли непосредственно исследовать предназначенные для них численные схемы. Поэтому исследование численных методов проводится на *модельных задачах* (в англоязычной литературе – *toy problem*). Модельная задача должна, с одной стороны, быть проще и понятнее исходной задачи, но, с другой стороны, отражать некоторые её свойства. Именно это и позволяет частично перенести требования, предъявляемые к численным схемам для решения промышленных задач, на модельные задачи. Исследуя численный метод на модельной задаче, нужно представлять, как наблюдаемые на ней и доказываемые для неё свойства численных схем будут отражаться на решении сложной задачи.

Предположим, например, что нашей целью является создание схемы для решения системы уравнений Эйлера для идеального газа (которая сама является упрощённой моделью для трёхмерных задач газовой динамики):

$$\begin{aligned} &\left(\frac{\partial}{\partial t} + u\frac{\partial}{\partial x}\right)\rho + \rho\frac{\partial u}{\partial x} = 0; \\ &\left(\frac{\partial}{\partial t} + u\frac{\partial}{\partial x}\right)u + \frac{1}{\rho}\frac{\partial p}{\partial x} = 0; \\ &\left(\frac{\partial}{\partial t} + u\frac{\partial}{\partial x}\right)p + \gamma p\frac{\partial u}{\partial x} = 0. \end{aligned} \tag{7.6}$$

Неизвестными в этих уравнениях являются поля физических величин: ρ – плотность, u – скорость, p – давление. Показатель адиабаты γ является параметром уравнения и для воздуха при нормальных условиях обычно принимается равным 1.4. Будем считать, что эти уравнения заданы на произведении интервалов $(x_{\min}, x_{\max}) \times (0, t_{\max})$, где $x_{\max} > x_{\min}$ и $t_{\max}$ – некоторые величины. При моделировании конкретного физического процесса систему (7.6) необходимо дополнить начальными и граничными условиями.

Систему (7.6) можно переписать в матричном виде

$$\frac{\partial U}{\partial t} + A(U)\frac{\partial U}{\partial x} = 0,$$

где $U = (\rho, u, p)^T$. Матрица $A(U)$ имеет собственные значения u, $u+c$ и $u-c$, где $c = \sqrt{\gamma p/\rho}$. Отношение давления к плотности должно быть положительным;

невыполнение этого условия делает начально-краевую задачу для уравнений Эйлера физически бессмысленной и математически некорректной.

Если положить $u(t,x) = \mu = const$ и $p(t,x) = const$, то последние два уравнения в (7.6) выполнятся автоматически, а первое сведётся к

$$\frac{\partial \rho}{\partial t} + \mu \frac{\partial \rho}{\partial x} = 0.$$

Это не что иное, как уравнение переноса, численные методы для решения которого мы рассматривали в нашем курсе. Если уравнение переноса рассматривается именно как частный случай уравнений Эйлера, то начальные условия для ρ строго положительные.

Пусть для решения задачи Коши для уравнения переноса мы применяем схему вида (7.3)–(7.4). Теорема 7.1 утверждает, что решение сходится в $\|\cdot\|_2$ к точному. Но эта сходимость не гарантирует ни того, на любой сетке численное решение будет положительным, ни даже того, что мы сможем добиться положительности численного решения измельчением сетки! Рассмотрим, например, задачу

$$u_0(x) = \begin{cases} \delta, & x \in [-\pi, 0), \\ 1, & x \in [0, \pi), \end{cases}$$

где $\delta > 0$. В силу линейности легко заметить, что амплитуда "выбросов" около разрыва пропорциональна величине самого разрыва. Поэтому для любой схемы, дающей такие "выбросы", найдётся такое $\delta > 0$, что решение не будет положительным. Следовательно, те подходы, которыми получены такие схемы, приходится считать неприменимыми к задачам с разрывами большой амплитуды.

Заметим, что сходимость в $\|\cdot\|_2$ с порядком выше $1/2$ (что для задач с гладкими решениями обычно имеет место) автоматически влечёт сходимость в $\|\cdot\|_\infty$ (с некоторым меньшим порядком). А, значит, $\|\cdot\|_\infty$-норма ошибки при измельчении сетки рано или поздно станет меньше минимума плотности в точном решении, что в силу неравенства треугольника обеспечивает положительность численного решения.

