УДК 519.6
ББК 22.193
К 17

Издание осуществлено при поддержке Российского фонда фундаментальных исследований по проекту 04-01-14012д

Калиткин Н. Н., Альшин А. Б., Альшина Е. А., Рогов Б. В. **Вычисления на квазиравномерных сетках.** — М.: ФИЗМАТЛИТ, 2005. — 224 с. — ISBN 5-9221-0565-5.

Монография посвящена изучению квазиравномерных сеток и их приложений. Сгущение таких сеток позволяет получать апостериорную асимптотическую оценку погрешности и повышать порядок точности так же, как это делается на равномерных сетках. Квазиравномерные сетки можно строить в неограниченных областях. Все это позволяет предложить методы решения широкого круга задач: квадратурные формулы, разностные схемы для разных типов уравнений в частных производных, интегральные уравнения, задачи на собственные значения и т.п. Изложение иллюстрировано большим количеством примеров.

Книга адресована математикам, научным сотрудникам и инженерам, ведущим прикладные расчеты, а также аспирантам и студентам старших курсов соответствующих специальностей.

© ФИЗМАТЛИТ, 2005
© Н. Н. Калиткин, А. Б. Альшин, Е. А. Альшина, Б. В. Рогов, 2005

ISBN 5-9221-0565-5

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

Эту книгу мы решили написать по трем причинам.

Во-первых, в практике вычислений метод сгущения сеток является мощным средством для численного решения широкого круга задач с гарантированной точностью. Однако даже математики-прикладники не всегда знают возможности этого метода и как им грамотно пользоваться. Большинство же инженеров, рассчитывающих свои задачи на компьютерах, почти не слышали о нем.

Во-вторых, среди сеток есть один важный класс — квазиравномерные сетки. Они легко адаптируются ко многим задачам, и при этом позволяют использовать все возможности метода сгущения сеток. Это дает возможность добиваться высокой точности при малом объеме вычислений даже для весьма сложных задач. Но о такой возможности знает сейчас лишь узкий круг наиболее квалифицированных специалистов.

В-третьих, квазиравномерные сетки с *конечным* числом интервалов можно строить даже в неограниченной области: последний узел сетки будет бесконечно удаленной точкой. Это позволило нам в последние годы предложить эффективные методы решения многих классов задач в неограниченной области: несобственных интегралов, краевых и начально-краевых задач для дифференциальных уравнений. Метод естественно переносится на интегральные и интегро-дифференциальные уравнения, а также на любое число измерений.

Все это открывает широчайшие возможности, о которых пока почти никто не подозревает. Например, можно естественно решать аэродинамические задачи дозвукового обтекания, где граничным условием является невозмущенность набегающего потока на бесконечном удалении от обтекаемого тела (сейчас такие задачи решают громоздкими искусственными приемами).

Насколько нам известно, впервые квазиравномерные сетки (и сам этот термин) предложил А.А. Самарский около 1952 г. Его коллектив выполнял тогда расчет мощности взрыва первой советской термоядерной бомбы. Задача была крайне сложная, а расчеты велись вручную, на электроарифмометрах “Мерседес”: до пуска первой приличной отечественной ЭВМ “Стрела” оставалось еще больше года. Необходимо было предельно экономизировать вычисления и брать сетку из небольшого числа узлов, иначе расчеты не уложились бы в сроки. Но и точность требовалась высокая, а для этого в некоторых участках сле-

довало брать малый шаг. Как же удовлетворить этим противоречивым требованиям?

Тогда А.А. Самарский предложил неравномерную сетку с такими шагами h_n, чтобы разность соседних шагов была много меньше самих шагов: $h_{n+1} - h_n = O(h_n^2)$. Но далекие интервалы могут при этом сильно различаться. Это позволяло строить сетки подробные в наиболее важных участках и редкие вдали от этих участков. Общее число интервалов и объем расчетов оказывались небольшими, а точность вычислений оставалась высокой. Он назвал эти сетки квазиравномерными.

В результате важные расчеты были проведены в срок, а взрыв первой советской термоядерной бомбы в 1953 г. блестяще подтвердил их точность.

Опубликована была эта идея много позднее. Но один из авторов (Н.Н. Калиткин) работает в коллективе А.А. Самарского с 1958 г. и познакомился с ней "из первых рук". Он систематически использовал ее в своей работе и включил в курс лекций, которые читал студентам физического факультета МГУ в 1970-е годы, и книгу "Численные методы" (1978 г.).

Для этого пришлось дать строгое математическое определение квазиравномерных сеток. Оказалось, что это определение обобщается на неограниченную область. Это позволило тогда же построить новые методы вычисления несобственных интегралов, а позднее — решения интегральных уравнений. Но лишь около 2000 г. удалось понять, как распространить эту идею на краевые задачи в неограниченной области. Сразу возникло очень много приложений к задачам, которые ранее с трудом решали искусственными приемами или вообще не умели решать.

Эту книгу мы писали для двух категорий читателей. Одна — это вычислители-практики, физики и инженеры, самостоятельно изучавшие численные методы, а также студенты и аспиранты, которым книга может служить учебным пособием. Для них мы подробно изложили метод сгущающихся сеток и практические приемы его использования. Все изложение построено просто, не требует углубленного знания математики и иллюстрировано большим количеством примеров. При этом подробно рассказывается и показывается, как надо строить алгоритм и программу, чтобы они давали ответ с гарантированной точностью и проводили диагностику ошибок.

Другая категория — это опытные математики-прикладники. Общие аспекты, описанные выше, им знакомы, но даже они могут найти полезное в той достаточно полной и тщательной методике решения прикладных задач, которая изложена в книге. Но наиболее интересными для них будут новые перспективы метода квазиравномерных сеток, особенно для случая неограниченной области. Здесь они найдут ряд новейших результатов, не все из которых даже успели попасть

в журналы, и применить эти подходы к решению своих собственных задач.

Книга разделена на главы, параграфы и пункты. Роль традиционного введения играет глава I. В каждой главе своя нумерация формул, рисунков, таблиц, примеров, теорем и определений (не по параграфам, а сплошная). Ссылки на формулы, рисунки и т. п. данной главы имеют следующий вид: формула (32), рис. 15. Если нужно сослаться на другую главу, то пишут: формула (IV, 32), рис. III.15. Если надо сослаться на текст в данном параграфе, то пишут: см. п. 3, на другой параграф этой же главы ссылаются так: см. § 1, п. 3; ссылка на другую главу имеет вид: гл. IV, § 1, п. 3. При ссылках на литературу пишутся фамилии всех или первого из авторов и год, например [Хайрер и др.; 1990].

В конце книги приведен список литературы. Для удобства он составлен в алфавитном порядке по фамилии первого автора. Список содержит учебные пособия и монографии, в которых есть главы, специально посвященные сгущению сеток. Из них только [Марчук, Шайдуров; 1979] целиком посвящена этой теме. Однако квазиравномерные сетки рассмотрены только в одной книге — [Калиткин; 1978]. Кроме того, в список литературы вошли статьи в журналах, где опубликован ряд оригинальных результатов, как уже ставших классическими, так и новейших. Основные методы и идеи, изложенные в книге, мы "обкатывали" в лекционных курсах, которые читали студентам физического факультета Московского государственного университета, Московского физико-технического института и Московского института электронной техники.

Некоторые интересные начально-краевые задачи в неограниченной области предложили нам А.Г. Свешников и Ю.Д. Плетнер. Совместно с авторами работали над приложениями и участвовали в расчетах примеров аспиранты и студенты Московского физико-технического института и Московского института электронной техники А.А. Болтнев, О.А. Качер, А.Б. Корягина, П.В. Корякин, К.И. Луцкий, А.С. Малистов, И.А. Панин, С.Л. Панченко, Н.М. Шляхов. В оформлении книги ощутимую помощь оказали сотрудники ИММ РАН Т.Г. Ермакова и Л.В. Кузьмина. Всем им мы выражаем искреннюю благодарность.

Многие методы, изложенные в этой книге, были разработаны в ходе выполнения проектов, поддержанных грантами РФФИ №№ 93-01-00861, 96-01-00305, 97-01-00005, 99-01-00082, 00-01-00151, 02-01-00066, 02-10-00253, 03-01-00439, Президентской программой поддержки научных школ и молодых кандидатов наук НШ-1918.2003.1, МК-1907.2004.9 и Фондом содействия отечественной науке.

Авторы

Глава I
О ЧИСЛЕННОМ АНАЛИЗЕ

Это вводная глава. В ней рассматриваются стадии решения естественнонаучных и технических задач: построение математической модели, разработка численного метода, расчет и обоснование его точности. Детально рассматривается природа погрешности и различные причины ее возникновения. Изложение сопровождается кратким историческим очерком.

§ 1. Математическое моделирование

1. Немного истории. Уже в глубокой древности считалось, что наука занимается непреложными и неизменными истинами: естественные науки открывают законы природы, а математика придает этим законам совершенную форму. Уже Архимед открыл закон рычага и выталкивающую силу ("Архимед открыл народу: тело, впернутое в воду, выпирает оттуды масса выпертой воды"). Сильнейшее подтверждение этой точки зрения дал Исаак Ньютон, опубликовавший в 1687 г. свой труд "Математические начала натуральной философии". Он сформулировал основные законы механики. Он же создал математический аппарат — дифференциальное и интегральное исчисления (одновременно с ним и даже чуть раньше то же сделал Готфрид Вильгельм Лейбниц), что позволило дать строгие математические формулировки задач и многие из них точно решить. Триумфом механики стали выведенные еще Ньютоном законы движения небесных тел.

С этого момента началось бурное развитие механики. Формулировались все новые частные задачи. Многие из них точно решались; для иных создавались численные методы решения (но об этом позднее). Стали развиваться другие разделы физики; для описания задач сплошной среды (теплопроводность, газодинамика и т. п.) был разработан аппарат уравнений в частных производных. Блестящим завершением этого периода стали уравнения Максвелла для электромагнитного излучения (их долго не хотели признавать, но эксперименты Генриха Герца по обнаружению электромагнитных волн и Петра Николаевича Лебедева по измерению давления света поставили здесь точку).

Так длилось до начала XX века. А затем...

Затем Хендрик Антон Лоренц, Анри Пуанкаре и Альберт Эйнштейн создали специальную теорию относительности. Классическая

ньютоновская механика оказалась частным случаем релятивисткой механики, справедливым при скоростях много меньших скорости света. То, что три века считалось точным законом природы, оказалось лишь неким приближением.

Через два десятилетия последовал новый удар. Эрвин Шредингер и Вернер фон Гейзенберг заложили основы квантовой механики. Ньютоновская механика оказалась частным случаем квантовой, когда размеры тел становятся много больше атомных. Наконец, Поль Дирак создал еще более общую теорию —релятивистскую квантовую механику, также немедленно подтвержденную экспериментами исключительно высокой точности. Кроме того, теория Дирака предсказала, что помимо электрона должна существовать частица с той же массой, но положительным зарядом — позитрон. И вскоре позитрон был обнаружен в экспериментах. Это убедило всех сомневающихся.

В результате ученые осознали, что абсолютно точных фундаментальных законов — истин в последней инстанции — нет. Любой закон является лишь приближенным описанием природы, частным случаем более общего закона (быть может, еще не открытого). Он достаточно точно выполняется лишь в определенных условиях, например: не слишком больших скоростях, не слишком малых размерах тел и т. п. Надо только четко сформулировать эти условия применимости и следить за тем, чтобы не выходить за их границы.

Но ведь приближениями пользовались намного раньше. Например, уравнения газодинамики (уравнения Эйлера) достаточно сложные; точно они не решаются, а решать их численно довольно трудоемко. Но если нас интересуют не любые движения масс газа, а лишь небольшие колебания давления, то удается приближенно заменить эти уравнения гораздо более простым уравнением акустики. Его не трудно точно решить, и оно превосходно описывает многие явления, такие как распространение звука. Но для описания сильного взрыва уравнение акустики уже не пригодно, т. к. здесь изменение давления велико, и нарушены условия применимости.

Такие приближения строились с XVIII века, неспешными в те годы темпами, проверялись и становились классическими, т. е. входили в золотой фонд науки и всесторонне изучались. Однако с 1940-х годов стремительное развитие техники, основанной на сложных физических и химических принципах, потребовало аккуратного проектирования и тщательного расчета новых конструкций. Пришлось разрабатывать специальное приближение для каждой конкретной задачи или явления. Такие приближения стали называть моделями данного явления.

2. Математическая модель. Различают две стадии построения модели. Сначала обсуждают разные стороны и процессы данного явления. При этом оценивают, какие процессы и факторы обязательно надо учесть, а какие — пренебрежимо малы и могут быть отброшены. Отобранные факторы составляют *предметную* (механическую,

физическую, химическую, биологическую, социологическую и т. п.) модель явления. Затем отобранные факторы описывают математическими уравнениями (алгебраическими, дифференциальными, интегральными и т. п.). Эту совокупность уравнений называют *математической* моделью.

Поясним это на примерах баллистических задач.

Пример 1. Пусть камень брошен со скоростью v_0 под углом α к горизонту с высоты y_0 (рис. 1). Учтем только силу притяжения,

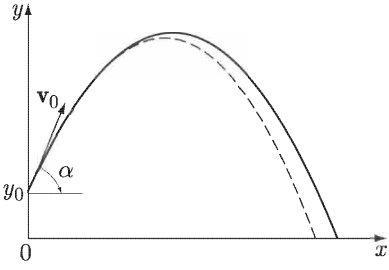

Рис. 1. Полет молота (сплошная линия) и волана (пунктир)

действующую вертикально вниз, это и есть физическая модель. Тогда согласно ньютоновской механике движение по горизонтали будет равномерным, а по вертикали — равноускоренным:

$$x = v_0 t \cos\alpha, \quad y = y_0 + v_0 t \sin\alpha - \frac{gt^2}{2}; \tag{1}$$

здесь t — время, $g \approx 10$ м/с^2 — ускорение свободного падения. Уравнения (1) являются математической моделью и одновременно дают решение в параметрической форме (роль параметра играет t). Можно исключить t из (1) и получить траекторию полета, которая оказывается параболой:

$$y = y_0 + x \operatorname{tg}\alpha - \frac{gx^2}{2v_0^2 \cos^2\alpha}. \tag{2}$$

Полагая $y = 0$, найдем дальность броска:

$$x_{\text{fin}} = v_0 \cos\alpha \frac{v_0 \sin\alpha + \sqrt{v_0^2 \sin^2\alpha + 2gy_0}}{g} \underset{y_0 \to 0}{\longrightarrow} v_0^2 \frac{\sin 2\alpha}{g}. \tag{3}$$

Дальность броска зависит от угла α. Можно даже явно найти оптимальный угол, обеспечивающий наибольшую дальность броска:

$$\sin\alpha_{\text{opt}} = \frac{1}{\sqrt{2(1 + gy_0/v_0^2)}} \underset{y_0 \to 0}{\longrightarrow} \frac{1}{\sqrt{2}}, \quad x_{\text{opt}} = \frac{v_0\sqrt{v_0^2 + 2gy_0}}{g} \underset{y_0 \to 0}{\longrightarrow} \frac{v_0^2}{g}; \tag{4}$$

при $y_0 = 0$ оптимальный угол $\alpha_{\text{opt}} = 45^\circ$ (что знают даже школьники), а при $y_0 > 0$ он меньше.

Эта модель очень проста. Но где она применима? Она хорошо описывает полет сравнительно небольших массивных тел с умеренными скоростями (бросок камня, толкание ядра, метание молота). Обработка спортивных киносъемок показывает, что лучшие метатели молота выпускают снаряд под углом 42–43° (бросок идет с высоты плеча $y_0 \approx 1.5$ м), а начальная скорость $v_0 \approx 20$ м/с обеспечивает дальность $x_{\text{opt}} \approx 41$–$42$ м.

Пример 2. Попробуем применить модель (1) к другим объектам. Круглая пуля времен нашествия французов имела скорость $v_0 > 100$ м/с и, согласно формуле (4), могла бы лететь на 1 км и более. Однако дальнобойность ружей тогда не превышала 200 м. Очень нагляден полет волана в бадминтоне: с какой скоростью его ни посылай, он дальше ~ 36 м не полетит; вдобавок хорошо видно, что траектория его полета не параболическая, а имеет более крутое снижение (пунктир на рис. 1).

Причина легко угадывается — надо учесть сопротивление воздуха. Его сила $\mathbf{F}$ направлена обратно скорости $\mathbf{v}$, а ее величина при средних (дозвуковых) скоростях примерно пропорциональна квадрату скорости, т. е. $F \approx -k(v)v$, где $k(v) \approx k_0 v$. Коэффициент k_0 зависит от размеров и формы тела и свойств воздуха (температуры и плотности). Примем эту физическую модель и запишем математическую модель — ньютоновские уравнения движения для координат x, y и компонент скоростей v_x, v_y:

$$\frac{dx}{dt} = v_x, \quad \frac{dy}{dt} = v_y, \quad \frac{dv_x}{dt} = -\frac{k(v)}{m} v_x, \quad \frac{dv_y}{dt} = -g - \frac{k(v)}{m} v_y, \tag{5}$$

где m — есть масса тела,

$$k(v) = k_0 v, \quad v = \left(v_x^2 + v_y^2\right)^{1/2}. \tag{6}$$

Уравнение надо дополнить начальными условиями при $t = 0$:

$$x(0) = 0, \quad y(0) = y_0, \quad v_x(0) = v_0 \cos\alpha, \quad v_y(0) = v_0 \sin\alpha. \tag{7}$$

Новая модель существенно сложнее, и решить задачу (5)–(7) явно уже не удается. Однако она существенно точнее. Нетрудно численно решить ее на компьютере и увидеть все те эффекты (несимметричность траектории и уменьшение дальности), о которых говорилось выше. Для малых начальных скоростей $k(v) \approx 0$, и модель (5)–(7) переходит в (1).

Заметим, что для модели (5)–(7) удается найти частный случай, где строится точное решение. Рассмотрим полет вертикально брошенного тела: $\alpha = 90°$. Тогда $v_x t \equiv 0$ и $xt \equiv 0$, а оставшиеся уравнения принимают следующий вид:

$$\frac{dy}{dt} = v, \quad \frac{dv}{dt} = -g \mp \frac{k_0}{m} v^2, \quad y(0) = y_0, \quad v(0) = v_0 > 0. \tag{8}$$

Здесь знак "−" соответствует стадии полета вверх, а знак "+" — стадии падения. Уравнение для скорости интегрируется точно: на стадии подъема

$$v(t) = \frac{1}{\sigma} \cdot \frac{\sigma v_0 - \operatorname{tg}(\sigma g t)}{1 + \sigma v_0 \operatorname{tg}(\sigma g t)}, \quad \sigma = \sqrt{\frac{k_0}{mg}}, \quad 0 \leqslant t \leqslant t_{up}, \tag{9}$$

где время подъема

$$t_{up} = \frac{1}{\sigma g} \operatorname{arctg}(\sigma v_0) \underset{v_0 \to \infty}{\to} \frac{\pi}{2\sigma g}. \tag{10}$$

На стадии падения

$$v(t) = -\frac{1}{\sigma} \operatorname{th}[\sigma g (t - t_{up})], \quad t \geqslant t_{up}, \quad v(+\infty) = -\frac{1}{\sigma}. \tag{11}$$

Время подъема оказалось конечным, сколь бы большой ни была начальная скорость, а скорость падения не превышает некоторой предельной, что кардинально отличается от модели (1). Высота подъема находится интегрированием скорости:

$$y(t) = y_0 + \int_0^t v(\tau) d\tau. \tag{12}$$

Этот интеграл также точно берется, но формулы различны для стадий подъема $0 \leqslant t \leqslant t_{up}$ и спуска $t_{up} \leqslant t$; первая из этих формул довольно громоздка. Значение $y_m = y(t_{up})$ определяет максимальную высоту подъема.

Пример 3. Современные винтовки и орудия изготавливают с высокой точностью. Обычная снайперская винтовка может поразить цель на 800 м, крупнокалиберная — на 1.5–2 км, а дальнобойное морское орудие при стрельбе на 30 км дает рассеивание ± 3 м. Однако для точного попадания в цель надо правильно установить прицел, т. е. верно определить угол α. Расчет этих углов делают по моделям типа (5), но дополненных еще рядом эффектов:

— начальные скорости v_0 достигают сейчас для пуль ~ 1 км/с, а для снарядов ~ 2 км/с, поэтому зависимость $k(v)$ становится гораздо более сложной;

— коэффициент k_0 зависит от плотности и температуры воздуха, которые могут меняться вдоль трассы полета, что требует внесения поправок;

— надо учитывать движение цели;

— надо вносить поправки на скорость и направление ветра;

— и еще много других поправок; так в морском бое вносят даже поправку на вращение Земли!

Раньше такие поправки заранее рассчитывались и печатались книжечкой, которую артиллеристы носили с собой. Но уже в первую мировую войну у моряков появились механические устройства для ав-

томатизации наводки. Сейчас соответствующие программы включены в компьютерные системы управления огнем. Но снайперы по-прежнему держат таблицу стрельб в голове.

Общая тенденция видна из рассмотренных примеров. Если учесть лишь немногие важнейшие эффекты, то может получиться достаточно простая модель. Ее нетрудно будет численно рассчитать, а то и получить явное решение. Однако модель будет грубой и применимой лишь к сильно ограниченному кругу явлений.

Если учесть слишком много эффектов, то модель окажется весьма точной, но очень сложной; неизвестно, сумеем ли мы провести по ней расчеты, т. е. найдутся ли подходящие алгоритмы и хватит ли мощности компьютера.

Поэтому хорошая модель — это разумный компромисс между требованиями к полноте и точности модели и нашими вычислительными возможностями. К счастью, быстрое развитие компьютеров позволяет постоянно улучшать модели.

Однако заметим, что это лишь тенденция, а не закон. Известны примеры, когда более простая модель оказывалась и более точной. Так, все средневековые астрономы предсказывали положение планет, солнечные и лунные затмения по геоцентрической модели Птолемея (планеты, Луна и Солнце движутся по малым кругам-эпициклам, центры которых вращаются вокруг Земли по большим кругам-циклам). Иоганн Кеплер предложил простую гелиоцентрическую модель с эллиптическими орбитами, оказавшуюся более точной. И только Ньютон объяснил, почему она правильна.

Так что построение хорошей модели — не просто наука, но и искусство.

Напрашивается вопрос: а зачем это нужно знать математику-прикладнику? Ведь он только решает предложенную математическую задачу, а формулируют модель другие. На этот вопрос ответим позже.

3. Модель–алгоритм–программа. После того, как математическая модель построена, ее уравнения надо решить. Для простейших моделей вроде (1) удается получить решение в явном виде. Большинство моделей требует численного решения, и надо выбрать или построить алгоритм расчета.

Для несложных моделей удается обойтись описанными в учебниках алгоритмами: одним или комбинацией нескольких. Так, для системы дифференциальных уравнений (5)–(7) хорошие результаты даст численное интегрирование по явным схемам Рунге–Кутты, желательно 4-го порядка точности. Можно воспользоваться стандартными программами, имеющимися во многих пакетах математических программ. Но при этом надо четко представлять, какова погрешность расчета, предусмотрен ли в программе контроль точности, и насколько можно ему доверять. Во многих случаях этот программный контроль оказывается иллюзией (примеры этого будут приведены в главе II).

Для сложных моделей имеющиеся алгоритмы могут оказаться непригодными или малоэффективными. Тогда приходится разрабатывать оригинальные алгоритмы, обосновывать их точность и отлаживать программу. Для контроля правильной работы алгоритмов и программ полезно использовать частное точное решение вроде (9)–(11), если его удается найти. Далее в книге будут показаны некоторые приемы такого контроля. Есть и способы проверки, не требующие знания точных решений.

Наконец программа надежно отлажена и начались расчеты. Но успокаиваться еще рано. Надо вернуться к исходному явлению и взять несколько экспериментов, проведенных в заметно различающихся условиях (например, броски с разными скоростями v_0 под разными углами α). Если все расчеты хорошо совпадут с экспериментально измеренными длинами или киносъемками траекторий, то мы справились с задачей. Если же оказались заметные расхождения...

Тогда надо заново проверять все. При этом заказчик уверен, что его модель правильна, а виноват математик, который неправильно написал алгоритм или программу ("В любой сколь угодно малой программе есть по меньшей мере одна ошибка"). На самом деле и модель может чего-то существенного не учесть. Даже в не такой уж сложной модели полета снаряда (пример 3) список факторов можно продолжить: вращение снаряда, прецессия оси вращения, зависимость ускорения свободного падения g от высоты y (меняется расстояние от центра Земли), падение плотности атмосферы с высотой. Поэтому математик должен разобраться и в модели. А возможно, контрольные эксперименты были проведены неаккуратно, в них тоже надо разобраться.

§ 2. Источники погрешности

Термины "численные методы" и "приближенный анализ" — синонимы. Всякий раз точная задача заменяется приближенной. Скажем, по заданной величине w нужно вычислить u. Символически запишем операцию так:

$$u = A(w). \tag{13}$$

Пример 4. $u = A(w) = \int\limits_a^b w(x)dx$. Интегралы даже от простых комбинаций элементарных функций далеко не всегда удается взять точно. Возможны следующие способы упрощения задачи.

1) $w(x) \approx \widetilde{w}(x) = P(x)$; по теореме Вейерштрасса всякую гладкую функцию можно приблизить полиномами, а интегралы от полиномов берутся точно.

2) Можно приближенно заменить интеграл интегральной суммой:

$$\int_a^b w(x)dx = A \approx \widetilde{A} = \sum_i w(x_i)h_i;$$

$$\widetilde{A} \to A \quad \text{при} \quad h \to 0.$$

3) Можно взять комбинацию способов $w(x) \approx \widetilde{w}(x)$ и $A \approx \widetilde{A}$.

Будем считать, что и решение приближенной задачи близко к точному $\widetilde{u} \approx u$. Но как оценить точность $\widetilde{u} - u$? Для этого понадобится понятие нормы.

1. Величины и нормы. Как исходные данные, так и решение могут быть величинами различных типов. Это могут быть числа u, w; векторы разной размерности $\mathbf{u} = \{u_p,\ 1 \leqslant p \leqslant P\}$, $\mathbf{w} = \{w_q,\ 1 \leqslant q \leqslant Q\}$; матрицы; функции одной переменной $u(x)$, $w(y)$ или многих переменных; вектор-функции и т. п. При этом аргумент (аргументы) функции может быть непрерывным $a \leqslant x \leqslant b$ или дискретным $x \in \Omega$, где $\Omega = \{x_n,\ 1 \leqslant n \leqslant N\}$ есть некоторая сетка.

Проиллюстрируем сказанное некоторыми примерами.

- Решается уравнение с одним неизвестным; u, w суть числа, вещественные или комплексные.
- Решается система N линейных или нелинейных уравнений с таким же числом неизвестных; u, w суть векторы одинаковой размерности N.
- Ищется определенный интеграл от $w(x)$; последняя есть функция непрерывного аргумента, u есть число.
- Строится сплайн-аппроксимация функции, табулированной на сетке Ω; исходные данные — функция дискретного аргумента $w(x_n)$, которую также можно рассматривать как вектор $\{w_n\}$; решение $u(x)$ есть функция непрерывного аргумента.
- Решается дифференциальное уравнение $du/dx = w(u, x)$; исходные данные — непрерывная функция двух аргументов $w(u, x)$, решение есть непрерывная функция одного аргумента $u(x)$. Однако для численного интегрирования вводится сетка $\{x_n\}$, т. ч. численное решение оказывается функцией дискретного аргумента $u(x_n)$.

Количественной мерой точности является норма погрешности. Укажем некоторые наиболее употребительные нормы. Для числа u есть единственная норма

$$\|u\| = |u|.$$

Для ограниченных функций $u(x)$, $x \in [a, b]$, вводится чебышевская норма

$$\|u\|_C = \max_{x \in [a,b]} |u(x)|, \tag{14}$$

а для функций, интегрируемых с квадратом с весом $\rho(x)$, — гильбертова норма

$$\|u\|_{L_2} = \left[\int_a^b u^2(x)\rho(x)dx\right]^{1/2}, \quad \rho(x) > 0. \tag{15}$$

Для функций дискретного аргумента $u(x_n)$ или векторов $\{u_n\}$ вводят дискретные аналоги норм (14), (15):

$$\|u\|_c = \max_{1\leqslant n\leqslant N} |u_n|, \quad \|u\|_{l_2} = \left(\sum_{n=1}^{N} \rho_n u_n^2\right)^{1/2}, \quad \rho_n > 0. \tag{16}$$

Для матриц употребляют еще большее число норм.

Видно, что для одной и той же величины могут использоваться разные нормы. Между ними существуют определенные соотношения. Для функций непрерывного аргумента это односторонние неравенства; так

$$\|u\|_C \cdot \left[\int_a^b \rho(x)dx\right]^{1/2} \geqslant \|u\|_{L_2}.$$

При этом из малости нормы, стоящей в левой части, следует малость нормы, стоящей в правой части, но не наоборот; то же относится к сходимости методов в данных нормах. Первую норму называют более *сильной*, чем вторую. Наглядное отличие между нормами (14) и (15) таково: малость нормы C означает, что $u(x)$ мала во всех точках $[a, b]$; малость нормы L_2 означает, что $u(x)$ мала почти во всех точках, но на незначительной части $[a, b]$ может быть не мала.

Для функций дискретного (конечномерного) аргумента неравенства между нормами двусторонние. Например, для (16) выполняется

$$\|u\|_c \cdot \left(\sum_{n=1}^{N} \rho_n\right)^{1/2} \geqslant \|u\|_{l_2} \geqslant \|u\|_c \cdot \sqrt{\min_{1\leqslant n\leqslant N} \rho_n}.$$

Поэтому из сходимости в одной норме следует сходимость в другой. Такие нормы называют *эквивалентными*. В конечномерных пространствах все нормы эквивалентны, но в бесконечномерном пространстве это не так.

Различают три источника погрешности решения: погрешность исходных данных, погрешность метода и погрешность округлений. Рассмотрим их подробнее.

2. Неустранимая погрешность. Та часть погрешности решения, которая обусловлена ошибками исходных данных δw задачи (13), равна

$$\delta u = \frac{dA}{dw} \cdot \delta w, \quad \|\delta u\| \leqslant \left\|\frac{dA}{dw}\right\| \cdot \|\delta w\|. \tag{17}$$

Она зависит только от исходной задачи (13) и ошибок начальных данных, т. е. никакое искусство вычислителя не может ее уменьшить. Поэтому ее называют неустранимой. Она тем больше, чем хуже обусловленность задачи (13) и чем больше погрешность начальных данных. Для оценки ее величины надо выяснить величины обоих факторов, т. е. $\|\delta w\|$ и $\|dA/dw\|$.

Задача (13) может быть чисто математической либо обработкой экспериментальных данных. Как ни удивительно, качественный характер исходных данных и их погрешности в обоих случаях схожи. Математические данные могут быть функцией $w(x)$ или таблицей w_n; последнее обычно, когда непосредственное вычисление $w(x)$ настолько трудоемко, что сделано заранее на некоторой сетке. Эксперимент дает не только таблицу w_n; существует много экспериментов, регистрирующих функцию $w(x)$, например осциллограммы.

Погрешность в каждой точке состоит из систематической и случайной ошибок. Для экспериментальных данных это общеизвестно. Но это справедливо и для математических задач, если w само является результатом трудоемкого вспомогательного математического расчета. Тогда роль систематической ошибки играет погрешность вспомогательного метода, а роль случайной ошибки — ошибка округления компьютерных вычислений (см. п. 3 и п. 4).

Величину погрешности удобно характеризовать отношением $\|\delta w\|/\|w\|$. Для простейших функций, математически вычисляемых на (80–64)-разрядных компьютерах, она может составлять 10^{-20}–10^{-16}. При сложных математических вычислениях относительная погрешность возрастает до 10^{-8}–10^{-3}. Для экспериментально измеряемых величин даже в астрономии и геодезии точность лучше 10^{-6} редко достигается; в технике она обычно составляет 10^{-4}–10^{-2}, а в передовых областях науки (физика плотной плазмы, химическая кинетика и т. п.) может ухудшаться до 0.1 и более.

Для оценки неустранимой погрешности (17) надо знать $\|dA/dw\|$. Ее не часто удается определить теоретически. На практике можно применять следующий способ. Решим задачу (13) несколько раз, искусственно прибавляя к w различные вариации δw_j. В результате решения получим соответствующие им вариации δu_j. Составим отношения $c_j = \|\delta u_j\|/\|\delta w_j\|$; если они близки по порядку величины, то их среднее значение можно принять за $\|dA/dw\|$.

Описанный способ трудоемок и нестрог, ибо в нем оцениваются одновременно устойчивости задачи (13) и алгоритма ее решения. Поэтому его используют редко. К сожалению, чаще всего ограничиваются интуитивной оценкой обусловленности, что ненадежно.

3. Погрешность метода. Многие алгоритмы строят так, чтобы у них были управляющие параметры. Например, у итерационного алгоритма это число итераций q, у разностного — шаг сетки h. Алгоритм строят так, чтобы при стремлении параметра к некоторому пределу

($q \to \infty$, $h \to 0$ и т. п.) численное решение стремилось бы к точному. Отличие численного решения от точного при конкретном значении параметра называют погрешностью метода.

Сам факт сходимости и скорость сходимости устанавливают теоретическими исследованиями для каждого метода отдельно. Есть и некоторые полуэмпирические способы исследования сходимости и ее скорости (например, по расчетам на сгущающихся сетках). Все это позволяет оценивать погрешность метода в конкретных расчетах и выбирать параметры метода для обеспечения заданной точности.

Параметры целесообразно выбирать так, чтобы погрешность метода была меньше неустранимой погрешности в ~ 10 раз. Заметно большая погрешность ухудшает общую точность; заметно меньшая — не улучшает общей точности, но увеличивает трудоемкость расчетов.

Есть методы, дающие точный ответ за конечное число действий. Например, это явное решение уравнения в элементарных функциях, или решение систем линейных уравнений методом Гаусса. В них отсутствует погрешность метода.

4. Погрешность округления. Все числа записываются и операции на ними производятся с конечным числом знаков, т. е. с ошибками. Число x с ошибкой записывают как

$$x \pm \Delta \quad \text{или} \quad x(1 \pm \delta),$$

где Δ называют абсолютной ошибкой, а δ — относительной. Разумеется, это не точные значения ошибок; ошибки являются случайными величинами (их распределения можно считать гауссовыми), а Δ, δ — их стандартными уклонениями (стандартами). При сложении или вычитании чисел складываются квадраты стандартов их абсолютных ошибок:

$$(x_1 \pm \Delta_1) \pm (x_2 \pm \Delta_2) \to (x_1 \pm x_2) \pm \sqrt{\Delta_1^2 + \Delta_2^2}\,, \tag{18}$$

а при умножении или делении то же касается относительных ошибок:

$$[x_1(1 \pm \delta_1)] \cdot [x_2(1 \pm \delta_2)]^{\pm 1} \to (x_1 \cdot x_2^{\pm 1})\bigl(1 \pm \sqrt{\delta_1^2 + \delta_2^2}\bigr). \tag{19}$$

Ошибка округления при компьютерной записи числа составляет половину последнего разряда мантиссы. Для компьютеров с 32-, 64- и 80-разрядными числами это составляет $\delta_0 \approx 10^{-8}$, 10^{-16} и 10^{-20} соответственно. При выполнении N последовательных умножений и делений относительная ошибка, согласно (19), увеличивается в $\sqrt{N}$ раз. Даже при огромных числах действий, которые выполняют современные компьютеры, $\delta_0 \cdot \sqrt{N}$ остается небольшой величиной. Казалось бы, ошибками округления в компьютерах можно пренебречь, и учитывать их лишь при “ручных” расчетах с малым числом знаков.

Однако при сложениях и вычитаниях (18) величина $x_1 \pm x_2$ может стать очень малой и сопоставимой с ошибкой. Это наверняка случится

при плохой обусловленности задачи (13), но может произойти и при хорошей обусловленности задачи, но неудачном построении алгоритма. Такие примеры будут приведены далее. Поэтому даже при вычислениях на многоразрядных компьютерах нельзя забывать об ошибках округления.

Основная рекомендация такова: суммарные ошибки округления должны быть менее погрешности метода, по крайней мере в ~ 10 раз. Для этого следует выбирать многоразрядный компьютер, проводить расчеты с двойной точностью и обращать внимание на тонкости математического обеспечения: не любое математическое обеспечение позволяет полностью использовать все разряды, имеющиеся в процессоре.

Глава II

СГУЩЕНИЕ РАВНОМЕРНЫХ СЕТОК

В этой главе рассмотрены принципы вычислений на сетках и показано, как можно оценить их погрешности и (или) повысить точность, если выполнить расчеты на нескольких равномерных сетках с разным числом узлов. Изложены практические приемы контроля точности и диагностики ошибок, полезные для составления и отладки программ.

§ 1. Точность сеточных методов

1. Вычисление на сетках. В задачах математического анализа возникают функции $u(x)$, с которыми требуется что-то сделать: проинтегрировать, продифференцировать, решить интегральное или дифференциальное уравнение, которому удовлетворяет эта функция. Если не удается точно решить до конца задачу, то обращаются к численному анализу.

В численном анализе мы производим расчеты, т. ч. в принципе не можем использовать значение функции во всех точках области ее определения (этих точек бесконечно много). Мы вынуждены ограничиться конечным числом точек, т. е. сеткой. Введем необходимые понятия, рассматривая для простоты функции одной переменной.

Пусть задана функция $u(x)$ на отрезке $a \leqslant x \leqslant b$. Это можно также записать так: $x \in G$, $G = [a, b]$. Введем в этой области сетку Ω_N с узлами x_n:

$$\Omega_N = \{x_n, 0 \leqslant n \leqslant N; a = x_0 < x_1 < x_2 < \ldots < x_N = b\}. \qquad (1)$$

Точки x_n назовем *узлами* сетки. *Шагом* сетки назовем величину

$$h_n = x_n - x_{n-1}, \quad 1 \leqslant n \leqslant N. \qquad (2)$$

Равномерной называют сетку, все шаги которой одинаковы:

$$h_n \equiv h = \frac{b-a}{N}. \qquad (3)$$

Сетку с неодинаковыми шагами называют неравномерной. Отрезок

$$x_{n-1} \leqslant x \leqslant x_n, \quad 1 \leqslant n \leqslant N, \qquad (4)$$

называют n-м *интервалом* сетки. Вводят также середину интервала:

$$x_{n-1/2} = \frac{x_n + x_{n-1}}{2}; \quad 1 \leqslant n \leqslant N. \tag{5}$$

Рассматриваемую функцию $u(x)$ при этом вычисляют в целых или полуцелых узлах сетки:

$$u_n \equiv u(x_n), \quad u_{n-1/2} \equiv u(x_{n-1/2}). \tag{6}$$

По аналогии можно ввести любую дробную долю интервала и значения функции в них, если это потребуется.

Замечание 1. Определение (5) полуцелых узлов (середин интервалов) используют на равномерных и произвольных неравномерных сетках. Но в главе III мы увидим, что для одного специального вида неравномерных сеток выгодно дать иное определение.

Замечание 2. Иногда узлы x_0 и (или) x_N ставят не на границах области $[a, b]$, а внутри ее, полагая $a < x_0$ или $x_N < b$. Обычно при этом размещают вне области дополнительные узлы $x_{-1} < a$ или $x_{N+1} > b$ и экстраполируют $u(x)$ вне области в эти узлы. Это бывает полезно для некоторых видов краевых задач.

Многомерность. Для функции многих переменных $u(\mathbf{x})$, заданной в многомерной области G, также вводят сетки. Если G есть многомерный прямоугольный параллелепипед, в нем нетрудно ввести прямоугольную сетку; для случая двух переменных она выглядит следующим образом:

$$\begin{gathered} G = [a \leqslant x \leqslant b, \quad \alpha \leqslant y \leqslant \beta], \\ \Omega_{NM} = \big\{ (x_n, y_m); 0 \leqslant n \leqslant N, \quad 0 \leqslant m \leqslant M \big\}, \\ a \leqslant x_0 < x_1 < \ldots < x_N \leqslant b, \quad \alpha \leqslant y_0 < y_1 < \ldots < y_M \leqslant \beta. \end{gathered} \tag{7}$$

Обобщением понятия интервала будет ячейка сетки:

$$[x_{n-1} \leqslant x \leqslant x_n, \quad y_{m-1} \leqslant y \leqslant y_m], \quad 1 \leqslant n \leqslant N, \quad 1 \leqslant m \leqslant M, \tag{8}$$

а целыми узлами сетки будут точки (x_n, y_m). Значение функции в узлах обозначают

$$u_{nm} = u(x_n, y_m). \tag{9}$$

Аналогично можно ввести полуцелые узлы.