Приведём ещё один пример. Представим, например, что переменная u, для которой мы решаем уравнение переноса, имеет смысл массовой доли пара в смеси пара и воды, то есть $u = 1$ означает, что в данной точке находится пар, $u = 0$ – что в данной точке находится вода, а $u \in (0,1)$ – что в данной точке есть и пар, и вода. Пусть начальные данные являются кусочно-постоянной функцией со значениями 0 и 1. Если при численном моделировании в сеточном узле возникает, скажем, значение $u = 1.2$, это будет означать, что в данной точке есть 120% пара и -20% жидкости. В зависимости от того, находится ли в ячейке пар или жидкость, для этой ячейки могут выполняться дополнительные

действия, и наличие массовой доли, выходящей за пределы отрезка $[0,1]$, может привести к непредсказуемым последствиям.

7.4. Линейные монотонные схемы. Рассмотрим теперь полностью дискретную схему вида

$$v_j^{n+1} = \sum_{k=-S}^{S} b_k v_{j+k}^n, \quad j = 0, \ldots, N_x - 1, \quad n = 0, \ldots, N_t - 1; \tag{7.7}$$

$$v^0 = \Pi_h u_0. \tag{7.8}$$

Коэффициенты b_k могут зависеть от τ, h, μ.

Определение 1. *Схема вида* (7.7)–(7.8) *называется монотонной, если для всех* $k = -S, \ldots, S$ *выполняется* $b_k \geqslant 0$.

Упражнение 5. *Доказать, что для схем вида* (7.7)–(7.8) *монотонность равносильна устойчивости в* $\| \cdot \|_\infty$ *с константой* $K = 1$.

Возникает вопрос, можно ли в классе монотонных схем вида вида (7.7)–(7.8) построить схемы высокого порядка аппроксимации. Ответ на этот вопрос отрицательный.

Теорема 7.2 (Годунов)**.** *Пусть* $\mu\tau/h \notin \mathbb{Z}$. *Тогда монотонная схема не является точной на квадратичной функции.*

Доказательство. Допустим противное. Пусть $u_j^n = (jh + \mu\tau)^2$. Тогда в силу точности на квадратичной функции $u_j^{n+1} = (jh)^2$. Но $\min\{u_j^n\} > 0$, поскольку $\mu\tau/h$ не является целым числом, а $u_0^{n+1} = 0$. Таким образом, минимум решения уменьшился со временем, что противоречит условию. □

Теорема Годунова показывает, что монотонность и высокий порядок точности являются противоречащими друг другу условиями, по меньшей мере, для схем вида (7.7).

Современный подход к решению этой дилеммы заключается в использовании *нелинейных* схем даже для линейного уравнения. Теорема Годунова никак не использует тот линейность схемы и, следовательно, остаётся справедливой. Тем не менее, на практике это позволяет существенно улучшить точность численных расчётов. Чтобы избежать дублирования текста, мы рассмотрим эти вопросы ниже, сразу для нелинейного уравнения.

8. Случай уравнения с переменным коэффициентом

8.1. Задача. В настоящем разделе рассмотрим задачу Коши для уравнения

$$\frac{\partial u(t,x)}{\partial t} + \frac{\partial}{\partial x}\left(\mu(x)u(t,x)\right) = 0, \quad 0 < t < t_{\max}, \quad x \in \mathbb{R}; \tag{8.1}$$

$$u(0,x) = u_0(x), \quad x \in \mathbb{R}, \tag{8.2}$$

где $\mu(x)$ и $u_0(x)$ – гладкие 2π-периодические функции, причём величина $\mu_{\min} = \min\limits_{x\in\mathbb{R}} \mu(x)$ строго положительна. Также обозначим $\mu_{\max} = \max\limits_{x\in\mathbb{R}} \mu(x)$.

Заменой переменных $\tilde{u}(t,x) = \mu(x)u(t,x)$ уравнение (8.1) приводится к уравнению с переменным коэффициентом:

$$\frac{\partial \tilde{u}(t,x)}{\partial t} + \mu(x)\frac{\partial \tilde{u}(t,x)}{\partial x} = 0. \tag{8.3}$$

Решение (8.3) постоянно вдоль характеристических кривых $x(t)$, определяемых уравнением $dx/dt = \mu(x)$. Поэтому, чтобы найти $\tilde{u}(t,x)$ в заданной точке (t,x), нужно найти проходящую через эту точку характеристику. Поскольку $\mu(x)$ гладкая и 2π-периодическая, она ограниченная, следовательно, эта характеристика пересечёт ось $t = 0$ в некоторой точке x_0. Значение $u_0(x_0)$ и будет равно искомой величине $\tilde{u}(t,x)$.

Лемма 8.1. *Решение* (8.1)–(8.2) *при любом* $0 < t < t_{\max}$ *удовлетворяет оценке* $\|u(t,\,\cdot\,)\|_2 \leqslant c\|u_0\|_2$, *где* $c = (\mu_{\max}/\mu_{\min})^{1/2}$.

Доказательство. Введём на пространстве 2π-периодических непрерывных функций норму

$$\|f\|_\mu = \left(\int\limits_0^{2\pi} \mu(x)|f^2(x)|dx\right)^{1/2}$$

и заметим, что

$$\frac{d}{dt}\|u(t,\,\cdot\,)\|_\mu^2 = \int\limits_0^{2\pi} \mu(x)2\overline{u(t,x)}\frac{\partial u(t,x)}{\partial x} =$$

$$= -\int\limits_0^{2\pi} 2\mu(x)\overline{u(t,x)}\frac{\partial}{\partial x}(\mu(x)u(t,x))dx = -\int\limits_0^{2\pi} \frac{\partial}{\partial x}\left(|\mu(x)u(t,x)|^2\right)dx = 0.$$

Следовательно, $\|u(t,\,\cdot\,)\|_\mu = \|u(0,\,\cdot\,)\|_\mu$. Поскольку для любой 2π-периодической непрерывной функции имеют место неравенства

$$\sqrt{\mu_{\min}}\|f\|_2 \leqslant \|f\|_\mu \leqslant \sqrt{\mu_{\max}}\|f\|_2,$$

имеем

$$\|u(t,\cdot)\|_2 \leqslant \mu_{\min}^{-1/2}\|u(t,\cdot)\|_\mu = \mu_{\min}^{-1/2}\|u_0\|_\mu \leqslant \mu_{\min}^{-1/2}\mu_{\max}^{1/2}\|u_0\|_2,$$

что и требовалось доказать. □

8.2. Пример хорошей схемы. Под сеточной функцией будем понимать N_x-периодическую последовательность комплексных чисел.

Для решения задачи (8.1)–(8.2) можно использовать полудискретную схему

$$\frac{dv_j}{dt} + \frac{v_{j+1}\mu(x_{j+1}) - v_{j-1}\mu(x_{j-1})}{2h} = 0, \quad j = 0,\dots,N_x-1, \tag{8.4}$$

$$v_j(0) = u_0(x_j). \tag{8.5}$$

Далее будем использовать сокращение $\mu_j \equiv \mu(x_j)$. Введём на пространстве сеточных функций норму

$$\|f\|_\mu = \left(\sum_{j=0}^{N_x-1} h\mu_j|f_j|^2\right)^{1/2}$$

Лемма 8.2. *Любое решение* (8.4) *удовлетворяет равенству* $\|v(t)\|_\mu = \|v(0)\|_\mu$ *и оценке* $\|v(t)\|_2 \leqslant c\|v(0)\|_2$, *где* $c = (\mu_{\max}/\mu_{\min})^{1/2}$.

Доказательство. Запишем

$$\frac{d\|v(t)\|_\mu^2}{dt} = \mathrm{Re}\sum_j 2h\lambda_j\bar v_j\frac{dv_j}{dt} = -\mathrm{Re}\sum_j \mu_j\bar v_j(v_{j+1}\mu_{j+1} - v_{j-1}\mu_{j-1}).$$

Разобьём сумму на две и во второй сдвинем индексы суммирования, пользуясь периодичностью v_j. Тогда

$$\frac{d\|v(t)\|_\mu^2}{dt} = -\mathrm{Re}\sum_j(\mu_j\bar v_j v_{j+1}\mu_{j+1} - \mu_{j+1}\bar v_{j+1}v_j\mu_j) = 0.$$

Первое утверждение доказано. Доказательство второго утверждения повторяет конец доказательства леммы 8.1. □

Упражнение 6. *Доказать, что для решения* $v(t)$ *по полудискретной схеме* (8.4)–(8.5) *справедлива оценка ошибки*

$$\left(\sum_{j=0}^{N_x-1} h|v_j(t) - u(t,x_j)|^2\right)^{1/2} \leqslant cth^2,$$

где c *зависит от* u_0 *и* μ, *но не зависит от* t *и* h.