Эти выражения естественно обобщаются на любое число измерений. Многие области более сложной формы (цилиндр, сферу, эллипсоид и т. п.) можно преобразовать в прямоугольный параллелепипед соответствующей заменой переменных и в новых координатах ввести сетку (7). Это эквивалентно введению криволинейных координат в области G.

Узлы сетки типа (7) образованы пересечением семейств координатных линий (в многомерном случае — гиперплоскостей); такие сетки называются *регулярными*. Ячейкой регулярной сетки может быть не

только криволинейный прямоугольник, но и треугольник или любой многоугольник (многогранник в многомерном случае). К таким сеткам в полной мере относятся те соображения о сгущении сеток и оценках точности, которые рассмотрены в данной книге.

Но в многомерном случае есть и такие ситуации, когда узлы сетки — это как угодно выбранные точки области G. Тогда ближайшие точки каким-то образом соединяют, разбивая область G на ячейки-многогранники (это само является непростой задачей). Такие сетки называют *нерегулярными*; их мы не будем рассматривать в данной книге.

Методы вычисления на сетках стали интенсивно развиваться с конца XVII века. Уже Исаак Ньютон и его современники предложили ряд формул для интерполирования, численного интегрирования и дифференцирования. Леонард Эйлер написал первую схему численного решения обыкновенных дифференциальных уравнений (схему ломаных) и первую разностную схему для уравнения в частных производных (явная схема Эйлера для уравнения теплопроводности), которую вычислители использовали вплоть до 1940-х годов. Первые методы были несложными, но математики продолжали конструировать более сложные схемы, добиваясь высокой точности.

Однако как проконтролировать эту точность? С конца 1790-х годов математики стали уделять много внимания оценке погрешности вычислений. Жозеф Луи Лагранж построил оценку погрешности ряда Тейлора, а Огюстен Коши обосновал сходимость схемы ломанных Эйлера. Полученные оценки были априорными, в них погрешность выражалась через некоторую производную $u(x)$. Это было не слишком удобно при практическом применении.

Практикам больше подошел способ, предложенный Карлом Давидом Тольме Рунге в 1985 г. для численного интегрирования обыкновенных дифференциальных уравнений по его новой схеме. Рекомендовалось провести два расчета: на сетке с шагом h и затем с вдвое меньшим шагом. Оба расчета сравнивались, и совпадающие знаки считались верными. Льюис Фрай Ричардсон, занимавшийся уравнениями в частных производных, в 1910 г. усовершенствовал этот способ (но полное объяснение дал лишь в 1927 г.). Он показал, как по двум расчетам с шагами h и $h/2$ по схеме p-го порядка точности можно получить хорошую оценку погрешности, а также повысить точность результата. Эти способы не требовали знания производных $u(x)$ и давали апостериорную оценку.

Позднее такие процедуры были построены для произвольного числа расчетов на последовательно сгущающихся сетках. Они позволяют получать результат, по погрешности эквивалентный расчету по схеме гораздо более высокого порядка точности, чем исходная схема. Общие формулы такого расчета сравнительно громоздки. Но Ромберг в 1955 г. нашел простой вариант этих формул, если при каждом очередном сгущении шаг сетки уменьшается ровно вдвое.

Эти методы применимы практически к любым задачам с сеточным представлением функции: к интерполяции, численному дифференцированию, численному интегрированию, решению задач Коши и краевых задач для обыкновенных дифференциальных уравнений, задач для уравнений в частных производных, для интегральных и интегродифференциальных уравнений. Более того, до сих пор это единственный способ получить для сеточных решений асимптотически точную оценку погрешности. Они служат надежной основой для построения программ с контролем точности. В этой главе будет показано, как это делается.

Однако классическая формулировка этого метода рассчитана на равномерные сетки. Практические же расчеты зачастую требуют неравномерных сеток, адаптированных к конкретным задачам. В главе III будет рассмотрен важный класс неравномерных сеток, на которых метод сгущения полностью распространяется.

2. Погрешность сеточных методов. Рассмотрим основные типичные сеточные формулы и их погрешности на примерах простейших квадратурных формул. Пусть надо вычислить определенный интеграл

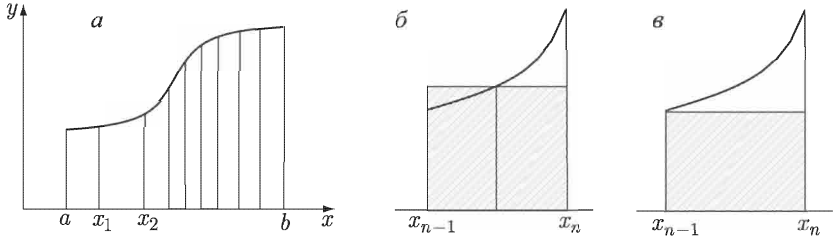

Рис. 1. Численное интегрирование: а) введение сетки, б) формула левых прямоугольников, в) формула средних

от некоторой функции $u(x)$, которая предполагается интегрируемой (более детальные требования к ней сформулируем позже). Для этого введем на отрезке $[a, b]$ некоторую произвольную неравномерную сетку Ω_N (1) и представим интеграл как сумму интегралов по интервалам сетки (рис. 1, $а$):

$$U = \int_a^b u(x)dx = \sum_{n=1}^{N} \int_{x_{n-1}}^{x_n} u(x)dx. \tag{10}$$

Длина каждого интервала равна h_n. Выберем произвольно в каждом интервале по точке $\xi_n \in [x_{n-1}, x_n]$ и заменим интеграл (10) интегральной суммой:

$$U_N = \sum_{n=1}^{N} u(\xi_n)\, h_n, \quad \xi_n \in [x_{n-1}, x_n], \quad U_N \underset{\max h_n \to 0}{\longrightarrow} U. \tag{11}$$

Как известно, для любой интегрируемой функции (в частности, для кусочно-непрерывной) интегральная сумма сходится к интегралу, если длины всех интервалов стремятся к нулю. Таким образом, (11) есть пример сходящегося сеточного метода.

Но как оценить скорость сходимости, т. е. дать оценку погрешности? Для произвольной интегрируемой функции $u(x)$ это, вообще говоря, нельзя сделать. Надо наложить на $u(x)$ более жесткие ограничения, например потребовать непрерывности или кусочной непрерывности самой $u(x)$ и некоторого количества ее производных. Для упрощения всех выводов мы обычно будем начинать с жесткого ограничения: $u(x)$ *есть достаточно гладкая функция*, т. е. имеет столько непрерывных и ограниченных производных, сколько понадобится по ходу изложения. Мы будем также конкретизировать это требование: $u(x)$ *есть* p *гладкая функция*, т. е. она и ее p производных непрерывны и ограничены. Математики говорят: "p раз непрерывно дифференцируемая функция", но это длиннее. А затем мы будем ослаблять требования, а именно уменьшать p и заменять непрерывность на кусочную непрерывность и выяснять, к чему это приведет.

Далее будем использовать обозначения:

$$M_q = \|u^{(q)}\|_C = \max_{x\in[a,b]} |u^{(q)}(x)|, \quad q = 0, 1, 2, \ldots; \tag{12}$$

$q = 0$ соответствует самой функции.

Левые прямоугольники. Выберем в качестве ξ_n левую границу интервала: $\xi_n = x_{n-1}$. Это соответствует заштрихованному прямоугольнику на рис. 1, *б*, а интегральную сумму (11) в этом случае называют формулой левых прямоугольников:

$$U_N = \sum_{n=1}^{N} u_{n-1} h_n; \tag{13}$$

аналогично можно написать формулу правых прямоугольников. Исследуем погрешность формулы (13), предполагая $u(x)$ достаточно гладкой.

Для этого в каждом интервале сетки разложим функцию по формуле Тейлора, беря левый конец интервала за центр разложения:

$$u(x) = \sum_{q=0} \frac{1}{q!} u^{(q)}(x_{n-1})(x - x_{n-1})^q; \tag{14}$$

количество членов суммы (14) определяется гладкостью функции. Если проинтегрировать ряды (14) в своих интервалах, т. е. от x_{n-1} до x_n, то интегрирование главных членов ($q = 0$) дает формулу левых прямоугольников (13). Значит, интегрирование остальных членов дает

погрешность этой формулы:

$$R_N \equiv U_N - U = -\sum_{n=1}^{N} \sum_{q=1} \frac{h_n^{q+1}}{(q+1)!} u_{n-1}^{(q)}. \tag{15}$$

Напомним, что это выражение погрешности справедливо для произвольной неравномерной сетки.

Запишем формулы (13) и (15) для случая равномерной сетки. Меняя в (15) порядок суммирования, получим:

$$U_N = h \sum_{n=1}^{N} u_{n-1}, \quad R_N = \sum_{q=1} A_q h^q, \quad h = \text{const}, \tag{16}$$

где коэффициенты A_q сами являются интегральными суммами и приближенно заменяются интегралами, которые точно берутся, а именно

$$A_q = -\frac{h}{(q+1)!} \sum_{n=1}^{N} u_{n-1}^{(q)} = -\frac{1}{(q+1)!} \int_a^b u^{(q)}(x)dx + O(h) =$$

$$= \frac{u^{(q-1)}(a) - u^{(q-1)}(b)}{(q+1)!} + O(h). \tag{17}$$

Таким образом, на равномерной сетке погрешность формулы левых прямоугольников (13) раскладывается в ряд по степеням шага сетки h. Этот ряд содержит *все* степени $q \geqslant 1$ шага h. Главный член ошибки пропорционален h, т. е. $R_N = O(h)$, а формула (13) имеет первый порядок точности.

Оценка (16) записывается до выполнения вычислений, поэтому ее называют априорной. Запишем главный член ошибки:

$$R_N \approx A_1 h \approx \frac{h}{2}[u(a) - u(b)]. \tag{18}$$

Когда $h \to 0$, остальные члены погрешности (16) гораздо меньше этого члена, и ими можно пренебречь. Поэтому оценку (18) называют асимптотически точной или просто *асимптотической*. Чтобы вывод этой оценки был справедлив, ряды (16) и (14) должны содержать член с $u^{(1)}(x)$, т. е. $u(x)$ должна иметь непрерывную и ограниченную 1-ю производную.

Построение априорной оценки (18) позволяет проводить вычисления с гарантированной погрешностью ε. Для этого достаточно выбрать такой шаг h, чтобы выполнялось $|R_N| \leqslant \varepsilon$ согласно формуле (18). Поскольку эта оценка асимптотическая, то требуя $|R_N| \approx \varepsilon$, мы получим оптимальное значение шага: больший шаг приведет к превышению уровня погрешности, меньший шаг неоправданно увеличит объем вычислений.

Сделаем существенное замечание. Приближенная замена интегральных сумм в (17) интегралами правильна. Однако нельзя просто так подставлять эти интегралы в (16): ведь замены делались приближенно с точностью $O(h)$, и каждая такая поправка имеет тот же порядок, что и следующий член в формуле (16).

Формула средних. Рисунок 1, *б* показывает, что формула прямоугольников несимметрична. Построим симметричную квадратурную формулу. Для этого в качестве ξ_n возьмем середину интервала: $\xi_n = x_{n-1/2}$. Это по-прежнему дает формулу средних на произвольной неравномерной сетке:

$$U_N = \sum_{n=1}^{N} h_n u_{n-1/2}. \tag{19}$$

Для определения ее погрешности используем в каждом интервале локальные разложения с центрами в серединах интервалов:

$$u(x) = \sum_{q=0} \frac{1}{q!} u^{(q)}_{n-1/2} (x - x_{n-1/2})^q, \quad x_{n-1} \leqslant x \leqslant x_n. \tag{20}$$

Подстановка главных членов рядов (20) в интегралы по-прежнему дает выбранную квадратурную формулу. Однако далее все члены с нечетными q обращаются в нуль при интегрировании благодаря симметрии, и выражение для погрешности содержит только четные члены:

$$R_N = -\sum_{n=1}^{N} \sum_{q=1} \frac{h_n^{2q+1}}{2^{2q}(2q+1)!} u^{(2q)}_{n-1/2}. \tag{21}$$

На равномерной сетке оценка погрешности принимает следующий вид:

$$R_N = \sum_{q=1} B_q h^{2q}, \quad h = \text{const}. \tag{22}$$

Здесь коэффициенты также суть интегральные суммы, приближенно заменяемые интегралами:

$$B_q = -\frac{h}{2^{2q}(2q+1)!} \sum_{n=1}^{N} u^{(2q)}_{n-1/2} = -\frac{1}{2^{2q}(2q+1)!} \int_a^b u^{(2q)}(x)dx + O(h^2) =$$

$$= \frac{u^{(2q-1)}(a) - u^{(2q-1)}(b)}{2^{2q}(2q+1)!} + O(h^2). \tag{23}$$

Если ограничиться первым членом ряда (22), то получим априорную асимптотическую оценку погрешности:

$$R_N \approx \frac{h^2}{24} [u'(a) - u'(b)] = O(h^2). \tag{24}$$

Формула средних имеет второй порядок точности, что существенно лучше, чем для формулы прямоугольников. Другое важное отличие заключается в том, что ряд (16) содержит все степени h подряд, а ряд (22) только четные степени.

Априорную оценку (24) также можно использовать для выбора шага h, обеспечивающего заранее заданную точность расчета. Правда, это немного менее удобно, чем для формулы прямоугольников: там в оценку входили только значения $u(x)$, которые мы все равно должны вычислять; здесь же надо дополнительно вычислить $u'(x)$, что не всегда желательно. Это намек на то, что в других случаях априорные оценки могут оказаться еще более сложными и менее удобными для практического использования.

Итоги. Мы рассмотрели хотя и простейшие, но типичные образцы сеточных формул и априорных оценок их погрешностей. На равномерных сетках при достаточно гладких функциях погрешность разлагается в ряд по степеням шага h. Если главный член ряда погрешности есть $O(h^p)$, то сеточная формула имеет p-й порядок точности. Ряд для погрешности может содержать все степени h подряд, начиная с p-й, но может содержать лишь степени той же четности, что и p (первое типично для несимметричных сеточных формул, второе — для симметричных). Число членов ряда погрешности (точнее, максимальная степень шага) зависит от гладкости функции.

3. Гладкость и насыщение. Как влияет степень гладкости функции на точность сеточных формул и оценку погрешности? Рассмотрим это подробно на примере формулы прямоугольников (13).

Если $u(x)$ имеет ограниченную и непрерывную 1-ю производную, то член с этой производной остается в ряде Тейлора (14) и остаточном члене (16), причем интеграл от $u'(x)$ в (17) можно точно взять, что и было сделано. В итоге оказывается справедливой априорная асимптотическая оценка (18), означающая погрешность $O(h)$.

Откажемся от требования непрерывности 1-й производной, но сохраним требование ограниченности. Тогда первым членом ряда Тейлора (14) можно пользоваться только в тех интервалах, где $u'(x)$ непрерывна. В тех интервалах, где $u'(x)$ разрывна, приходится ограничиться лишь более грубой оценкой:

$$|u(x) - u_{n-1}| \leqslant (x - x_{n-1}) \max |u'(x)|, \quad x_{n-1} \leqslant x \leqslant x_n. \tag{25}$$

Теперь, во-первых, после подстановки (25) в остаточный член (16), (17) в соответствующих интегралах суммы (18) для A_1 вместо u'_{n-1} возникает $\max |u'(x)|$ в этом интервале. Во-вторых, на практике не часто удается установить, в каких именно интервалах u' разрывна. Поэтому фактически приходится во всех интервалах заменять u'_{n-1} на $M_1 = \|u'\|_C$. Тогда вместо асимптотической оценки (18) получаем

мажорантную, хотя тоже априорную оценку:

$$|R_N| \leqslant \frac{h}{2}(b-a)M_1 = O(h), \quad M_1 = \max_{x\in[a,b]} |u'(x)|. \tag{26}$$

Хотя точность по-прежнему есть $O(h)$, но мажорантная оценка хуже асимптотической. Во-первых, обычно она существенно завышена, т. ч. априорный выбор h для получения требуемой точности ε на основании (26) приводит к существенному уменьшению шага, т. е. к неоправданному увеличению объема вычислений. Вторую же причину обсудим далее, когда рассмотрим метод сгущения сеток.

Еще ослабим требования. Будем считать $u(x)$ ограниченной и непрерывной, но недифференцируемой функцией. Тогда пользоваться рядом Тейлора вообще нельзя, и оценить остаточный член невозможно. Сумма (13) сходится к интегралу, поскольку она интегральная, но скорость сходимости хуже $O(h)$ и неизвестна. Далее на численном примере это будет проиллюстрировано.

Рассмотрим обратную ситуацию. Пусть $u(x)$ имеет большее число непрерывных производных. Тогда в рядах (14) и (16) имеется больше членов, но оценка (18) остается прежней. Более высокого порядка точности мы не получаем.

Таким образом, для формулы прямоугольников (13) есть некоторая оптимальная гладкость $p = 1$ функции $u(x)$, обеспечивающая предельную точность $O(h)$. Если гладкость $u(x)$ хуже оптимальной, то фактический порядок точности формулы хуже предельного. Но если гладкость лучше оптимальной, то порядок точности не повышается. Это свойство сеточной формулы (13) называют *насыщением*.

Формула средних также обладает насыщением, но для оптимальной гладкости $p = 2$, когда ее точность $O(h^2)$. Каждое ухудшение гладкости на единицу уменьшает порядок тоже на единицу. Свойство насыщения имеется у большинства сеточных методов.

Избыточная гладкость на первый взгляд кажется ненужной. Но далее увидим, что она полезна, т. к. позволяет улучшить точность методом сгущения сеток.

§ 2. Сгущение равномерных сеток

1. Апостериорная оценка точности. Рассмотрим общую постановку задачи. Пусть надо численным расчетом найти некоторую величину, точное значение которой равно U; оно нам не известно. Расчет выполняется на равномерной сетке Ω_N с числом интервалов N и шагом $h = (b-a)/N$; вычисления по сеточной формуле дают величину U_N, которая является приближением к U.

Будем считать, что для сеточной формулы проведено теоретическое исследование и получена априорная оценка погрешности в виде ряда

по степеням шага h:

$$R_N \equiv U_N - U = \sum_{q=0} A_q h^{p+sq} \approx A_0 h^p. \tag{27}$$

Очевидно, точность этой сеточной формулы есть $O(h^p)$; для произвольных несимметрично построенных формул обычно величина $s = 1$, для симметричных формул $s = 2$. Число членов суммы в (27) определяется "запасом гладкости"; если гладкость оптимальна (в точности обеспечивает насыщение), то сумма содержит только один член $q = 0$. Если превышает оптимальную на l единиц, то появляются члены с $sq \leqslant l$. Первый член ряда $A_0 h^p$ есть асимптотическая оценка погрешности.

Проведем расчеты на двух разных сетках с числом узлов $N_1 = N$ и $N_2 = rN$, т. е. с шагами $h = (b-a)/N$ и h/r. Числа N_1 и N_2 целые, но их отношение r может быть нецелым; для определенности полагаем $r > 1$, т. е. вторая сетка более подробна. Соответственно, расчет на ней даст более точный результат.

Для обоих расчетов соотношения (27) перепишем в виде

$$R_N \equiv U_N - U, \quad R_{rN} \equiv U_{rN} - U, \tag{28}$$

причем при учете только главного члена ошибки справедливо приближенное соотношение

$$R_N \approx A_0 h^p = r^p \cdot A_0 \left(\frac{h}{r}\right)^p \approx r^p R_{rN}. \tag{29}$$

Вычтем второе равенство (28) из первого; неизвестное точное значение U при этом сократится. Исключая далее R_N с помощью (29), получим соотношение

$$R_{rN} \approx \frac{U_N - U_{rN}}{r^p - 1}. \tag{30}$$

Формула (30) показывает, что по результатам расчета на двух сетках можно получить оценку погрешности. Эта оценка является *апостериорной*, т. к. проводится не до, а после выполнения расчетов. Кроме того, она не мажорантная, а *асимптотическая*.

Таким образом, выполнив расчет на двух сетках, мы получаем в качестве ответа значение на более густой сетке U_{rN}, а в качестве его погрешности — оценку (30). Поскольку это асимптотическая оценка, то при достаточно малом шаге подробной сетки эта оценка будет очень близка к фактической погрешности. Если погрешность оказалась больше заданного ε, то берут еще более густую сетку. Этот способ проведения расчетов с гарантированной точностью является одним из самых надежных практических приемов. Оценку (30) нетрудно включить в любую программу.

Однако эта оценка правильна лишь при соблюдении двух условий: 1) из теоретических исследований известен порядок точности p сеточной формулы; 2) сеточная функция $u(x)$ имеет гладкость не хуже

оптимальной, обеспечивающей насыщение формулы. При нарушении этих условий оценка становится ошибочной.

Выбор шага. Пусть требуется получить погрешность не более ε. Как выбрать необходимый шаг (число интервалов) сетки? Начинают с того, что на глазок выбирают не слишком большие, но разумные числа интервалов N и rN. На этих сетках производят расчеты с вычислением апостериорной оценки погрешности (30). Если оказалось, что $|R_{rN}| \leqslant \leqslant \varepsilon$, то требуемая точность достигнута и расчет окончен.

В противном случае надо сгустить сетку, выбрав новое $\widetilde{r} > r$. Поскольку погрешность убывает как $h^p \sim r^{-p}$, для получения $|R_{\widetilde{r}N}| = \varepsilon$ следует выбрать

$$\widetilde{r} \approx r \left(\frac{|R_{rN}|}{\varepsilon} \right)^{1/p} > r; \tag{31}$$

а число узлов требуемой сетки будет $\widetilde{r}N$ (разумеется, округленное до целого).

Формулой (31) разумно пользоваться, если отношение $\widetilde{r}/r$ получилось небольшим. Если же $\widetilde{r} \gg r$, это означает, что первые сетки были слишком грубыми и дали плохую точность. Прогнозировать нужный шаг по ним рискованно, и желательно проверить точность еще одним расчетом. А это уже дает запутанную "кухню" в построении программы, не говоря о лишних расчетах.

Обсудим еще выбор r. Формально любое r позволяет оценить точность. Например, чтобы минимизировать объем расчетов, можно взять сетки с N и $N+1$ интервалами; тогда $r = 1 + 1/N$. Но при этом значения U_N и U_{rN} весьма близки, и нахождение их разности в (30) приводит к значительной ошибке округления. Значит, значения r должны быть заметно больше 1.

Практики предпочитают простую и легко программируемую процедуру. Выбирают довольно грубую начальную сетку Ω_N, полагают $r = 2$ и проводят вычисления на этих двух сетках. Если требуемая точность ε не достигнута, первую сетку отбрасывают, а вторую снова сгущают вдвое. Повторяют процедуру до тех пор, пока на последней сетке не получится заданная точность.

Пример 1. Рунге в 1895 г. проводил расчеты по схеме точности $O\left(h^2\right)$, сгущая сетку вдвое и считая верными совпадающие знаки, т. е. полагая $R_{2N} \approx U_N - U_{2N}$. Но при $p = 2$ и $r = 2$ согласно (30) погрешность примерно втрое меньше этой разности, т. ч. Рунге несколько перестраховывался.

2. Повышение точности. Из расчетов на двух сетках можно извлечь еще большую пользу. Величина R_{rN} является оценкой погрешности для U_{rN} не по порядку величины, а асимптотической, т. е. она совпадает с истинной ошибкой с точностью до малых более высокого

порядка. Значит, прибавляя ее к U_{rN} (точнее, вычитая), получим более точное приближение к U:

$$U_{rN}^{(1)} = U_{rN} - R_{rN} = U_{rN} + \frac{U_{rN} - U_N}{r^p - 1}. \tag{32}$$

Оценим погрешность этого уточненного значения. Подставляя в правую часть (32) разложение (27) для шагов h и h/r, получим

$$R_{rN}^{(1)} = U_{rN}^{(1)} - U = \sum_{q=0} B_q \left(\frac{h}{r}\right)^{p+s+sq},$$
$$B_q = -A_{q+1}\frac{r^p[r^{s(q+1)} - 1]}{r^p - 1}. \tag{33}$$

Ряд для погрешности аналогичен (27), только шаг его h/r соответствует второй сетке, и начинается ряд с $(p+s)$-й степени шага. Следовательно, улучшенный результат (32) имеет более высокую точность $O(h^{p+s})$, т. е. прибавление асимптотической апостериорной погрешности улучшает порядок точности сеточной формулы на величину s.

Однако надо помнить, что для справедливости представления (33) необходимо, чтобы в сумму входил хотя бы один член, т. е. выполнялось $B_0 \sim A_1 \neq 0$. А для этого функция $u(x)$ должна быть на s единиц более гладкой, чем для насыщения исходной сеточной формулы. Можно сказать иначе: вся процедура (32) построения улучшенного решения есть некая новая сеточная формула точности $O(h^{p+s})$, насыщение которой наступает при более высокой гладкости $u(x)$.

Вот почему избыточная гладкость $u(x)$ в исходной сеточной формуле не бесполезна: она позволяет уточнять решение методом сгущения сеток.

Выражение (33) есть априорная оценка погрешности уточненного результата типа (32). Как можно получить апостериорную оценку, будет рассказано в п. 4. Обычно же, выполняя расчет по двум сеткам, берут $U_{rN}^{(1)}$ в качестве окончательного результата и считают, что его погрешность еще меньше, чем R_{rN}.

Пример 2. Пусть число интервалов равномерной сетки N четное, а ее шаг равен h. Объединим попарно каждый нечетный интервал с правым четным, т. е. $[x_{n-1}, x_n]$ с $[x_n, x_{n+1}]$, где n — нечетное. Тогда объединенные интервалы образуют сетку с шагом $2h$ и числом интервалов $N/2$ (вместо сгущения сетки получилось разряжение). Применим формулу левых прямоугольников к паре исходных интервалов и получим приближенный интеграл

$$U_N = hu_{n-1} + hu_n; \tag{34}$$

если применить ту же формулу к сдвоенному интервалу, то получим

$$U_{N/2} = 2hu_{n-1} \tag{35}$$

(здесь для простоты записи употребляется буква U, принятая для интегралов от a до b, хотя тогда бы суммировались все пары интервалов).

Точность формул (34), (35) есть $O(h)$, т.ч. $p = 1$ и $r = 2$. Тогда формула (32) дает

$$U_N^{(1)} = U_N + (U_N - U_{N/2}) = 2hu_n. \tag{36}$$

Нетрудно заметить, что получилась формула средних для объединенного интервала $[x_{n-1}, x_{n+1}]$, а ее порядок точности $p = 2$ действительно на единицу выше, чем у формулы прямоугольников.

Запомним еще, что вместо сгущения сетки можно пользоваться разряжением; здесь употреблялось особенно удобное разряжение вдвое, когда из сетки выбрасываются все нечетные узлы.

3. Рекуррентное сгущение. В п. 2 уже упоминалось, что при расчетах с заданной точностью приходится несколько раз сгущать сетку. Покажем, что такие расчеты позволяют кардинально улучшить точность, если сеточная функция достаточно гладкая, т.е. разложение погрешности по степеням шага (27) содержит достаточно много членов.

Рассмотрим случай, когда каждая следующая сетка в одно и то же число раз r гуще предыдущей; при этом формулы примут особенно простой вид, что важно для прикладных программ. Для упрощения записи немного изменим обозначения. Число интервалов начальной сетки обозначим $N_0 \equiv N$, тогда k-я сетка имеет $N_k = r^k N$ интервалов ($k = 0, 1, 2, \ldots$); все эти числа должны быть целыми (r может быть нецелым). Результат расчета по основной сеточной формуле на k-й сетке обозначим U_{0k} вместо $U_{r^k N}$. Далее построим следующий алгоритм, иллюстрируемый таблицей 1.

Таблица 1. Рекуррентное сгущение сетки в r раз

k	N_k	Порядок точности							
		p		$p+s$		$p+2s$		$p+3s$	
0	N	U_{00}	—	—	—	—	—	—	
1	rN	U_{01}	R_{11}	U_{11}	—	—	—	—	
2	r^2N	U_{02}	R_{12}	U_{12}	R_{22}	U_{22}	—	—	
3	r^3N	U_{03}	R_{13}	U_{13}	R_{23}	U_{23}	R_{33}	U_{33}	
…	…	…	…	…	…	…	…	…	…

Во-первых, берем пару соседних сеток: $(k-1)$-ю и k-ю, $k = 1, 2, \ldots$ Отношение их шагов равно r. Расчет по исходной формуле дает на них величины $U_{0,k-1}$ и U_{0k}, имеющие точность $O(h^p)$. Расчет по формулам (30) и (32) для каждой пары дает априорную оценку погрешности значения величины U_{0k} и уточненное значение; обозначим их через R_{1k} и U_{1k} соответственно и запишем в k-й строке. Так получим два столбца таблицы 1; первая строка в них остается незаполненной.

Во-вторых, столбец U_{1k} можно рассматривать как расчеты по новой сеточной формуле точности $O(h^{p+s})$, поскольку для этих величин априорная оценка точности имеет тот же вид (33), что и (27); отличие только в начальной степени шага. Отношение шагов соседних сеток по-прежнему равно r. Следовательно, для этих величин $U_{1,k-1}$ и U_{1k} $(k = 2, 3, \ldots)$ также можно вычислить апостериорную оценку погрешности R_{2k} и уточненное значение $U_{2,k}$ по формулам (30) и (32), но только подставляя в них r^{p+s} вместо r^p. Найденные величины образуют два следующих столбца таблицы 1, и в них отсутствуют уже две первые строки. Новое уточненное значение U_{2k} имеет точность $O(h^{p+2s})$.

Аналогично строятся следующие столбцы, и получается треугольная таблица. В ее k-й строке результат наивысшего порядка точности $p + sk$ есть U_{kk}, стоящий на правом конце строки. Таким образом, рекуррентное сгущение сетки эквивалентно построению схем высокого порядка точности.

Замечание 3. Число сгущений сетки (строчек в таблице 1) может быть большим; методические расчеты проводят с 10–15 сгущениями. Однако число столбцов ограничено гладкостью используемой функции $u(x)$. Повышать порядок точности можно лишь до тех пор, пока он не превышает степени h в последнем члене разложения исходной погрешности (27). Остальные столбцы таблицы заполнять нельзя.

Например, пусть 8 раз непрерывно дифференцируемая функция $u(x)$ интегрируется по формуле средних (19). Ее порядок точности $p = 2$, разложение (21), (27) содержит $u''(x)$ и имеет приращение степени $s = 2$, т. ч. последний член разложения содержит $u^{(8)}h^8$. Видно, что только трижды можно повышать точность. Последним столбцом будет U_{3k}.

Замечание 4. Возможна нестандартная ситуация. Для формулы левых прямоугольников (13) погрешность (18) есть $O(h)$ и разлагается в ряд по всем степеням h, т. е. $s = 1$. В примере 2 показано, что первое сгущение дает формулу средних точности $O(h^2)$, как и ожидалось. Но у формулы средних $s = 2$. Значит, второе уточнение повышает точность сразу до $O(h^4)$, вместо ожидаемого $O(h^3)$. Однако стандартная процедура с одним и тем же $s = 1$ этого не учитывает, т. ч. третье уточнение исключает несуществующий член $O(h^3)$.

Однако ошибочных результатов при этом не получается. Просто можно было более экономно организовать вычисления, если заранее теоретически выяснить такое поведение погрешности. Но это редко удается сделать.

4. Многомерность и наборы сеток. Задача может содержать много независимых переменных. Например, функция распределения частиц в газах $u(t, x, y, z, v_x, v_y, v_z)$ зависит от времени t, трех координат x, y, z и трех компонент скорости. По каждой из переменных вводится своя сетка со своим числом интервалов, т. е. своим τ, h_x,

h_y, h_z, γ_x, γ_y, γ_z. Погрешность основного численного метода имеет зачастую разный порядок малости по каждому из этих шагов. Опишем типичные ситуации, ограничиваясь для простоты двумя переменными t, x и шагами τ, h. Они собраны в таблице 2, а главный член погрешности имеет вид:

$$R \approx A\tau^p + Bh^q = O(\tau^p + h^q). \tag{37}$$

Порядок точности p, q по каждой переменной определяется выбранным методом и предполагается известным.

Таблица 2. Типичные погрешности многомерных задач

№	Задача	Численный метод	Погрешность	$p:q$	r_t, r_x
1.	Уравнение переноса	Чисто неявная схема (безусловно устойчивая и монотонная)	$O(\tau + h)$	1:1	$r_t = r_x$
2.	Уравнение тепло-проводности	Чисто неявная схема (безусловно устойчивая и монотонная)	$O(\tau + h^2)$	1:2	$r_t = r_x^2$
3.	Уравнение тепло-проводности	Схема с полусуммой (безусловно устойчивая, но немонотонная)	$O(\tau^2 + h^2)$	1:1	$r_t = r_x$
4.	Уравнение тепло-проводности	Комплексная схема Розенброка (безусловно устойчива и почти монотонна)	$O(\tau^2 + h^2)$	1:1	$r_t = r_x$
5.	Уравнение тепло-проводности	Схема повышенной точности (безусловно устойчива)	$O(\tau^2 + h^4)$	1:2	$r_t = r_x^2$
6.	Уравнение тепло-проводности	Двухстадийная схема Розенброка (А-устойчивая)	$O(\tau^3 + h^2)$	3:2	$r_t^3 = r_x^2$

Дальнейшие члены разложения погрешности содержат все более высокие степени τ и h, либо идущие подряд, либо той же четности, что и главные, и это зависит от симметричности численного метода относительно данной переменной, т. е. по каждой переменной характер разложения может быть свой.

Возникает вопрос, которого не было в одномерных задачах: во сколько раз сгущать сетки по разным переменным?

Для однократного сгущения ответ очевиден. Апостериорное нахождение R основано на том, что для схемы p-го порядка при сгущении сетки в r раз погрешность убывает в известное число раз, а именно r^p. Главный член погрешности (37) состоит из двух равноправных слагаемых. Значит надо, чтобы при сгущении сеток они убывали в одно

и то же число раз. Для этого коэффициенты сгущения каждой сетки r_t и r_x должны удовлетворять соотношению

$$r_t^p = r_x^q. \tag{38}$$

Тогда для оценки погрешности и уточнения решения по-прежнему можно пользоваться формулами (30) и (32), подставляя в них в качестве r^p величину (38). Например, для 1-й, 3-й и 4-й схем таблицы 2 выполняется $p = q$, т. ч. $r_t = r_x$ и сетки по обеим переменным следует сгущать в одинаковое число раз; это очень просто.

Однако для 2-й и 5-й схем выполняется $q = 2p$, откуда $r_t = r_x^2$; если мы захотим уменьшить шаг по пространству вдвое ($r_x = 2$), то шаг по времени надо уменьшить вчетверо ($r_t = 4$). Еще сложнее 6-я схема, где следует брать $r_t = r_x^{2/3}$.

При большом числе переменных главный член погрешности (37) содержит больше слагаемых, и в каждом может стоять своя степень — порядок точности. Но принцип остается тот же: коэффициенты сгущения всех сеток выбирают так, чтобы каждое слагаемое в (37) уменьшалось в одинаковое число раз.

Трудоемкость метода сгущения сеток сильно возрастает в многомерных задачах. В самом деле, в одномерном случае при сгущении сетки в r раз трудоемкость возрастает обычно тоже в r раз; это хорошо видно на примере квадратурных формул прямоугольников или средних. Однако в двумерном случае даже при $p = q$ и $r_t = r_x \equiv r$ число узлов двумерной сетки возрастает в $r_t r_x = r^2$ раз; во столько же раз, если не больше, возрастает трудоемкость (т. е. в 4 раза при $r = 2$). Еще хуже ситуация у схем с $2p = q$, где $r_t = r_x^2$, т. ч. число узлов и трудоемкость возрастает как r_x^3 (в 8 раз при $r_x = 2$). А для задач с еще большим числом измерений трудоемкость катастрофически нарастает.

Таким образом, сгущение многомерных сеток оказывается трудоемким процессом. Чтобы уменьшить трудоемкость, желательно выбирать r_x как можно ближе к единице. Но здесь мы сталкиваемся с неприятным ограничением: число интервалов сетки по каждой переменной должно быть целым. Если мы хотим многократно сгущать сетки, то целым должно быть $r_x^k N_0$, $k = 0, 1, 2, \ldots$, где N_0 — начальное число интервалов. Наименьшее число, которое это обеспечивает, есть $r_x = 2$.

Наборы сеток. Существуют некоторые наборы целых чисел, которые очень удобно использовать для построения сгущающихся сеток. Они представлены в таблице 3 и обладают следующими свойствами.

Первый набор тривиален — числа интервалов увеличиваются ровно вдвое, а в качестве начального числа интервалов можно брать любое число N_0. Этот набор пригоден для любых целей, включая рекуррентное повышение на много порядков путем многократного сгущения сетки. Но выше говорилось, что $r = 2$ приводит к большой трудоемкости для многомерных задач. Для одномерных задач трудоемкость остается

приемлемой, но точки на графиках погрешности (далее в §2) лежат довольно редко.

В остальных наборах отношение чисел интервалов (или величин шагов) соседних сеток почти одинаково. Оно с хорошей точностью равно дробной степени двойки; в таблице 3 указано максимальное отличие фактического r от этого среднего значения (в %). Такие наборы полезны даже для одномерных задач, ибо позволяют гуще поставить точки на графиках погрешности. В многомерных задачах они дают многократное уменьшение трудоемкости расчетов.

Таблица 3. Рекомендуемые наборы сеток

№	r	Набор N								
1	2 точно	N_0	$2N_0$	$4N_0$	$8N_0$	$16N_0$	$32N_0$	$64N_0$	...	...
2	$2^{1/2} \pm 0.2\%$	12	17	24	34	48	68	...	...	...
3	$2^{1/2} \pm 1.0\%$	5	7	10	14	20	28	40	...	...
4	$2^{1/3} \pm 0.8\%$	12	15	19	24	30	38	48	...	...
5	$2^{1/4} \pm 2.1\%$	10	12	14	17	20	24	28	34	...

В самом деле, возьмем 1-ю или 3-ю схему из таблицы 2. Для них $p = q$ и $r_t = r_x$. Выбор первого набора сеток из таблицы 3 дает четырехкратное увеличение трудоемкости при каждом сгущении. Если же взять 2-й или 3-й наборы $r_t = r_x \approx 2^{1/2}$, то трудоемкость возрастает всего в $r_x r_t \approx 2$ раза при каждом очередном сгущении. Это вдвое выгоднее, чем при 1-м наборе сеток. Использование 4-го или 5-го набора дает еще большую экономию.

Для 2-й и 5-й схем $q = 2p$ и надо брать $r_t = r_x^2$. Если взять $r_x = 2^{1/2}$ (т. е. 2-й или 3-й наборы сетки) и $r_t = 2$ (1-й набор сеток), то трудоемкость возрастает в $2^{3/2} = 2.8$ (вместо 8 раз, как было бы при $r_x = 2$). Если же взять $r_x = 2^{1/4}$ и $r_t = 2^{1/2}$, то трудоемкость возрастает лишь в $2^{3/4} \approx 1.7$ раз при каждом сгущении сеток.

Возможны и другие комбинации. Так для 6-й схемы точности $O(\tau^3 + h^2)$ надо $r_t = r_x^{2/3}$. Возьмем для времени 4-й набор с $r_t = 2^{1/3}$, а для пространства — 2-й или 3-й наборы с $r_x \approx 2^{1/2}$. Их соотношение такое, какое нужно.

Ограничения. Из всех наборов таблицы 3 только первый начинается с произвольного числа узлов. Все остальные надо начинать только с одного из чисел соответствующего ряда (хотя не обязательно с первого числа).

Значение r для каждой пары сеток набора немного отличается от среднего r. Поэтому значение погрешности (она же поправка), определенное по средним r, несет дополнительную ошибку (примерно в p раз больше, чем относительная ошибка r, что может составлять 1–8%). Это позволяет хорошо провести первое экстраполяционное уточнение

(R_{1k} и U_{1k} в табл. 1); но нахождение погрешностей-поправок более высоких порядков уже вряд ли возможно. Только для очень удачного 2-го набора есть надежда использовать второе уточнение, но уже не третье.

Однако увеличивать количество строк в таблице 1, т. е. проводить много сгущений до достижения заданной точности, все эти наборы позволяют.