8.3. Пример плохой схемы. Устойчивость схемы (8.4)–(8.5) заключается в том, что решение (8.4) удовлетворяет равенству

$$\|u(t,\,\cdot\,)\|_\mu = \|u(0,\,\cdot\,)\|_\mu. \tag{8.6}$$

Это свойство схемы отражает аналогичное свойство решений (8.1). Но для более сложных уравнений (например, если в (8.1) допустить зависимость μ также и от t) следовать этой же стратегии может быть затруднительно или невозможно. Чтобы не усложнять изложение, мы останемся в рамках задачи Коши (8.1)–(8.2), но рассмотрим схему, не использующую свойство (8.6).

Запишем для решения (8.1)–(8.2) полудискретную схему

$$\frac{dv_j}{dt} + \mu_j \frac{v_{j+1} - v_{j-1}}{2h} + v_j \frac{\mu_{j+1} - \mu_{j-1}}{2h} = 0, \tag{8.7}$$

$$v_j(0) = u_0(x_j). \tag{8.8}$$

Здесь и далее используется сокращение $\mu_j \equiv \mu(x_j)$.

Для анализа этой схемы нам понадобится неравенство Коши – Буняковского (скалярное произведение векторов не превосходит произведения их длин) в следующем частном случае: для любых сеточных функций v и w выполняется

$$\left|\sum_j h v_j w_j\right| \leqslant \left(\sum_j h|v_j|^2\right)^{1/2} \left(\sum_j h|w_j|^2\right)^{1/2} = \|v\|_2\,\|w\|_2.$$

Проанализируем устойчивость схемы (8.7)–(8.8).

$$\frac{d\|v\|_2^2}{dt} = \mathrm{Re}\sum_j 2h\bar v_j \frac{dv_j}{dt} = -\mathrm{Re}\sum_j \bar v_j\left(\mu_j(v_{j+1} - v_{j-1}) + v_j(\mu_{j+1} - \mu_{j-1})\right).$$

Пользуясь периодичностью и очевидным тождеством $\mathrm{Re}(v_j\bar v_{j+1}) = \mathrm{Re}(\bar v_j v_{j+1})$,

$$\begin{aligned}\frac{d\|v\|_2^2}{dt} &= -\mathrm{Re}\sum_j \bar v_j\mu_j v_{j+1} - \bar v_{j+1}\mu_{j+1}v_j + |v_j|^2(\mu_{j+1} - \mu_{j-1}) = \\ &= -\mathrm{Re}\sum_j h\bar v_j v_{j+1}\frac{\mu_j - \mu_{j+1}}{h} + 2h|v_j|^2\frac{\mu_{j+1} - \mu_{j-1}}{2h}.\end{aligned}$$

Введём $L = \sup|\partial\mu/\partial x|$. Тогда имеем

$$\frac{d\|v\|_2^2}{dt} \leqslant L\sum_j h|v_j v_{j+1}| + 2L\sum_j h|v_j|^2 \leqslant 3L\|v\|_2^2.$$

Последнее неравенство записано в силу неравенства Коши – Буняковского для сеточных функций v и w, где w имеет компоненты $w_j = v_{j+1}$. Сокращая одну степень $\|v\|_2$, получаем $d\|v\|_2/dt \leqslant (3L/2)\|v\|_2$, что даёт

$$\|v(t)\|_2 \leqslant \exp(3Lt/2)\|v(0)\|. \tag{8.9}$$

Таким образом, мы доказали, что схема является устойчивой в $\|\cdot\|_2$ с константой $K(t) = \exp(3Lt/2)$.

Устойчивость с константой $K(t) = \exp(ct)$, где $c \geqslant 0$, вместе с условием p-го порядка аппроксимации доказывает оценку ошибки вида $\|\varepsilon(t)\| \sim t\exp(ct)$, где c не зависит от h. Важно отличать этот эффект от линейной неустойчивости схем с постоянными коэффициентами, которая обычно приводит к росту ошибки со скоростью $\|\varepsilon(t)\| \sim t\exp(ct/h)$ и, как следствие, отсутствию сходимости.