§ 3. Контроль и диагностика

1. Отладка программ. Хорошая программа должна содержать не только расчет величины U по заданной сеточной формуле, но и процедуру рекуррентного сгущения сетки в r раз. В начальных данных при этом задают число интервалов N_0 начальной сетки, максимальное число сгущений (исходя из мощности компьютера) и коэффициент сгущения r. По умолчанию берут $r = 2$, ибо это наименьшее число, которое при любом N_0 обеспечивает целочисленность всех $N_k = r^k N_0$ (о других выборах r поговорим позже). Задается также требуемая погрешность ε для выбранной нормы решения.

Обязательным приемом проверки программ является тестирование на задаче (лучше нескольких задачах) с известным точным решением, причем достаточно гладким; чем выше гладкость, тем лучше. В первую очередь проверяют сходимость сеточного решения U_{0k} (столбец табл. 1) к точному U при увеличении k ($k \to \infty$). Однако выполнять это визуально по таблице 1 не очень удобно. Лучше вычислять истинную погрешность на каждой сетке

$$\overline{R}_{0k} = U_{0k} - U, \quad k = 0, 1, 2, \dots, \tag{39}$$

выдавать ее выбранную норму и визуально проверять $\|\overline{R}_{0k}\| \to 0$ при $k \to \infty$; так легче контролировать сходимость в далеких знаках. Эта погрешность оценена по точному решению без применения сгущения сеток, поэтому в величине $\overline{R}_{lk}$ выбрано значение индекса $l = 0$, а не $l = 1$.

Однако этого недостаточно. Если в формуле средних точности $O(h^2)$ будет сделана описка в индексе (взято u_{n-1} вместо $u_{n-1/2}$), то получится формула левых прямоугольников точности $O(h)$. Сходимость все равно будет, только с худшим порядком. Значит, надо проверять еще порядок сходимости: выполняется ли соотношение $\|\overline{R}\| \approx Ah^p$ с теоретическим p. Для этого строят график в двойных логарифмических переменных $\lg \|\overline{R}_{0k}\|$ от $\lg N_k$ (или $-\lg h$). При точном степенном законе убывания этот график будет прямой линией с наклоном $\operatorname{tg} \alpha = -p$. Поскольку степенной закон есть лишь асимптотика при $N \to \infty$, реальный график будет кривой, асимптотически переходящей в нужную прямую (рис. 2). Неправильный наклон этой линии свидетельствует об ошибке в программе.

Еще больше информации дает расчет эффективного порядка точности — среднего наклона каждого участка кривой на рисунке 2 (от одной расчетной точки до другой):

$$\overline{p}_{0k} = \frac{\lg(\|\overline{R}_{0,k-1}\|/\|\overline{R}_{0k}\|)}{\lg r}, \quad r = \frac{N_k}{N_{k-1}} = \text{const}, \quad k = 1, 2, \ldots \tag{40}$$

При $k \to \infty$ эффективный наклон $\overline{p}_{0k} \to p$. Это стремление при больших k монотонно, причем монотонность обычно сохраняется и в начале столбца. Это удобно визуально контролировать, выдавая столбец $\overline{p}_{0k}$ (табл. 4).

Рекомендуется усилить проверку, используя уточнения U_{lk} из таблицы 1. Для них также вычисляют точные значения погрешностей, вычитая точное решение:

$$\overline{R}_{lk} = U_{lk} - U, \quad l = 1, 2, 3, \ldots, \quad k = 1, 2, 3, \ldots; \tag{41}$$

здесь l — индекс столбца, k — индекс строки; для $l = 0$ расчет уже проведен в (39). Каждый l-й столбец погрешности $\|\overline{R}_{lk}\| \sim h^{p+ls}$ согласно априорным оценкам, поэтому соответствующая кривая в двойном логарифмическом масштабе также асимптотически выходит на прямую с наклоном $\operatorname{tg}\alpha = -(p + ls)$, как показано на рисунке 2. Аналогично для более тщательного анализа вычисляются столбцы эффективных показателей степени:

$$\overline{p}_{lk} = \frac{\lg(\|\overline{R}_{l,k-1}\|/\|\overline{R}_{lk}\|)}{\lg r}, \quad l = 1, 2, \ldots, \quad k = 1, 2, \ldots \tag{42}$$

Ожидается, что в l-м столбце $\overline{p}_{lk} \to p + ls$ при $k \to \infty$; эти столбцы собирают в таблицу 4.

Если все эффективные порядки точности быстро сходятся к теоретическим пределам и $U_{lk} \to U$, то с достаточной уверенностью можно считать программу отлаженной.

Если предел эффективных порядков точности определенно меньше теоретического, то либо тестовая функция недостаточно гладкая, либо имеется ошибка в программе. Если эффективные порядки точности оказались больше теоретических, то программа правильна, а теоретическое исследование было недостаточно полным, и его можно улучшить.

Пример 3. Рассмотрим формулу средних на равномерной сетке (19) и выберем в качестве тестовой функции $u(x) = \exp(x)$, $0 \leqslant x \leqslant 4$, все производные которой существуют и непрерывны (тем самым ограничены). Тогда точный ответ

$$U = \int_0^4 u(x)dx = e^4 - 1 = 53.59815003, \quad u(x) = e^x. \tag{43}$$

Результаты расчетов представлены на рисунке 2 и в таблице 4. Видно, что все эффективные порядки точности соответствуют ожидаемым, сеточный расчет сходится к точному. Рисунок 2 наглядно показывает, насколько улучшается точность при рекуррентном уточнении решения. Поскольку сами формулы уточнения (32) имеют ничтожно малую трудоемкость по сравнению с вычислениями по основной сеточной формуле, трудоемкость получения всех величин на одной горизонтальной линии таблицы 2 практически одинакова. Видно, что выгодно проводить много уточнений, сколько позволяет гладкость функции.

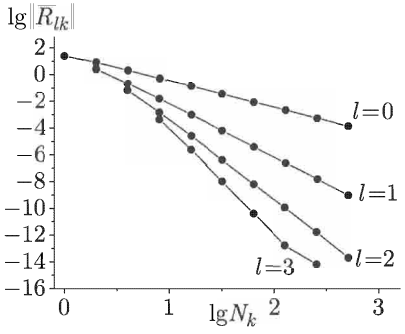

Рис. 2. Погрешности решения и рекуррентных уточнений для примера 3

Таблица 4. Эффективные порядки точности (по точному решению) для примера 3

k	N_k	U_{0k}	$\lg\|\overline{R}_{0k}\|$	$\overline{p}_{0k}$	$\overline{p}_{1k}$	$\overline{p}_{2k}$	$\overline{p}_{3k}$	$\overline{p}_{4k}$
0	1	29.55622	1.38097	—	—	—	—	—
1	2	45.60764	0.90258	1.58919	—	—	—	—
2	4	51.42836	0.33642	1.88073	3.52365	—	—	—
3	8	53.04388	−0.25657	1.96890	3.86434	5.51208	—	—
4	16	53.45883	−0.85693	1.99214	3.96480	5.86149	7.50979	—
5	32	53.56327	−1.45907	1.99803	3.99112	5.96409	7.86092	9.50928
6	64	53.58943	−2.06167	1.99951	3.99777	5.99094	7.96405	9.86045
Теория		53.59815	$-\infty$	2	4	6	8	10

Пример 4. Произведем расчет того же интеграла (43) по формуле левых прямоугольников на равномерной сетке (16). Тестирование проведено аналогично примеру 3, его результаты представлены на рисунке 3 и в таблице 5. Видно, что здесь первые две линии погрешностей и первые два столбца эффективных порядков точности ведут себя в соответствии с простейшими априорными оценками, но третья кривая

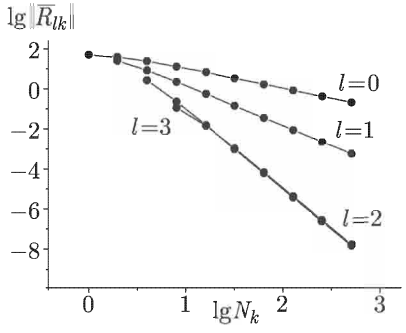

Рис. 3. Погрешности решения и рекуррентных уточнений для примера 4

Таблица 5. Эффективные порядки точности (по точному решению) для примера 4

k	N_k	U_{0k}	$\lg\|\|\overline{R}_{0k}\|\|$	$\overline{p}_{0k}$	$\overline{p}_{1k}$	$\overline{p}_{2k}$	$\overline{p}_{3k}$	$\overline{p}_{4k}$
0	1	4.00000	1.69737	—	—	—	—	—
1	2	16.77811	1.56784	0.42979	—	—	—	—
2	4	31.19288	1.35187	0.71665	1.58919	—	—	—
3	8	41.31062	1.09069	0.86664	1.88073	3.52365	—	—
4	16	47.17725	0.80850	0.93635	1.96890	3.86434	2.95741	—
5	32	50.31804	0.51647	0.96903	1.99214	3.96480	3.75067	5.44757
6	64	51.94065	0.21970	0.98474	1.99803	3.99112	3.93767	5.84578
7	128	52.76504	−0.07962	0.99243	1.99951	3.99777	3.98441	5.96019
8	256	53.18050	−0.37962	0.99623	1.99988	3.99944	3.99610	5.99015
9	512	53.38905	−0.68042	0.99812	1.99997	3.99988	3.99901	5.97297
Теория		53.59815	$-\infty$	1	2	3	4	5

и столбец — нет: они показывают аномальную точность. Причина этого обсуждалась в замечании 4, §2. Вывод: программа правильна, а теоретические исследования были не полными.

2. Контроль расчетов. Когда программа отлажена, по ней производят расчеты новых задач, где точный ответ неизвестен. Как найти приближенное решение с погрешностью не более заданного ε? Для этого используют рекуррентное сгущение сетки и рассматривают таблицу 1 апостериорных оценок погрешности. Но для полной уверенности рассчитывают еще таблицу эффективных порядков точности, используя близость апостериорных оценок погрешности R_{lk} и точных погрешностей $\overline{R}_{l-1,k}$ (разница в индексах l подчеркивает, что погрешности R_{lk} оценены при сгущении сетки и не требуют знания точного решения, а погрешности $\overline{R}_{l-1,k}$ оценены по точному решению без

сгущения сетки):

$$p_{lk} = \frac{\lg(\|R_{l,k-1}\|/\|R_{lk}\|)}{\lg r}, \quad l = 1, 2, \ldots, \quad k = 1, 2, \ldots \tag{44}$$

Их сводят в таблицу, аналогичную таблице 4, а величины $\lg \|R_{lk}\|$ представляют на графиках (аналогично рис. 2). Затем анализируют полученные результаты.

Если столбцы p_{lk} (наклоны l-х линий) стремятся к теоретическим пределам $p + sl - s$, то можно уверенно сказать, что функция $u(x)$ имеет достаточную гладкость, а все расчеты достоверны. Тогда в треугольной таблице погрешностей двигаются слева направо сначала по 2-й строке (в ней только одна погрешность), потом по 3-й и т.д. Останавливаются тогда, когда впервые выполнится условие $\|R_{lk}\| \leqslant \varepsilon$. Соответствующее значение U_{lk} считают искомым ответом.

Если в каких-то столбцах эффективный порядок точности оказался выше теоретического, действуют так же.

Если начиная с какого-то столбца эффективный порядок точности определенно ниже теоретического, то $u(x)$ недостаточно гладкая. Достоверны только те столбцы, где p_{lk} стремятся к теоретическому пределу. Все столбцы, лежащие правее их, надо отбросить и далее не использовать. По оставленным столбцам производится выбор ответа U_{lk} с погрешностью $\|R_{lk}\| \leqslant \varepsilon$, как описано выше.

Порядок вычислений выгоден построчный. Основная величина U_{0k} рассчитывается не на всех сетках сразу, т. к. это неоправданно увеличило бы трудоемкость. После расчета на первой и второй сетках начинается составление треугольника апостериорных погрешностей, уточненных значений и эффективных порядков точности. Затем на единицу увеличивают номер сетки k, вычисляют U_{0k} и очередную строку треугольника. Одновременно анализируют поведение и выясняют количество достоверных столбцов. Расчет прекращают на той сетке, на которой достигнута заданная точность.

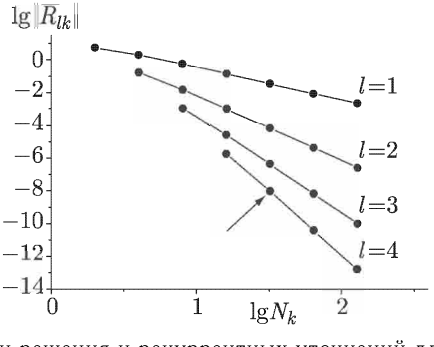

Рис. 4. Погрешности решения и рекуррентных уточнений для примера 5 оценены по сгущению сетки и не требуют знания точного решения. Стрелкой указан расчет, в котором впервые была достигнута заданная точность $\varepsilon = 10^{-8}$

Таблица 6. Эффективные порядки точности для примера 5

k	N_k	U_{0k}	$\lg R_{1k}$	p_{1k}	p_{2k}	p_{3k}	p_{4k}	p_{5k}
0	1	29.55622	—	—	—	—	—	—
1	2	45.60764	1.67719	—	—	—	—	—
2	4	51.42836	0.66281	1.46343	—	—	—	—
3	8	53.04388	−0.61895	1.84919	3.49504	—	—	—
4	16	53.45883	−1.97822	1.96101	3.85721	5.50514	—	—
5	32	53.56327	−3.35770	1.99017	3.96302	5.85976	7.50807	—
6	64	53.58943	−4.74229	1.99754	3.99067	5.96366	7.86050	9.50885
Теория		53.59815	$-\infty$	2	4	6	8	10

Пример 5. Проведем расчет с контролем для условий примера 3, т. е. интеграла (43) по формуле средних, требуя $\varepsilon = 10^{-8}$. Эффективные порядки точности, представленные в таблице 6, очень близки к таблице 4. График апостериорных погрешностей также очень близок к линиям истинных погрешностей (рис. 4). В таблице 7 приведены значения интеграла, полученные при рекуррентном сгущении, и апостериорно оцененные погрешности. Легко видеть, что заданная точность впервые достигается в строке k =5 и в столбце $l = 4$. Таким образом, окончательный результат получен на сетке $N = 32$ по уточнению с погрешностью $O(h^8)$.

Таблица 7. Значения интеграла в примере 5, полученные по основной сеточной формуле U_{0k}, при рекуррентном сгущении сетки U_{lk} и апостериорные оценки их точности R_{lk}. Впервые заданная точность $\varepsilon = 10^{-8}$ достигается при $k = 5$, $l = 4$, $N_k = 32$

k	N_k	U_{0k}	U_{1k}	R_{1k}	U_{2k}	R_{2k}
0	1	29.55622	—	—	—	—
1	2	45.60764	50.95811	5.35047	—	—
2	4	51.42836	53.36860	1.94024	53.52929	0.1607
3	8	53.04388	53.58239	0.53851	53.59664	0.01425
4	16	53.45883	53.59714	0.13832	53.59812	9.8348E−4
5	32	53.56327	53.59809	0.03482	53.59815	6.3063E−5

k	N_k	U_{3k}	R_{3k}	U_{4k}	R_{4k}
0	1	—	—	—	—
1	2	—	—	—	—
2	4	—	—	—	—
3	8	53.59771	0.00107	—	—
4	16	53.59815	2.3537E−5	53.59815	1.715E−6
5	32	53.59815	4.0531E−7	53.59815	**9.423E**−9

Пример 6. Попробуем вычислить по формуле средних следующий интеграл:

$$U = \int_{-2}^{4.5} u(x)dx = 12.125, \quad u(x) = |x|. \tag{45}$$

Возьмем начальную сетку $N_0 = 1$ и коэффициент сгущения $r = 2$. Проведем расчет, составим таблицу погрешностей и эффективных порядков (табл. 8), а также рисунок 5 с графиками погрешностей. Видно, что уже линия погрешности основной формулы $l = 0$ не является плавной кривой, а соответствующий эффективный порядок точности p_{0k} немонотонен и не стремится к теоретическому пределу $p = 2$, хотя колеблется около него. Поведение погрешностей "уточненных" решений показывает, что реального уточнения нет.

Такие результаты свидетельствуют о недостаточной гладкости функции $u(x)$. В самом деле, она непрерывна, но $u'(x)$ имеет разрыв при $x = 0$. Эта единственная особая точка существенно ухудшает

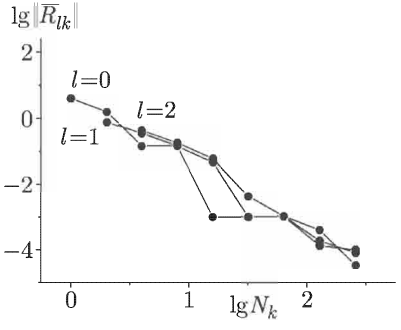

Рис. 5. Погрешности решения и рекуррентных уточнений для примера 6

Таблица 8. Эффективные порядки точности (по точному решению) для примера 6

k	N_k	U_{0k}	$\lg\|\|\overline{R}_{0k}\|\|$	$\overline{p}_{0k}$	$\overline{p}_{1k}$	$\overline{p}_{2k}$	$\overline{p}_{3k}$	$\overline{p}_{4k}$
0	1	8.12500	0.60273	—	—	—	—	—
1	2	10.56250	0.19404	1.35614	—	—	—	—
2	4	11.98438	−0.85290	3.47393	1.16993	—	—	—
3	8	11.98438	−0.85290	0.00000	1.24511	1.23563	—	—
4	16	12.12402	−3.01368	7.16993	1.62560	1.57049	1.55718	—
5	32	12.12402	−3.01368	0.00000	5.54432	3.82911	3.60523	3.55471
6	64	12.12402	−3.01368	0.00000	0.00000	2.06274	2.44946	2.54758
7	128	12.12462	−3.42238	1.35614	2.41504	2.90689	2.98932	3.00367
8	256	12.12497	−4.46931	3.47393	1.16993	0.39514	0.18597	0.13412
Теория		12.12500	$-\infty$	2	4	6	8	10

сходимость, а оценка погрешности из асимптотической превращается в мажорантную (отсюда неплавность кривых на рис. 5).

Однако если расположить узлы сетки так, чтобы особая точка $x = 0$ стала узлом основной и всех последующих сеток, то положение кардинально меняется. Интеграл (45), по существу, разбивается на два интеграла по отрезкам $[-2, 0]$ и $[0, 4.5]$, на каждом из которых функция неограниченно дифференцируема. Отсюда следует рекомендация: *поставь в каждую особую точку узел сетки.*

Такие сетки называют *специальными*; но их не всегда удается сделать равномерными.

Пример 7. Вычислим по формуле средних интеграл

$$U = \int_0^4 u(x)dx = 8, \quad u(x) = \frac{3}{2}\sqrt{x}\,. \tag{46}$$

Здесь имеется особая точка $x = 0$, в которой $u'(0) = \infty$. Она граничная, т. ч. автоматически будет узлом сетки. Выберем $N_0 = 1$ и $r = 2$ и проведем расчет по стандартной методике; его результаты даны в таблице 9 и

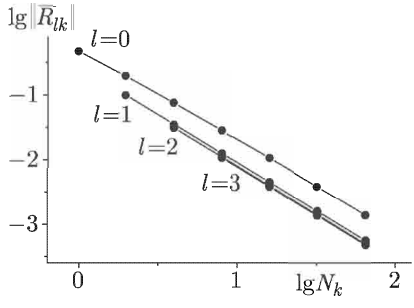

Рис. 6. Погрешности решения и рекуррентных уточнений для примера 7

Таблица 9. Эффективные порядки точности (по точному решению) для примера 7

k	N_k	U_{0k}	$\lg\|\|\overline{R}_{0k}\|\|$	$\overline{p}_{0k}$	$\overline{p}_{1k}$	$\overline{p}_{2k}$	$\overline{p}_{3k}$	$\overline{p}_{4k}$
0	1	8.48528	−0.31436					
1	2	8.19615	−0.70820	1.30685				
2	4	8.07573	−1.12200	1.37306	1.48732			
3	8	8.02839	−1.54863	1.41566	1.49736	1.49950		
4	16	8.01044	−1.98352	1.44305	1.49951	1.49997	1.49999	
5	32	8.00379	−2.42383	1.46102	1.49991	1.50000	1.50000	1.5000
6	64	8.00137	−2.8515	1.47305	1.49998	1.50000	1.50000	1.5000
Теория		8.00000	$-\infty$	2	4	6	8	10

на рисунке 6. Здесь погрешность основного расчета R_{0k} изображается плавной кривой, и первый столбец p_{0k} плавно стремится к пределу, но этот предел не только отличается от теоретического $p = 2$, но и не является целым числом! Это связано с тем, что бесконечность производной носит характер $u'(x) \sim x^{-1/2}$, нецелой степени.

Видно, что при таких особенностях даже специальная сетка не спасает от неприятностей. Но данный метод контроля позволяет их выявить.

3. Ошибки округления. Сгущая сетки и рекуррентно повышая порядок точности, мы не можем неограниченно уменьшать погрешность, ведь рано или поздно мы выходим на ошибки округления. Попробуем априорно оценить этот предел — наилучшую достижимую точность и те сетки, на которых она реализуется. Сделаем это для одномерных задач, решаемых на компьютере с относительной точностью δ (для 64-разрядных чисел с плавающей точкой $\delta \approx 10^{-16}$).

Пусть основной оператор задачи содержит вычисление q-й производной $u(x)$. Это фактически требует вычисления разности q-го порядка для соседних сеточных значений u_n. Согласно главе I при этом вносится ошибка округления $\sim \delta/h^q \sim \delta N^q$. Кроме того общее число операций во всех узлах сетки $\sim N$; по правилам статистики это приводит к увеличению полной ошибки округления еще в $\sim N^{1/2}$ раз. Поэтому полная ошибка округления есть $\delta C_0 N^{q+1/2}$, где C_0 — константа, зависящая от задачи ($q = 0$ соответствует задаче интегрирования).

Рекуррентное повышение точности эквивалентно построению некоторого нового численного метода, имеющего порядок точности $P > p$. Его погрешность есть $O(h^P) = cN^{-P}$, где константа c зависит от задачи.

Полная погрешность ε есть сумма погрешностей метода и округления. При малых N погрешность метода велика, а погрешность округления пренебрежимо мала. При увеличении N погрешность метода уменьшается, а погрешность округления возрастает. Примерно там, где они сравниваются, достигается минимум полной погрешности. Отсюда получаются следующие значения оптимальных сеток и точности:

$$\lg N_{\text{opt}} \approx -\frac{\lg(\delta C_0/c)}{P+q+1/2}, \quad \lg \varepsilon_{\text{opt}} \approx \lg c + \frac{\lg(\delta C_0/c)}{1+(q+1/2)/P}. \tag{47}$$

Коэффициенты c, C_0 лишь для плохо обусловленных задач могут сильно отличаться от единицы; поэтому для большинства оценок можно положить $c \sim C_0 \sim 1$, $\lg c \sim \lg C_0 \sim 0$. Типичными являются значения $q = 0$ (задачи интегрирования) и $q = 2$, т. е. теплопроводности, акустики и т. п. Для 64-разрядных вычислений $\lg \delta = -16$. В таблице 10 приведены оценки оптимальных сеток и точности.

Из таблицы и формулы (47) видно, что для методов невысокого порядка точности $P \leqslant 3$ (а это, например, схема точности $O(h)$ плюс два повышения точности по сгущениям сетки) разумно использовать очень подробные сетки. Для методов высокого порядка точности $P \geqslant$

Таблица 10. Оптимальные сетки для 64-разрядных вычислений

	q \ P	1	2	3	4	5	6
$\lg N$	0	10.7	6.4	4.6	3.6	2.9	2.5
	2	4.6	3.6	2.9	2.3	1.9	1.7
$-\lg\varepsilon$	0	10.7	12.8	13.7	14.2	14.5	14.8
	2	4.6	7.1	8.7	9.9	10.7	11.3

$\geqslant 4$ оптимум достигается уже на скромных сетках $N \sim 100$–1000, а дальнейшее сгущение сетки уже ухудшает точность за счет ошибок округления. Чем больше q, тем меньше оптимальное N и хуже предельная точность; легче получить высокую точность в задачах численного интегрирования, чем при решении уравнений в частных производных.

Если же проводить вычисления с 32-разрядными числами $\lg\delta \approx -8$, то все числа в таблице надо уменьшить вдвое. Оптимальные сетки для хороших методов будут иметь $N \approx 10$–30, предельная точность составит 4–6 верных знаков (и то лишь на хорошо обусловленных задачах). Поэтому еще раз напомним, что в расчетах всегда следует использовать максимальную разрядность, предоставляемую компьютером и математическим обеспечением.

В практических вычислениях особенно распространены: для уравнений в частных производных — симметричные схемы точности $O(h^2)$ с разложением ошибки по степеням h^2; для задач Коши для обыкновенных дифференциальных уравнений — несимметричные схемы Рунге–Кутта точности $O(h^4)$ с разложением ошибки по всем степеням h. В обоих случаях двукратное сгущение сетки с повышением порядка точности дает в итоге $O(h^6)$. Это чаще всего и оказывается оптимальным.

Диагностика. В практических вычислениях необходимо обнаруживать ошибки округления и своевременно прерывать расчет. Для этого следующим образом анализируют графики погрешностей или таблицы R_{lk} и локальных наклонов p_{lk} (эффективных порядков точности).

Пока сетки достаточно грубы, ошибки округления пренебрежимо малы. Тогда ломаные $p_{lk}(N_k)$ монотонно стремятся к своим теоретическим пределам p_l (рис. 7, *а*), ломаные $\lg R_{lk}$ также монотонно стремятся к своим теоретическим асимптотам. Отклонение $p_{lk}(N_k)$ от монотонного стремления к теоретическому пределу может быть либо возрастанием $|p_{lk} - p_l|$, либо сменой знака величины $(p_{lk} - p_l)$ при очередном увеличении номера сетки на единицу. Первое же такое отклонение означает, что ошибки округления становятся заметными. Безопаснее всего прекращать расчеты именно в этот момент. Этот критерий продолжения счета удобно записать в виде

$$0 < \frac{p_{l,k+1} - p_l}{p_{l,k} - p_l} < 1. \tag{48}$$

Если одно из неравенств нарушается, расчет прекращается.

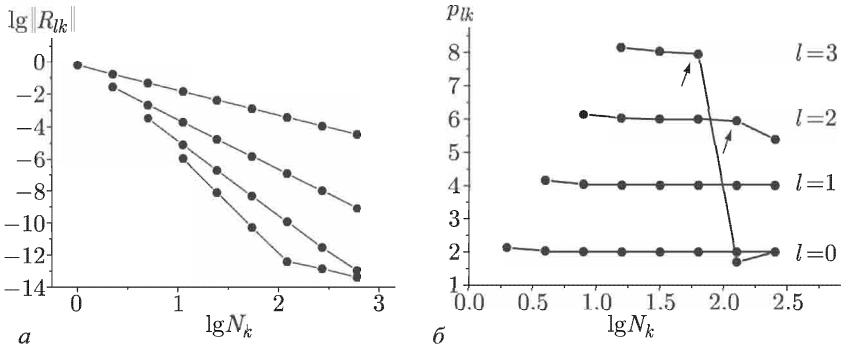

Рис. 7. Погрешности (*а*) и эффективные порядки точности (*б*) при рекуррентном сгущении. Стрелки указывают на заметные ошибки округления

Из таблицы 8 видно, что для каждого l-го повышения точности критерий остановки выполняется при своем k. Это k тем меньше, чем больше l, т. е. дальние столбцы таблиц погрешностей и локальных порядков точности обрезаются на более ранних строчках (более грубых сетках). В столбцах с меньшим l можно использовать больше строк-сеток; но, судя по той же таблице, предельная точность ε_l в каждом столбце будет тем меньше, чем меньше его номер l (поскольку при этом меньше порядок результата P).

Однако последние рассуждения полностью справедливы лишь для достаточно хорошо обусловленных задач. Для очень плохо обусловленных задач (например, очень медленно сходящихся несобственных интегралов) может оказаться, что в первых столбцах таблицы уточнений будет достижима лучшая предельная точность, чем в следующих столбцах. Поэтому рекомендуется продолжать сгущение сеток до тех пор, пока во всех столбцах не выполнится критерий остановки. После этого сравнивают последние $|R_{lk}|$ во всех столбцах и выбирают наименьшую величину. Она и принимается за наименьшую достижимую погрешность, а соответствующее значение U_{lk} — за наилучшее приближение к точному решению.

Разумеется, если еще раньше будет достигнута заданная точность (см. п. 2), то расчет можно сразу прекратить.

Заметим, что критерий (48) — некоторая перестраховка. Даже после его выполнения погрешность может еще некоторое число сгущений убывать (рис. 7, *б*). Однако выйти на большие ошибки округления настолько вероятно, что лучше остановиться раньше, чем позже.

4. Многомерность. В многомерных задачах всегда можно использовать сгущение сеток, находить оценку главного члена погрешности и проводить однократное повышение порядка точности; но не для всех методов можно выполнять рекуррентное многократное повышение порядка точности. Поясним это на примере двумерных схем.

Согласно (37), (38), главный член погрешности есть $O(\tau^p + h^q)$, и для его нахождения и исключения надо сгущать сетки по t и x в соотношении $r_t^p = r_x^q$. В одномерных задачах погрешность разлагается в ряд (27), содержащий степени шага $p + ks$, $k = 0, 1, \ldots$ (обычно $s = 1$ для несимметрично построенных схем, $s = 2$ — для симметричных). Если имеются 2 переменные t и x, то по каждой из них схема имеет свои порядки точности p и q и свои разложения по степеням шага s и σ. Поэтому после исключения главного члена погрешности остается следующий член $O(\tau^{p+s} + h^{q+\sigma})$. Очевидно, для его исключения надо сгущать сетки в отношениях $r_t^{p+s} = r_x^{q+\sigma}$. Это соотношение совместимо с предыдущим сгущением (38) только при выполнении условия

$$\frac{p}{s} = \frac{q}{\sigma}. \tag{49}$$

Если метод таков, что для разложения его погрешности выполнено условие (49), то одно и то же согласованное сгущение сеток по обеим переменным (38) позволяет рекуррентно исключать главный и все следующие члены ряда погрешности. Если же условие (49) не выполнено, то сгущение (38) позволяет исключить только главный член погрешности (37), но более высокие уже не удается.

В качестве примера проанализируем схемы, перечисленные в таблице 2.

— Чисто неявная схема для уравнения переноса имеет $p = q = 1$. Она несимметрична по каждой переменной , т. ч. $s = \sigma = 1$. Соотношение (49) выполнено, и возможно рекуррентное повышение порядка точности при соотношении сгущения $r_t = r_x$.

— Чисто неявная схема для уравнения теплопроводности имеет $p = 1$, $q = 2$. Она несимметрична по t и симметрична по x, т. ч. $s = 1$, а $\sigma = 2$. Соотношение (49) выполнено, и возможно рекуррентное повышение порядка точности при соотношении сгущения $r_t = r_x^2$.

— Схема с полусуммой для уравнения переноса имеет $p = q = 2$. Она симметрична по каждой переменной , т. ч. $s = \sigma = 2$. Соотношение (49) выполнено, и возможно рекуррентное повышение порядка точности при соотношении сгущения $r_t = r_x$.

— Схема повышенной точности для уравнения теплопроводности (5-я схема) имеет $p = 2$, $q = 4$. Она симметрична по обеим переменным, т. ч. $s = \sigma = 2$. Соотношение (49) не выполнено, и возможно лишь однократное повышение точности до $O(\tau^4 + h^6)$.

Отладка программы. Как и в одномерных случаях, она проводится на тестовых задах с достаточно гладкими решениями; обычно такие тесты нетрудно построить. Обязательно проводится исследование сходимости к точному решению при согласованном сгущении сеток по обеим (всем) переменным в соотношении $r_t^p = r_x^q$. Если теоретическое исследование доказало выполнимость соотношения (49), то проверяют и рекуррентное повышение порядка точности.

Рекомендуется дополнительная проверка, т. е. необходимо фиксировать сетку по одной из переменных и многократно сгущать по другой. При этом численное решение должно сходиться к пределу, но этот предел будет точным решением не исходной (дифференциальной и т. п.) задачи, а дифференциально-разностной задачи. В подавляющем большинстве тестов для этого предела трудно построить точное выражение. Поэтому исследование такой сходимости проводят не вычитанием точного решения (здесь мы его не знаем), а сравнением численных решений на двух соседних сетках, как описано в п. 2.

Такое исследование важно, т. к. позволяет проверить, действительно ли разложение погрешности по данному шагу содержит те степени, которые предсказывает теория.

Контроль. Когда программа отлажена, в стационарные расчеты включают сгущение сеток одновременно по всем переменным в теоретических соотношениях. Всегда предусматривают расчет апостериорной погрешности и однократное повышение порядка точности. Если для схемы выполнено соотношение (49), то следует предусмотреть и рекуррентное повышение порядка точности. Однако на практике многократное повышение трудно осуществить, ведь сильное сгущение многомерных сеток приводит к трудоемким расчетам.

§ 4. Произвольные наборы сеток

1. Схема многократного сгущения. Данный параграф посвящен технике проведения расчетов с контролем точности на произвольном наборе сеток.

Пусть некоторое решение $u(x)$ вычисляется сеточным методом, имеющим p-й порядок точности. Если точное решение имеет достаточно высокие непрерывные производные, то сеточное решение $y(x;h)$ на сетке с шагом h разлагается в ряд

$$y(x;h) = u(x) + \sum_{r=p} v_r(x) h^r. \tag{50}$$

Здесь x — узлы сетки; $v_r(x)$ — некоторая комбинация r-х частных производных, зависящая от выбранного метода (в случае, если x — не скаляр, а вектор, когда $u(x)$ есть функция нескольких переменных); число членов в сумме определяется количеством непрерывных производных $u(x)$ и предполагается достаточно большим. Выберем последовательность сеток, для простоты равномерных, с шагами соответственно h_n $(1 \leqslant n \leqslant N)$, где $h_1 > h_2 > \ldots > h_N$. Тогда разложение (50) можно записать в форме

$$y_{n1}(x) = u(x) + \sum_{r=p} v_r(x) h_n^r, \quad 1 \leqslant n \leqslant N, \tag{51}$$

где $y_{n1}(x) \equiv y(x; h_n)$ обозначает численное решение, полученное исходным сеточным методом при шаге h_n.

Возьмем участок последовательности сеток длиной m с шагами h_k, $n-m+1 \leqslant k \leqslant n$ $(m \leqslant n)$. Найдем коэффициенты γ_{knm}, являющиеся решением системы линейных уравнений

$$\sum_{k=n-m+1}^{n} \gamma_{knm} = 1;$$
$$\sum_{k=n-m+1}^{n} \gamma_{knm} h_k^r = 0 \quad \text{при} \quad p \leqslant r \leqslant p+m-2. \tag{52}$$

Составим линейную комбинацию

$$y_{nm}(x) = \sum_{k=n-m+1}^{n} \gamma_{knm} y_{k1}(x). \tag{53}$$

Подставляя (51) в (53) и учитывая (52), нетрудно убедиться, что первые $m-1$ членов разложения (51) сокращаются, и

$$y_{nm}(x) = u(x) + \sum_{r=p+m-1} v_r(x) \sum_{k=n-m+1}^{n} \gamma_{knm} h_k^r \equiv$$
$$\equiv u(x) + O(h^{p+m-1}). \tag{54}$$

Таким образом, линейная комбинация (53) фактически является сеточным расчетом повышенного $(p+m-1)$-го порядка точности, причем объем дополнительных вычислений для ее нахождения пренебрежимо мал по сравнению с тем, который необходим для нахождения $y_{n1}(x)$.

Опишем традиционную процедуру оценки погрешности (рис. 7). Выполним расчеты на первой, второй и т.д. сетках и вычислим $y_{n1}(x)$, имеющие точность $O(h^p)$. Затем по каждой паре соседних сеток, т. е. значений $y_{n-1,1}(x)$ и $y_{n1}(x)$, найдем значения $y_{n2}(x)$, имеющие точность $O(h^{p+1})$. По каждой тройке соседних сеток найдем значения $y_{n3}(x)$ точности $O(h^{p+2})$ и т.д. Из (54) видно, что разложение погрешности $y_{nm}(x)$ начинается с члена $O(h^{p+m-1})$, но этот член отсутствует в погрешности $y_{n,m+1}(x)$. Поэтому *апостериорной оценкой погрешности* $y_{nm}(x)$ *является величина*

$$\Delta_{nm}(x) = y_{n,m+1}(x) - y_{nm}(x). \tag{55}$$

Это не мажорантная оценка, а точная асимптотическая. Это значит, что при $h \to 0$ фактическая погрешность становится сколь угодно близкой к величине $\Delta_{nm}(x)$.

Изложенный способ требует многократного решения линейных систем (52). Далее будут получены как несложные явные, так и очень простые рекуррентные формулы их решения. Это, во-первых, заметно упрощает и удешевляет весь алгоритм экстраполяции; во-вторых, это позволяет надежно контролировать появление ошибок округления, что особенно существенно при расчетах на 16-разрядных персональных компьютерах. В целом это позволяет в определенном смысле оптимизировать процедуру сгущения сеток.

2. Явное решение. Сначала приведем некоторые вспомогательные соотношения. Напомним известное выражение для определителя Вандермонда:

$$V(h_1,\dots,h_m)=\begin{vmatrix} 1 & h_1 & h_1^2 & \dots & h_1^{m-1}\\ 1 & h_2 & h_2^2 & \dots & h_2^{m-1}\\ \dots & \dots & \dots & \dots & \dots\\ 1 & h_m & h_m^2 & \dots & h_m^{m-1}\end{vmatrix}=\prod_{1\leqslant j<i\leqslant m}(h_i-h_j).$$

Из него немедленно следует выражение для определителя несколько более общего вида:

$$V_p(h_1,\dots,h_m)=\begin{vmatrix} h_1^p & h_1^{p+1} & \dots & h_1^{p+m-1}\\ \dots & \dots & \dots & \dots\\ h_m^p & h_m^{p+1} & \dots & h_m^{p+m-1}\end{vmatrix}=\prod_{q=1}^{m}h_q^p\prod_{1\leqslant j<i\leqslant m}(h_i-h_j). \tag{56}$$

Потребуются еще некоторые соотношения. Пусть индекс k заключен в пределах $m\leqslant k\leqslant n$. Тогда сразу видно, что

$$\prod_{q=m,q\neq k}^{n}h_q^p=h_k^{-p}\prod_{q=m}^{n}h_q^p. \tag{57}$$

Нетрудно также убедиться, что

$$\prod_{m\leqslant j<i\leqslant n(i,j\neq k)}(h_i-h_j)=\frac{\displaystyle\prod_{m\leqslant j<i\leqslant n}(h_i-h_j)}{\displaystyle\prod_{k<i\leqslant n}(h_i-h_k)\cdot\prod_{m\leqslant j<k}(h_k-h_j)}=$$

$$=(-1)^{n-k}\frac{\displaystyle\prod_{m\leqslant j<i\leqslant n}(h_i-h_j)}{\displaystyle\prod_{m\leqslant j\leqslant n,j\neq k}(h_k-h_j)}. \tag{58}$$

Теперь можно написать явное выражение для системы линейных уравнений (52). Оно выражается через отношение определителей по

правилу Крамера:

$$\gamma_{\substack{knm\\ n-m+1\leqslant k\leqslant n,\\ 1\leqslant m\leqslant n}} = \frac{\begin{vmatrix} 1 & \dots & 1 & 1 & 1 & \dots & 1 \\ h^p_{n-m+1} & \dots & h^p_{n-m+k-1} & 0 & h^p_{n-m+k+1} & \dots & h^p_n \\ h^{p+1}_{n-m+1} & \dots & h^{p+1}_{n-m+k-1} & 0 & h^{p+1}_{n-m+k+1} & \dots & h^{p+1}_n \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ h^{p+m-2}_{n-m+1} & \dots & h^{p+m-2}_{n-m+k-1} & 0 & h^{p+m-2}_{n-m+k+1} & \dots & h^{p+m-2}_n \end{vmatrix}}{\begin{vmatrix} 1 & \dots & 1 \\ h^p_{n-m+1} & \dots & h^p_n \\ h^{p+1}_{n-m+1} & \dots & h^{p+1}_n \\ \dots & \dots & \dots \\ h^{p+m-2}_{n-m+1} & \dots & h^{p+m-2}_n \end{vmatrix}}.$$

Разложим оба определителя по элементам первой строки. В числителе только алгебраическое дополнение k-го элемента оказывается ненулевым (все остальные миноры содержат нулевой столбец). Кроме того, все миноры являются определителями, транспонированными к (56); индексы при h у них идут не подряд, пропущен индекс, соответствующий номеру элемента строки. С учетом этого и соотношений (57), (58) получим:

$$\gamma_{knm} = (-1)^k \frac{\prod\limits_{\substack{s=n-m+1\\ s\neq k}}^{n} h_s^p \prod\limits_{\substack{n-m+1\leqslant j<i\leqslant n\\ i,j\neq k}} (h_i - h_j)}{\sum\limits_{q=n-m+1}^{n} (-1)^q \prod\limits_{\substack{s=n+m-1\\ s\neq q}}^{n} h_s^p \prod\limits_{\substack{n-m+1\leqslant j<i\leqslant n\\ i,j\neq q}} (h_i - h_j)} = \tag{59}$$

$$= \frac{h_k^{-p} \prod\limits_{\substack{n-m+1\leqslant j<i\leqslant n\\ i,j\neq k}} (h_i - h_j)}{\sum\limits_{q=n-m+1}^{n} (-1)^{q-k} h_q^{-p} \prod\limits_{\substack{n-m+1\leqslant j<i\leqslant n\\ i,j\neq q}} (h_i - h_j)} = \tag{60}$$

$$= \frac{h_k^{-p} \prod\limits_{\substack{n-m+1\leqslant j\leqslant n\\ j\neq k}} (h_k - h_j)^{-1}}{\sum\limits_{q=n-m+1}^{n} h_q^{-p} \prod\limits_{\substack{n-m+1\leqslant j\leqslant n\\ j\neq q}} (h_k - h_j)^{-1}}, \tag{61}$$

$$n-m+1\leqslant k\leqslant n, \qquad 1\leqslant m\leqslant n.$$

Формула (61) требует наименьшего количества операций и является наиболее пригодной для непосредственного вычисления коэффициентов γ_{knm}. Однако далее будет построен еще более экономный способ вычисления.

3. Рекуррентные формулы. Покажем, что решение повышенной точности можно строить рекуррентно. Пусть уже найден столбец решений $y_{nm}(x)$, имеющих $(p+m-1)$-й порядок точности. Тогда повысить

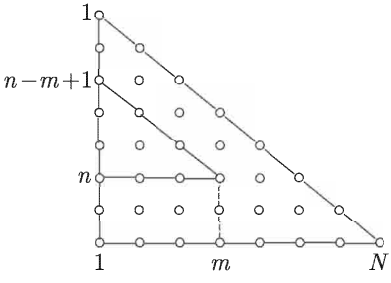

Рис. 8. Рекуррентное сгущение

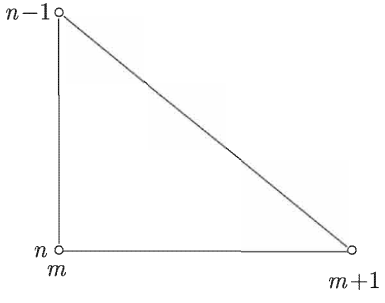

Рис. 9. Треугольник погрешностей

порядок точности еще на единицу можно с помощью линейной комбинации двух таких соседних решений (см. рис. 8 и 9)

$$y_{n,m+1}(x) = \frac{a_{nm}y_{nm}(x) - y_{n-1,m}(x)}{(a_{nm}-1)} \tag{62}$$

при соответствующем выборе коэффициента a_{nm}, не зависящего от x. В самом деле, подставим (54) с соответственно сдвинутыми индексами в (62) и получим

$$y_{n,m+1}(x) = u(x) +$$

$$+\frac{1}{a_{nm}-1}\sum_{r=p+m-1} v_r(x)\left[a_{nm}\sum_{k=n-m+1}^{n}\gamma_{knm}h_k^r - \sum_{k=n-m}^{n-1}\gamma_{k,n-1,m}h_k^r\right].$$

Если положить

$$a_{nm} = \frac{\displaystyle\sum_{k=n-m}^{n-1}\gamma_{k,n-1,m}h_k^{p+m-1}}{\displaystyle\sum_{k=n-m+1}^{n}\gamma_{knm}h_k^{p+m-1}}, \tag{63}$$

то обращается в нуль квадратная скобка, являющаяся множителем при $v_{p+m-1}(x)$. Тогда в сумме остаются только члены более высокого порядка малости, и

$$y_{n,m+1}(x) = u(x) + O(h^{p+m})$$

становится искомым решением повышенной точности.

Формулы (59)–(61), (63) можно преобразовать к очень несложной рекуррентной процедуре. Соответствующие преобразования используют соотношения (57), (58) и очень громоздки, хотя и не слишком трудны. Окончательный результат таков:

$$c_{n1} = 1, \quad c_{n+1,m+1} = (h_n - h_{n-m})c_{nm} \quad \text{при} \quad 1 \leqslant m \leqslant n-1; \tag{64}$$

$$d_{n1} = 1, \quad d_{n,m+1} = (h_n - h_{n-m})d_{nm} \quad \text{при} \quad 1 \leqslant m \leqslant n-2; \tag{65}$$

$$b_{n1} = h_n^{-p}, \quad b_{n,m+1} = c_{nm}b_{nm} - d_{nm}b_{n-1,m} \quad \text{при} \quad 1 \leqslant m \leqslant n-1; \tag{66}$$

$$a_{nm} = \frac{c_{nm}b_{nm}}{d_{nm}b_{n-1,m}} \quad \text{при} \quad 1 \leqslant m \leqslant n-1. \tag{67}$$

В частности, отсюда легко получить

$$a_{n1} = \left(\frac{h_{n-1}}{h_n}\right)^p, \qquad a_{n2} = \frac{h_{n-1} - h_{n-2}}{h_n - h_{n-1}} \cdot \frac{h_n^{-p} - h_{n-1}^{-p}}{h_{n-1}^{-p} - h_{n-2}^{-p}},$$

$$a_{n3} = \frac{(h_{n-1} - h_{n-3})(h_{n-2} - h_{n-3})}{(h_n - h_{n-1})(h_n - h_{n-2})} \times$$

$$\times \frac{(h_{n-1} - h_{n-2})(h_n^{-p} - h_{n-1}^{-p}) - (h_n - h_{n-1})(h_{n-1}^{-p} - h_{n-2}^{-p})}{(h_{n-1} - h_{n-3})(h_{n-1}^{-p} - h_{n-2}^{-p}) - (h_{n-1} - h_{n-2})(h_{n-2}^{-p} - h_{n-3}^{-p})}.$$

Рассчитаем по формулам (64)–(67) четыре треугольные матрицы c_{nm}, d_{nm}, b_{nm} и a_{nm}, зависящие только от набора сеток, но не от x. Запомним матрицу a_{nm}. С ее помощью уже при любых значениях x можно вычислять уточненные значения $y_{nm}(x)$ и их априорные оценки погрешности $\Delta_{nm}(x)$ по формулам (62) и (55).

Для полноты изложения отметим два важных частных случая (известных ранее), когда матрица a_{nm} записывается особенно просто. В первом случае исходная схема имеет первый порядок точности; тогда (см. [Хайрер и др., 1990])

$$a_{nm} = \frac{h_{n-m}}{h_n} \quad (p = 1). \tag{68}$$

Во втором случае шаги последовательных сеток убывают в геометрической прогрессии со знаменателем $1/l$; тогда (см., например, [Марчук, Шайдуров, 1979])

$$a_{nm} = l^{p+m-1} \quad (h_n = h_{n-1}/l). \tag{69}$$

Наиболее часто используют $l = 2$, ибо это наименьшее число, когда каждая предыдущая сетка вложена в последующую (нетрудно понять, что совпадение узлов последовательных сеток существенно упрощает экстраполяцию).

В [Хайрер и др., 1990] приведены также некоторые менее важные случаи, когда коэффициенты γ_{knm} выражаются особенно просто.

Оптимизация стратегии расчета заключается в следующем. Матрицы коэффициентов (64)–(67) находят для несколько большего N,

чем реально рассчитывают использовать. Расчеты проводят сначала только на первых двух сетках, вычисляют уточнение и погрешность. Если достигнутая точность достаточна, вычисления прекращают.

Если погрешность оказалась больше заданной величины, то выполняют расчет на третьей сетке. Это позволяет провести уже два уточнения. При неудовлетворительном результате подключают следующую сетку и т.д. Таким образом объем вычислений получается минимальным, необходимым для получения заданной точности (при данной априорно выбранной последовательности сеток; вопрос оптимального закона построения сеток очень сложен и здесь не рассматривается).

Замечание. В m-м столбце матрицы стоят численные решения $y_{nm}(x)$, имеющие один и тот же $(n+m-1)$-й порядок точности. Чем больше n, тем меньшим шагам соответствует это решение, т. е. тем меньше фактическая погрешность. Таким образом, наивысшая точность достигается в нижней N-й строке матрицы. В этой строке наивысший порядок точности имеет значение.

Разность $y_{NN}(x) - y_{N,N-1}(x) = \Delta_{N,N-1}(x)$ есть оценка погрешности величины $y_{N,N-1}(x)$. Погрешность же величины $y_{NN}(x)$ остается не оцененной. Однако обычно в качестве наилучшего приближения к $y(x)$ берут именно величину $y_{NN}(x)$ и предполагают, что ее погрешность не превосходит $\Delta_{N,N-1}(x)$.

4. Диагностика. Сформулированные выше стратегия оптимизации и замечание не учитывают влияния двух факторов, а именно еще более высоких членов разложения (50) и ошибок округления. Последние приводят к тому, что при уменьшении шага сетки точность сначала улучшается, а потом начинает ухудшаться. Это явление хорошо продемонстрировано в [Марчук, Шайдуров, 1979] на примерах с точными решениями. Однако остается открытым вопрос: как обнаружить такую ситуацию в прикладном расчете, где точное решение неизвестно, и как при этом добиться заданной точности?

Предположим пока, что указанные два фактора несущественны. Тогда погрешность (55) с хорошей точностью равна первому члену разложения (54):

$$\Delta_{nm}(x) \approx \delta_{nm}(x), \tag{70}$$

где

$$\delta_{nm}(x) = v_{p+m-1}(x) \sum_{k=n-m+1}^{n} \gamma_{knm} h_k^{p+m-1}. \tag{71}$$

Здесь и далее знак приближенного равенства означает указание на роль этих двух факторов. Сравнивая выражение (71) с (63), можно увидеть, что

$$a_{nm} = \frac{\delta_{n-1,m}(x)}{\delta_{nm}(x)}.$$

Перемножая такие соотношения для первых индексов n, $n-1$, $n-2, \ldots, m+1$, получим

$$\delta_{nm}(x) = \frac{\delta_{mm}(x)}{\prod\limits_{k=m+1}^{n} a_{km}}. \tag{72}$$

Подставляя (62) в (55), найдем

$$\Delta_{nm}(x) = \frac{y_{nm}(x) - y_{n-1,m}(x)}{a_{nm} - 1}. \tag{73}$$

Соотношения (70), (72) и (73) позволяют осуществить диагностику расчета. Рассмотрим, как это следует делать (см. [Калиткин, 1978]).

Зафиксируем m (т. е. выберем столбец на рис. 8) и рассмотрим зависимость $\Delta_{nm}(x)$ от $\delta_{nm}(x)$. Если бы не было ни старших членов погрешности, ни ошибок округления, то эти величины были бы строго равны. Это особенно удобно изобразить на графике, где по осям ординат и абсцисс отложены соответственно

$$\eta_n = \lg|\Delta_{nm}(x)| \quad \text{и} \quad \xi_n = \lg \prod_{k=m+1}^{n} a_{km} \equiv -\lg|\delta_{nm}(x)| + æ_m(x),$$
$$æ_m(x) = \lg|\delta_{mm}(x)|, \tag{74}$$

тогда строгому равенству в (70) соответствовала бы прямая с наклоном -1 (рис. 9).

При малых n, т. е. больших h_n, роль старших членов заметна, а ошибки округления пренебрежимо малы; при больших n картина обратная. Поскольку старшие члены регулярно зависят от шага, а погрешность округления — стохастически, то качественное поведение будет подобно ломаной, изображенной на рисунке 10. В ее левой части наклоны последовательных звеньев будут монотонно приближаться к -1, т. е. ломаная является вогнутой или выпуклой. Нарушение монотонности производной (звездочка на графике) указывает на появление ошибок округления; дальнейшее сгущение сетки надо продолжать крайне осторожно. Нарушение монотонности самой ломаной (две звездочки) означает преобладание ошибок округления; дальнейшее сгущение сетки недопустимо.

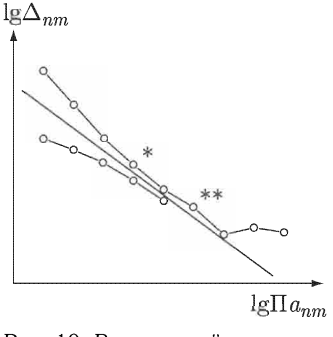

Рис. 10. Визуальный контроль влияния ошибок округления

Наилучшую точность, достижимую на выбранной последовательности сеток при вычислениях с данной разрядностью, примерно характеризует первый минимум ломаной на рисунке 10. Очевидно, чем

больше разрядность, тем глубже этот минимум. Влияние выбора последовательности сеток хорошо исследовано в [Марчук, Шайдуров, 1979].

Замечание. Данный анализ легко проводится в режиме диалога "человек–компьютер" с использованием графического изображения. Алгоритмизация его несколько более сложная.

Полная диагностика. Выше описан анализ, позволяющий выбрать наилучший отрезок последовательности сеток для фиксированного m. Полный анализ должен включать оптимизацию m, ибо многократное повышение порядка точности может оказаться невыгодным для решений, у которых $\max|y^{(r)}(x)|$ быстро возрастает с увеличением r. Опишем блок-схему соответствующего алгоритма, разработанного в [Калиткин, Кузьмина, 1975] для решения довольно трудной краевой задачи.

1. Задаем требуемую погрешность ε, выбираем закон построения последовательности сеток $\{h_n\}$ и рассчитываем матрицу a_{nm} при $1 \leqslant \leqslant n \leqslant N$, где N достаточно велико (реально $N \approx 6$–10).

2. Пусть выполнены расчеты решения $y_{k1}(x)$ при $1 \leqslant k \leqslant n$. Рассчитываем уточненные решения $y_{km}(x)$, $1 \leqslant m \leqslant k \leqslant n$.

3. Для начала анализа выбираем $n = 3$ или 4.

4. При каждом m проводим описанный выше анализ точности. Если при каком-нибудь m достигнута погрешность $|\Delta_{km}(x)| < \varepsilon$, то расчет заканчивается, и в качестве ответа выдаются значение $y_{k,m+1}(x)$ и погрешность $\Delta_{km}(x)$. Если ни при каком m не удается добиться заданной точности, то увеличиваем n на единицу и повторяем анализ.

5. Если при некотором $m = M$ на ломаной (рис. 10) достигнут минимум ε_m, который больше ε, то дальнейшее повышение порядка точности бессмысленно из-за ошибок округления. Надо увеличивать n, но рассматривать только $m < M$.

6. Если ни при каком m не удается достичь точности ε, то выбирается такое $\overline{m}$, при котором $\varepsilon_{\overline{m}} = \min \varepsilon_m$. В качестве ответа выдаются соответствующее значение $y_{k,\overline{m}}(x)$ и $\varepsilon_{\overline{m}}$ с примечанием "ошибки округления".

Разумеется, значения выбранных k, m, M, $\overline{m}$ зависят от x. Их тоже полезно выдавать для информации.

Частный случай. Особенно удобно в расчетах сгущение сеток вдвое, т. е. $h_n = h_{n-1}/l$ при $l = 2$. Тогда каждая предыдущая сетка вложена в последующую, а наличие совпадающих узлов существенно упрощает экстраполяцию. Они использовались в [Калиткин, Кузьмина, 1975], а в [Марчук, Шайдуров, 1979] показано их преимущество при вычислениях с не слишком большой разрядностью. В этом случае подстановка (69) в (74) дает

$$\varepsilon_n = (n - m)(p + m - 1)\lg l, \quad l = 2,$$

т. е. абсциссы ломаной на рисунке 10 будут эквидистантны.

5. Дополнение. Рассмотрим несколько существенных для практики ситуаций, на которые данные результаты обобщаются полностью или частично.

Симметричные схемы — это схемы, которые при одновременной смене знака шага h и направления нумерации индексов переходят сами в себя (см., например, [Хайрер и др., 1990]). Примерами таких схем являются формулы трапеций и Симпсона для численного интегрирования, или схема Кранка–Никольсон для параболического уравнения. На равномерных сетках погрешность таких схем разлагается в ряд, содержащий только четные степени шага. Тем самым вместо (51) получаем

$$y_{n1}(x) = u(x) + \sum_{r=p} v_{2r}(x) h_n^{2r}. \tag{75}$$

Нетрудно видеть, что в этом случае сохраняются все предыдущие рассуждения, только в формулах (52)–(74) всюду вместо h_n следует поставить h_n^2. Здесь каждое сгущение сетки повышает порядок точности не на единицу, а на двойку, т. е. экстраполяция особенно эффективна.

Квазиравномерные сетки. Пусть имеется последовательность сеток $\omega_n = \{x_i^{(n)}; i = 0, 1 \dots, I\}$, $n = 1, 2, \dots$ $(I_n < I_{n+1})$, составленная из неравномерных сеток, т. е. $h_i^{(n)} = x_{i+1}^{(n)} - x_i^{(n)} \neq \text{const}$, а зависит от i. Для произвольных неравномерных сеток описанная выше экстраполяция, очевидно, неприменима. Однако есть один важный случай, когда она возможна.

Возьмем некоторую функцию $x = \psi(t)$, строго возрастающую (т. е. $\psi'(t) > 0$) и имеющую достаточное число непрерывных производных. Выберем по переменной t последовательность равномерных сеток $\Omega_n = \{t_i^{(n)}; i = 0, 1, \dots, I_n\}$, у которых $\tau^{(n)} = t_{i+1}^{(n)} - t_i^{(n)} = \text{const} = (T/I_n)$ не зависит от i; здесь T — полная длина отрезка. Тогда сетки $\omega_n = \{x_i^{(n)}; i = 0, 1, \dots, I_n\}$, где $x_i^{(n)} = \psi(t_i^{(n)})$, называются *квазиравномерными*. Подчеркнем, что понятие квазиравномерности относится не к одной неравномерной сетке, а к их последовательности (насколько нам известно, квазиравномерные сетки были впервые предложены А.А. Самарским).

Если сетки квазиравномерны, то обычно нетрудно доказать, что разложение ошибки (51) сохраняется, только вместо шага по x надо ставить равномерный шаг по t или, что удобнее, обратную величину числа интервалов:

$$y_{n1}(x) = u(x) + \sum_{r=p} W_r(t) I_n^{-r}. \tag{76}$$

Разумеется, коэффициенты в членах разложения при этом другие и содержат производные не только $u(x)$, но и $\psi(t)$.

Доказательство такого разложения проводится для каждой конкретной схемы отдельно (см., например, [Калиткин, 1978]), причем

чем большее количество членов разложения требуется обосновать (т. е. чем большее число сгущений сетки предполагается использовать), тем большее число непрерывных производных должно иметь преобразование $x = \psi(t)$. Очевидно, в этом случае описанная выше экстраполяция на сгущающихся сетках сохраняется, если в формулы (52)–(74) подставить I_n^{-1} вместо h_n.

Поведение симметричных схем на квазиравномерных сетках более сложно. Исследование конкретных схем показывает, что разложение (76) для некоторых схем (например, квадратурной формулы трапеций) содержит только четные степени I_n, т. е. в формулы (52)–(74) следует подставлять I_n^{-2} вместо h_n, а каждое сгущение сетки повышает порядок точности на двойку. Для других же схем (например, схемы Кранка–Никольсон) нечетные члены разложения (76) не исчезают, и приходится пользоваться общей процедурой экстраполяции.

Особенности решения. Если некоторые производные решения $u(x)$ разрывны и/или неограниченны, то разложение погрешности не имеет вида (51), и описанную процедуру экстраполяции на сгущающихся сетках проводить нельзя.

Однако в ряде случаев (см., например, [Г.И. Марчук, В.В. Шайдуров, 1979]) для конкретных схем, определенных типов особенностей решения и специально подобранных последовательностей сеток удается установить справедливость разложений более общего вида:

$$y_{n1}(x) = u(x) + \sum_{r=1} v_r(x) h_n^{\alpha_r}, \quad 0 < \alpha_1 < \alpha_2 < \ldots \tag{77}$$

Здесь α_r — известные числа, не обязательно целые, а интервалы $\alpha_{r+1} - \alpha_r$ не обязательно одинаковые. В этом случае по аналогии с (52)–(54) нетрудно построить экстраполяцию

$$y_{nm}(x) = \sum_{k=n-m+1}^{n} \overline{\gamma}_{knm} y_{n1}(x), \tag{78}$$

где коэффициенты линейной комбинации (78) являются решением системы линейных уравнений

$$\sum_{k=n-m+1}^{n} \overline{\gamma}_{knm} = 1; \qquad \sum_{k=n-m+1}^{n} \overline{\gamma}_{knm} h_n^{\alpha_r} \equiv u(x) + O(h^{\alpha_m}), \quad 1 \leqslant r \leqslant m-1. \tag{79}$$

Погрешность экстраполяции (78) получаем подстановкой (78), (79) в (77):

$$y_{nm}(x) = u(x) + \sum_{r=m} v_r(x) \sum_{k=n-m+1}^{n} \overline{\gamma}_{knm} h_k^{\alpha_r} \equiv u(x) + O(h^{\alpha_m}), \tag{80}$$

т. е. подключение каждой новой сетки позволяет исключить очередной член разложения (77).

Дальнейшее упрощение формул (78)–(80), аналогичное выполненному для регулярных решений, в общем случае не удается сделать. Однако, несмотря на это, и для решений с особенностями метод экстраполяции на сгущающихся сетках является эффективным способом повышения точности при несложных исходных схемах и умеренном объеме вычислений.

6. Приложение. Приведем доказательства некоторых написанных выше формул. Рекуррентные соотношения (64), (65) последовательным умножением легко приводятся к явному виду:

$$c_{n1} = 1, \quad c_{n,m} = \prod_{i=n-m+1}^{n-1} (h_i - h_{n-m}) \ \text{ при } \ 2 \leqslant m \leqslant n-1, \tag{81}$$

$$d_{n1} = 1, \quad d_{n,m} = \prod_{i=n-m+1}^{n-1} (h_n - h_i) \ \text{ при } \ 2 \leqslant m \leqslant n-1. \tag{82}$$

Схемы первого порядка. Если $p = 1$, то из формул (64)–(66) нетрудно получить первые столбцы матрицы b_{nm}:

$$b_{n1} = \frac{1}{h_n}, \quad b_{n2} = -\frac{h_n - h_{n-1}}{h_n h_{n-1}},$$
$$b_{n3} = \frac{(h_n - h_{n-1})(h_n - h_{n-2})(h_{n-1} - h_{n-2})}{h_n h_{n-1} h_{n-2}}.$$

Отсюда легко угадывается общий вид этой матрицы:

$$b_{nm} = (-1)^{m-1} \frac{\displaystyle\prod_{n-m+1 \leqslant j < i \leqslant n} (h_i - h_j)}{\displaystyle\prod_{q=n-m+1}^{n} h_q}. \tag{83}$$

Выражение легко доказывается методом полной математической индукции. В самом деле, оно верно при $m \leqslant 3$. Пусть оно верно вплоть до некоторого m. Тогда, подставляя (81) и (83) в (66), получим

$$(-1)^{m-1} b_{n,m+1} =$$
$$= \frac{\displaystyle\prod_{r=n-m+1}^{n-1} (h_r - h_{n-m}) \cdot \prod_{n-m+1 \leqslant j < i \leqslant n} (h_i - h_j)}{\displaystyle\prod_{q=n-m+1}^{n} h_q} - \frac{\displaystyle\prod_{r=n-m+1}^{n-1} (h_n - h_r) \cdot \prod_{n-m \leqslant j < i \leqslant n-1} (h_i - h_j)}{\displaystyle\prod_{q=n-m}^{n-1} h_q}. \tag{84}$$

Заметим, что

$$\prod_{q=n-m+1}^{n} h_q = \prod_{q=n-m}^{n} \frac{h_q}{h_{n-m}},$$

$$\prod_{r=n-m+1}^{n-1} (h_r - h_{n-m}) = \prod_{r=n-m+1}^{n} \frac{h_r - h_{n-m}}{h_n - h_{n-m}},$$

$$\prod_{r=n-m+1}^{n} (h_r - h_{n-m}) \cdot \prod_{n-m+1\leqslant j<i\leqslant n} (h_i - h_j) = \prod_{n-m\leqslant j<i\leqslant n} (h_i - h_j). \tag{85}$$

С учетом этих равенств первое слагаемое в (84) преобразуется к виду

$$\frac{h_{n-m}}{h_n - h_{n-m}} \frac{\prod\limits_{n-m\leqslant j<i\leqslant n} (h_i - h_j)}{\prod\limits_{q=n-m}^{n} h_q}.$$

Аналогично преобразуем второе слагаемое (84) к таким же произведениям, только общим множителем в этом случае будет $h_n/(h_n - h_{n-m})$. Подстановка этих выражений в (84) дает

$$b_{n,m+1} = (-1)^m \frac{\prod\limits_{n-m\leqslant j<i\leqslant n} (h_i - h_j)}{\prod\limits_{q=n-m}^{n} h_q}.$$

Сравнивая его с (83), завершаем доказательство.

Теперь подставим выражения (81), (82) и (83) в формулу (67), что дает

$$a_{nm} = \frac{\prod\limits_{r=n-m+1}^{n-1} (h_r - h_{n-m}) \cdot \prod\limits_{n-m+1\leqslant j<i\leqslant n} (h_i - h_j) \cdot \prod\limits_{q=n-m}^{n-1} h_q}{\prod\limits_{r=n-m+1}^{n-1} (h_n - h_r) \cdot \prod\limits_{n-m\leqslant j<i\leqslant n} (h_i - h_j) \cdot \prod\limits_{q=n-m+1}^{n} h_q}.$$

Применяя к числителю и знаменателю преобразования типа (85), легко получим формулу (68):

$$a_{nm} = \frac{h_{n-m}}{h_n},$$

что и требовалось доказать.

Общий случай. Приведем основные этапы доказательства справедливости формул (64)–(67) для произвольного p. Из (66) и (81), (82) легко вычислить

$$b_{n2} = h_n^{-p} - h_{n-1}^{-p},$$

$$b_{n3} = (h_{n-1} - h_{n-2})h_n^{-p} - (h_n - h_{n-2})h_{n-1}^{-p} + (h_n - h_{n-1})h_{n-2}^{-p}.$$

Отсюда угадывается общая не рекуррентная запись

$$b_{nm} = \sum_{q=n-m+1}^{n} (-1)^{n-q} h_q^{-p} \cdot \prod_{n-m+1 \leqslant j < i \leqslant n;\, i,j \neq q} (h_i - h_j). \tag{86}$$

Она доказывается аналогично предыдущему случаю, методом полной математической индукции; при этом используются преобразования типа (85), среди которых отметим одно:

$$\prod_{r=n-m+1}^{n-1} (h_r - h_{n-m}) \cdot \prod_{n-m+1 \leqslant j < i \leqslant n;\, i,j \neq q} (h_i - h_j) =$$

$$= \frac{h_q - h_{n-m}}{h_n - h_{n-m}} \cdot \prod_{n-m \leqslant j < i \leqslant n;\, i,j \neq q} (h_i - h_j). \tag{87}$$

Из (67), (81), (82) и (86) следует

$$a_{nm} = -\frac{\displaystyle\sum_{q=n-m+1}^{n} (-1)^q (h_q - h_{n-m}) h_q^{-p} \cdot \prod_{n-m \leqslant j < i \leqslant n,\, i,j \neq q} (h_i - h_j)}{\displaystyle\sum_{q=n-m}^{n-1} (-1)^q (h_n - h_q) h_q^{-p} \cdot \prod_{n-m \leqslant j < i \leqslant n,\, i,j \neq q} (h_i - h_j)}, \tag{88}$$

причем обе суммы можно брать в одинаковых пределах $q \leqslant n - m \leqslant n$, ибо добавляемые при этом слагаемые равны нулю.

Далее заметим, что формула (63) — не единственное представление a_{nm} через γ_{knm}. Величины $y_{n,m+1}(x)$ и $y_{n-1,m}(x)$ суть линейные комбинации величин $y_{k1}(x)$ с диапазоном индексов $n - m \leqslant k \leqslant n$. При этом $y_{n1}(x)$ *не входит* в $y_{n-1,m}(x)$, а $y_{n-m,1}(x)$ *не входит* в $y_{n,m}(x)$. Подставляя выражения типа (53) в (62) и сравнивая коэффициенты при $y_{n1}(x)$ и $y_{n-m,1}(x)$, получим более удобные представления:

$$a_{nm} = \frac{\gamma_{nn,m+1}}{\gamma_{nn,m+1} - \gamma_{nnm}} \quad \text{или} \quad a_{nm} = 1 - \frac{\gamma_{n-m,n-1,m}}{\gamma_{n-m,n,m+1}}. \tag{89}$$

Из (60) с учетом того, что $k = n$, получаем:

$$\gamma_{nnm} = h_n^{-p} \frac{\displaystyle\prod_{n-m+1 \leqslant j < i \leqslant n-1} (h_i - h_j)}{\displaystyle\sum_{q=n-m+1}^{n} (-1)^{q-n} h_q^{-p} \prod_{n-m+1 \leqslant j < i \leqslant n,\, i,j \neq q} (h_i - h_j)}.$$

Увеличив здесь индекс m на единицу, запишем $y_{nn,m+1}$. После этого с учетом преобразования (87) найдем

$$\frac{\gamma_{nnm}}{\gamma_{nn,m+1}} = \frac{(h_n - h_{n-m}) \sum\limits_{q=n-m}^{n} (-1)^q h_q^{-p} \prod\limits_{n-m\leqslant j<i\leqslant n, i,j\neq q} (h_i - h_j)}{\sum\limits_{q=n-m+1}^{n} (-1)^q (h_q - h_{n-m}) h_q^{-p} \prod\limits_{n-m\leqslant j<i\leqslant n, i,j\neq q} (h_i - h_j)},$$

причем нижнюю сумму можно брать в тех же пределах, что и верхнюю, ибо слагаемое с $q = n - m$ в ней равно нулю. Подставляя последнее выражение в первую из формул (89), убеждаемся в совпадении найденного a_{nm} с (88). Тем самым формулы (64)–(67) доказаны.

Аналогично проводится преобразование второй из формул (89) к форме (88); попутно это доказывает эквивалентность обеих формул (89).

Глава III

КВАЗИРАВНОМЕРНЫЕ СЕТКИ

В этой главе введено определение квазиравномерных сеток, даны примеры одномерных сеток на конечном отрезке, полупрямой и прямой, а также в задачах слоистых сред. Показано, как строить многомерные регулярные квазиравномерные сетки. Изложены основные принципы вычисления интегралов и производных на квазиравномерных сетках. Показано, что метод сгущения сеток полностью применим для квазиравномерных сеток. Установлена связь между квазиравномерными и адаптивными сетками, а также методом замены переменных.

§ 1. Построение квазиравномерных сеток

1. Семейства сеток. Одна сетка является либо равномерной, либо неравномерной; никаких альтернатив здесь нет. Но при сгущении сеток рассматривают не одну сетку, а некоторое семейство сеток Ω_N с разными числами интервалов N. Тогда можно построить какие-то семейства неравномерных сеток, обладающие нужными нам свойствами. Очень важным является семейство квазиравномерных сеток.

Пусть нас интересуют сетки по x с числом интервалов N на отрезке $a \leqslant x \leqslant b$; обозначим их через $\Omega_N[x] = \{a = x_0 < x_1 < x_2 < \ldots < x_N = b\}$. Введем вспомогательную переменную ξ, $\alpha \leqslant \xi \leqslant \beta$. Рассмотрим некоторое преобразование $x(\xi)$, обладающее на отрезке $\alpha \leqslant \xi \leqslant \beta$ следующими тремя свойствами:

1) достаточно гладко, т. ч. существует достаточно много непрерывных ограниченных производных

$$\left|x^{(q)}(\xi)\right| \leqslant M_q, \quad q = 0, 1, \ldots, p, \quad p \gg 1; \quad \alpha \leqslant \xi \leqslant \beta; \tag{1}$$

2) строго монотонно:

$$x'(\xi) \geqslant m > 0; \quad \alpha \leqslant \xi \leqslant \beta; \tag{2}$$

3) преобразует отрезок $[\alpha, \beta]$ в отрезок $[a, b]$:

$$a = x(\alpha), \quad b = x(\beta). \tag{3}$$

Построим по переменной ξ на $[\alpha, \beta]$ равномерные сетки $\omega_N[\xi]$ со всевозможными числами интервалов $N = 1, 2, 3, \ldots$:

$$\xi_{nN} = \alpha + \frac{(\beta - \alpha)n}{N}, \quad n = 0, 1, \ldots, N; \tag{4}$$

индекс N в ξ_{nN} обычно будем опускать и писать просто ξ_n. Каждой сетке $\omega_N[\xi]$ преобразование $x(\xi)$ ставит в соответствие некоторую сетку $\Omega_N[x]$:

$$x_{nN} \equiv x_n = x(\xi_{nN}) \equiv x(\xi_n), \quad n = 0, 1, \ldots, N. \tag{5}$$

Таким образом, семейству равномерных сеток $\omega_N[\xi]$ сопоставлено некоторое семейство сеток $\Omega_N[x]$. Какими свойствами оно обладает?

Если преобразование $x(\xi)$ обладает свойствами (1)–(3), то семейство сеток $\Omega_N[x]$, порожденное семейством равномерных сеток $\omega_N[\xi]$, называется *квазиравномерным*.

Для краткости слово "семейство" часто опускают и говорят просто о квазиравномерных сетках. Семейство содержит сетки со всевозможными $N = 1, 2, 3, \ldots$; для практического использования из него обычно выбирают какие-то последовательности N, например сгущающиеся сетки с $N_k = r^k N_0$. На одном и том же отрезке $x \in [a, b]$ существует бесконечно много различных семейств квазиравномерных сеток, порожденных различными преобразованиями $x(\xi)$; те или иные преобразования подбирают в зависимости от задачи, которую предстоит решать на этих сетках.

У порождающей равномерной сетки $\omega_N[\xi]$ середина и вообще любая дробная часть интервала $[\xi_{n-1}, \xi_n]$ определяется по обычным линейным формулам (см. гл. II):

$$\begin{gathered} \xi_{n-1/2} = \frac{\xi_{n-1} + \xi_n}{2} = \alpha + \frac{(\beta - \alpha)(n - 1/2)}{N}; \\ \xi_{n-\gamma} = \gamma \xi_{n-1} + (1 - \gamma)\, \xi_n, \quad 0 \leqslant \gamma \leqslant 1. \end{gathered} \tag{6}$$

Однако для квазиравномерной сетки середина и любая дробная часть интервала $[x_{n-1}, x_n]$ строится с помощью того же нелинейного порождающего преобразования:

$$x_{n-1/2} = x(\xi_{n-1/2}), \quad x_{n-\gamma} = x(\xi_{n-\gamma}); \tag{7}$$

поэтому для нее линейные соотношения (6) неверны:

$$x_{n-1/2} \neq \frac{x_{n-1} + x_n}{2}, \quad x_{n-\gamma} \neq \gamma x_{n-1} + (1 - \gamma) x_n.$$

Все это удобно иллюстрировать с помощью графика функции, обратной к порождающему преобразованию $x(\xi)$. Отложим x по оси абсцисс, ξ — по оси ординат (рис. 1). На отрезке $[\alpha, \beta]$ оси ординат изображена равномерная сетка с целыми и полуцелыми узлами. По ней с помощью графика $x(\xi)$ на отрезке $[a, b]$ оси абсцисс построена квазиравномерная сетка.

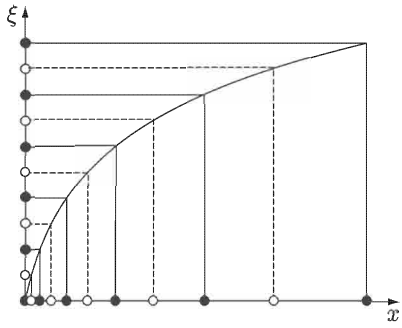

Рис. 1. Квазиравномерная сетка (пример 1). Точки — целые узлы сетки, кружки — полуцелые

Чтобы сгустить сетку $\Omega_N[x]$ в r раз, надо во столько же раз сгустить исходную сетку $\omega_N[\xi]$. Особенно наглядно сгущение вдвое ($r = 2$). Из формул (6), (7) и рисунка 1 видно, что при этом все целые узлы начальной сетки становятся четными узлами удвоенной сетки, а полуцелые точки начальной сетки — нечетными узлами удвоенной сетки. Это подтверждает разумность данного способа введения полуцелых (дробных) узлов квазиравномерной сетки.

Равномерная сетка $\omega_N[\xi]$ имеет постоянный шаг:

$$\Delta \equiv \xi_n - \xi_{n-1} = \frac{\beta - \alpha}{N} = \text{const} = O(N^{-1}). \tag{8}$$

Длины шагов h_n квазиравномерной сетки $\Omega_N[x]$ неодинаковы. Используя разложение $x(\xi)$ в ряд Тейлора–Маклорена с центром в полуцелой точке $\xi_{n-1/2}$, получим для них следующее выражение:

$$h_n \equiv x_n - x_{n-1} = x'(\xi_{n-1/2})\Delta + \frac{1}{6}x'''(\xi_{n-1/2})\Delta^3 + \ldots = O(N^{-1}); \tag{9}$$

благодаря симметрии этой формулы, разложение содержит степени Δ только одинаковой четности, а число членов в нем определяется гладкостью преобразования $x(\xi)$. Найдем из (9) отношение двух шагов с разными номерами:

$$\frac{h_m}{h_n} = \left[\frac{x'(\xi_{m-1/2})}{x'(\xi_{n-1/2})}\right] \cdot \left[1 + O(\Delta^2)\right]. \tag{10}$$

Вторая квадратная скобка здесь почти всегда равна 1 с точностью до $O(N^{-2})$. Но если интервалы соседние, т. е. $m = n + 1$, то первая квадратная скобка тоже почти равна 1. Действительно, $x'(\xi_{n\pm1/2}) = x'(\xi_n) \pm 0.5x''(\xi_n)\Delta + O(\Delta^2)$ и (10) превращается в

$$\frac{h_{n+1}}{h_n} = 1 + \left[\frac{x''(\xi_n)}{x'(\xi_n)}\right]\Delta + O(\Delta^2) = 1 + O(N^{-1}). \tag{11}$$

Отношение соседних шагов стремится к 1 при $N \to \infty$; легко убе-

диться, что разность соседних шагов $h_{n+1} - h_n = O(N^{-2}) = O(h_n^2)$, как и писалось в предисловии. Вот почему такие сетки были названы квазиравномерными. В то же время из (10) видно, что отношение далеких шагов может быть любым, большим или маленьким, и не стремится к 1 при увеличении числа узлов сетки. На рисунке 1 хорошо видно, что первый и последний шаги h_n сильно различаются.

Разложением $x(\xi)$ в ряд Тейлора нетрудно также проверить, что традиционное определение середины интервала и новое (7) весьма близки:

$$\frac{x_n + x_{n-1}}{2} - x_{n-1/2} = \frac{x''(\xi_{n-1/2})\Delta^2}{4} = O(N^{-2}) = O(h_n^2). \qquad (12)$$

Однако напомним, что несмотря на малую разницу, определение (7) имеет принципиальное преимущество — точное совмещение при сгущении сетки вдвое; далее увидим, что в неограниченной области появляется дополнительное улучшение.

Заметим, что преобразование $x(\xi)$ вводит пользователь, ориентируясь на особенности конкретной задачи или класса задач. Удобно для несимметричных задач выбирать $\alpha = 0$ и $\beta = 1$, а для симметричных — $\alpha = -1$ и $\beta = 1$; тогда $\Delta = 1/N$ или $\Delta = 2/N$ соответственно.

Отметим также интересные соотношения между тройками соседних шагов, которые легко выводятся разложением в ряд Тейлора:

$$h_n = \frac{1}{2}(h_{n-1} + h_{n+1})\left[1 + O(N^{-2})\right] = (h_{n-1} \cdot h_{n+1})^{1/2}\left[1 + O(N^{-2})\right]; \qquad (13)$$

центральный шаг с относительной точностью $O(N^{-2})$ равен среднеарифметическому или среднегеометрическому своих левого и правого соседа. Поскольку $O(N^{-2})$ есть достаточно малая величина при реальных $N \sim 100$ и более, точность этих соотношений велика.