Избавиться от экспоненциального множителя в оценке нормы решения невозможно: схема (8.7)–(8.8) действительно допускает экспоненциальный рост численного решения. Покажем это в численном эксперименте. Выберем

$$\mu(x) = 4 + 2\sin x - \sin(2x) + \frac{2}{3}\sin(3x) - \frac{1}{2}\sin(4x),$$

что является первыми членами ряда Фурье для функции $4+x$ на $[-\pi, \pi]$, продолженной периодическим образом, и положим $u_0(x) = 1/\mu(x)$. Тогда решением (8.1)–(8.2) будет функция $u(t, x) = u_0(x)$.

Проведём расчёт по схеме (8.7)–(8.8) до достижения $t_{\max} = 500$. Будем использовать сетки с числом узлов на период $N_x = 25, 50, 100$. Для интегрирования по времени будем использовать 3-стадийный метод Рунге – Кутты 3-го порядка с шагом $\tau = h/80$. Результаты расчётов приведены на рис. 9. Уменьшение τ принципиально не меняет результата, поэтому можно приближённо считать полученный результат решением по полудискретной схеме.

Представим схему (8.7)–(8.8) в виде

$$\frac{dv}{dt} + A_h v = 0, \quad v(0) = \Pi_h u_0. \tag{8.10}$$

Полученные результаты показывают быстрый рост ошибки на сетках с $N_x = 25$ и $N_x = 50$; при больших t наблюдается экспоненциальный рост. Это свидетельствует о том, что матрица A_h имеет собственное значение с отрицательной действительной частью. Также численные результаты, что скорость роста ошибки