Сгущение. Еще любопытнее соотношения шагов при сгущении сетки ровно вдвое. Все узлы первой сетки становятся четными узлами новой. Нечетные же узлы новой сетки делят каждый интервал h_n на левый h'_n и правый h''_n. Аналогичными разложениями можно получить такие приближенные соотношения:

$$h'_n \approx \frac{3}{8}h_n + \frac{1}{8}h_{n-1} \approx \frac{1}{2}(h_n^3 h_{n-1})^{1/4} \approx \frac{5}{8}h_n - \frac{1}{8}h_{n+1} \approx \frac{1}{2}\left(\frac{h_n^5}{h_{n+1}}\right)^{1/4};$$
$$h''_n \approx \frac{3}{8}h_n + \frac{1}{8}h_{n+1} \approx \frac{1}{2}(h_n^3 h_{n+1})^{1/4} \approx \frac{5}{8}h_n - \frac{1}{8}h_{n-1} \approx \frac{1}{2}\left(\frac{h_n^5}{h_{n-1}}\right)^{1/4}; \qquad (14)$$

их относительная точность также $O(N^{-2})$.

Пример 1. Пусть рассматриваемая функция $u(x)$ быстро меняется вблизи левой границы отрезка $[a, b]$, но медленно — вдали от нее. Тогда надо сделать сетку $\Omega_N[x]$ густой вблизи левой границы, но вблизи

правой границы сетка может быть достаточно редкой. Можно взять, например, такое порождающее преобразование:

$$x(\xi) = a + (b-a)\frac{e^{c\xi}-1}{e^c-1}, \quad c>0, \quad 0 \leqslant \xi \leqslant 1. \tag{15}$$

Легко проверить, что преобразование (15) удовлетворяет всем требуемым условиям (1)–(3). Величина c есть управляющий параметр; чем он больше, тем сильнее сгущаются шаги h_n у левой границы.

Для преобразования (15) справедливы соотношения:

$$\begin{gathered} x'(\xi) = c(b-a)\frac{e^{c\xi}}{e^c-1}, \quad \frac{h_{n+1}}{h_n} = e^{c/N} = \text{const} \approx 1 + \frac{c}{N}, \\ h_1 \approx c(b-a)\frac{1}{(e^c-1)N}, \quad \frac{h_N}{h_1} \approx e^c. \end{gathered} \tag{16}$$

Отношения соседних шагов одинаковы, т. е. шаги образуют геометрическую прогрессию. Отношение наибольшего шага h_N к наименьшему h_1 уже при $c>3$ становится очень большим, а первый шаг при этом очень мал: в $(e^c-1)/c$ раз меньше шага равномерной сетки $h=(b-a)/N$.

Именно этот случай (при $c=2$) изображен на рисунке 1. Заметим, что если нужна сетка, густая вблизи правой границы и редкая вблизи левой, то можно также воспользоваться преобразованием (15), но взяв $c<0$.

Пример 2. Построим другое преобразование, порождающее сетку $\Omega_N[x]$ также с малыми интервалами вблизи левой границы и гораздо большими — вблизи правой:

$$x(\xi) = a + (b-a)\frac{(c-1)^m\xi}{(c-\xi)^m}, \quad c>1, \quad m>0, \quad 0 \leqslant \xi \leqslant 1; \tag{17}$$

здесь есть два управляющих параметра c и m. Выполнение условий (1)–(3) легко проверить. Качественно эта сетка похожа на пример 1. Величины шагов также монотонно увеличиваются слева направо (но, разумеется, не в геометрической прогрессии).

Для преобразования (17) выполняются следующие соотношения:

$$\begin{gathered} x'(\xi) = (b-a)(c-1)^m\frac{c+(m-1)\xi}{(c-\xi)^{m+1}}, \\ h_1 \approx (b-a)\frac{(1-1/c)^m}{N}, \quad h_N \approx (b-a)\frac{c+m-1}{(c-1)N}, \\ \frac{h_N}{h_1} \approx c^m\frac{c+m-1}{(c-1)^{m+1}} > (1-1/c)^{-m}. \end{gathered} \tag{18}$$

Видно, что величина наименьшего шага h_1 составляет $(1-1/c)^m$-ю долю от шага равномерной сетки $h=(b-a)/N$ и становится очень малой при $c \to 1$ или $m \gg 1$. В этом случае отношение максимального шага h_N к минимальному еще больше, но сам максимальный шаг

остается не столь большим — всего лишь в $1 + m/(c-1)$ раз больше равномерного h.

Наиболее целесообразно полагать в (17), (18) значение $m = 1$, оставляя только один управляющий параметр c. Чем ближе c к 1, тем меньше начальные интервалы.

Сравнение примеров 1 и 2 показывает, что качественно сходные квазиравномерные сетки можно строить с помощью заметно отличающихся по виду преобразований.

Пример 3. Попробуем вместо (15) или (17) применить преобразование

$$x(\xi) = a + (b-a)\xi^2, \quad 0 \leqslant \xi \leqslant 1,$$

для построения сетки, подробной вблизи левой границы. Если построить график, аналогичный рисунку 1, то хорошо видно, что первые шаги сетки $\Omega_N[x]$ действительно будут очень малыми. Однако для этой сетки $h_1 = (b-a)/N^2$ и $h_2 = 3(b-a)/N^2$, т. ч. отношение $h_2/h_1 = 3$ и не стремится к 1 при $N \to \infty$. Сетка оказалась не квазиравномерной.

Причина в том, что условие монотонности (2) не вполне соблюдено. Вместо строгого неравенства выполнено лишь нестрогое $x'(\xi) \geqslant 0$, обращающееся в равенство при $\xi = 0$. Именно вблизи этой точки отношение соседних шагов не стремится к 1. Такое недопустимо для многих приложений.

Пример 4. Нередко приходится решать задачи о процессах в слоистой среде с чередованием толстых и тонких слоев (например, о прохождении звука через оконный пакет из тонких стекол и довольно больших промежутков). Для правильной разностной аппроксимации уравнений сетки должны быть квазиравномерными и притом специальными (т. е. границы слоев должны быть узлами сетки), а число интервалов в тонких слоях — не малым. Но для экономичности надо, чтобы в толстых слоях интервалов не было слишком много — примерно столько же, сколько и в тонких. Построим пример такой сетки.

Для простоты выберем трехслойную симметричную конфигурацию общей толщины $2b$ с двумя граничными тонкими слоями (их толщина $a \ll b$) и толстым средним слоем. Границы этих слоев — точки $x = -b,\ -b+a,\ b-a,\ b$. Рассмотрим следующее преобразование:

$$x(\xi) = \frac{c\xi}{(1+d\xi^2)^{1/2}}, \quad x'(\xi) = \frac{c}{(1+d\xi^2)^{3/2}}, \qquad (19)$$
$$c > 0, \quad d > 0, \quad -1 \leqslant \xi \leqslant 1.$$

Оно удовлетворяет требованиям (1)–(3).

Одно условие для подбора параметров c, d очевидно: $x(1) = b$. В качестве второго условия возьмем $x(\xi_*) = b - a$, где ξ_* — некоторое выбранное нами число $(0 < \xi_* < 1)$. Это означает, что общее число интервалов сетки распределится так: доля ξ_* — в центральном толстом

слое, $(1-\xi_*)/2$ — в каждом тонком. Подстановка (19) в эти условия дает параметры

$$d = \frac{(b-a)^2 - b^2\xi_*^2}{a\,(2b-a)\,\xi_*^2} \approx \frac{b}{2a}(\xi_*^{-2}-1) \gg 1,$$
$$c = \frac{b(b-a)}{\xi_*}\left[\frac{1-\xi_*^2}{a\,(2b-a)}\right]^{1/2} \approx \left[b^3\frac{(\xi_*^{-2}-1)}{2a}\right]^{1/2} \gg b; \tag{20}$$

здесь приближенные равенства относятся к случаю $b \gg a$, точные — к общему.

Разумеется, величину ξ_* и полное число интервалов N надо выбирать так, чтобы $N(1-\xi_*)/2$ было целым числом. Только в этом случае узел сетки всегда будет попадать на границу раздела сред. На рисунке 2 показан пример такой сетки для $a=1$, $b=6$, $\xi^*=0.5$, $N=12$.

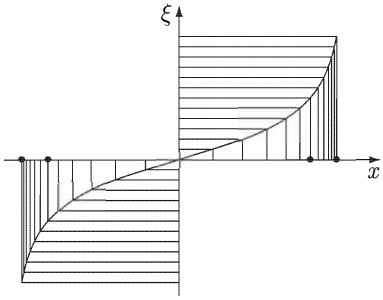

Рис. 2. Сетка (19) для трехслойной задачи. Точки — границы областей, а вертикальные линии соответствуют узлам сетки

Во всех рассмотренных примерах использовались преобразования $x(\xi)$, имеющие бесконечно много непрерывных производных, даже в многослойных задачах. Это оптимальная ситуация.

2. Неограниченная область. Выше неявно предполагалось, что $[a,b]$ — ограниченный отрезок. Однако существует немало задач в неограниченной области. Простейшим примером служат несобственные интегралы от $u(x)$ на полупрямой и прямой. В краевых задачах для дифференциальных уравнений также возможна постановка краевых условий в бесконечно удаленной точке; в задачах квантовой механики об энергетических уровнях в заданном поле, например, это условия быстрого затухания волновой функции на бесконечности.

Очевидно, в неограниченной области невозможно ввести равномерную сетку с конечным числом интервалов. Это до сих пор существенно ограничивало применение сеточных методов к подобным задачам. Однако оказалось, что квазиравномерную сетку в неограниченной области нетрудно построить. Покажем это.

Пусть задана полупрямая $[a, +\infty)$, т. е. область $a \leqslant x < +\infty$. Выберем такое преобразование $x(\xi)$ для конечного отрезка $\alpha \leqslant \xi \leqslant \beta$, чтобы оно удовлетворяло условиям строгой монотонности (2) и отображения границ отрезка (3). Последнее означает, что $x(\beta) = +\infty$. Условие достаточной гладкости придется изменить, ибо теперь все производные $x^{(q)}(\xi)$ будут обращаться в бесконечность при $\xi = \beta$. Поэтому вместо (1) потребуем, чтобы в области $\alpha \leqslant \xi < \beta$ существовало достаточно много непрерывных производных $x^{(q)}(\xi)$.

Эта идея полностью переносится на случай левой полупрямой $-\infty < x \leqslant b$ или случай прямой $-\infty < x < +\infty$. Напомним, что мы всегда можем выбрать такой масштаб по ξ, чтобы было $\alpha = 0$, $\beta = 1$ или $\alpha = -1$, $\beta = 1$. Проиллюстрируем сказанное примерами.

Пример 5. Пусть надо на полупрямой $a \leqslant x < +\infty$ построить квазиравномерную сетку $\Omega_N[x]$, наиболее густую вблизи левой границы. Возьмем преобразование

$$x(\xi) = a + \frac{c\xi}{(1-\xi)^m}, \quad x'(\xi) = c\frac{1 + (m-1)\xi}{(1-\xi)^{m+1}}, \quad 0 \leqslant \xi < 1; \tag{21}$$

здесь $c > 0$, $m > 0$ — управляющие параметры. Такая сетка построена на рисунке 3. Ее последний узел оказывается бесконечно удаленной

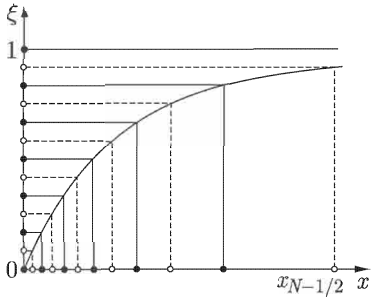

Рис. 3. Квазиравномерная сетка на полупрямой. Точки — целые узлы сетки, кружки — полуцелые

точкой, т. е. $x_N = +\infty$. Соответственно, последний интервал $[x_{N-1}, x_N)$ неограничен; разумеется, все остальные интервалы ограничены. Однако середина $x_{N-1/2}$ последнего неограниченного интервала оказывается конечной точкой, и то же относится к любой дробной точке $x_{N-\gamma}$, $0 < \gamma \leqslant 1$! Это еще раз демонстрирует разумность определения середины интервала (7).

Опишем простейший априорный способ подбора параметров преобразования (21). Целесообразное число интервалов сетки N определяется мощностью нашего компьютера. Если мы строим сетку, густую вблизи левой границы, то неявно предполагаем, что она наиболее важна для аккуратного решения задачи. Первому шагу соответствует $\xi_1 =$

$= 1/N$, а первой половине всех шагов — $\xi_{N/2} = 1/2$; подставляя их в (21), получаем:

$$h_1 \approx \frac{c}{N}, \quad x_{N/2} - a = 2^{m-1}c. \tag{22}$$

Величина первого шага практически не зависит от m; поэтому c подбирают так, чтобы первый шаг оказался достаточно малым и обеспечивал нужную точность расчета. Найдя c, подбираем m из условия, чтобы половина интервалов сетки охватила наиболее важную область полупрямой; чем шире эта область, тем большее m приходится брать.

Пример 6. Пусть на полупрямой в примере 5 подбор параметров дает $m < 1$. Это вряд ли целесообразно. Лучше предложить другое преобразование с похожими качествами и одним управляющим параметром:

$$x(\xi) = a - c\ln(1-\xi), \quad x'(\xi) = \frac{c}{1-\xi}, \quad 0 \leqslant \xi < 1. \tag{23}$$

Первый шаг для него точно такой же, как (22), но полсетки имеет иную протяженность:

$$h_1 \approx \frac{c}{N}, \quad x_{N/2} - a = c\ln 2. \tag{24}$$

Качественно сетка выглядит так же, как на рисунке 3.

Примеры 5 и 6 легко обобщаются на полупрямую $-\infty < x \leqslant a$.

Пример 7. Построим на прямой $-\infty < x < +\infty$ квазиравномерную сетку, наиболее густую вблизи точки $x = a$. Для этого возьмем преобразование, аналогичное (21) из примера 5, но имеющее два полюса:

$$x(\xi) = a + \frac{c\xi}{(1-\xi^2)^m}, \quad x'(\xi) = c\frac{1+(2m-1)\xi^2}{(1-\xi^2)^{m+1}}, \quad -1 < \xi < 1; \tag{25}$$

здесь $c > 0$, $m > 0$ — параметры. Значения параметров подбирают аналогично примеру 5. Так, наименьшим шагом будет центральный

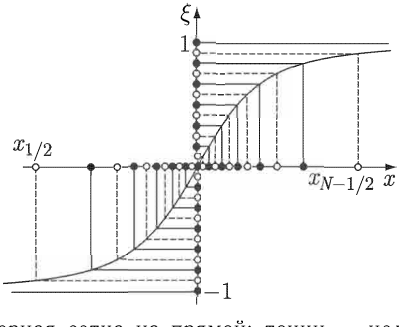

Рис. 4. Квазиравномерная сетка на прямой; точки — целые узлы, кружки — полуцелые

(при N нечетном он один, при четном N их два); с учетом того, что $\beta - \alpha = 2$, величина наименьшего шага равна $h_{N/2} \approx 2c/N$. Это позволяет подобрать c. Качественно процедура построения этой сетки изображена на рисунке 4; теперь бесконечно удаленными точками являются узлы x_0 и x_N.

Сделаем два замечания. Во-первых, мы неявно предполагали, что узлы ξ_n и x_n имеют нумерацию $n = 0, 1, \ldots, N$; это позволяет брать любое N, четное и нечетное. Но на интервале $-1 < \xi < 1$ можно также естественно ввести другую нумерацию узлов: $-N \leqslant n \leqslant N$; тогда узел ξ_0 оказывается центральным, а узлы x_{-N} и x_N — бесконечно удаленными точками. Такая симметрия представляет определенное удобство, но полное число интервалов при этом обязательно четное.

Во-вторых, если взять преобразование (25) только на отрезке $0 \leqslant \xi < 1$, то получим квазиравномерную сетку $\Omega_N[x]$ на правой полупрямой $a \leqslant x < +\infty$; она очень похожа на пример 5. Если же взять отрезок $-1 < \xi \leqslant 0$, то (25) даст квазиравномерную сетку на левой полупрямой $-\infty < x \leqslant a$.

Пример 8. Квазиравномерную сетку на прямой можно построить и аналогично примеру 6, с помощью логарифмически бесконечного преобразования:

$$x(\xi) = a - \frac{c}{\xi}\ln(1-\xi^2), \quad x'(\xi) = \frac{2c}{1-\xi^2} + \frac{c}{\xi^2}\ln(1-\xi^2), \quad -1 < \xi < 1. \tag{26}$$

Эта сетка также качественно похожа на рисунок 4, и ее центральный интервал $h_{N/2} \approx 2c/N$. Если взять преобразование (26) на отрезках $-1 < \xi \leqslant 0$ или $0 \leqslant \xi < 1$, оно дает квазиравномерные сетки на левой или правой полупрямой.

Пример 9. Приведем еще одно красивое преобразование, дающее квазиравномерную сетку на прямой:

$$x(\xi) = a + c \cdot \operatorname{tg}\xi, \quad x'(\xi) = \frac{c}{\cos^2\xi}, \quad -\frac{\pi}{2} < \xi < \frac{\pi}{2}. \tag{27}$$

Оно очень близко к преобразованию (25) с $m = 1$. Но его нагляднее можно иллюстрировать на рисунке 5, где ξ есть угол, а радиус окружности равен c; проекции равных дуг окружности на прямую дают интервалы сетки. Если ограничиться отрезками $-\pi/2 < \xi \leqslant 0$ или $0 \leqslant \xi < \pi/2$, то преобразование (27) дает сетки на полупрямых.

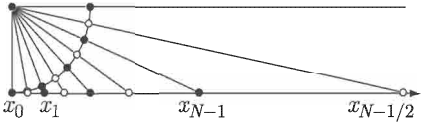

Рис. 5. Квазиравномерная сетка (27) на полупрямой; точки — целые узлы, кружки — полуцелые

Во всех рассмотренных примерах преобразования $x(\xi)$ имели бесконечно много непрерывных на $\alpha < \xi < \beta$ производных. Это важно для метода сгущения сеток. Надо по возможности избегать преобразований с разрывами даже высоких производных.

3. Адаптивность. В современных вычислениях, особенно в задачах для уравнений в частных производных, очень популярны так называемые *адаптивные* сетки. Это сетки, передающие не только общие особенности некоторого класса задач, но и характер конкретного искомого решения. Их строят обычно не до решения задачи, а в ходе ее решения, одновременно с искомым точным решением. Алгоритмы их построения разнообразны и довольно сложны.

Простейшим примером является автоматический выбор шага при решении задачи Коши для обыкновенных дифференциальных уравнений (см., например, [Хайрер и др.; 1990] или [Хайрер, Ваннер; 1999]). При этом задают допустимую ошибку ε и с помощью некоторого критерия (по локальному сгущению шага или по так называемым вложенным схемам) определяют такую величину очередного шага h_n, при которой вносимая на нем погрешность не превышает ε. В результате шаг h_n выбирается большим там, где искомое решение $u(x)$ меняется медленно, и малым в районах быстрого изменения решения.

На рисунке 6 качественно показан автоматический выбор шага в задаче о горении пороха. Процесс начинается медленно и считается

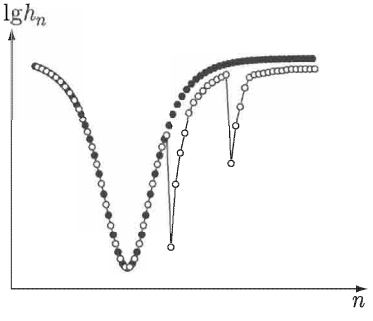

Рис. 6. Автоматический выбор шага в задаче теплового взрыва; • — надежно работающий алгоритм выбора, ○ — ненадежный

с крупным шагом. Затем горение ускоряется почти до взрыва и шаг уменьшается в сотни и тысячи раз. Когда большая часть реакции протекла, идет медленное догорание, и шаг сильно увеличивается.

Сгущение. Адаптивные сетки позволяют добиться хорошей точности расчета при умеренном числе интервалов сетки N. Однако контроль точности при этом можно проводить лишь на уровне интуитивного приема, предлагавшегося Рунге в 1895 г.: провести второй расчет, уменьшив все шаги в $r \geqslant 2$ раз, сравнить результаты и считать верными

совпадающие знаки. Впрочем, сравнить результаты будет непросто, ибо все узлы этих двух сеток могут не совпадать.

Заметим, что выбранные шаги (сетка) адаптированы к решению $u(x)$ и численному методу его нахождения. Если сменить метод численного интегрирования, то для того же решения $u(x)$ будет выбрана другая сетка.

Контроль. Адаптивные сетки во многом похожи на квазиравномерные. Однако в расчетах на адаптивных сетках есть принципиальная трудность: как получить апостериорную оценку точности? Разумеется, надо уменьшать ε, но по какому закону — неизвестно. Расчеты на тестовых примерах показывают, что на одних классах задач фактическая погрешность может быть больше ε, на других — меньше. Вдобавок адаптивные сетки для разных ε не вполне подобны друг другу: если на рисунке 6 ввести по оси абсцисс единый для всех сеток отрезок $0 \leqslant \xi \leqslant 1$, где $\xi = n/N$, то они станут близкими, но все же будут несколько отличаться.

Такое отсутствие четких закономерностей не позволяет использовать для адаптивных сеток те приемы сгущения сеток, которые хорошо работают для равномерных сеток (как показано в гл. II) и квазиравномерных сеток (как будет показано в § 2).

Преодолеть данную трудность можно следующим образом. Выберем достаточно (но не слишком) малое ε. Построим соответствующую адаптивную сетку $\Omega_N[x]$ с числом узлов $N(\varepsilon)$ и шагами $h_n(\varepsilon)$. Ее узлами и областью определения будут

$$x_0 = a, \quad x_n = x_0 + \sum_{k=1}^{n} h_k, \quad 1 \leqslant n \leqslant N; \quad x_N = b. \tag{28}$$

Аппроксимируем адаптивную сетку (28) некоторым достаточно гладким и строго монотонным преобразованием $x \approx \Phi(\xi)$, $\xi \in [0, 1]$, так, чтобы точно передавались концы отрезка:

$$a = x_0 = \Phi(0), \quad b = x_N = \Phi(1), \tag{29}$$

и приближенно, но с хорошей относительной точностью передавались шаги сетки:

$$\frac{\Phi(\xi_n) - \Phi(\xi_{n-1})}{h_n} \approx 1, \quad 1 \leqslant n \leqslant N, \quad \xi_n = \frac{n}{N}. \tag{30}$$

Квазиравномерные сетки, порожденные преобразованием $\Phi(\xi)$, сгущаются по обычным правилам п. 1; на них, как будет показано в § 2, можно получать апостериорные оценки точности аналогично равномерным сеткам. В то же время они очень близки к адаптивным сеткам.

Разряжение. Можно предложить другой способ. Выберем заведомо малое ε, чтобы обеспечить достаточно высокую точность. Слегка поварьируем его, чтобы число интервалов N нацело делилось на 2^k.

На полученной сетке $\Omega_N[x]$ выбросим все нечетные узлы и получим сетку $\Omega_{N/2}[x]$. В ней также выбросим все нечетные узлы, получив сетку $\Omega_{N/4}[x]$, и так k раз. Наиболее подробная сетка может быть точно описана некоторым строго монотонным преобразованием $\Phi(\xi)$: достаточно построить монотонную достаточно гладкую интерполяцию. Тогда вся описанная совокупность сеток будет порождена этим преобразованием, т. е. являться квазиравномерной последовательностью. Сравнение $u(x)$ на этих сетках производить удобно, ибо они имеют много совпадающих узлов.

Аппроксимация. Целесообразен следующий способ построения аппроксимации $\Phi(\xi)$, удовлетворяющей условиям (29), (30). Выберем на отрезке $0 \leqslant \xi \leqslant 1$ некоторую систему линейно независимых функций $\varphi_m(\xi)$ и построим обобщенный многочлен:

$$\Phi(\xi) = \sum_{m=0}^{M} c_m \varphi_m(\xi), \tag{31}$$

где c_m — свободные параметры. Подставляя (31) в (29), получим два линейных уравнения, позволяющие исключить два параметра. Условие (30) будем трактовать методом наименьших квадратов:

$$\sum_{n=1}^{N} \left[\frac{\Phi(\xi_n) - \Phi(\xi_{n-1})}{h_n} - 1 \right]^2 = \min. \tag{32}$$

Приравнивая нулю производные от квадратичной формы (32) по оставшимся независимым параметрам, получим систему линейных уравнений для этих параметров. Отсюда видно, что должно выполняться $M \leqslant N + 1$, т. е. обобщенный многочлен не может содержать слишком много слагаемых. На практике желательно $M \ll N + 1$ (если при этом соотношения (30) выполняются с приличной точностью); тогда это естественное требование для метода наименьших квадратов.

Построив преобразование $\Phi(\xi)$, надо проверить, будет ли оно строго монотонным (сам по себе описанный метод этого не гарантирует).

Сплайны высокой степени p могут оказаться особенно простыми для описанной аппроксимации. Введем на отрезке $0 \leqslant \xi \leqslant 1$ небольшое число так называемых узлов сплайна η_m:

$$0 = \eta_0 < \eta_1 < \eta_2 < \ldots < \eta_M = 1, \quad M \ll N. \tag{33}$$

Граничные узлы η_0, η_M совпадают с граничными узлами сетки ξ_0, ξ_N, но внутренние могут не совпадать. Полиномиальный сплайн $S(\xi)$ степени p дефекта 1 имеет, как известно, $(p-1)$-ю непрерывную производную, а его p-я производная кусочно постоянна и разрывна в узлах η_m.

Запишем этот сплайн в так называемой глобальной форме:

$$S(\xi) = \sum_{k=0}^{p} a_k(\xi - \eta_0)^k + \sum_{m=1}^{M-1} b_m(\xi - \eta_m)_+^p, \quad \eta_0 \leqslant \xi \leqslant \eta_M, \tag{34}$$

где введена усеченная степенная функция

$$\xi_+^p = \begin{cases} \xi^p, & \xi \geqslant 0, \\ 0, & \xi \leqslant 0. \end{cases} \tag{35}$$

При такой записи все коэффициенты сплайна a_k, b_m являются независимыми свободными параметрами, а их число равно $p + M$ (не $M + 1$, как было выше). Первое из условий (29) сразу дает $a_0 = a$, а остальные находятся описанным выше алгоритмом.

Удвоение. Однократное сгущение адаптивной сетки в $r = 2$ раза можно провести гораздо более простым способом. Воспользуемся формулами (14) деления шага "пополам" для квазиравномерной сетки. Они приближенные, т. ч. при расчете прямо по ним получится $h'_n + h''_n \neq \neq h_n$. Надо ввести нормировку так, чтобы получить равенство. Кроме того, для каждого полушага есть четыре варианта записи; какой из них выбрать?

Поскольку адаптивные сетки не всегда бывают высокого качества (h_{n+1}/h_n может заметно отличаться от 1), безопаснее выбирать величины с дробными степенями; для каждой половины шага предпочтительней ориентироваться на вариант с ближайшим к ней соседом. С учетом этого получим для внутренних интервалов сетки:

$$h''_n + h'_n = h_n, \quad \frac{h''_n}{h'_n} = \left(\frac{h_{n+1}}{h_{n-1}}\right)^{1/4}, \tag{36}$$

откуда

$$h'_n = \frac{h_n}{1 + (h_{n+1}/h_{n-1})^{1/4}}, \quad h''_n = \frac{h_n}{1 + (h_{n-1}/h_{n+1})^{1/4}}, \quad 2 \leqslant n \leqslant N - 1. \tag{37}$$

Для первого и последнего интервала приходится пользоваться одним соседом, что дает

$$\begin{aligned} h'_1 &= \frac{h_1}{1 + (h_2/h_1)^{1/2}}, \quad h''_1 = \frac{h_1}{1 + (h_1/h_2)^{1/2}}; \\ h'_N &= \frac{h_N}{1 + (h_N/h_{N-1})^{1/2}}, \quad h''_N = \frac{h_N}{1 + (h_{N-1}/h_N)^{1/2}}. \end{aligned} \tag{38}$$

Формулы (37), (38) пригодны для однократного сгущения сетки. Применимость их к рекуррентному сгущению неясна и требует дополнительного исследования. Однако даже однократное сгущение позволяет получить апостериорную оценку погрешности.

Сбои алгоритмов. Заметим, что если соседние шаги адаптивной сетки h_n сильно отличаются, то условие монотонности любой аппроксимации $\Phi(\xi)$, в том числе сплайновой, может нарушаться. В ряде случаев удается добиться монотонности, используя иные базисные функции, в частности гиперболические сплайны. Но вообще это свидетельствует об огромных градиентах точного решения, а в этих условиях обычные численные методы не могут обеспечивать высокой точности.

Другой причиной сбоев может быть неудачный алгоритм построения самой адаптивной сетки (кружки на рис. 6). Наша практика расчетов показала, что даже известные зарубежные пакеты прикладных программ нередко дают такие сбои в совершенно безобидных участках точного решения.

При сбоях любого типа правильное сгущение сетки становится невозможным, ибо в этом случае адаптивная сетка неэквивалентна квазиравномерной.

4. Многомерность. Пусть имеется многомерная область G. Введем в ней координаты $x, y, \ldots$ Координатные линии (поверхности, гиперповерхности) $x = \text{const}$, $y = \text{const}, \ldots$ могут быть кривыми (неплоскими) и неортогональными. Но пересечения этих линий или поверхностей $x = x_n$, $y = y_m, \ldots$ образуют регулярную сетку $r_{nm\ldots} = (x_n, y_m, \ldots)$. В этом случае можно построить квазиравномерные сетки.

Для этого по каждой координате вводим свое порождающее преобразование $x(\xi)$, $y(\eta), \ldots$ и по переменным ξ, $\eta, \ldots$ берем равномерные сетки. Особенно просто это делается, если область G в координатах x, $y, \ldots$ есть криволинейный четырехугольник (параллелепипед). Тогда в переменных ξ, $\eta, \ldots$ эта область становится прямоугольником (прямоугольным параллелепипедом). Вдоль каждой координатной линии ξ меняется в одних и тех же пределах. Можно считать, что $0 \leqslant \xi \leqslant 1$, и ввести одномерную сетку $\xi_n = n/N$, $0 \leqslant n \leqslant N$. Аналогичную сетку со своим числом интервалов M можно ввести по второй переменной: $\eta_m = m/M$, $0 \leqslant m \leqslant M$, и т.д.

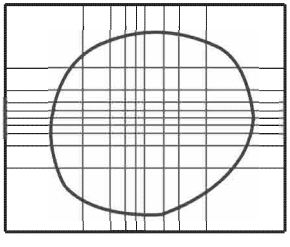

Рис. 7. Двумерная квазиравномерная сетка

Если область G не является таким прямоугольником, то ее можно поместить внутрь некоторого прямоугольника (рис. 7) и применить к нему описанную процедуру. При этом отрезки координатных линий, попавшие внутрь G, будут содержать не все возможные значения индексов $n, m, \ldots$ ... Но многомерная сетка все равно будет квазиравномерной, а ее сгущение выполняется сгущением равномерных сеток по ξ, $\eta, \ldots$

Нетрудно видеть, что сгущение двумерной сетки ровно в $r = 2$ раза по каждой переменной аналогично одномерному случаю. Каждая

координатная линия исходной сетки становится четной линией новой, а каждая "полуцелая" линия исходной сетки становится нечетной линией новой. Таким образом, каждая ячейка исходной сетки делится на четыре новые ячейки.

Однако во многих задачах коэффициенты сгущения по разным координатам бывает выгодно брать различными. Такие примеры показаны в следующих главах.

Отметим также, что можно строить квазиравномерные сетки не только с четырехугольными ячейками, но также с треугольными или шестиугольными. Треугольные получаются одинаковым диагональным делением квазиравномерных четырехугольных, а шестиугольные — объединением треугольных (рис. 8).

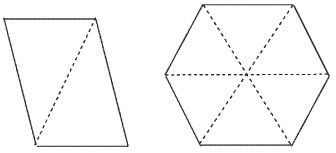

Рис. 8. Примеры построения сеток с треугольными и шестиугольными ячейками

Неограниченная область не представляет принципиальных затруднений. Пусть, например, надо построить сетку на полуплоскости $y \geqslant 0$. Возьмем прямоугольные координаты в диапазонах $-\infty < x < +\infty$, $0 \leqslant y < +\infty$. Подберем соответствующие одномерные преобразования, описанные в п. 2. Получим квазиравномерную регулярную сетку, качественно изображенную на рисунке 9.

Можно поступить иначе. Перейдем от декартовых координат (x, y)

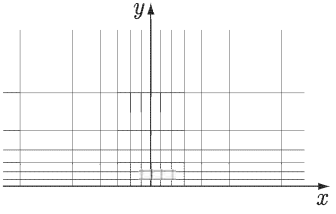

Рис. 9. Квазиравномерная сетка в полуплоскости

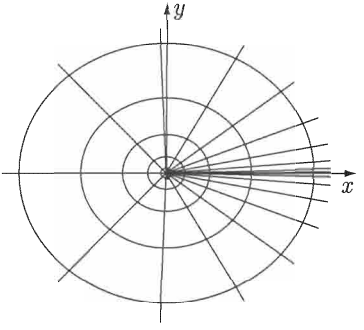

Рис. 10. Квазиравномерная сетка в плоскости

к полярным (ρ, φ) (в трехмерном случае — к сферическим). Неограниченна лишь радиальная координата $0 \leqslant \rho < +\infty$, и область ее определения есть полупрямая. Угловая координата ограничена на полуплоскости отрезком $0 \leqslant \varphi \leqslant \pi$, а на плоскости — отрезком $0 \leqslant \varphi \leqslant 2\pi$.

Отметим только одну тонкость для плоскости: зависимость всех функций от угла является 2π-периодической. Поэтому преобразование для угла $\varphi(\eta)$, $0 \leqslant \eta \leqslant 1$, должно быть достаточно гладким вместе со своим периодическим продолжением (рис. 10). Нарушение этого условия может катастрофически ухудшить точность многих расчетов.

Для полуплоскости последнего ограничения не существует.

Адаптивность. В многомерных задачах также часто строят адаптивные сетки. Свести их к квазиравномерным в общем случае нельзя. Но если эти сетки регулярны, то можно построить процедуру их однократного сгущения вдвое. Проиллюстрируем это на примере двумерной сетки (рис. 11).

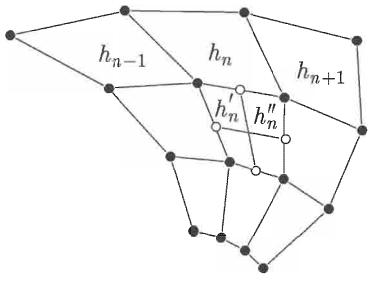

Рис. 11. Сгущение регулярной адаптивной сетки

Пусть для определенности регулярная сетка состоит из четырехугольников (узлов и соединяющих их прямых отрезков), которые образуют ломаные линии сетки (гладких кривых линий в адаптивных сетках нет). Вдоль такой линии берем три соседних шага h_{n-1}, h_n, h_{n+1}, равных длинам отрезков на рисунке 11. Затем по одномерным формулам (37) делим центральный отрезок на два: левый h'_n и правый h''_n. Если рассматриваемый отрезок опирается на границу области и имеет только одного соседа, то используют формулы (38).

Разделив таким образом каждую из четырех сторон ячейки на две части, соединяют полученные точки прямыми и делят исходную четырехугольную ячейку на четыре четырехугольные (пунктирные линии на рис. 11). Сгущенная сетка построена.

Такая процедура применима к регулярной сетке (рис. 11). Также сгущается регулярная треугольная адаптивная сетка.

Аналогично сгущается трехмерная регулярная сетка, построенная из параллелепипедов или тетраэдров: каждое ребро делится на две части по формулам (37), в граничных ячейках — (38); полученные точки соединяют плоскостями, делящими ячейку на 8 ячеек той же формы.

Применять данный метод не следует в трех случаях. Во-первых, для рекуррентного сгущения (сгустить этим методом можно многократно, но извлечь пользу из этого трудно, т. к. повышение точности на несколько порядков может и не произойти).

Во-вторых, для нерегулярных сеток (обычно это треугольные сетки). В этом случае неясно, какие отрезки, исходящие из одного узла, следует считать продолжением друг друга.

В-третьих, если исходная адаптивная сетка построена неудачно и содержит продолжающие друг друга отрезки с отношениями h_{n+1}/h_n существенно больше или меньше 1.

§ 2. Аппроксимация интегралов и производных

1. Аппроксимация интегралов. В главе II были написаны простейшие квадратурные формулы левых прямоугольников (II.13), а именно

$$U_N = \sum_{n=1}^{N} u_{n-1} h_n, \quad R_N \approx \frac{1}{2} \sum_{n=1}^{N} u'_{n-1} h_n^2, \quad h_n = x_n - x_{n-1}, \tag{39}$$

и средних (II.19):

$$U_N = \sum_{n=1}^{N} u_{n-1/2} h_n, \quad R_N \approx \frac{1}{24} \sum_{n=1}^{N} u''_{n-1/2} h_n^3. \tag{40}$$

Они составлены для произвольной сетки, в том числе и для квазиравномерной. Вместе с ними приведены главные члены их погрешностей (II.15) и (II.21). Напомним, что для достаточно гладких функций $u(x)$ погрешность в каждом n-м интервале сетки разлагается в ряд по степеням шага h_n; для несимметричной формулы (39) этот ряд содержит все степени h_n, а для симметричной формулы (40) — только четные степени.

Как ведут себя погрешности формул (39), (40) при увеличении числа интервалов сетки N, в общем случае сказать нельзя. Однако для квазиравномерных сеток это можно сделать. В самом деле, если квазиравномерная сетка $\Omega_N[x]$ порождена достаточно гладким преобразованием $x(\xi)$, то для ее шагов справедливо соотношение

$$\begin{gathered} h_n = x_n - x_{n-1} = x'_{n-1}\Delta[1 + O(\Delta)] = x'_{n-1/2}\Delta[1 + O(\Delta^2)], \\ \Delta = \xi_n - \xi_{n-1} = \frac{\beta - \alpha}{N} = \text{const}. \end{gathered} \tag{41}$$

Подставляя эти соотношения в погрешность формулы прямоугольников (39), получим

$$R_N \approx \Delta^2 \sum_{n=1}^{N} u'_{n-1} (x'_{n-1})^2 \approx \Delta \int_{\alpha}^{\beta} \frac{du}{dx} \left(\frac{dx}{d\xi}\right)^2 d\xi \approx \text{const} \cdot \Delta = \frac{\text{const}}{N}. \tag{42}$$

Аналогичная подстановка (41) в (40) дает погрешность формулы средних на квазиравномерной сетке:

$$R_N \approx \frac{\Delta^3}{24} \sum_{n=1}^{N} u''_{n-1/2} (x'_{n-1/2})^3 \approx \frac{\Delta^2}{24} \int_{\alpha}^{\beta} \frac{d^2u}{dx^2} \left(\frac{dx}{d\xi}\right)^3 d\xi \approx \text{const} \cdot \Delta^2 = \frac{\text{const}}{N^2}. \tag{43}$$

Для равномерных сеток погрешности этих формул имели вид $R_n \approx$ $\approx \text{const} \cdot h$ или $\approx \text{const} \cdot h^2$ соответственно. Видно, что для квазиравномерных сеток их вид оказался точно таким же, с очевидной заме-

ной $h \to \Delta$, и закон убывания при возрастании N тот же. Это значит, что при сгущении квазиравномерных сеток можно находить асимптотическую апостериорную оценку погрешности, а также повышать порядок точности по формулам (II. 31) и (II. 29), где коэффициент r тот же самый — отношение чисел интервалов N двух сеток.

Если провести разложение функций $u(x)$ и $x(\xi)$ в ряды по степеням Δ, можно получить соответствующие ряды для погрешностей, уточняющие главные члены (42) и (43). В этом случае для формулы прямоугольников к (42) добавляются члены ряда со всеми степенями $\Delta \sim N^{-1}$ подряд. Для формулы средних к (43) благодаря симметрии добавляются степени Δ только с одинаковыми четностями. Поэтому на квазиравномерных сетках можно проводить многократное сгущение сеток с рекуррентным повышением порядка точности и соответствующими апостериорными оценками погрешности, как и на равномерных сетках. По-прежнему для формулы средних $s = 2$, т. е. каждая лишняя сетка позволяет повысить порядок точности на 2.

Псевдоравномерные сетки. В главе II рассматривалось интегрирование функции $u(x)$, которая в одной точке (или в конечном числе точек) разрывна или имеет разрыв некоторой производной, но во всех остальных точках непрерывна и ограничена вместе с достаточным числом своих производных. Было показано, что для численного интегрирования таких функций целесообразно вводить специальные сетки, у которых все упомянутые точки разрыва являются узлами. Как сгущать такие сетки?

Построить квазиравномерную сетку, одновременно являющуюся специальной, в принципе можно, но достаточно трудно. Но для целей численного интегрирования достаточен гораздо более простой способ. Возьмем первую специальную сетку, выбрав в качестве узлов все точки разрыва $u(x)$ и ее производных и добавив к ним любое количество других узлов. Вторую сетку получим делением каждого интервала 1-й сетки на две равные части. Так же строим 3-ю сетку из 2-й сетки и т.д. Легко видеть, что для численного интегрирования на такой последовательности сгущающихся сеток экстраполяционный метод Ричардсона применим в полном объеме (коэффициент сгущения $r = 2$). Можно многократно повышать порядок точности и вычислять апостериорную оценку погрешности.

Однако такие сетки принципиально отличны от квазиравномерных: l-я сетка будет содержать группы по 2^{l-1} равных интервалов, полученных двоичным делением интервала 1-й сетки. Зато при переходе от одной такой группы интервалов к соседней возникает скачок: h_{n+1}/h_n будет таким же, как для 1-й сетки, и не будет стремиться к 1 при $N \to$ $\to \infty$.

Поэтому такие сетки будем называть *псевдоравномерными*. Они дают хорошие результаты только для численного интегрирования, но для вычисления производных они фактически непригодны.

2. Симметричные аппроксимации производных. Наиболее часто приходится вычислять на сетках первые и вторые производные $u(x)$. Задачи с более высокими производными встречаются много реже (например, задачи упругого бруска, где фигурирует 4-я производная). Рассмотрим простейшие сеточные аппроксимации этих производных на произвольной сетке (рис. 12).

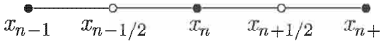

Рис. 12. Аппроксимация производных на сетке

При составлении разностных схем нам обычно нужны производные не в любых точках, а только в узлах x_n или полуцелых точках $x_{n-1/2}$. Начнем с первой производной в n-м узле и составим для ее определения простейшее симметричное выражение:

$$u'_n \approx \frac{u_{n+1}-u_{n-1}}{x_{n+1}-x_{n-1}} = \frac{u_{n+1}-u_{n-1}}{h_{n+1}+h_n}. \tag{44}$$

Исследуем его точность. Для этого запишем разложение в ряды Тейлора:

$$\begin{gathered} u_{n+1} = u_n + h_{n+1}u'_n + \frac{1}{2}h^2_{n+1}u''_n + \frac{1}{6}h^3_{n+1}u'''_n + \frac{1}{24}h^4_{n+1}u''''_n + O(h^5_{n+1}), \\ u_{n-1} = u_n - h_n u'_n + \frac{1}{2}h^2_n u''_n - \frac{1}{6}h^3_n u'''_n + \frac{1}{24}h^4_n u''''_n - O(h^5_n). \end{gathered} \tag{45}$$

Подставляя (45) в (44), получим

$$\begin{aligned} \frac{u_{n+1}-u_{n-1}}{h_{n+1}+h_n} = u'_N + \frac{h_{n+1}-h_n}{2}u''_n + \frac{1}{6}(h^2_{n+1}+h_{n+1}h_n+h^2_n)u'''_n + \\ + \frac{1}{24}(h_{n+1}-h_n)(h^2_{n+1}+h^2_n)u''''_n + O(h^4). \end{aligned} \tag{46}$$

Видно, что если сетка произвольная (в частности, псевдоравномерная), то главный член погрешности есть $(h_{n+1}-h_n)u''_n/2 = O(h)$, т. е. формула (44) имеет лишь первый порядок точности. Но для квазиравномерной сетки положение существенно меняется: $h_n = x'_{n-1/2}\Delta + O(\Delta^3)$ и $h_{n+1} = x'_{n+1/2}\Delta + O(\Delta^3)$. Подстановка этих выражений в (46) дает

$$\frac{u_{n+1}-u_{n-1}}{h_{n+1}+h_n} = u'_n + \frac{1}{2}\Delta^2[x''_n u''_n + (x'_n)^2 u'''_n] + O(\Delta^4). \tag{47}$$

Таким образом, на квазиравномерной сетке главный член погрешности есть $O(\Delta^2) = O(N^{-2})$, а следующий за ним есть $O(\Delta^4)$, т. ч. погрешность имеет второй порядок малости и разлагается в ряд по четным степеням Δ.

Простейшая аппроксимация первой производной в полуцелой точке пишется по аналогии с (44):

$$u'_{n-1/2} \approx \frac{u_n - u_{n-1}}{h_n}. \tag{48}$$

Если сетка квазиравномерная, то при ее сгущении вдвое все узлы и полуцелые точки старой сетки становятся соответственно четными и нечетными узлами новой сетки. Поэтому на сгущенной вдвое сетке (48) переходит в (44), и погрешность аппроксимации (48) можно оценивать согласно (47), подставляя туда шаг $\Delta/2$. Видно, что аппроксимация (48) имеет второй порядок точности, а погрешность разлагается в ряд по четным степеням Δ.

Вторую производную в n-й точке можно рассматривать как первую производную от первой производной. Тогда ее нетрудно выразить через значения u' в соседних полуцелых точках:

$$u''_n \approx \frac{u'_{n+1/2} - u'_{n-1/2}}{x_{n+1/2} - x_{n-1/2}} \approx \frac{2}{h_n + h_{n+1}} \left(\frac{u_{n+1} - u_n}{h_{n+1}} - \frac{u_n - u_{n-1}}{h_n} \right). \tag{49}$$

Исследуем погрешность этой аппроксимации, подставляя в правую часть (49) разложение (45). Получим

$$\frac{2}{h_n + h_{n+1}} - \left(\frac{u_{n+1} - u_n}{h_{n+1}} - \frac{u_n - u_{n-1}}{h_n} \right) =$$

$$= u''_n + \frac{h_{n+1} - h_n}{3} u'''_n + \frac{1}{12}(h_n^2 + h_n h_{n+1} + h_{n+1}^2) u''''_n + \dots \tag{50}$$

Опять видно, что на произвольной неравномерной сетке главным членом погрешности является

$$\frac{(h_{n+1} - h_n) u'''_n}{3} = O(h),$$

т. е. формула (49) имеет лишь первый порядок точности. На квазиравномерной сетке (50) принимает вид

$$\frac{2}{h_n + h_{n+1}} \left(\frac{u_{n+1} - u_n}{h_{n+1}} - \frac{u_n - u_{n-1}}{h_n} \right) =$$

$$= u''_n + \Delta^2 \left[\frac{1}{3} x''_n u'''_n + \frac{1}{4} (x'_n)^2 u''''_n \right] + O(\Delta^4). \tag{51}$$

Таким образом, на квазиравномерной сетке аппроксимация (49) имеет второй порядок точности, а погрешность разлагается в ряд по четным степеням Δ.

Все написанные здесь выражения для первой и второй производных симметричны. По аналогии можно написать аппроксимации для старших производных, но они потребуют привлечения большего числа

узлов. Например, для аппроксимации u''''_n надо добавить еще $u_{n\pm2}$, т. е. использовать 5 узлов. Погрешность всех этих аппроксимаций на квазиравномерной сетке имеет точность $O(\Delta^2) = O(N^{-2})$, а для апостериорной оценки погрешности и повышения порядка точности можно использовать сгущение сеток, причем многократное.

3. Несимметричные аппроксимации. Описанные в п. 2 симметричные аппроксимации производных удобны для внутренних узлов сетки. Но в записи граничных условий для дифференциальных уравнений могут содержаться производные, а для производной в граничном узле симметричную разностную запись составить нельзя. Как при этом добиться второго порядка точности? Рассмотрим этот вопрос, ограничиваясь для простоты вычислением только первой производной (т. к. в краевые условия для наиболее распространенных дифференциальных уравнений второго порядка не могут входить более высокие производные).

Аппроксимация $u(x)$ интерполяционным многочленом Ньютона, построенным по трем узлам — это многочлен второй степени, точно передающий значения функции в этих трех узлах:

$$u(x) = u_{n-1} + (x - x_{n-1})\frac{u_n - u_{n-1}}{h_n} + \\ + (x - x_{n-1})(x - x_n)\frac{1}{h_{n+1} + h_n}\left(\frac{u_{n+1} - u_n}{h_{n+1}} - \frac{u_n - u_{n-1}}{h_n}\right) + O(h^3). \tag{52}$$

Сам этот многочлен обеспечивает третий порядок точности в любой целой или нецелой точке на произвольной неравномерной сетке. Его первая производная также в любой точке имеет второй порядок точности:

$$u'(x) = \frac{u_n - u_{n-1}}{h_n} + \frac{2x - x_{n-1} - x_n}{h_n + h_{n+1}}\left(\frac{u_{n+1} - u_n}{h_{n+1}} - \frac{u_n - u_{n-1}}{h_n}\right) + O(h^2). \tag{53}$$

Подставляя в (53) значения $x = x_{n-1}$, x_n, x_{n+1}, получим производные в узлах сетки:

$$u'_{n-1} = \frac{(u_n - u_{n-1})(2 + h_{n+1}/h_n) - (u_{n+1} - u_n)h_n/h_{n+1}}{h_n + h_{n+1}} + O(h^2), \tag{54}$$

$$u'_n = \frac{(u_{n+1} - u_n)h_n/h_{n+1} + (u_n - u_{n-1})h_{n+1}/h_n}{h_n + h_{n+1}} + O(h^2), \tag{55}$$

$$u'_{n+1} = \frac{(u_{n+1} - u_n)(2 + h_n/h_{n+1}) - (u_n - u_{n-1})h_{n+1}/h_n}{h_n + h_{n+1}} + O(h^2). \tag{56}$$

Подчеркнем, что формулы (54)–(56) имеют второй порядок точности на произвольной неравномерной сетке. На квазиравномерных сетках, как было показано выше (44), для производных в центральной точке можно написать гораздо более простую формулу точности $O(h^2)$:

$$u'_n = \frac{u_{n+1} - u_{n-1}}{h_n + h_{n+1}} + O(h^2). \tag{57}$$

Однако это упрощение нетривиально. Для производных в крайних точках не удается заметно упростить формулы (54) и (56) в случае квазиравномерной сетки.

Если использовать выражения (54)-(56) на квазиравномерных сетках, то погрешность будет $\mathrm{const} \cdot \Delta[1 + O(\Delta^2)]$; это можно показать разложением в ряды Тейлора (аналогично п. 2). Тем самым к ним можно применять сгущение сетки и процедуру Ричардсона (разумеется, на произвольных сетках этого нельзя делать).

§ 3. Случай неограниченной области

1. Замена шага. Как показано в § 1, квазиравномерные сетки можно строить даже в неограниченной области. Однако в этом случае формальное применение выражений, полученных в § 1–2, становится невозможным. В самом деле, пусть для определенности задана полупрямая $0 \leqslant x < +\infty$. Тогда $x_N = +\infty$ и $h_N = +\infty$, т. к. последний интервал сетки бесконечен. При вычислении интеграла по формулам прямоугольников (39) или средних (40) значения u_{N-1} или $u_{N-1/2}$ отличны, вообще говоря, от нуля (хотя малы, поскольку интеграл подразумевается сходящимся). Эти значения должны умножаться на h_N, которое бесконечно велико. Таким образом, интегральная сумма оказывается бесконечной при любом N, что бессмысленно.

Аналогична ситуация с производными. Расчет производной в примыкающем к границе интервале оказывается неверным. Например, по формуле (57) получим в приграничном узле

$$u'_{N-1} \approx \frac{u_N - u_{N-2}}{h_{N-1} + h_N} = 0 \quad (h_N = \infty), \tag{58}$$

каковы бы ни были конечные значения u_n в узлах. Значит, правильно передать значение u'_{N-1} формулой (57) в неограниченной области не удается.

Однако можно так видоизменить все выражения, что они будут разумно распространяться на бесконечную область. Есть два таких способа. Во-первых, можно воспользоваться дробными точками:

$$h_n = x_n - x_{n-1} \approx \frac{1}{1-2\gamma}(x_{n-\gamma} - x_{n-1+\gamma}) \underset{\gamma \to 1/4}{\longrightarrow} 2(x_{n-1/4} - x_{n-3/4}). \tag{59}$$

Во-вторых, это замена

$$h_n = x_n - x_{n-1} \approx x_\xi(\xi_{n-1/2}) \cdot \Delta. \tag{60}$$

На квазиравномерных сетках в конечной области обе замены имеют относительную погрешность $O(\Delta^2)$. В самом деле, разлагая $x(\xi)$ в ряды Тейлора с центром в точке $\xi_{n-1/2}$, легко получим

$$h_n - x'_{n-1/2}\Delta = \frac{1}{24}x'''_{n-1/2}\Delta^3 + O(\Delta^5). \tag{61}$$

Поскольку $h_n = O(\Delta)$, это означает абсолютную ошибку $O(\Delta^3)$ и относительную ошибку $O(\Delta^2)$. Аналогичное разложение нетрудно получить и для замены (59).

Значение γ в (59) формально можно брать любым в разумных пределах $0 < \gamma < 1/2$. Наиболее простой и удобный вид дальнейшие формулы приобретают при $\gamma = 1/4$.

Выражения (59), (60) аппроксимируют шаги h_n лишь в ограниченной области. Для неограниченной области преобразование $x(\xi)$ таково, что само $x(\xi)$ и его производные обращаются в бесконечность, т. ч. правая часть равенства (61) может стать много больше самой величины $x'_{n-1/2} \cdot \Delta$. Но мы имеем право ввести правые части (59), (60) как самостоятельные определения шага, пригодные в неограниченной области и очень близкие к традиционному определению в случае ограниченной области.

2. Аппроксимация. Рассмотрим, какой вид принимают простейшие формулы численного интегрирования и дифференцирования при заменах шага (60) или (59) с $\gamma = 1/4$. Формула левых прямоугольников примет вид

$$U_N = \sum_{n=1}^{N} u_{n-1}x'_{n-1/2}\Delta \quad \text{или} \quad U_N = 2 \cdot \sum_{n=1}^{N} u_{n-1}(x_{n-1/4} - x_{n-3/4}). \tag{62}$$

Для формулы средних получим

$$U_N = \sum_{n=1}^{N} u_{n-1/2}x'_{n-1/2}\Delta \quad \text{или} \quad U_N = 2 \cdot \sum_{n=1}^{N} u_{n-1/2}(x_{n-1/4} - x_{n-3/4}). \tag{63}$$

Видно, что теперь все члены интегральных сумм и сами суммы остаются конечными при любых конечных N (вопрос о сходимости при $N \to \infty$ надо исследовать особо).

Аналогично для симметричных производных: вместо первой производной (48) получаем

$$u'_{n-1/2} \approx \frac{u_n - u_{n-1}}{x'_{n-1/2}\Delta} \quad \text{или} \quad u'_{n-1/2} \approx \frac{u_n - u_{n-1}}{2(x_{n-1/4} - x_{n-3/4})}, \tag{64}$$

а для второй производной вместо (49) имеем

$$u''_n \approx \frac{1}{x'_n\Delta^2}\left(\frac{u_{n+1} - u_n}{x'_{n+1/2}} - \frac{u_n - u_{n-1}}{x'_{n-1/2}}\right) \tag{65}$$

или

$$u''_n \approx \frac{1}{2(x_{n+1/2} - x_{n-1/2})} \left(\frac{u_{n+1} - u_n}{x_{n+3/4} - x_{n+1/4}} - \frac{u_n - u_{n-1}}{x_{n-1/4} - x_{n-3/4}} \right). \qquad (66)$$

Даже в неограниченной области здесь не возникает бесконечностей в знаменателях, поскольку все дробные точки граничного бесконечного интервала остаются конечными. Поэтому различные производные на границах не обращаются в нули и реально зависят от значений u_n в бесконечно удаленном узле и соседних с ним. Это позволяет рассчитывать на разумную аппроксимацию граничных условий (хотя этот вопрос надо специально исследовать для каждого конкретного класса задач).

Для односторонней аппроксимации первой производной (54) или (56) соответственно получим:

$$u'_{n-1} = \frac{(u_n - u_{n-1})\left(2 + \dfrac{x'_{n+1/2}}{x'_{n-1/2}}\right) - (u_{n+1} - u_n)\dfrac{x'_{n-1/2}}{x'_{n+1/2}}}{2x'_n\Delta} + O(\Delta^2) =$$

$$= \frac{(u_n - u_{n-1})\left[2 + \dfrac{x_{n+3/4} - x_{n+1/4}}{x_{n-1/4} - x_{n-3/4}}\right] - (u_{n+1} - u_n)\dfrac{x_{n-1/4} - x_{n-3/4}}{x_{n+3/4} - x_{n+1/4}}}{2(x_{n+1/2} - x_{n-1/2})} +$$

$$+ O(\Delta^2); \qquad (67)$$

$$u'_{n+1} = \frac{(u_{n+1} - u_n)\left(2 + \dfrac{x'_{n-1/2}}{x'_{n+1/2}}\right) - (u_n - u_{n-1})\dfrac{x'_{n+1/2}}{x'_{n-1/2}}}{2x'_n\Delta} + O(\Delta^2) =$$

$$= \frac{(u_{n+1} - u_n)\left[2 + \dfrac{x_{n-1/4} - x_{n-3/4}}{x_{n+3/4} - x_{n+1/4}}\right] - (u_n - u_{n-1})\dfrac{x_{n+3/4} - x_{n+1/4}}{x_{n-1/4} - x_{n-3/4}}}{2(x_{n+1/2} - x_{n-1/2})} +$$

$$+ O(\Delta^2). \qquad (68)$$

Для первой производной в средней точке проще всего написать аналог (57):

$$u'_n = \frac{u_{n+1} - u_{n-1}}{2x'_n\Delta} + O(\Delta^2) = \frac{u_{n+1} - u_{n-1}}{2(x_{n+1/2} - x_{n-1/2})} + O(\Delta^2); \qquad (69)$$

можно написать также аналог (55), но это дает более громоздкие выражения.

Все формулы для производных (64)–(69) имеют второй порядок аппроксимации.

3. Сравнение с методом замены переменной. Первый вариант формулы средних (63) можно интерпретировать следующим образом. Произведем в исходном интеграле замену переменных $x \to \xi$. Тогда интеграл будет браться по ξ от функции $u(x(\xi))x'(\xi)$. Применив к нему классическую формулу средних, получим в точности написанную первую формулу (63).

Аналогичные соображения верны для части формул вычисления производных.

Однако метод квазиравномерных сеток — более общий подход, чем замена переменных. Например, уже вторая формула (63) и обе формулы (62) не эквивалентны замене переменных с последующим применением классических квадратурных формул.

§ 4. Аппроксимация функций в двумерной неограниченной области

1. Скалярное произведение. Задача о приближении функции двух переменных по ее значениям в дискретном наборе точек надежно решена лишь в случае регулярной равномерной сетки. Если же сетка сильно неравномерна, то стандартные методы аппроксимации становятся непригодными. Практикам хорошо известно, что ни один стандартный графопостроитель не дает удовлетворительного качества при построении поверхностей по сведениям, заданным на сильно неравномерной или нерегулярной двумерной сетке.

При расчетах в неограниченных областях мы используем регулярные квазиравномерные двумерные сетки. Такая сетка, имея конечное число узлов, покрывает неограниченную область. Соседние интервалы такой сетки почти одинаковы, хотя далекие могут сильно отличаться. Двумерная аппроксимация на такой сетке необходима для восстановления численного решения в любой точке неограниченной области по результатам сеточных расчетов.

Используем взвешенный метод наименьших квадратов. Применим следующую нумерацию узлов регулярной прямоугольной квазиравномерной сетки:

$$x_n, \quad -N \leqslant n \leqslant N; \qquad y_m, \quad -M \leqslant m \leqslant M.$$

Определим скалярное произведение двумерных сеточных функций по формуле

$$(f,g) = \sum_{n=-N+1}^{N-1} \sum_{m=-M+1}^{M-1} f_{nm} g_{nm} (x_{n+1/2} - x_{n-1/2})(y_{m+1/2} - y_{m-1/2}). \tag{70}$$

Это выражение приближает двумерный интеграл

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)g(x,y)dxdy \tag{71}$$

с точностью $O(N^{-1}+M^{-1})$, поскольку не учитывает вклада тех половинок ячеек, которые примыкают к границам. Для большинства задач в ограниченной области такой точности недостаточно.

В неограниченной области часто приходится решать задачи, в которых функции обращаются в нуль на бесконечности, причем достаточно быстро. В таких задачах вклад граничных ячеек пренебрежимо мал, и сумма (70) приближает интеграл (71) с точностью $O(N^{-2}+M^{-2})$. Именно этот случай будет рассмотрен ниже.

С другой стороны, если трактовать множители $(x_{n+1/2}-x_{n-1/2})\times$ $\times\,(y_{m+1/2}-y_{m-1/2})$ в выбранном нами скалярном произведении как вес, то в нашей модификации метода наименьших квадратов точка имеет тем больший вес, чем большая площадь у соответствующей ей ячейки квазиравномерной сетки. Это оправдано, когда данные во всех точках заданы с примерно одинаковой точностью, но сетка сильно неравномерная.

2. Базисные функции. В качестве базисных функций выберем функции следующего вида:

$$\begin{gathered}\varphi_j(x,y)=H_k(x)H_l(y)e^{-x^2/2-y^2/2};\\ 0\leqslant k\leqslant K-1,\quad 0\leqslant l\leqslant L-1,\quad 1\leqslant j\leqslant J=K\cdot L.\end{gathered} \tag{72}$$

Здесь

$$H_k(x)=k!\sum_{q=0}^{[k/2]}(-1)^q\frac{(2x)^{k-2q}}{q!(k-2q)!} \tag{73}$$

— полиномы Эрмита, которые, как известно, ортогональны:

$$\int_{-\infty}^{\infty} H_k(x)H_l(x)e^{-x^2}dx=0\quad\text{при}\quad k\neq l. \tag{74}$$

Базисные функции (72) экспоненциально затухают на ∞. Из (74) следует, что они ортогональны в смысле интеграла (71). Для них сумма (70) аппроксимирует интеграл (71) с точностью $O(N^{-2}+M^{-2})$. Поэтому базис (72) будет почти ортогонален в смысле скалярного произведения (70): недиагональные элементы матрицы скалярных произведений будут очень малыми $O(N^{-2}+M^{-2})$.

Пусть выбрано J базисных функций, и исходные данные $u(x_n,y_m)$, $-N+1\leqslant n\leqslant N-1$, $-M+1\leqslant m\leqslant M-1$, заданы во внутренних

узлах квазиравномерной сетки с числом узлов $(2N-1)(2M-1) \gg J$. Найдем линейную аппроксимацию по выбранной системе базисных функций

$$\varphi(x,y) = \sum_{j=1}^{J} a_j \varphi_j(x,y) \tag{75}$$

из условия наилучшего среднеквадратичного приближения в смысле нормы, порожденной скалярным произведением (70):

$$\sum_{n=-N+1}^{N-1} \sum_{m=-M+1}^{M-1} [u(x_n,y_m) - \varphi(x_n,y_m)]^2 \times$$

$$\times (x_{n+1/2} - x_{n-1/2})(y_{m+1/2} - y_{m-1/2}) = \min.$$

Коэффициенты такой линейной аппроксимации являются решением системы линейных алгебраических уравнений (СЛАУ) с матрицей Грама, составленной из скалярных произведений базисных функций:

$$\Gamma_{ij} = (\varphi_i, \varphi_j).$$

Для достижения высокой точности важна хорошая обусловленность матрицы Г. Проведем следующие тесты.

Тест 1. Помножим матрицу на единичный столбец и полученный результат используем в качестве правой части СЛАУ. Норма отклонения полученного решения СЛАУ от единичного столбца будет мерой обусловленности $\|\Delta\|$.

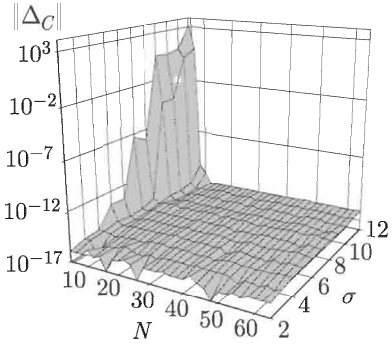

Рис. 13. Обусловленность матрицы скалярных произведений диагональной от числа узлов сетки и максимальной суммарной степени полиномов

Результат такого теста приведен на рисунке 13. Обусловленность матрицы зависит от числа базисных функций и числа точек сетки. В тесте использовано равное число узлов сетки по обеим пространственным переменным $N = M$. Число базисных функций удобно характери-

зовать $\sigma = L + K - 2$ — максимальной суммарной степенью полиномов в формуле (75). На рисунке 13 приведена погрешность $\|\Delta\|_C$ в тесте 1 в зависимости от σ и N. (В этой и последующих иллюстрациях полагаем $N = M$, т. ч. полное число узлов сетки равно $(2N-1)^2$.) В широком диапазоне этих параметров ошибка в проведенном тесте не превосходит 10^{-9}–10^{-7}. Ошибка экспоненциально нарастает, когда число точек сетки становится сравнимо с числом базисных функций. Это подтверждает хорошую обусловленность матрицы Грама в рабочем диапазоне параметров. Хорошая обусловленность матрицы связана, прежде всего, с удачным выбором базисных функций (72) и скалярного произведения (70).

Тест 2. Близость системы базисных функций к ортогональной в смысле выбранного скалярного произведения отражается на структуре матрицы Грама. Матрица ортогональной системы была бы диагональной. Мерой ортогональности системы может служить отношение недиагональной части сферической нормы матрицы Г к диагональной:

$$\frac{S_1}{S_2} = \frac{\sum\limits_{j=1}^{J} \sum\limits_{i=1, i\neq j}^{J} (\varphi_i, \varphi_j)^2}{\sum\limits_{j=1}^{J} (\varphi_j, \varphi_j)^2}.$$

Несмотря на то, что числитель этой дроби содержит $J(J-1)$ неотрицательных слагаемых, а знаменатель — J слагаемых, величина S_1/S_2 остается малой в рабочем диапазоне параметров. График зависимости величины S_1/S_2 от максимальной суммарной степени полиномов σ и параметра сетки $N = M$ приведен на рисунке 14.

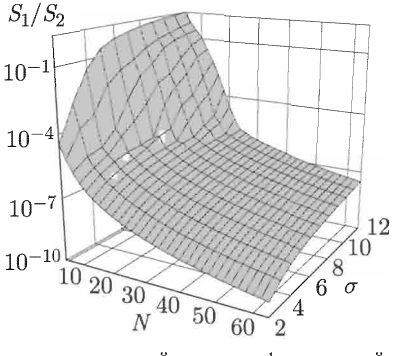

Рис. 14. Отношение недиагональной части сферической нормы матрицы скалярных произведений к диагональной от числа узлов сетки и максимальной суммарной степени полиномов

Описанный подход был успешно применен для построения функции двух переменных, аппроксимирующей результаты двумерных расчетов

на квазиравномерной сетке. Для оценки точности результат аппроксимации сравнивался с исходными данными — результатами численного решения начально-краевой задачи. Максимальное по модулю отклонение аппроксимирующей функции от исходных данных, например для числа узлов сетки $N = 128$, было порядка 10^{-10}. Исходными данными для задачи двумерной аппроксимации в описанном случае служили результаты сеточных расчетов, которые были проведены с контролем точности, и их погрешность составила примерно 10^{-6}. Таким образом, двумерная среднеквадратичная аппроксимация по предложенному алгоритму позволяет найти численное решение задачи в произвольной точке неограниченной области по результатам расчетов на квазиравномерной сетке с очень высокой точностью. Без помощи построенной аппроксимационной функции невозможно было бы визуализировать результаты расчетов.

Полуплоскость. Изложенный метод был применен также и для задач в полуплоскости $x \in (-\infty, +\infty)$, $y \in [0, +\infty)$. В этом случае удобно выбрать следующие базисные функции:

$$\varphi_j(x, y) = H_l(x) L_k(y) e^{-x^2/2 - y/2}, \tag{76}$$

где

$$L_k(y) = (k!)^2 \sum_{q=0}^{k} (-1)^q \frac{y^q}{(q!)^2 (k-q)!}$$

— полиномы Лаггера.

Базисные функции точно ортогональны в смысле скалярного произведения

$$(f, g) = \int_{-\infty}^{\infty} dx \int_{0}^{\infty} f(x, y) g(x, y) dy$$

и приближенно ортогональны в смысле сеточного скалярного произведения

$$(f, g) = \sum_{n=-N+1}^{N-1} \sum_{m=0}^{M-1} f_{nm} g_{nm} (x_{n+1/2} - x_{n-1/2})(y_{m+1/2} - y_{m-1/2});$$

чтобы правильно передать слагаемые с $m = 0$, здесь следует полагать $y_{-1/2} = 0$.

Все поверхности на рисунках глав VI, VII, иллюстрирующих результаты двумерных расчетов, построены с применением описанного алгоритма аппроксимации.

3. Нелинейная среднеквадратичная аппроксимация. Было замечено, что существенно повысить точность аппроксимации в случае неограниченного интервала, не увеличивая размерность базиса, можно

перейдя к нелинейной среднеквадратичной аппроксимации. Поясним этот прием на примере одномерной задачи. Пусть функция $f(x)$ задана своими значениями во внутренних узлах квазиравномерной сетки x_n, $-N+1 \leqslant n \leqslant N-1$, покрывающей всю прямую. Будем использовать базисные функции вида (72), но введем масштабирующий множитель в аргумент базисных функций:

$$\varphi_k(x) = H_k(\mu x) \cdot \exp\Big(-\frac{(\mu x)^2}{2}\Big).$$

Геометрический смысл такого приема в следующем: меняя масштабирующий множитель, мы совмещаем области активного изменения базисных функций и приближаемой функции. В случае бесконечной области это особенно важно.

Целесообразность такого подхода иллюстрирует рисунок 15. На нем изображена зависимость ошибки аппроксимации от масштабирующего коэффициента μ в полулогарифмическом масштабе. Характер зависимости указывает на то, что разность этих двух величин меняет знак, а норма погрешности точно проходит через ноль. Кроме того, типична ситуация, когда минимум ошибки очень глубокий и узкий. Значит, повышение качества аппроксимации требует определения μ с максимально возможным числом знаков.

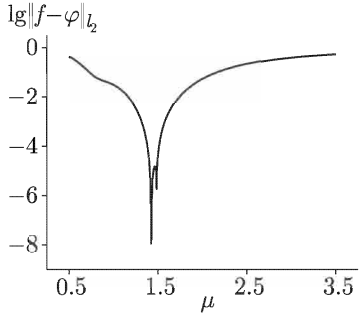

Рис. 15. Погрешность аппроксимации в зависимости от масштабирующего множителя в аргументе базисной функции

Рассмотрим отклонение аппроксимации от исходных данных в сеточной норме L_2 как нелинейную функцию, зависящую не только от коэффициентов линейной аппроксимации, но и от масштабирующего множителя в аргументе базисных функций:

$$I(a_0, a_1, \ldots, a_M, \mu) = \|f(x) - \sum_{k=0}^{M} a_k \varphi_k(\mu x)\|_{L_2}, \tag{77}$$

где $f(x)$ — заданная сеточная функция.

В качестве численного метода поиска минимума нелинейной функции (77) рекомендуется использовать квазиньютоновские итерации [Дэннис мл., Шнабель, 1988]. Этот алгоритм минимизации реализован, например, в стандартной библиотеке IMSL языка Фортран [Бартеньев, 2001].

Применение такой нелинейной аппроксимации позволило существенно (на 3–4 порядка) улучшить точность аппроксимации, не меняя размер базиса и, следовательно, не увеличивая объем вычислений.

Глава IV

КВАДРАТУРЫ НА КВАЗИРАВНОМЕРНЫХ СЕТКАХ

В главе мы выводим различные квадратурные формулы на квазиравномерных сетках: простейшие аналоги формул трапеций и средних, имеющие второй порядок точности, и коллокационно-сеточные формулы высоких порядков точности, сходные с формулами Гаусса–Кристоффеля. Для всех этих формул применим экстраполяционный метод Ричардсона, что позволяет выполнять расчеты с гарантированной точностью. Особенно выгодны эти методы для вычисления несобственных интегралов на прямой и полупрямой, поскольку они дают хорошие результаты даже для слабо убывающих функций (при степенных законах убывания), где метод Гаусса–Кристоффеля практически неприменим. Это позволяет решить ряд актуальных физических и технических задач.

§ 1. Простейшие формулы

1. Формула трапеций. Эта формула не упоминалась в главе II, но ее вывод тривиален. Интеграл по одному интервалу сетки заменяется площадью трапеции с основаниями u_{n-1} и u_n и высотой h_n (рис. 1). Это дает на произвольной неравномерной сетке

$$U_N = \sum_{n=1}^{N} h_n \frac{u_n + u_{n-1}}{2}. \qquad (1)$$

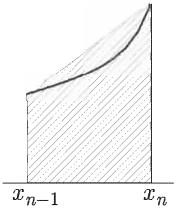

Рис. 1. Квадратурная формула трапеций

Формула (1) построена симметрично, поэтому ее погрешность в каждом интервале разлагается в ряд по степеням h_n^2. Главный член погрешности содержит $u''_{n-1/2}$, поскольку члены разложения с $u'_{n-1/2}$ сокращаются в силу симметрии. Интеграл и его погрешность имеют размерность $[hu]$, поэтому главный член погрешности должен (с точностью до численного множителя) состоять из слагаемых $h_n^3 u''_{n-1/2}$. Этот численный коэффициент можно определить, например, подстановкой $u(x) = x^3$, что дает

$$R_N \approx \frac{1}{12} \sum_{n=1}^{N-1} h_n^3 u''_{n-1/2}. \qquad (2)$$

Эта погрешность вдвое больше, чем для формулы средних, и имеет противоположный знак. На равномерной сетке формула (2) переходит в следующую асимптотическую оценку:

$$R_n \approx \frac{1}{12}h^2\sum_{n=1}^{N} hu''_{n-1/2} \approx \frac{h^2}{12}\int_a^b u''(x)dx = \frac{h^2}{12}[u'(b)-u'(a)], \quad h=\text{const}. \tag{3}$$

Таким образом, формула трапеций имеет второй порядок точности.

Напомним, что формулы (1), (2) относятся к произвольным неравномерным сеткам, но в (2) за точку $x_{n-1/2}$ принимается середина интервала в обычном смысле (полусумма его концов). Но на произвольных сетках говорить об общем порядке точности при увеличении N невозможно.

Квазиравномерные сетки. Пока ограничимся конечной областью $[a,b]$. Тогда все интервалы имеют конечные длины h_n, т. ч. вид формулы трапеций (1) пригоден для аппроксимации интеграла. Надо лишь аккуратно получить априорную оценку погрешности.

Для этого далее будем понимать $(n-1/2)$-ю точку $x_{n-1/2}$ в смысле квазиравномерной сетки. Введем обозначение $v(\xi)=u(x(\xi))$ и заменим в интеграле $dx=x'(\xi)d\xi$. Тогда точное выражение погрешности формулы трапеций (1) примет следующий вид:

$$R_N = U_N - U = \sum_{n=1}^{N}\left[\frac{(v_{n-1}+v_n)h_n}{2} - \int_{\xi_{n-1}}^{\xi_n} v(\xi)x'(\xi)d\xi\right]. \tag{4}$$

Предполагая пока функцию $v(\xi)$ и порождающее преобразование $x(\xi)$ неограниченно дифференцируемыми, разложим все величины в каждом интервале в ряды Тейлора–Маклорена с центром в точке $\xi_{n-1/2}$. Учитывая симметрию и сокращение нечетных членов, легко получим:

$$x'(\xi) = \sum_{m=0}^{\infty}\frac{1}{m!}(\xi-\xi_{n-1/2})^m x^{(m+1)}(\xi_{n-1/2}), \quad |\xi-\xi_{n-1/2}| \leqslant \frac{\Delta}{2}; \tag{5}$$

$$h_n = x_n - x_{n-1/2} = \Delta\sum_{m=0}^{\infty}\frac{(\Delta/2)^{2m}}{(2m+1)!}x^{(2m+1)}(\xi_{n-1/2}), \quad \Delta=\frac{\beta-\alpha}{N}; \tag{6}$$

$$v(\xi) = \sum_{m=0}^{\infty}\frac{1}{m!}(\xi-\xi_{n-1/2})^m v^{(m)}(\xi_{n-1/2}), \quad |\xi-\xi_{n-1/2}| \leqslant \frac{\Delta}{2}; \tag{7}$$

$$\frac{v_{n-1}+v_n}{2} = \sum_{m=0}^{\infty}\frac{(\Delta/2)^{2m}}{(2m)!}v^{(2m)}(\xi_{n-1/2}); \tag{8}$$

напомним, что $[\alpha,\beta]$ есть отрезок изменения ξ. В формулах (5)–(8) фигурируют производные $v^{(m)}(\xi)\equiv d^m v/d\xi^m$, которые вычисляются по

обычным правилам:

$$v'(\xi) = u'(x)x'(\xi), \quad v''(\xi) = u'(x)x''(\xi) + u''(x)[x'(\xi)]^2 \tag{9}$$

и т.д. В случае, если существует лишь ограниченное число производных, то число членов в суммах (5)–(8) определяется порядками существующих производных. Но функцию $u(x)$ нам задают, и ее свойства от нас не зависят, а преобразование $x(\xi)$ мы выбираем сами. Отсюда следует очевидная рекомендация: *выбирать $x(\xi)$ надо так, чтобы его гладкость была не хуже, чем у $u(x)$*.

Напомним, что все примеры порождающих преобразований в главе III были неограниченно дифференцируемыми.

Подставим (5)–(8) в (4) и произведем все интегрирования. Благодаря симметрии формул все нечетные степени шага Δ сократятся. Это дает

$$R_N = \sum_{m=0}^{\infty} A_m \Delta^{2m+2} = O(\Delta^2) = O(N^{-2}); \tag{10}$$

здесь главный член погрешности с учетом (9) имеет коэффициент

$$A_0 = \frac{1}{12} \sum_{n=1}^{N} [v''(\xi_{n-1/2})x'(\xi_{n-1/2}) - v'(\xi_{n-1/2})x''(\xi_{n-1/2})] = \\ = \frac{1}{12} \sum_{n=1}^{N} u''(x_{n-1/2})[x'(\xi_{n-1/2})]^3. \tag{11}$$

Нетрудно также заметить, что для (10) есть простая асимптотическая оценка

$$R_N \approx A_0 \Delta^3 \approx \frac{\Delta^2}{12} \int_{\alpha}^{\beta} \frac{d^2u}{dx^2} \left(\frac{dx}{d\xi}\right)^3 d\xi = \frac{\text{const}}{N^2}. \tag{12}$$

Таким образом, на квазиравномерных сетках формула трапеций (1) имеет второй порядок точности, а ее погрешность разлагается в ряд по четным степеням $\Delta^2 \sim N^{-2}$. Следовательно, при расчетах на сгущающихся сетках можно пользоваться методом Ричардсона как для апостериорной асимптотической оценки погрешности, так и для повышения порядка точности. Для этого составляют такие же треугольные таблицы, как в главе II, а под коэффициентом сгущения r подразумевают отношение чисел интервалов сеток N.

Можно проводить многократное сгущение сеток и рекуррентное повышение порядка точности. Поскольку погрешность (10) разлагается в ряд по четным степеням N, то каждое сгущение позволяет повысить порядок точности на $s = 2$.

Неограниченная область. Если рассматривается несобственный интеграл на прямой ($a = -\infty$, $b = +\infty$), то можно построить квазиравномерную сетку, но длины ее граничных интервалов h_1 и h_N

бесконечны. Если берется интеграл по полупрямой, то бесконечен один из этих интервалов. Поэтому непосредственно формулу трапеций (1) применять нельзя.

Однако можно воспользоваться одной из замен h_n, имеющих второй порядок точности на квазиравномерных сетках, описанных в главе III, § 3. Так, замена $h_n \approx x'(\xi_{n-1/2})\Delta$ дает следующий вариант формулы трапеций:

$$U_N = \sum_{n=1}^{N} (u_{n-1} + u_n) x'_{n-1/2} \frac{\Delta}{2}, \quad \Delta = \frac{\beta - \alpha}{N}; \tag{13}$$

заменяя $h_n = x_n - x_{n-1} \approx 2(x_{n-1/4} - x_{n-3/4})$, получим другой вариант формулы трапеций:

$$U_N = \sum_{n=1}^{N} (u_{n-1} + u_n)(x_{n-1/4} - x_{n-3/4}). \tag{14}$$

Обе формулы дают конечную величину суммы в случае неограниченной области, т. е. формально они применимы. Остается исследовать их погрешность.