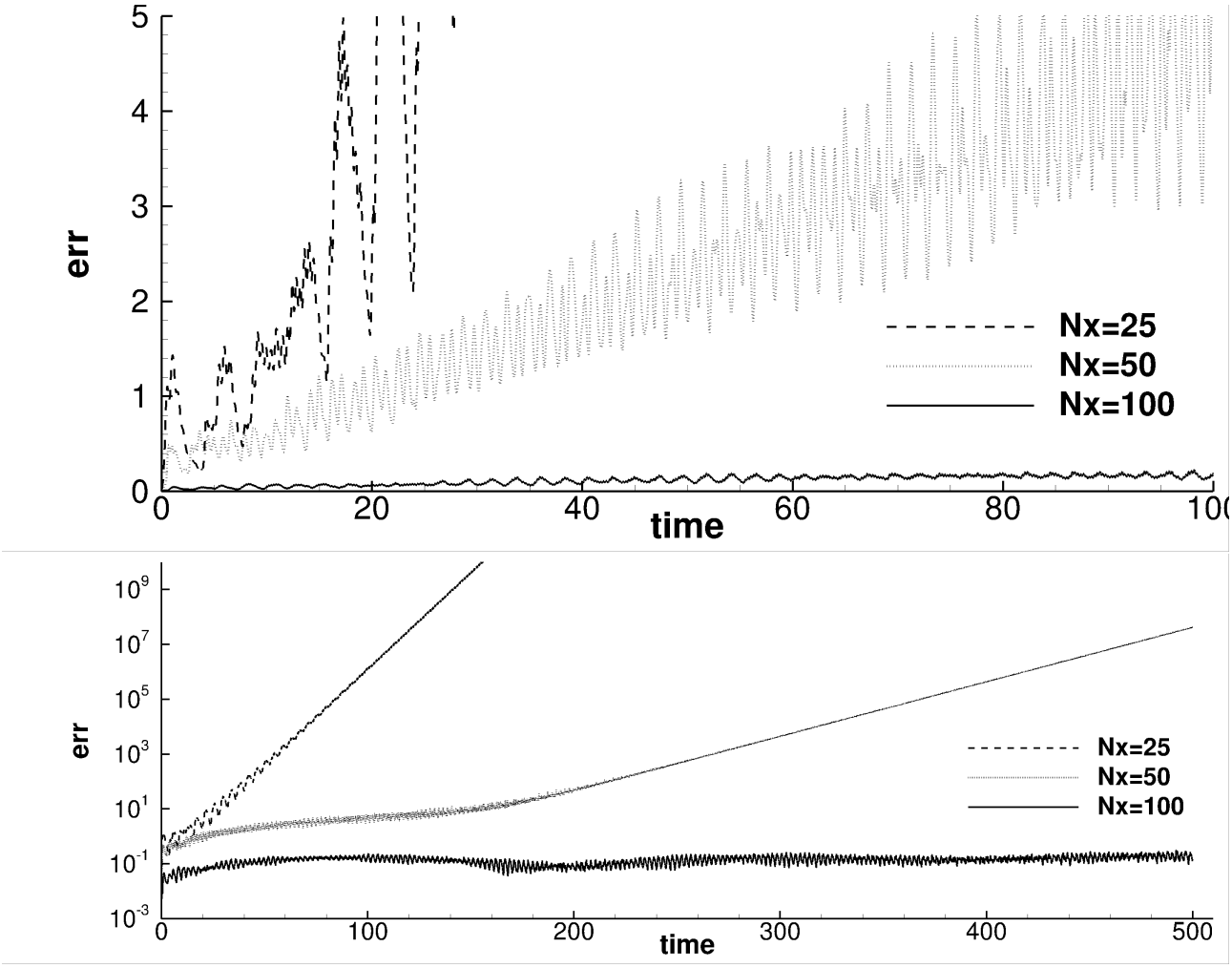

Рис. 9. Норма ошибки численного решения уравнения с переменным коэффициентом. Верхнее и нижнее изображения отличаются только масштабами по осям

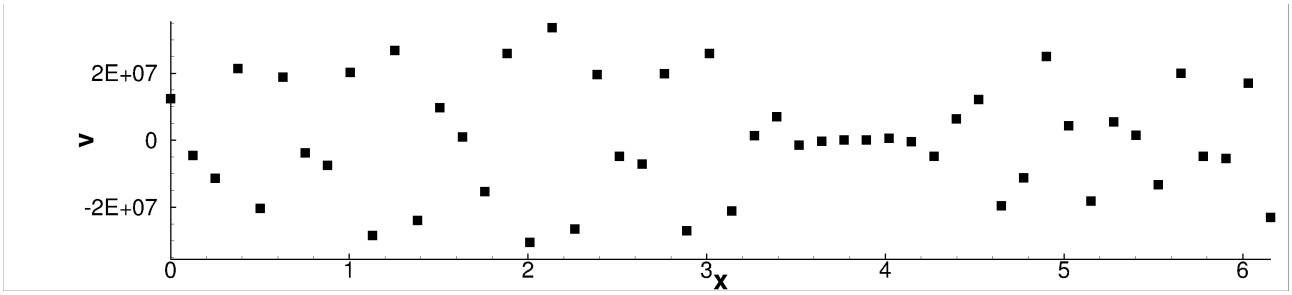

Рис. 10. Численное решение на сетке с $N_x = 50$ на момент $t = 500$

не увеличивается с измельчением сетки, что следует из оценки (8.9) и свидетельствует, что наблюдаемый эффект не является линейной неустойчивостью схемы постоянными коэффициентами.

Численное решение на сетке с $N_x = 25$ на момент $t = 500$ приведено на рис. 10. Поразмышляем над тем, чем именно является это решение.

Пусть $v(t) \in \mathbb{C}^N$ – решение некоторого ОДУ

$$\frac{dv}{dt} + Av = 0.$$

Предположим для простоты, что матрица A диагонализуемая, то есть представима в виде $A = S\Lambda S^{-1}$, где Λ – диагональная матрица. Тогда $v(t) = S\text{diag}\{\exp(-\lambda_j t)\}S^{-1}v(0)$, где λ_j – элементы матрицы Λ. Будем считать, что собственные значения упорядочены в порядке возрастания их действительной части:

$$\text{Re}\,\lambda_0 \leqslant \ldots \leqslant \text{Re}\,\lambda_{N_x-1}.$$

Рассмотрим величину $v(t)/\|v(t)\|$:

$$\frac{v(t)}{\|v(t)\|} = \frac{S\text{diag}\{\exp(-\lambda_j t)\}S^{-1}v(0)}{\|S\text{diag}\{\exp(-\lambda_j t)\}S^{-1}v(0)\|} = \frac{S\text{diag}\{\exp((\lambda_0 - \lambda_j)t)\}S^{-1}v(0)}{\|S\text{diag}\{\exp((\lambda_0 - \lambda_j)t)\}S^{-1}v(0)\|}.$$

Предположим вначале, что нулевая (т. е. соответствующая λ_0) компонента $Sv(0)$ отлична от нуля. При $t \to \infty$ все величины вида $\exp((\lambda_0 - \lambda_j)t)$ стремятся к нулю, за исключением значений j, для которых $\text{Re}\lambda_j = \text{Re}\lambda_0$. Поэтому при больших t значение $v(t)/\|v(t)\|$ оказывается близко к сумме собственных векторов, соответствующих этим значениям j. Именно эту сумму (с точностью до $\bar{\bar{o}}(\|v(t)\|)$) мы и видим на рис. 10. Если кроме $j = 0$ таких значений нет, то $v(t)/\|v(t)\| \to e_0$, где e_0 – собственный вектор, соответствующий собственному значению λ_0 (определённый с точностью до множителя).