Для этого из сумм (13) или (14) вычитают интеграл аналогично (4) и подставляют разложения (5), (7), (8); вместо (6) пользуются разложением

$$x_{n-1/4} - x_{n-3/4} = \frac{\Delta}{2} \sum_{m=0}^{\infty} \frac{(\Delta/4)^{2m}}{(2m+1)!} x^{(2m+1)}(\xi_{n-1/2}). \tag{15}$$

Опять в силу симметрии все нечетные степени Δ сокращаются, т. ч. погрешность оказывается рядом по степеням Δ^2 вида (10). Разумеется, коэффициенты этого ряда будут иными. Так, для главного члена погрешности варианта (13) получаем выражение

$$R_N \approx A_0 \Delta^3 = \frac{\Delta^3}{24} \sum_{n=1}^{N} [3 v_{\xi\xi} x_\xi - (v x_\xi)_{\xi\xi}]_{n-1/2} \approx$$
$$\approx \frac{\Delta^2}{24} \int_{\alpha}^{\beta} \left\{ v_{\xi\xi}(\xi) x_\xi(\xi) - \frac{d^2}{d\xi^2} [v(\xi) x_\xi(\xi)] \right\} d\xi = \frac{\text{const}}{N^2}, \tag{16}$$

а для варианта (14)

$$R_N \approx A_0 \Delta^3 = \Delta^3 \sum_{n=1}^{N} \left[\frac{1}{8} v_{\xi\xi} x_\xi + \frac{1}{96} v x_{\xi\xi\xi} - \frac{1}{24} (v x_\xi)_{\xi\xi} \right]_{n-1/2} \approx$$
$$\approx \Delta^2 \int_{\alpha}^{\beta} \left\{ \frac{1}{8} v_{\xi\xi}(\xi) x_\xi(\xi) + \frac{1}{96} v(\xi) x_{\xi\xi\xi}(\xi) - \frac{1}{24} \frac{d^2}{d\xi^2} [v(\xi) x_\xi(\xi)] \right\} d\xi = \frac{\text{const}}{N^2}. \tag{17}$$

Здесь индекс ξ означает дифференцирование по ξ. Из (16), (17) видно, что в общем случае от преобразования $x(\xi)$ надо требовать порядка гладкости на единицу выше, чем имеет $u(x)$.

Очевидно следующее достаточное условие: если все интегралы в оценках (16), (17) абсолютно сходящиеся, то варианты формулы трапеций в неограниченной области (13), (14) сходятся со вторым порядком точности.

Это условие зависит от поведения не только $u(x)$, но и преобразования $x(\xi)$. Например, рассмотрим интеграл на полупрямой $0 \leqslant x < +\infty$ для функции $u(x)$, убывающей на бесконечности как $u(x) \approx \approx \mathrm{const} \cdot x^{-\gamma}$; сам интеграл абсолютно сходится при $\gamma > 1$. Выберем преобразование $x(\xi) = c\xi/(1-\xi)^{\delta}$, $\delta > 0$, необязательно целое. Тогда указанное достаточное условие сходимости формулы трапеций (13) или (14) выполняется при

$$\gamma > 1 + \frac{2}{\delta}. \tag{18}$$

Наиболее часто выбираемое преобразование $x(\xi) = c\xi/(1-\xi)$ соответствует $\delta = 1$; это гарантирует сходимость квадратурной формулы лишь при $\gamma > 3$ (кстати, преобразование $x = c \cdot \mathrm{tg}\,\xi$ тоже соответствует $\delta = 1$). Если же взять преобразование с достаточно большим δ, то можно гарантировать сходимость при любом $\gamma > 1$.

Пример 1. Невыполнение условия (18) приводит к снижению порядка точности метода. Для вычисления интеграла

$$\int_0^{\infty} \frac{dx}{1+x^2} = \frac{\pi}{2}, \quad \gamma = 2,$$

используем квадратурную формулу трапеций на квазиравномерной сетке, покрывающей полупрямую. Используем два разных преобразования, порождающих такую сетку:

1) $x(\xi) = c \cdot \mathrm{tg}\,\xi$ ($\delta = 1$),

2) $x(\xi) = c \cdot \xi/(1-\xi)^3$ ($\delta = 3$).

Вторая сетка удовлетворяет условию (18), тогда как для первой это условие нарушено. Убывание погрешности численного решения с ростом числа узлов сетки в двойном логарифмическом масштабе показано на рисунке 2. Кружки соответствуют расчетам на тангенциальной сетке, а точки — на степенной сетке. В первом случае порядок точности падает до перво-

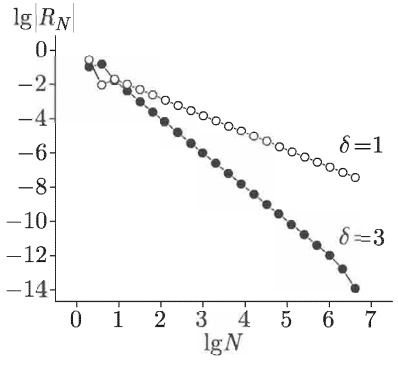

Рис. 2. Нарушение условия (18) приводит к снижению порядка точности квадратурной формулы

го, тогда как при выполнении условия (18) формула трапеций имеет теоретический порядок точности $R_N = O(N^{-2})$.

Заметим также, что формулы (13), (14) пригодны как в неограниченной, так и в ограниченной области; в последнем случае вопрос об их сходимости тривиален.

2. Формула средних. По аналогии с формулой трапеций, напишем три варианта формулы средних на квазиравномерной сетке:

$$U_N = \sum_{n=1}^{N} u_{n-1/2}(x_n - x_{n-1}), \tag{19}$$

$$U_N = \sum_{n=1}^{N} u_{n-1/2} \cdot 2(x_{n-1/4} - x_{n-3/4}), \tag{20}$$

$$U_N = \sum_{n=1}^{N} u_{n-1/2} \cdot x_\xi(\xi_{n-1/2})\Delta; \tag{21}$$

здесь дробные доли интервала понимаются в смысле квазиравномерной сетки. Вариант (19) пригоден только для ограниченной области, а варианты (20), (21) справедливы как для ограниченной области, так и для неограниченной.

Погрешности этих формул находят также подстановкой разложений (5)–(8) и (15). Опять в силу симметрии формул остаются ряды типа (10) по четным степеням Δ, а главные члены погрешности принимают соответственно следующий вид:

$$R_N \approx -\frac{\Delta^3}{24}\sum_{n=1}^{N}[v_{\xi\xi}x_\xi + 2v_\xi x_{\xi\xi}]_{n-1/2} \approx$$

$$\approx -\frac{\Delta^2}{24}\int_\alpha^\beta [v_{\xi\xi}(\xi)x_\xi(\xi) + 2v_\xi(\xi)x_{\xi\xi}(\xi)]d\xi, \tag{22}$$

$$R_N \approx \frac{\Delta^3}{24}\sum_{n=1}^{N}\left[\frac{1}{4}vx_{\xi\xi\xi} - (vx_\xi)_{\xi\xi}\right]_{n-1/2} \approx$$

$$\approx \frac{\Delta^2}{24}\int_\alpha^\beta \left\{\frac{1}{4}v(\xi)x_{\xi\xi\xi}(\xi) - \frac{d^2}{d\xi^2}[v(\xi)x_\xi(\xi)]\right\}d\xi, \tag{23}$$

$$R_N \approx -\frac{\Delta^3}{24}\sum_{n=1}^{N}[(vx_\xi)_{\xi\xi}]_{n-1/2} \approx$$

$$\approx -\frac{\Delta^2}{24}\int_\alpha^\beta \frac{d^2}{d\xi^2}[v(\xi)x_\xi(\xi)]d\xi = -\frac{\Delta^2}{24}\left[\frac{d}{d\xi}(vx_\xi)\right]_{\xi=\alpha}^{\xi=\beta}. \tag{24}$$

Последнее выражение можно получить из асимптотической оценки (II.24) заменой $h_n \approx x'(\xi_{n-1/2})\Delta$.

В неограниченной области достаточное условие сходимости со вторым порядком точности полностью аналогично формуле трапеций.

Замена переменных. Если выполнено преобразование $x(\xi)$, то исходный интеграл, который мог браться по неограниченной области, переходит в интеграл по конечному отрезку:

$$U = \int_a^b u(x)dx = \int_\alpha^\beta w(\xi)d\xi, \quad w(\xi) = u(x(\xi))x_\xi(\xi). \tag{25}$$

Не проще ли непосредственно вычислить последний интеграл с помощью квадратурной формулы трапеции или средних на равномерной сетке по переменной ξ? Не окажется ли метод квазиравномерных сеток эквивалентным такому подходу?

Видно, что для формулы средних вариант (21) действительно точно эквивалентен вычислению второго интеграла (25) на равномерной сетке. Однако варианты (19) и (20) эквивалентными этому способу уже не являются. А для формулы трапеций ни один из вариантов п. 1 не эквивалентен такому способу.

Возможно также, что $w(\xi)$ на концах отрезка $[\alpha, \beta]$ окажется неограниченной или будет иметь неограниченные производные. Первое означает, что новый интеграл окажется также несобственным, хотя по другой причине. Это осложнит его вычисление. Неограниченность производных также осложнит вычисления на равномерных по ξ сетках, ибо позволит пользоваться лишь формулами невысокого порядка точности. Поэтому метод замены переменных не имеет заметных преимуществ перед методом квазиравномерных сеток.

Формула Эйлера. Остаточный член (24) имеет особенно простой вид, и его нетрудно вычислить априорно. Тогда его можно ввести как поправку и получить формулу повышенной точности, лишь незначительно отличающуюся от формулы средних (21):

$$U_N = \Delta \sum_{n=1}^{N} (vx_\xi)_{n-1/2} - \frac{\Delta^2}{24}\left[\frac{d}{d\xi}(vx_\xi)\right]_{\xi=\alpha}^{\xi=\beta}.$$

Погрешность этой формулы есть $O(\Delta^4)$, т. е. она имеет четвертый порядок точности. Это обобщение известной квадратурной формулы Эйлера на случай квазиравномерной сетки. Далее можно явно вычислить главный член погрешности этой формулы, учесть его как следующую поправку и т.д. Этот процесс даст цепочку формул Эйлера–Маклорена высоких порядков точности.

Эти формулы эффективны для тех несобственных интегралов, у которых $w(\xi) = u(x(\xi))x_\xi(\xi)$ имеет ограниченные высокие производные. Особенно они выгодны, если эти производные обращаются в нуль на концах отрезка, т. к. тогда дополнительные члены пропадают, и сам вариант формулы средних (21) обеспечивает высокий порядок точности.

Формулы Гаусса–Кристоффеля часто рекомендуют для вычисления несобственных интегралов на полупрямой или прямой. Это формулы наивысшей алгебраической точности, т. к. формула с N узлами точна для многочлена степени $2N-1$ (поэтому иногда говорят, что эти формулы имеют порядок точности $2N-1$, хотя это нестрогое выражение).

Однако реально формулами Гаусса–Кристоффеля можно пользоваться только в тех случаях, когда табулированы их узлы и веса. Они имеются в литературе лишь для функций, убывающих на бесконечности как $\exp(-x)$ на полупрямой или как $\exp(-x^2)$ на прямой. Это лишь малая часть всех практически интересных случаев. А простейшие формулы метода квазиравномерных сеток легко позволяют интегрировать функции со степенным и любыми другими законами убывания.

Пример 2. Рассмотрим несобственный интеграл

$$\int_0^\infty \frac{dx}{1+x^2} = \frac{\pi}{2}.$$

Введем квазиравномерную сетку, покрывающую полупрямую $[0,\infty)$, пригодную для вычисления несобственного интеграла для функции, слабо убывающей на бесконечности ($\gamma = 2$):

$$x(\xi) = c \cdot \frac{\xi}{(1-\xi)^3}.$$

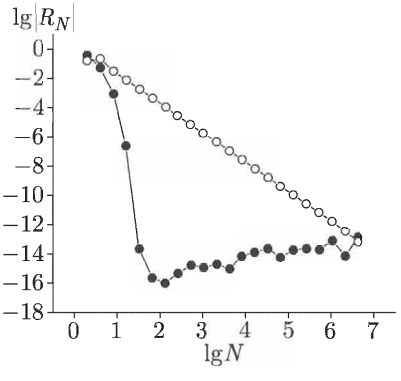

Рис. 3. Тесты на сгущающихся квазиравномерных сетках для формулы средних: ○ — (20), • — (21)

На примере вычисления этого несобственного интеграла сравним два варианта записи формулы средних на квазиравномерной сетке (20) и (21). Проведем тесты на сгущающихся сетках и нанесем на график зависимость погрешности квадратурных формул от числа узлов сетки в двойном логарифмическом масштабе (рис. 3).

Характер убывания погрешности квадратурной формулы (20) соответствует степенной зависимости погрешности от числа узлов сетки (прямая в двойном логарифмическом масштабе). Наклон прямой

подтверждает второй порядок точности метода $|R_N| \sim N^{-2}$. Но результаты для квадратурной формулы (21) приятно удивляют: характер зависимости погрешности от числа узлов не степенной, а экспоненциальный $|R_N| \sim e^{-N}$.

§ 2. Коллокационно-сеточные формулы

1. Коллокация. Пусть требуется найти значение интеграла

$$U = \int_a^b u(x)\rho(x)dx, \quad \rho(x) \neq 0, \tag{26}$$

где весовая функция $\rho(x)$ непрерывна на интервале (a, b). Чаще всего функцию $u(x)$ приближают многочленом, чтобы уже от него вычислить интеграл. Такая аппроксимация позволяет получить формулу численного интегрирования для вычисления значения U, которое линейно зависит от значения функции $u(x)$ в узлах:

$$U = \sum_{m=1}^{M} c_m u(x_m) + R_M,$$

где величины x_m называют узлами, c_m — весами, а R_M — погрешностью.

Формулы Гаусса–Кристоффеля. Одним из наиболее интересных аспектов формулы является выбор узлов и весов. Задача заключается в подборе $2M$ параметров так, чтобы квадратурная формула

$$U = \int_a^b u(x)\rho(x)dx \approx \sum_{m=1}^{M} c_m u(x_m)$$

была точна для любого многочлена степени не выше $2M - 1$. Оказывается, это можно сделать. Полученная квадратурная формула называется формулой Гаусса–Кристоффеля. Можно также доказать, что узлами формулы Гаусса–Кристоффеля являются нули многочленов соответствующей степени $P_M(x)$, ортогональных на $[a, b]$ с весом $\rho(x)$. Сложность задачи заключается в том, что для каждого нового веса нам необходимо знать свою систему ортогональных многочленов и значения их нулей. Лишь для некоторых функций $\rho(x)$ эти наборы известны.

Классическая формула Гаусса. Впервые формулу такого типа построил Гаусс для случая $\rho(x) = 1$. Линейным преобразованием можно перейти от отрезка $[a, b]$ к отрезку $[-1, 1]$. На нем ортогональны с единичным весом многочлены Лежандра. Нули этих многочленов симметричны относительно середины отрезка, и при маленьких степенях

для них известна точная формула в радикалах. Приведем узлы и веса простейших формул для отрезка $[-1,1]$:

$$M=1,\quad x_1=0,\quad c_1=2; \tag{27}$$

$$M=2,\quad -x_1=x_2=\sqrt{1/3},\quad c_1=c_2=1; \tag{28}$$

$$M=3,\quad -x_1=x_3=\sqrt{3/5},\quad x_2=0,\quad c_1=c_3=5/9,\quad c_2=8/9; \tag{29}$$

$$M=4,\quad -x_1=x_4=\sqrt{\frac{15+2\sqrt{30}}{35}},\quad -x_2=x_3=\sqrt{\frac{15-2\sqrt{30}}{35}},$$
$$c_1=c_4=\frac{18-\sqrt{30}}{36},\quad c_2=c_3=\frac{18+\sqrt{30}}{36}; \tag{30}$$

$$M=5,\quad -x_1=x_5=\sqrt{\frac{35+2\sqrt{70}}{63}},\quad -x_2=x_4=\sqrt{\frac{35-2\sqrt{70}}{63}},$$
$$x_3=0,\quad c_1=c_5=\frac{322-13\sqrt{70}}{900},\quad c_2=c_4=\frac{322+13\sqrt{70}}{900},\quad c_3=\frac{128}{225}. \tag{31}$$

Видно, что узлы и веса (27)–(31) симметричны относительно середины отрезка интегрирования.

Приведем мажорантную оценку погрешности классической формулы Гаусса, причем не для преобразованного отрезка $[-1,1]$, а непосредственно для исходного отрезка $[a,b]$:

$$|R_M|\leqslant\frac{(b-a)^{2M+1}(M!)^4}{(2M+1)[(2M)!]^3}\|u^{(2M)}\|_C\approx\frac{(b-a)}{2.5\sqrt{M}}\left(\frac{b-a}{3M}\right)^{2M}\|u^{(2M)}\|_C; \tag{32}$$

последнее приближенное равенство получено заменой факториалов по формуле Стирлинга. В оценке (32) нет такого малого параметра, как шаг h. Его роль играет величина $(b-a)/(3M)$. При достаточно больших M эта величина в степени $2M$ становится малой, на чем и основана высокая точность формулы Гаусса (разумеется, если $u^{(2M)}$ невелика). У быстро меняющихся функций (например, сильно осциллирующих) высокие производные не малы, и для них формулы Гаусса–Кристофеля не обеспечивают хорошей точности.

Обычно в литературе указывается, что расчет по формуле Гаусса–Кристофеля устойчив. Это не вполне верно. Результаты расчета чувствительны к точности, с которой заданы узлы и веса формулы. Если они заданы точными формулами вроде (27)–(31), то компьютер вычисляет их с машинной точностью, что обеспечивает устойчивость (это вычисление стоит проводить с максимальной разрядностью компьютера). Если же брать узлы и веса из справочной литературы, где они не всегда приведены с достаточным количеством знаков, точность расчета зачастую резко ухудшается. Поэтому явные выражения (27)–(31) особенно ценны для практики.

Метод Левина. Подойдем к задаче с другой стороны. Вместо поиска аппроксимации для подынтегральной функции, попробуем приблизить первообразную для нее. Рассмотрим интеграл

$$U = \int_a^b u(x)e^{p(x)}dx, \tag{33}$$

где $u(x)$ и $p(x)$ могут быть, вообще говоря, комплексными. Право выбора функции $p(x)$ остается за нами. Выберем ее так, чтобы $u(x)$ менялась медленно.

Возьмем систему линейно-независимых функций $\varphi_m(x)$, $1 \leqslant m \leqslant M$, и на отрезке $[a,b]$ положим

$$u(x)e^{p(x)} \approx \frac{d}{dx}[w(x)e^{p(x)}], \tag{34}$$

где $w(x) = \sum\limits_{m=1}^{M} c_m\varphi_m(x)$.

Введем на $[a,b]$ M точек коллокации $\xi_k \in [a,b]$, $1 \leqslant k \leqslant M$. Потребуем, чтобы в точках коллокации приближенное равенство (34) выполнялось точно:

$$u(x)e^{p(x)} = \frac{d}{dx}[w(x)e^{p(x)}], \quad x = \xi_k, \quad 1 \leqslant k \leqslant M.$$

Тогда для определения коэффициентов c_m получим систему линейных уравнений порядка M:

$$\sum_{m=1}^{M} c_m[\varphi'_m(\xi_k) + p'(\xi_k)\varphi_m(\xi_k)] = u(\xi_k), \quad 1 \leqslant k \leqslant M. \tag{35}$$

Матрица системы уравнений (35) должна быть невырожденной, что накладывает дополнительные условия на выбор точек коллокации и функции $\varphi_m(x)$. Из-за увеличения числа уравнений M обусловленность системы (35) может ухудшаться. Целесообразно решать ее методом Гаусса с выбором главного элемента. Вычислив коэффициенты c_m, мы получим приближенное значение интеграла:

$$U = \int_a^b u(x)e^{p(x)}dx \approx \sum_{m=1}^{M} [c_m\varphi_m(x)e^{p(x)}]\Big|_a^b.$$

В частности, когда есть всего одна точка коллокации, система состоит из единственного уравнения. Это дает нам

$$c_1 = \frac{u(\xi_1)}{\varphi'_1(\xi_1) + p'(\xi_1)\varphi_1(\xi_1)}.$$

Если в качестве ξ_1 выбрать середину интервала $\overline{x} = (a+b)/2$, а в качестве функции $\varphi_1(x)$ выбрать 1 (можно любую константу, т. к.

это не изменит результата), то квадратурная формула для вычисления интеграла будет

$$U \approx \frac{u(\overline{x})(e^{p(b)} - e^{p(a)})}{p'(\overline{x})}.$$

Условие разрешимости системы становится простым: $p'(\overline{x}) \neq 0$. Оно накладывает дополнительное условие на выбор показателя экспоненты. Однако мы заранее не знаем, что $p'(x) \neq 0$ именно в середине отрезка. Было бы разумным потребовать выполнения этого условия во всех точках $[a, b]$. Так как мы имеем дело с непрерывными функциями, это влечет $p'(x) > 0$ или $p'(x) < 0$ на всем интервале. Следовательно, показатель экспоненты $p(x)$ должен быть функцией монотонной. В частности,

$$U = \int_a^b u(x)e^{-\alpha x}dx \approx u(\overline{x})\frac{e^{-\alpha a} - e^{-\alpha b}}{\alpha}.$$

Система (35) линейных уравнений будет всегда иметь единственное решение в том и только в том случае, когда

$$\det[\varphi_m'(\xi_k) + p'(\xi_k)\varphi_m(\xi_k)] \neq 0, \quad 1 \leqslant m \leqslant M, \quad 1 \leqslant k \leqslant M.$$

В общем случае определить, какие системы функций $p(x)$ и $\varphi_m(x)$ удовлетворяют такому условию, затруднительно.

Коллокация для произвольного веса. Рассмотрим интеграл

$$U = \int_a^b u(x)\rho(x)dx, \tag{36}$$

где вес $\rho(x)$ выбирается так, чтобы выделить быстроменяющуюся часть подынтегрального выражения. При $\rho(x) = e^{p(x)}$ получаем метод Левина. Также мы предполагаем, что

$$u(x)\rho(x) \approx \frac{d}{dx}[w(x)\rho(x)] \tag{37}$$

на интервале $[a, b]$. Как и в методе Левина, представляем $w(x)$ обобщенным многочленом по первым M функциям системы $\{\varphi_m\}$:

$$w(x) = \sum_{m=1}^{M} c_m\varphi_m(x)$$

на всем интервале, и вводим M точек коллокаций $\xi_k \in [a, b]$, $1 \leqslant k \leqslant M$. Рассматривая в этих точках приближенное равенство как точное, т. е.

$$u(\xi_k)\rho(\xi_k) = \frac{d}{dx}[w(\xi_k)\rho(\xi_k)], \quad 1 \leqslant k \leqslant M, \tag{38}$$

получаем линейную систему уравнений для коэффициентов:

$$\sum_{m=1}^{M} c_m \left[\varphi'_m(\xi_k) + \varphi_m(\xi_k) \frac{\rho'(\xi_k)}{\rho(\xi_k)} \right] = u(\xi_k), \quad 1 \leqslant k \leqslant M. \tag{39}$$

Полученная система уравнений (39) похожа на систему (34) в методе Левина. Действительно, если положить $p(x) = \ln(\rho(x))$, то она полностью с ней совпадет. Приближенное значение интеграла в случае произвольного веса есть

$$U = \int_a^b u(x)\rho(x)dx \approx \sum_{m=1}^{M} \left[c_m \varphi_m(x)\rho(x) \right] \Big|_a^b. \tag{40}$$

Однако все равно возникает проблема выбора веса. Почему мы не можем взять его просто единицей? Предлагаемый здесь метод выделяет быстропеременную составляющую первообразной для того, чтобы оставшуюся часть $w(x)$ можно было хорошо приблизить линейной комбинацией выбранных функций $\varphi_m(x)$, которые меняются слабо.

Однако мы не знаем ни точную первообразную, ни ее быстроменяющуюся составляющую. Но мы делаем предположение, что из первообразной можно выделить тот же самый вес, что и у подынтегральной функции, оставив только медленно меняющуюся $w(x)$. А это уже дополнительное условие, которое мы накладываем на выбор такого веса. Вернемся к формуле (37). Продифференцируем правую часть и разделим на $\rho(x)$обе части выражения:

$$u(x) \approx w'(x) + w(x)\frac{\rho'(x)}{\rho(x)}.$$

Мы хотим, чтобы функции $u(x)$, $w'(x)$ и $w(x)$ были медленно меняющимися. Следовательно, должно не сильно изменяться и отношение ρ'/ρ (или $p(x)$ в методе Левина). Конечно, это не является достаточным условием выбора веса. Например, для интеграла $\int_0^\infty P(x)e^{-x-x^2}dx$, где $P(x)$ — слабо меняющаяся функция, в качестве веса лучше выбрать всю экспоненту, нежели только $\rho(x) = e^{-x^2}$, для которой отношение ρ'/ρ также слабо изменяется.

В принципе, мы не накладывали особых условий на систему функций $\varphi_m(x)$, кроме разрешимости системы (39). Метод коллокаций должен работать для любых линейно-независимых функций. Одним из вариантов выбора таких функций является система многочленов $\varphi_m(x) = x^{m-1}$, $1 \leqslant m \leqslant M$. Однако "центрированная" система $\varphi_m(x) = (x-\overline{x})^{m-1}$, $1 \leqslant m \leqslant M$, где $\overline{x} = (b+a)/2$ — середина интервала, может улучшить обусловленность системы.

Равномерная сетка. Разобьем наш интервал на N равных частей. Получим $\{x_n\}$ — сетку на $[a,b]$, где $x_0 = a$, $x_N = b$. Представим интеграл (36) в виде суммы

$$U = \int_a^b u(x)\rho(x)dx = \sum_{n=1}^{N} \int_{x_{n-1}}^{x_n} u(x)\rho(x)dx. \tag{41}$$

На каждом из полученных отрезков $[x_{n-1}, x_n]$ воспользуемся полученной нами квадратурной формулой (40). При этом функция $w(x)$ на каждом интервале будет, вообще говоря, своя; обозначим ее через $w_n(x)$, $x \in [x_{n-1}, x_n]$. Такая суммарная функция $w(x)$ оказывается разрывной во внутренних узлах x_n, $1 \leqslant n \leqslant N-1$. В результате получим

$$U = \sum_{n=1}^{N} [w_n(x)\rho(x)]\Big|_{x_{n-1}}^{x_n} \equiv w_N(x_M)\rho(x_N) + \\ + \sum_{n=1}^{N-1} \rho(x_n)[w_n(x_n) - w_{n+1}(x_n)] - w_0(x_0)\rho(x_0). \tag{42}$$

Разрыв функции $w(x)$ в узлах сетки дает дополнительный вклад — слагаемые в последней сумме равенства (42). Именно это отличает коллокационно-сеточную формулу (42) от обычной коллокационной формулы (40).

Опишем подробно коллокационно-сеточный алгоритм. На каждом n-м интервале вводится свой набор точек коллокации $\xi_{kn} \in \in [x_{n-1}, x_n]$; число точек коллокации одинаково во всех интервалах $1 \leqslant \leqslant k \leqslant M$, и строятся они по одинаковым законам. Системы базисных функций также строятся по общему закону, но формально могут оказаться различными в разных интервалах: $\varphi_{mn}(x)$, $1 \leqslant m \leqslant M$, $x \in [x_{n-1}, x_n]$; примером является система степеней, центрированных относительно середин сеточных интервалов:

$$\varphi_{mn}(x) = (x - x_{n-1/2})^{m-1}, \quad 1 \leqslant m \leqslant M, \quad x \in [x_{n-1}, x_n]. \tag{43}$$

Разумеется, вес $\rho(x)$ одинаков во всех интервалах.

В каждом интервале подбирается своя первообразная

$$w_n(x) = \sum_{m=1}^{M} c_{mn}\varphi_{mn}(x), \quad x \in [x_{n-1}, x_n]. \tag{44}$$

Коэффициенты c_{mn} формулы (44) определяются из условия коллока-

ции, т. е. решением системы линейных уравнений:

$$\sum_{m=1}^{M}\left[\varphi'_{mn}(\xi_{kn})+\varphi_{mn}(\xi_{kn})\frac{\rho'(\xi_{kn})}{\rho(\xi_{kn})}\right]c_{mn}=u(\xi_{kn}),\quad 1\leqslant k\leqslant M. \tag{45}$$

Затем на каждом интервале $[x_{n-1},x_n]$ интеграл вычисляется как разность первообразных на его границах, а полный интеграл определяется суммированием по интервалам:

$$U_N=\sum_{n=1}^{N}\left[\sum_{m=1}^{M}c_{mn}\varphi_{mn}(x)\rho(x)\right]\Bigg|_{x_{n-1}}^{x_n}. \tag{46}$$

Именно алгоритм (43)–(46) используется в практических вычислениях.

Как целесообразно выбирать узлы коллокации? Если $\rho(x)=1$, а сетка x_n равномерная ($h=x_n-x_{n-1}=\mathrm{const}$), то представляется целесообразным использовать нули многочленов Лежандра (27)–(31), приводимые к каждому отдельному интервалу. В оценке погрешности (32) при этом появляется множитель $(h/3M)^{2M}$. Тогда приведенные в (27)–(31) наборы узлов обеспечивают очень высокую точность до $O(h^{10})$.

Ниже будет показано, что аналогичный выбор точек коллокации возможен при достаточно произвольных $\rho(x)$ и квазиравномерных сетках.

Квазиравномерная сетка. График подынтегральной функции дает вычислителю информацию о поведении функции и участках отрезка интегрирования, вносящих основной вклад в интеграл. Разумно сгустить сетку на таких участках. Использование равномерной сетки в этом случае для достижения хорошей точности требует неприемлемо большого объема вычислений. В случае же бесконечного интервала равномерная сетка вовсе непригодна. Произвольная неравномерная сетка не позволяет контролировать точность расчетов. Этой проблемы позволяет избежать использование в расчетах квазиравномерных сеток. В частности, квазиравномерную сетку можно построить в неограниченной области.

Бесконечный интервал. Особую осторожность надо проявлять в случае неограниченной области интегрирования. Граничный интервал квазиравномерной сетки в этом случае неограничен. Пренебречь им нельзя, т. к. заранее нам неизвестен его вклад в суммарный интеграл. Если функция не является экспоненциально затухающей, эта добавка может достигать порядка $1/N$, что вряд ли является удовлетворительным. Применение метода коллокации к бесконечному интервалу имеет свои особенности. Рассмотрим вклад этого участка:

$$\int_{x_{N-1}}^{x_N}u(x)\rho(x)dx=w_N(x_N)\rho(x_N)-w_N(x_{N-1})\rho(x_{N-1}).$$

Если приближать $w(x)$ многочленами, то при большом количестве точек коллокаций, а следовательно, и при большой степени многочлена, $w(x)$ может расти быстрее, чем затухает вес. При таких условиях результаты расчета не могут оказаться верными. Отсюда получаем дополнительные ограничения на выбор системы $\varphi_m(x)$ и количества точек коллокации:

$$\lim_{x\to\infty} \varphi_m(x)\rho(x) = 0, \quad 1 \leqslant m \leqslant M.$$

Функции, имеющие несколько особенностей. Рассмотрим несобственный интеграл

$$U = \int_0^{+\infty} \frac{e^{-x}}{\sqrt{x}} dx.$$

Он имеет две особенности — на бесконечности и в нуле. В подобных случая далеко не всегда удается ввести единую квазиравномерную сетку, т. к. трудно придумать явное бесконечно гладкое выражение для преобразования $x(\xi)$, хорошо сгущающее сетку вблизи каждой сингулярности и разреживающее ее на бесконечности. Недостаточно удачный выбор преобразования может существенно ухудшить точность расчета. Поэтому такие интегралы следует разбить на части, рассматривая каждую из них как отдельный интеграл со своей квазиравномерной сеткой, учитывающей нужную особенность.

2. Погрешность коллокационно-сеточных формул.

Равномерная сетка. Рассмотрим произвольный интервал $[x_{n-1}, x_n]$. Далее будем оставаться в этом интервале, поэтому индекс n часто будем опускать. Введем середину интервала $\overline{x} = (x_n + x_{n-1})/2$ и его длину $h = x_n - x_{n-1}$.

Сделаем замену переменных $x = \overline{x} + \zeta$, $-h/2 \leqslant \zeta \leqslant h/2$, и запишем отрезок ряда Тейлора

$$u(x) = \sum_{q=0}^{\infty} \frac{\zeta^q}{q!} \overline{u}^{(q)} = \overline{u} + \zeta\overline{u}_x + \frac{1}{2}\zeta^2\overline{u}_{xx} + \frac{1}{6}\zeta^3\overline{u}_{xxx} + \ldots,$$

где за центр разложения принята середина интервала $\overline{x}$. В случае конечного количества непрерывных и ограниченных производных ряд можно оборвать в соответствующем месте, получив остаточный член порядка $O(h^s)$. Поэтому далее везде будем записывать сумму бесконечного ряда, считая, что он обрывается, если производных недостаточно. Обозначим $F(x) = u(x)\rho(x)$ и $\Psi(x) = \dfrac{d}{dx}[w(x)\rho(x)]$. Истинное значение интеграла по этому интервалу S и его приближенное значение σ

оказываются равны соответственно

$$S = \int\limits_{x_{n-1}}^{x_n} F(x)dx = \int\limits_{-h/2}^{h/2} F(\overline{x}+\zeta)d\zeta = h\sum_{q=0}^{\infty} \frac{(h/2)^q F^{(q)}(\overline{x})}{(q+1)!}\delta_q,$$

$$\sigma = \int\limits_{x_{n-1}}^{x_n} \Psi(x)dx = \int\limits_{-h/2}^{h/2} \Psi(\overline{x}+\zeta)d\zeta = h\sum_{q=0}^{\infty} \frac{(h/2)^q \Psi^{(q)}(\overline{x})}{(q+1)!}\delta_q,$$

где $\delta_q = \begin{cases} 1 \text{ при } q \text{ четном} \\ 0 \text{ при } q \text{ нечетном} \end{cases}$, т. к. нечетные степени выпадают в силу симметрии.

Тогда условие коллокации будет выглядеть так:

$$F(\overline{x}+\zeta_k) = \Psi(\overline{x}+\zeta_k), \quad 1 \leqslant k \leqslant M;$$

здесь ζ_k — точки коллокации в локальной системе координат интервала.

Запишем его в виде

$$\sum_{q=0}^{\infty} \frac{\xi_k^q}{q!}(F^{(q)}(\overline{x}) - \Psi^{(q)}(\overline{x})) = 0. \tag{47}$$

Умножим каждое из M равенств (47) на произвольный коэффициент α_k и сложим:

$$\sum_{k=1}^{M}\sum_{q=0}^{\infty} \alpha_k \frac{\xi_k^q}{q!}(F^{(q)}(\overline{x}) - \Psi^{(q)}(\overline{x})) = 0. \tag{48}$$

Тогда выражение (48) можно добавить к $(S-\sigma)/h$, не изменив значения:

$$\frac{S-\sigma}{h} = \sum_{q=0}^{\infty} \frac{F^{(q)}(\overline{x}) - \Psi^{(q)}(\overline{x})}{q!} \left[\frac{(h/2)^q}{q+1}\delta_q - \sum_{k=1}^{M} \alpha_k \xi_k^q\right]. \tag{49}$$

Получение порядка точности M. Формула (49) верна при любых α_k. Если потребовать равенства нулю первых M квадратных скобок в сумме (49), то для определения α_k, $0 \leqslant k \leqslant M-1$, получим систему линейных алгебраических уравнений, определитель матрицы которой есть определитель Вандермонда, отличный от нуля. Поэтому такие α_k, $0 \leqslant k \leqslant M-1$, найдутся; это гарантирует погрешность квадратурной формулы $S - \sigma = O(h^M)$, поскольку $|\xi_k| \leqslant h/2$. Следовательно, для произвольно расположенных точек коллокации погрешность квадратурной формулы есть $O(1/N^M)$.

Выбор узлов. Также мы можем выбирать и расположение M точек коллокации. Используя эту степень свободы, можно обратить в нуль еще M квадратных скобок в выражении для погрешности (49). Таким образом, погрешность станет $O(1/N^{2M})$, т. е. мы повышаем порядок точности вдвое. При этом указанная точность получается независимо от $F(x)$ и $\Psi(x)$ (а следовательно, и для любого выбора веса $\rho(x)$). Требуется только наличие непрерывных и ограниченных производных вплоть до $2M$-го порядка, чтобы в сумме (48) можно было учесть $2M + 1$ слагаемых. Итак, для *любых функций* $u(x)$ *и* $\rho(x)$ *набор оптимальных точек коллокации один и тот же.*

Например, возьмем вес $\rho(x) = 1$ и базисную систему $\varphi_m(x) = (x - \overline{x})^{m-1}$. Тогда $\Psi(x)$ оказывается многочленом степени $2M - 1$. Задача свелась к классической формуле Гаусса. А для нее квадратурные узлы задания функции, обеспечивающие максимальную точность $O(h^{2M})$, суть нули многочленов Лежандра степени M. Следовательно, они же являются оптимальными точками коллокации для произвольных функций и весов.

Особо отметим отличие полученных точек от узлов в формулах Гаусса–Кристоффеля: расположение последних зависит от веса, который мы выбрали. В коллокационном же методе нули полиномов Лежандра и только они годны для всех весов. На практике достаточно знать нули только первых 5 многочленов: они приведены в (27)–(31). Дальнейшее увеличение количества уравнений в системе (39) ухудшает ее обусловленность. А для первых 5 многочленов Лежандра нули известны в радикалах. Следует позаботиться о достаточно точном вычислении положения этих нулей, поскольку результат столь же чувствителен к точности их задания, как и для формул Гаусса–Кристоффеля. Рекомендуется использовать наибольшую разрядность чисел, достижимую на ЭВМ.

Квазиравномерная сетка. Пусть $x = x(\eta)$: $[\alpha, \beta] \to [a, b]$ — производящая функция квазиравномерной сетки. Гладкость функции $x(\eta)$ позволяет нам сделать в интеграле замену переменных:

$$U = \int_a^b F(x)dx = \int_\alpha^\beta F(x(\eta))x'(\eta)d\eta = \int_\alpha^\beta \widetilde{F}(\eta)d\eta, \tag{50}$$

при этом нам неважно, конечен или нет интервал (a, b). Заметим, что мы свели задачу к вычислению квадратуры с использованием равномерной сетки на конечном интервале, но с другой функцией $\widetilde{F}(\eta)$. Действительно, мы приближаем (50) интегралом

$$U \approx \int_a^b \Psi(x)dx = \int_\alpha^\beta \Psi(x(\eta))x'(\eta)d\eta = \int_\alpha^\beta \widetilde{\Psi}(\eta)d\eta \tag{51}$$

и условие коллокации для интервала $[a, b]$

$$F(\eta_k) = \Psi(\eta_k), \quad 1 \leqslant k \leqslant M,$$

равносильно условию коллокации для интервала $[\alpha, \beta]$

$$\widetilde{F}(\tau_k) = \widetilde{\Psi}(\tau_k),$$

где τ_k — точки коллокации по переменной η, соответствующие точкам коллокации ξ_k по переменной x в смысле преобразования $x(\eta)$.

Следовательно, порядок точности квадратурной коллокационно-сеточной формулы $O(N^{-2M})$ сохраняется. Нужно лишь помнить, что меняется условие ограниченности производных, т. к. функции несколько модифицируются.

Ограниченность производных. Разумеется, если сама функция $F(x)$ или ее нужные производные неограничены, то формула коллокаций непригодна. В случае бесконечного интервала это особенно актуально, т. к. производные должны быть ограничены не только у $\widetilde{F}(\eta)$, но и у функции $\widetilde{\Psi}(\eta)$, для первообразной которой использовалось приближение многочленами, неограниченными на бесконечности.

Погрешность решения линейной системы. При использовании коллокационно-сеточного метода для каждого интервала сетки приходится решать систему линейных уравнений (39), обусловленность которой с ростом числа узлов коллокации M резко ухудшается. На практике опасно использовать более 5 узлов коллокации, но этого обычно вполне достаточно для достижения нужной точности.