Может оказаться так, что первая компонента $S^{-1}v_0$ равна нулю. Тогда она не входит ни в числитель, ни в знаменатель, и с формальной точки зрения можно её отбросить, а её роль займёт следующая компонента. Но если мы имеем дело с машинными вычислениями, то эта компонента может появиться за счёт ошибок округления.

8.4. Принцип замороженных коэффициентов. Оценка устойчивости, аналогичная полученной нами, справедлива для широкого класса схем. В случае полностью дискретных схем она устанавливается следующей теоремой. В слегка иной формулировке она была доакзана в работе P. Lax и L. Nirenberg[9].

[9] P. Lax, L. Nirenberg. On Stability for Difference Schemes; a Sharp Form of Gårding's Inequality. Communications on Pure and Applied Mathematics. Vol. XIX, No. 4, 473–492 (1966).

Теорема 8.3. *Пусть $p_s(t,x)$, $s=-S,\dots,S$, – дважды непрерывно дифференцируемы по x и 2π-периодичны по x, причём p_s, её первая и вторая производные по x ограничены равномерно по $t\in[0,t_{\max}]$. Рассмотрим систему равенств*

$$f_j^{n+1}=\sum_{s=-S}^{S} p_s(n\tau,jh)\ f_{j+s}^n,\quad j\in\mathbb{Z}. \tag{8.11}$$

Пусть при всех $t\in[0,t_{\max}]$, $x\in\mathbb{R}$ и $N_x\in\mathbb{N}$ любое N_x-периодическое решение

$$\tilde f_j^{n+1}=\sum_{s=-S}^{S} a_s\tilde f_{j+s}^n,\quad j\in\mathbb{Z},$$

где $a_s=p_s(t,x)$, удовлетворяет $\|\tilde f^n\|_2\leqslant\|\tilde f^0\|_2$. Тогда существует такая функция $K(t)$, что любое решение (8.11) *удовлетворяет $\|f^n\|_2\leqslant K(t_n)\|f^0\|_2$.*

Сведение анализа устойчивости схемы с переменными коэффициентами к анализу устойчивости схемы с постоянными коэффициентами называется *принципом замороженных коэффициентов*.

Доказательство теоремы 8.3 слишком трудоёмкое, чтобы рассматривать его в настоящем курсе. Отметим, что в нём по существу используется равномерность сетки.

8.5. Рассуждения о роли численной диссипации. Несмотря на то, что сходимость численного решения имеет место (и доказана нами выше), экспоненциальный рост ошибки со временем обычно на практике является неприемлемым (за исключением случая, когда сама решаемая задача предусматривает экспоненциальный рост возмущений).

Вновь возвращаясь к рис. 10, мы приходим к выводу, что экспоненциально растущая составляющая численного решения представляет собой высокочастотный шум. Основным способом борьбы с ним является подавление высокочастотных возмущений (*filtering*). Если нам уже дана схема (8.10), то фильтрация заключается в добавлении положительно определённой матрицы C_h:

$$\frac{dv}{dt}+(A_h+C_h)v=0.$$

Например, в качестве C_h можно взять некоторую степень циркулянта с элементами $a_0=2$, $a_1=a_{-1}=-1$, умноженную на некоторый положительный коэффициент. Фактически мы уже встречали эту фильтрацию, когда рассматривали схемы для линейного уравнения переноса с постоянным коэффициентом на равномерной сетке: схемы максимального порядка на симметричном шаблоне

были бездиссипативными ($d\|v\|_2/dt = 0$), а использование шаблона, имеющем против направления переноса на один узел больше, чем по направлению переноса, обеспечивало затухание высокочастотных гармоник.

Недостаток численной диссипации может привести к экспоненциальному росту возмущений, а избыток численной диссипации увеличивает ошибку аппроксимации. Оптимального метода выбрать численную диссипацию (способ её введения и множитель) не существует. В дальнейшем мы будем учитывать необходимость численной диссипации и поэтому рассматривать только такие схемы, которые ей обладают.