Регулярный и нерегулярный режимы. При правильном выборе узлов коллокации погрешность квадратурной формулы есть $O(h^{2M})$. Следует ожидать, что при достаточно маленьком шаге h погрешность есть $\varepsilon = C\,h^{2N} + O(h^{2N+1})$, где второе слагаемое пренебрежимо мало по сравнению с первым. График зависимости $\lg \varepsilon$ от $\lg h$ асимптотически близок к прямой с коэффициентом наклона $-2M$. Такое убывание погрешности с ростом числа узлов сетки назовем *регулярным режимом*. Ясно, что при малых N выход на него еще не достигается в силу влияния второго слагаемого. При очень же малых шагах сетки, т. е. при большом количестве интервалов N, как говорилось выше, скажется влияние погрешности решения системы линейных уравнений и ошибок округления. То есть режим убывания погрешности с ростом числа узлов *нерегулярный*.

3. Примеры. Проиллюстрируем применение коллокационно-сеточных квадратурных формул на примерах.

Пример 3. Для начала выберем в качестве тестовой быстро затухающую подынтегральную функцию

$$\int_0^\infty e^{-p^2x^2}dx = \frac{\sqrt{\pi}}{2p}. \tag{52}$$

Достаточно разумным выбором веса будет все подынтегральное выражение. Тогда отношение $\rho'/\rho = -2px$ меняется не так значительно по сравнению с самим весом $\rho(x) = e^{-p^2x^2}$. На рисунке 4 представлены результаты теста на сгущающихся сетках. В двойном логарифмическом масштабе показано убывание погрешности численного решения при последовательном удвоении числа узлов сетки. Черные значки соответствуют выбору узлов коллокации в нулях полиномов Лежандра для разного числа точек коллокации M. Пустые значки для того же числа точек коллокации M показывают результаты при произвольном расположении узлов коллокации.

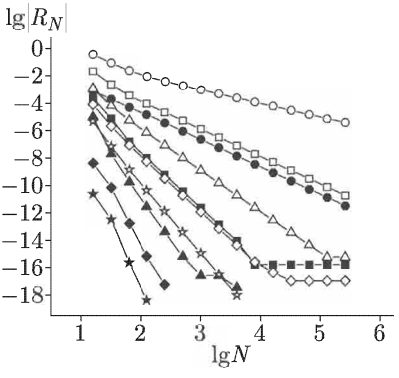

Рис. 4. Тест на сгущающихся сетках для примера 3. Черные значки — узлы коллокации выбраны в нулях полиномов Лежандра; пустые значки — произвольный выбор узлов коллокации. • — $M = 1$, ■ — $M = 2$, ▲ — $M = 3$, ♦ — $M = 4$, ★ — $M = 5$

Порядок точности соответствует теории. Если узлы расположены в нулях полиномов Лежандра, мы получаем наклон кривой, соответствующий $R_N = O(N^{-2M})$. Если же поставить узлы коллокации произвольно, то порядок точности падает вдвое.

Был проведен также следующий эксперимент. Для двух точек коллокации $M = 2$, узлы выбирались симметрично относительно концов интервала. Приближая положение этих двух точек к положению нулей полинома Лежандра второй степени и сохраняя при этом симметричное расположение этих точек, мы следили за порядком убывания погрешности. Точность везде имела второй порядок и резко изменяла его до 4-го лишь при близком совпадении узлов коллокации с нулями полиномов

Лежандра. Это говорит о необходимости предельно аккуратно задавать положение точек коллокации.

Выбор параметров сетки. В расчетах использовалось семейство квазиравномерных сеток, порожденное преобразованием $x(\eta) = c \cdot \mathrm{tg}(\pi\eta/2)$, $\eta \in [0,1]$, и покрывающее полупрямую. Характерный размер области, дающей основной вклад в интеграл (52), есть $3/p$. Разумно потребовать, чтобы в этой области оказалась половина узлов сетки. Откуда получаем $c = 3/p$. Действительно, при вычислении интеграла (52) для $p = 1$, изменив значение масштабного коэффициента сетки с $c = 10$ на $c = 3$, в эксперименте удалось быстрее выйти на заданный уровень точности.

Отметим, что горизонтальные участки кривых соответствуют выходу на ошибки округления; их уровень $\sim 10^{-16}$–10^{-18} слегка уменьшается при увеличении M.

Видно, что формулы 8–10-го порядка обеспечивают очень высокую точность уже при небольшом числе интервалов.

Пример 4. Теперь выберем более сложную подынтегральную функцию с медленным степенным затуханием

$$U = \int_0^\infty \frac{\mathrm{arctg}(x)}{\sqrt{x^4 + x^2 + 1}}\, dx.$$

Арктангенс — медленноменяющаяся функция, поэтому в качестве веса разумно выбрать $\rho(x) = \dfrac{1}{\sqrt{x^4 + x^2 + 1}} \underset{x\to\infty}{\longrightarrow} \dfrac{1}{x^2}$. Для $M = 1$ и точке коллокации $\xi_1 = 0$ согласно (27) мы получаем полное совпадение с теорией (рис. 5), т. е. наклон линии соответствует второму порядку точности. Однако с увеличением числа узлов коллокации точность даже ухудшается, т. е. при $M \geqslant 2$ все наклоны соответствуют первому порядку точности! Это связано с тем, что при $M > 1$ функция $w(x)$ растет на

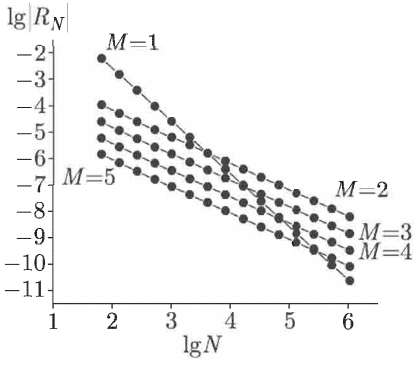

Рис. 5. Результаты теста на сгущающихся сетках для примера 4

бесконечности как полином степени M, а поэтому $\lim\limits_{x\to\infty} w(x)\rho(x) \not\to 0$ и вклад последнего интервала не может быть учтен верно, что соответствует общей точности расчета $O(N^{-1})$.

Пример 5. Рассмотрим интеграл от осциллирующей функции

$$\int\limits_0^\infty \frac{\cos px}{x^2+a^2}dx.$$

Если применять коллокационно-сеточный метод к интегралу, записанному в такой форме, то не удается получить сходимость даже при одной точке коллокации. Выбрав в качестве веса быстроизменяющуюся часть подынтегральной функции $\rho(x) = \cos(px)$, получим $\rho'/\rho = -p\,\mathrm{tg}(px)$. О медленном изменении этой дроби говорить не приходится; полученная функция оказывается даже неограниченной.

Эту проблему можно решить, учитывая соотношение $\cos(px) = \mathrm{Re}[e^{ipx}]$ и переходя к комплексному интегралу

$$\int\limits_{-\infty}^\infty \frac{e^{ipx}}{x^2+a^2}dx = \frac{\pi}{a}e^{-|ap|}.$$

В качестве веса лучше выбрать все подынтегральное выражение, тогда проблема последнего интервала снимается, по крайней мере, для одной точки. На рисунке 6 видно, что прямая убывания ошибки с ростом числа узлов сетки, соответствующая одной точке коллокации, имеет тангенс угла наклона -2, т. е. $R_N = O(N^{-2})$. При большем числе узлов коллокации увеличения порядка точности метода не происходит из-за невозможности правильно учесть вклад последнего интервала.

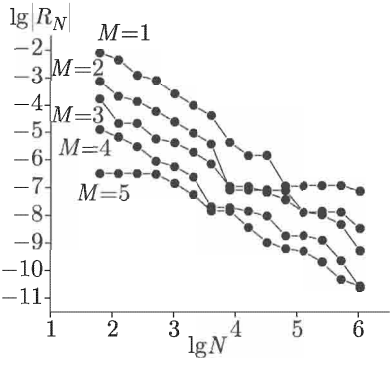

Рис. 6. Результаты теста на сгущающихся сетках для примера 5

Заметим, что даже при $M = 1$ кривая погрешности не плавная. Фактически результаты хуже, чем в примере 4, хотя подынтегральная функция убывает с той же скоростью, а сам интеграл приме-

ра 5 сходится даже быстрее из-за знакопеременности подынтегральной функции. Причина в том, что погрешность выражается через старшие производные подынтегральной функции; благодаря осциллирующему характеру подынтегральной функции в примере 5, ее производные медленно убывают на бесконечности.

§ 3. Вычисление плазменных микрополей

1. Модель плазменного микрополя. Распределение напряженности микроскопического электрического поля в плазме строится методами статистической физики через интеграл Фурье

$$f(E) = \frac{2}{\pi}\frac{E}{E_0^2}\int\limits_0^\infty \sin\left(\frac{E}{E_0}l\right) Q\,(l)\,ldl; \tag{53}$$

здесь E — напряженность микрополя. Для фурье-образа $Q(l)$ в [Грим, 1972] построен ряд модельных приближений. Эти приближения были усовершенствованы в кандидатской диссертации [Голосной, 1995], где получено следующее выражение:

$$\ln Q(l) = -\frac{1}{x_e}\sum_k \frac{(z_k l)^{3/2} x_k}{1 + \dfrac{6}{5}(z_k l)^{1/2} x_e^{-1/3} + 2z_0\Gamma(z_k/l)^{1/2}}. \tag{54}$$

Здесь z_k, x_k — заряды и концентрации ионов плазмы, x_e — концентрация электронной; z_0 — заряд, в окрестности которого ищется микрополе; Γ — параметр неидеальности; E_0 — характерный масштаб. В работе [Голосной, 1995] не был построен хороший метод вычисления данного интеграла. Наблюдались случаи, когда на "хвосте" функции распределения $f(E)$ появлялась немонотонность. Ниже показано, как применение коллокационно-сеточных методов позволяет вычислить этот интеграл с высокой точностью.

Нас интересует расчет интеграла (53), содержащего осцилляцию и экспоненциальное затухание. Решим задачу с упрощенной функцией $Q(l)$, сохраняющей особенности около 0 и ∞:

$$Q(l) = \exp\left(-\frac{al^2}{1+l}\right). \tag{55}$$

Сделаем замену $p = E/E_0$. Интеграл (53) с точностью до множителя равен

$$I(p) = \int\limits_0^\infty \sin(px)\exp\left(-\frac{ax^2}{1+x}\right)xdx. \tag{56}$$

Как мы уже знаем, в случае осцилляций следует перейти к комплексным функциям:

$$I(p) = \operatorname{Im} \int_0^\infty \exp\Big(- \frac{ax^2}{1+x} + ipx \Big) x dx. \tag{57}$$

На рисунке 7 показано убывание погрешности вычисления интеграла (57) с ростом числа узлов сетки при различном числе узлов коллокации $M = 1, 2, 3, 4, 5$. Наличие регулярного режима убывания погрешности говорит о высокой надежности результатов. Порядок точности метода везде соответствует теоретическому $R_N = O(N^{-2M})$.

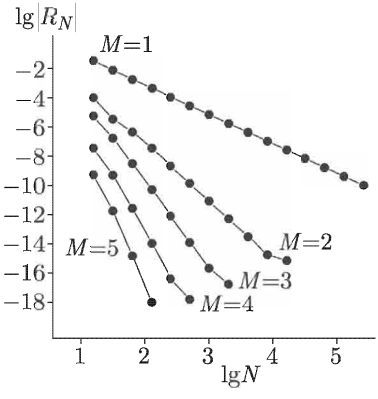

Рис. 7. Результаты тестов на сгущающихся сетках при вычислении плазменных микрополей. Зависимость логарифма погрешности от логарифма числа узлов сетки N при разном числе точек коллокации M

Для решения основной задачи — вычисления плазменного микрополя — рассчитаем значения интеграла (57) для различных аргументов p. Заметим, что с увеличением p подынтегральная функция все сильнее осциллирует и задача вычисления интеграла (57)

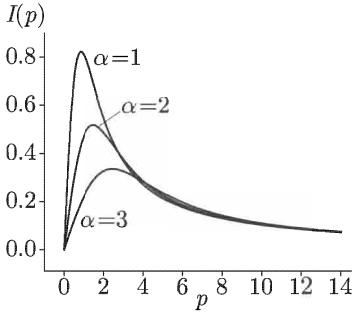

Рис. 8. Распределение плазменного микрополя

усложняется. Пример зависимости $I(p)$ для различных значений a приведен на рисунке 8. Можно обратить внимание, что при увеличении экспоненциального затухания кривая "расширяется", а ее максимум уменьшается. Также видно, что при увеличении осцилляций значение $I(p)$ быстро падает, и при достижении им порядка 10^{-13}–10^{-15} дальнейший расчет становится бессмысленным из-за ошибок округления, которые имеют такой же порядок. Немонотонность $I(p)$ не наблюдается вплоть до выхода на ошибки округления.

Глава V

СПЕКТРАЛЬНЫЕ ЗАДАЧИ

В этой главе построены численные методы нахождения спектров дифференциальных операторов в неограниченной области. Проведено сравнение метода дополненного вектора, метода обратных итераций и обратных итераций со сдвигом по скорости сходимости и устойчивости. Показана неприменимость фазового метода в случае неограниченной области.

§ 1. Известные методы

Подробнее остановимся на задаче вычисления спектра дифференциального оператора. Эти задачи могут быть линейными $Bv(x) = \mu v(x)$, где B — дифференциальный оператор (включающий граничные условия), или нелинейными $B(v(x), \mu) = 0$. Примером нелинейных задач являются интегро-дифференциальные уравнения Харти–Фока для многоэлектронного атома (см., например, [Ландау, Лифшиц, 1963]).

Сначала для численного решения таких задач использовали метод стрельбы. Однако возникали серьезные трудности: 1) сопутствующая задача Коши нередко оказывалась очень плохо обусловленной при наличии особых точек, а также в случае большой или полубесконечной, как в квантовой механике, области; 2) обычно собственных значений много, и стрельба могла сходиться неизвестно к какому из них, или вообще не сходиться.

Преодолеть первую трудность позволяет сеточный метод. Введем сетку $\omega = \{x_n,\ 0 \leqslant n \leqslant N\}$, дифференциальный оператор B аппроксимируем разностным оператором A и получим алгебраическую задачу $A(u, \lambda) = 0$, где $u = \{u_n\}$ есть сеточная функция; очевидно, $u_n \approx \approx v(x_n)$, $\lambda \approx \mu$. Эта краевая задача обычно хорошо обусловлена даже при плохой обусловленности задачи Коши. Для бесконечной или очень большой области применим адаптивную квазиравномерную сетку со специальной аппроксимацией производных (гл. III, § 2). Алгебраическую задачу $A(u, \lambda) = 0$ решим итерационным методом дополненного вектора, пригодным для нелинейных задач. Но вопрос о сходимости итераций остается.

Вторую трудность удачно преодолевают лишь для линейной задачи Штурма–Лиувилля; но именно к ней сводятся многие актуальные проблемы (в частности, квантовой механики). Это делают фазовым

методом [Никифоров и др., 2000]. Модификации этого метода сходятся к нужному собственному значению, причем за 2–4 итерации. Однако такие модификации узко специализированы и требуют адаптации нулевого приближения к задаче.

Ниже рассмотрены задачи с линейными операторами B, A. Для задачи Штурма–Лиувилля построен вариант, не требующий адаптации — метод дополненного вектора фазы. Однако показано, что все варианты фазового метода неприменимы в больших областях. Для оператора произвольного порядка улучшен метод обратных итераций со сдвигом, который сходится очень быстро, хотя заранее неизвестно, к какому собственному значению.

§ 2. Методы, пригодные в ограниченной области

1. Дополненный вектор фазы. Он строится для задачи Штурма–Лиувилля

$$\frac{d}{dx}\left[p(x)\frac{dv}{dx}\right] + q(x)v(x) = \lambda r(x)v(x),$$
$$p(x) > 0, \quad q(x) > 0, \quad r(x) > 0, \quad v(a) = v(b) = 0. \tag{1}$$

В 1950-х годах И.М. Соболь предложил перейти к фазовым переменным: ввести новые функции амплитуды $\rho(x)$ и фазы $\varphi(x)$ соотношениями (см., например, [Никифоров и др., 2000])

$$v(x) = \rho(x)\sin\varphi(x),$$
$$\frac{dv}{dx} = \frac{\rho(x)}{p(x)}\cos\varphi(x), \quad \rho(x) > 0. \tag{2}$$

Подстановка (2) в (1) показывает, что задача на собственные значения формулируется только для фазы:

$$\frac{d\varphi}{dx} = F(x,\varphi(x),\lambda) \equiv \frac{1}{p(x)}\cos^2\varphi(x) + [q(x) - \lambda r(x)]\sin^2\varphi(x),$$
$$\varphi(a) = 0, \quad \varphi(b) = \pi k, \tag{3}$$

где $k = 1, 2, \ldots$ — число полуволн фазы, т. е. номер собственного значения.

Теорема. *Задача* (3) *имеет единственное решение* λ_k, $\varphi_k(x)$; *величины* λ_k *монотонно убывают с ростом* k.

Поскольку для целых l справедливо $\varphi'(\pi l) = 1/p(\pi l) > 0$, то каждому k соответствует только одно решение этой задачи; при доказательстве надо учитывать, что производная $\partial[\varphi'(x)]/\partial\lambda = -r(x)\sin^2\varphi(x) < 0$ всюду, кроме отдельных точек $x = \pi l$, в которых она обращается в нуль. Поэтому $\varphi(x,\lambda)$ монотонно убывает по λ, и очевидна единствен-

ность решения задачи на собственные значения (3). Очевидно также, что λ_k монотонно убывает по k.

Таким образом, переход к фазе позволяет выделить из спектра задачи (1) нужное собственное значение. Решив задачу (3) и найдя $\varphi_k(x)$ и λ_k, получим амплитуду интегрированием соотношения

$$\frac{d}{dx}\ln\rho(x) = \left[\frac{1}{p(x)} - q(x) + \lambda r(x)\right]\sin\varphi(x)\cos\varphi(x)$$

с точностью до нормировочного множителя.

По-видимому, выбранная замена (2) при $p(x) \neq 1$ приводит к наиболее простым формулам для решаемых задач.

Для численной реализации используем идею дополненного вектора [Калиткин, 1965]. Обозначим $\varphi_n = \varphi(x_n)$ и аппроксимируем (3) разностной схемой точности $O(h^2)$:

$$\begin{gathered}\frac{\varphi_n - \varphi_{n-1}}{x_n - x_{n-1}} = F\left(x_{n-1/2}, \frac{\varphi_n + \varphi_{n-1}}{2}, \lambda\right), \\ 1 \leqslant n \leqslant N; \quad \varphi_0 = 0, \quad \varphi_N = \pi k.\end{gathered} \tag{4}$$

Введем дополненный вектор фазы с $N+2$ компонентами:

$$\Phi = \{\varphi_0, \varphi_1, \ldots, \varphi_N, \varphi_{N+1} \equiv \lambda\}.$$

Для него задача (4) есть система $N+2$ нелинейных уравнений с таким же числом неизвестных. Ее следует решать итерационным методом Ньютона или его непрерывным аналогом [Жидков, Пузынин, 1967], [Ермаков, Калиткин, 1981]. Вблизи решения итерации сходятся квадратично. Сходимость вдали от решения исследовалась на тестах.

Пример 1. Возьмем в качестве задачи (1) радиальную часть уравнения Шредингера для атома водорода в сферических координатах [Ландау, Лифшиц, 1963], выбрав атомные единицы. Этому соответствуют

$$\begin{gathered}p(x) \equiv 1, \quad r(x) \equiv 1, \quad q(x) = \frac{2}{x} - \frac{l(l+1)}{x^2}, \\ \lambda = -2E, \quad a = 0, \quad b = R;\end{gathered} \tag{5}$$

здесь R — радиус атома, E — энергия уровня, $l = 0, 1, \ldots$ есть орбитальное квантовое число, $k = 1, 2, \ldots$ есть радиальное квантовое число.

Для $R = \infty$ точные собственные значения $\lambda = (k+l)^{-2}$, а собственные функции экспоненциально затухают. Например, при $k = 1$ решением является

$$v(x) = \text{const} \cdot x^l \exp\left(-\frac{x}{l+1}\right). \tag{6}$$

Решения задачи (1), (5) близки к предельным при больших R; на практике достаточно $R > 20$. Это обеспечивает хороший тест.

Расчеты. В подобных сеточных задачах термин "сходимость" имеет два употребления. Один — это сходимость сеточного решения к точному при $N \to \infty$; она следует из аппроксимации и устойчивости, и будет проиллюстрирована позднее. Второй — сходимость итерационного процесса нахождения сеточного решения. Это сейчас наиболее интересно, ибо определяет эффективность алгоритма.

Сходимость итераций иллюстрируется на примере (5) при $R = 22$, $k = 1$, $l = 1$ для равномерной сетки с $N = 100$. Сеточное решение определялось с высокой точностью по сошедшимся итерациям и использовалось для нахождения погрешностей на первых итерациях. Нулевое приближение профиля $\varphi(x)$ бралось линейным (что естественно и универсально):

$$\varphi^{(0)}(x) = \pi k \frac{x-a}{b-a},$$

а $\lambda^{(0)}$ довольно далеко отстояло от точного.

Зависимость погрешности $|\lambda^{(s)} - \lambda|$ от числа итераций показана на рисунке 1. Видно, что первая итерация даже ухудшает точность (из-за плохого $\lambda^{(0)}$). Но далее процесс быстро приходит в область квадратичной сходимости, и достаточно шести итераций. Таким образом, метод дополненного вектора фазы по скорости практически не уступает лучшим специализированным методам [Никифоров и др., 2000], а по универсальности существенно превосходит их.

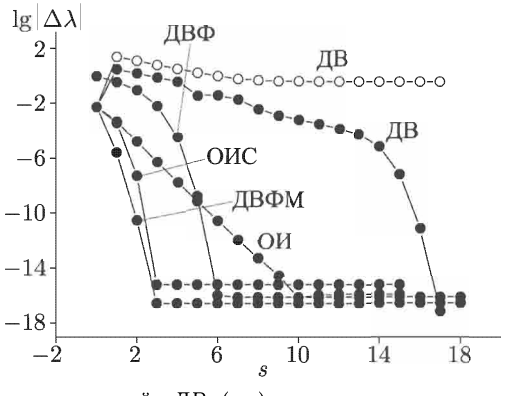

Рис. 1. Сходимость итераций: ДВ (-○-) — метод дополненного вектора при плохом начальном приближении, итерации сходятся к решению с другими квантовыми числами ($k = 2$, $l = 0$); ДВ (-●-) — метод дополненного вектора, сходимость к нужному собственному значению ($k = 3$, $l = 0$); ДВФ (-●-) — метод дополненного вектора фазы для равномерной сетки на конечном промежутке $R = 22$ ($k = 1$, $l = 1$); ДВФМ (-●-) — метод дополненного вектора фазы с выбором начального приближения по результатам расчетов на более грубой сетке; ОИ (-●-) — метод обратной итерации ($k = 3$, $l = 0$); ОИС (-●-) — метод обратной итерации со сдвигом ($k = 3$, $l = 0$); число узлов сетки во всех расчетах $N = 100$

Начальное приближение для итерационного процесса выше было выбрано линейным. Если же сначала провести расчет на грубой сетке и полученное решение взять в качестве начального приближения для итерационного процесса на более подробной сетке, то скорость сходимости возрастает [Марчук, Шайдуров, 1979] и для получения рекордной точности 10^{-16} достаточно всего трех итераций. Тем самым многосеточный метод дополненного вектора фазы по скорости превосходит все существующие [Никифоров и др., 2000].

2. Бесконечная область. Для спектроскопии разреженных газов следует брать $R \gg 1$. Однако расчеты приведенного примера при $R > 40$ и равномерной сетке $N = 100$ давали расходимость итераций. Сгущением сетки до $N = 1000$ удавалось отодвинуть границу расходимости итераций до $R \approx 55$, но дальнейшее продвижение было почти невозможным (рис. 2).

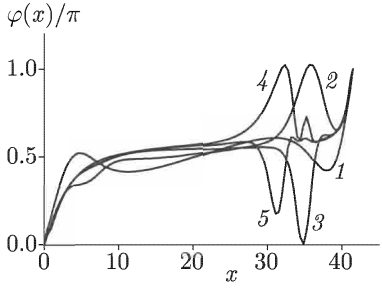

Рис. 2. Расчет при очень большом $R = 41.5$, $N = 100$, $k = 1$. Цифры около кривых — номера итераций. Видно, что итерации не сходятся

Причина этого оказалась фундаментальной. Положим $R = \infty$ и возьмем точное решение (6). Тогда из (2) легко получить точное решение для фазы:

$$\operatorname{ctg}\varphi(x) = \frac{l+1}{x} - \frac{1}{l+1}, \quad \operatorname{ctg}\varphi(\infty) = -\frac{1}{l+1}, \quad k = 1. \tag{7}$$

Таким образом, $\varphi(\infty) \neq \pi$. Аналогичные точные решения можно построить для произвольного k, причем опять $\varphi(\infty) \neq \pi k$. Следовательно, *фазовый метод не дает правильного предела при* $R \to \infty$.

Таким образом, фазовый метод применим лишь для не очень больших R. Ставить же при $R = \infty$ граничное условие, точно соответствующее (7), бессмысленно: оно будет привязано к специфическому виду $q(x)$, т. е. не будет универсальным.

Однако допустимые R довольно велики. Поэтому фазовый метод остается пригодным для расчетов спектров кристаллов или плотной плазмы, не говоря уже о колебаниях упругих мембран. Главное его преимущество — сходимость к нужному собственному значению, причем очень быстрая.

§ 3. Итерационные методы вычисления спектров в неограниченных областях

1. Обратные итерации. Линейную дифференциальную задачу произвольного порядка $Bv(x) = \mu v(x)$ можно аппроксимировать сеточной задачей $Au = \lambda u$. Это задача на собственные значения матрицы A порядка $N-1$ (если краевые условия $u_0 = u_N = 0$), причем обычно нужны лишь несколько первых собственных значений и векторов. Как их лучше найти?

Пусть известно хорошее приближение $\overline{\lambda}$ к одному из собственных значений λ_k, т. е. $|\overline{\lambda} - \lambda_k| \ll |\overline{\lambda} - \lambda_m|$, $m \neq k$. Тогда можно применить процесс обратных итераций (см., например, [Уилкинсон, 1970; Калиткин, 1978]):

$$(A - \overline{\lambda}E)\, u^{(s+1)} = u^{(s)}. \tag{8}$$

Вектор $u^{(s)}$ сходится *по направлению* к k-му собственному вектору. Очередное приближение к собственному значению определяют из соотношения

$$(\lambda^{(s+1)} - \overline{\lambda})\, u^{(s+1)} \approx u^{(s)}. \tag{9}$$

Сходимость линейная со знаменателем

$$q \approx \frac{|\overline{\lambda} - \lambda_k|}{\min\limits_{k \neq m} |\overline{\lambda} - \lambda_m|} < 1;$$

поэтому число итераций чувствительно к выбору $\overline{\lambda}$, но слабо зависит от выбора $u^{(0)}$. Если $\overline{\lambda}$ выбрано наудачу, то процесс сойдется к ближайшему λ_k.

Однако векторы $u^{(s+1)}$ и $u^{(s)}$ не параллельны, а в литературе нет никаких указаний, как конкретизировать понятие близости в (9). Потребуем для этого минимального расхождения правой и левой частей в (9):

$$\left\| (\lambda^{(s+1)} - \overline{\lambda})\, u^{(s+1)} - u^{(s)} \right\|_{L_2} = \min. \tag{10}$$

Для неэрмитовых матриц A (несамосопряженных операторов B) собственные векторы могут быть комплексными, поэтому норма и скалярные произведения равны:

$$\|u\|_{l_2}^2 = (u, u), \quad (u, w) = \sum_{n=1}^{N} \rho_n u_n^* w_n, \tag{11}$$

где для общности введены вещественные веса $\rho_n > 0$, а звездочка означает комплексно сопряженное число. Минимизируя (10) с учетом (11), получим

$$\lambda^{(s+1)} = \overline{\lambda} + \frac{(u^{(s+1)}, u^{(s)})}{(u^{(s+1)}, u^{(s+1)})}.$$

После вычисления $\lambda^{(s+1)}$ следует нормировать вектор $u^{(s+1)}$ на единицу во избежание переполнения, а затем выполнять следующую итерацию.

2. Обратные итерации со сдвигом. Сходимость итераций кардинально улучшается, если в (8), (10) вместо $\overline{\lambda}$ подставить значение $\lambda^{(s)}$, найденное на предыдущей итерации. Для произвольных матриц A сходимость становится квадратичной (как и для метода дополненного вектора), а для эрмитовых матриц — даже кубической [Уилкинсон, 1970].

Однако такая сходимость гарантирована лишь в малой окрестности решения. Скорость сходимости вдали от решения и область сходимости трудно установить теоретически. Необходимо исследование на тестах и сопоставление с конкурентными методами.

Пример 2. Рассматривается задача (1), (5) при $R = \infty$. Для нее используется разностная схема на квазиравномерной сетке с $x_N = \infty$, пригодная на полупрямой:

$$\frac{1}{x_{n+1/2} - x_{n-1/2}}\left(\frac{u_{n-1} - u_n}{x_{n-1/4} - x_{n-3/4}} - \frac{u_n - u_{n+1}}{x_{n+3/4} - x_{n+1/4}}\right) + \\ + 4\left(E + \frac{1}{x_n} - \frac{l(l+1)}{2x_n^2}\right)u_n = 0, \quad (12)$$

$$1 \leqslant n \leqslant N-1; \quad u_0 = u_N = 0, \quad \lambda = -2E.$$

Она дополняется условием нормировки

$$\|u\|_{l_2}^2 = \sum_{n=1}^{N-1} (x_{n+1/2} - x_{n-1/2})\, u_n^2 = 1. \quad (13)$$

Здесь также интересна сходимость итерационного процесса решения сеточных уравнений (12).

Результаты расчетов. Численные расчеты подтвердили, что метод обратных итераций мало чувствителен к нулевому приближению $u_n^{(0)}$. Число итераций было небольшим, если $u_n^{(0)}$ имела ограниченную сеточную норму (13); даже при очень плохом $u_n^{(0)} = \mathrm{const}$, когда эта норма неограничена, сходимость сохранялась.

При выборе $\lambda^{(0)}$ в интервале $(\lambda_{k-1} + \lambda_k) < 2\lambda^{(0)} < (\lambda_k + \lambda_{k+1})$ итерации сходились к λ_k. Чем сильней $\lambda^{(0)}$ отстояло от границ этого интервала, тем меньше было число итераций; но даже вблизи границ оно оставалось небольшим.

Сходимость итераций для теста с $k = 3$, $l = 0$ при $N = 100$ показана на рисунке 1. Видно, что обратные итерации без сдвига сходятся линейно (прямая в полулогарифмическом масштабе), причем предельная точность, определяемая ошибками округления, достигается за 10 итераций. Включение сдвига приводит к кубической сходимости,

и достаточно всего 3 итераций; ошибка округления при этом больше, но незначительно.

Здесь также целесообразно в качестве начального приближения для функции использовать результаты расчета на более грубой сетке.

3. Метод дополненного вектора. Для сравнения на рисунке 1 приведены расчеты по методу дополненного вектора. При том же нулевом приближении первая итерация удаляется от решения, затем итерации сходятся линейно и довольно медленно, и лишь вблизи решения сходимость становится квадратичной. С несколько худшего нулевого приближения итерации сходятся уже к решению с другими квантовыми числами. Сходимость оказалась очень чувствительной к выбору нулевого приближения. Поэтому метод дополненного вектора существенно менее надежен (правда, перенос начального приближения для функции с более грубой сетки существенно улучшает сходимость). Единственное преимущество этого метода в том, что его ошибки округления несколько меньше.

§ 4. Сравнение итерационных методов

1. Сходимость по шагу. Сходимость сеточного решения к точному при $N \to \infty$ доказывается при разумных ограничениях на коэффициенты уравнения. Разностные схемы (4) и (12) имеют аппроксимацию $O(N^{-2})$, и численные расчеты на тестах подтверждают такую точность, в том числе при $R = \infty$.

Иллюстрация приведена на рисунке 3, где для $k = 2$, $l = 1$ показана зависимость от N погрешностей λ и сеточных норм c и l_2 для $\{u_n\}$. В двойном логарифмическом масштабе линии прямые, что указывает

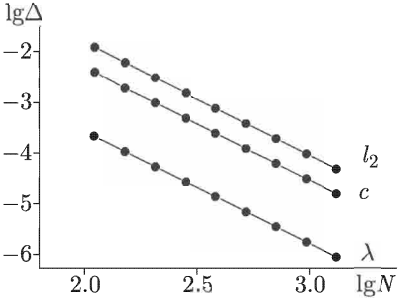

Рис. 3. Погрешность сеточного решения для теста при $k = 2$, $l = 1$, $R = \infty$; λ — для собственного значения, c и l_2 — для соответствующих сеточных норм собственных функций. Точки — расчеты, они соединены линиями

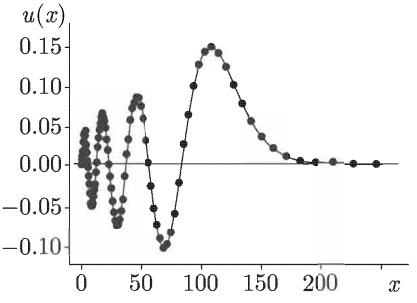

Рис. 4. Собственная функция для $k = 7$, $l = 1$, $R = \infty$; точки — сеточные значения при $N = 100$, кривая — точное решение

на степенную зависимость от N. Их наклон точно соответствует степени 2.

Само решение для $R = \infty$ изображено на рисунке 4; расчет здесь соответствует $N = 100$. Видно хорошее совпадение при небольшом числе точек сетки на каждые полволны решения. Это демонстрирует удачность предложенных разностных схем.

2. Границы сходимости. Границы сходимости предложенных методов определялись варьированием начального приближения для собственных функций и собственных значений. Метод дополненного вектора сходится даже при весьма далеком начальном приближении для функции $u^{(0)}(x) = 0.1$. Но начальное значение энергии в этом случае нужно задать очень точно (погрешность собственного значения не более нескольких процентов от длины интервала между соседними собственными значениями). Если же начальное приближение для функции задано убывающим как $u^{(0)}(x) = 0.1/x$, то границы сходимости для собственных значений существенно расширяются. Однако характер сходимости неудовлетворительный. При четных значениях l границы сходимости для собственных значений, соответствующих нечетным k, очень широкие. Сходимости же к собственному значению $k = 2$ получить не удается, даже если очень точно задать начальное приближение для энергии. К другим четным k сходимость наблюдается в очень узких пределах. Наоборот, при нечетных l сходимость хороша только для собственных значений, соответствующих четным значениям k.

Такой характер сходимости можно объяснить, исследовав отклонение численного решения от точного. На рисунке 1 заметна немонотонность погрешности метода дополненного вектора. При $l = 0$ для четного $k = 2$ первая итерация отбрасывает численное решение гораздо дальше от точного, чем для нечетного $k = 3$. Поэтому метод сходится только для второго случая.

Этот недостаток метода Ньютона хорошо известен для решения функциональных уравнений с одной переменной. Если начальное приближение выбрано с неудачной стороны от корня, то первая итерация

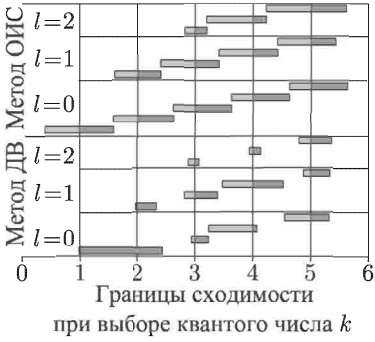

Рис. 5. Сравнение границ сходимости для методов ДВ и ОИС

отбрасывает процесс далеко от корня. В многомерном случае мы имеем похожую картину.

Если задать начальное приближение для функции экспоненциально убывающим $u^{(0)}(x) = 0.1e^{-x}$ (характер убывания соответствует асимптотике точного решения при $l = 0$), то границы сходимости для собственных чисел становятся еще шире, и сходимость наблюдается именно к ближайшему собственному значению. Такой компромисс при выборе начального приближения можно считать оптимальным для данного метода.

Границы сходимости метода обратной итерации со сдвигом шире, чем у метода дополненного вектора. При задании начального приближения для функции экспоненциально убывающим, интервалы сходимости практически полностью покрывают числовую ось. Сравнение границ сходимости для методов дополненного вектора и обратной итерации отражено на рисунке 5.

Глава VI

КЛАССИЧЕСКИЕ УРАВНЕНИЯ В НЕОГРАНИЧЕННОЙ ОБЛАСТИ

В главе предложены численные методы решения задач для классических уравнений математической физики в неограниченной области с использованием квазиравномерных сеток. Многие известные разностные схемы модифицированы для случая неограниченной области. Показано, что такие свойства разностных схем, как устойчивость, в этом случае определяются спектром оператора, аппроксимирующего производные на квазиравномерной сетке. Получены некоторые оценки на границы такого спектра.

§ 1. Параболические уравнения

1. Уравнение теплопроводности для прямой и полупрямой. Рассмотрим одномерное уравнение теплопроводности на прямой

$$u_t = u_{xx} + f(x,t), \quad x \in (-\infty, +\infty), \quad t \in [0,T], \qquad u(x,0) = f_0(x). \tag{1}$$

Следует также задать краевые условия при $x \to \pm\infty$.

Введем квазиравномерную сетку, покрывающую всю прямую, например тангенциальную (III. 27)

$$x_n = c \cdot \operatorname{tg}\left(\frac{\pi n}{2N}\right), \quad -N \leqslant n \leqslant N. \tag{2}$$

Применим метод прямых. В качестве неизвестного рассмотрим вектор значений функции $u(x,t)$ во всех узлах сетки $u_n t \equiv u(x_n,t)$. Аппроксимируем во всех внутренних узлах сетки вторую пространственную производную со вторым порядком, используя симметричную разделенную разность с дробными узлами:

$$\frac{du_n}{dt} = \frac{0.5}{x_{n+1/2} - x_{n-1/2}}\left(\frac{u_{n+1} - u_n}{x_{n+3/4} - x_{n+1/4}} - \frac{u_n - u_{n-1}}{x_{n-1/4} - x_{n-3/4}}\right) + f_n, \qquad -N+1 \leqslant n \leqslant N-1.$$

Для решения полученной системы обыкновенных дифференциальных уравнений вида

$$\frac{du}{dt} = G(u)$$

используем одностадийное семейство схем Розенброка [Rosenbrock, 1963]:

$$\widehat{u}_n = u_n + \tau \mathrm{Re} k_n, \quad (E - \alpha\tau G_u)k = G(u_n), \qquad -N+1 \leqslant n \leqslant N-1; \tag{3}$$

здесь $G_u \equiv \partial G/\partial u$ — матрица Якоби, E — единичная матрица, τ — шаг по времени, u_n — решение на текущем временной слое, $\widehat{u}_n$ — решение на новом временном слое, α — численный параметр, определяющий семейство одностадийных схем Розенброка. Таким образом, вектор $\mathbf{k} = \{k_n\}$ определяется из решения системы линейных уравнений с матрицей $E - \alpha\tau G_u$.

Неизвестными в левой части системы (3) являются значения k_n, $-N+1 \leqslant n \leqslant N-1$; соответственно, порядок линейной системы равен $2N-1$, а в матрицу Якоби G_u входят только производные по u_n с указанными значениями индекса n. Но в правую часть линейной системы (3) входят так же значения u_{-N} и u_N; они берутся из краевых значений задачи (1), т. е. на ∞. Возможность такой постановки была объяснена в главе III.

Решение линейной системы (3) следует выполнять прямыми методами (например, методом Гаусса с выбором главного элемента). При этом переход с исходного слоя на новый производится за конечное, заранее известное число операций, как в явных схемах. В то же время нахождение $\mathbf{k} = \{k_n\}$ из решения линейной системы вносит в схему (3) неявность. Поэтому такие схемы называют явно-неявными.

Необходимо отметить, что для линейных задач матрица G_u остается одной и той же на всех временных слоях. Поэтому нет необходимости решать линейную систему (3) на каждом временном слое; достаточно единственный раз найти обратную матрицу $(E - \alpha\tau G_u)^{-1}$. Такой подход существенно повышает экономичность предложенного численного метода.

При разных параметрах α схема (3) обладает разными свойствами. При $\alpha = 0$ это явная схема; она имеет погрешность $O(\tau)$ и условно устойчива при $2\tau \leqslant h^2$ (h — наименьший шаг пространственной сетки). Этот вариант схемы практически непригоден для расчета жестких задач, в том числе задач теплопроводности.

При $\alpha = 0.5$ для уравнения теплопроводности (1) получается известная схема “с полусуммой”. Она имеет точность $O(\tau^2)$ и безусловно устойчива; поэтому ее часто используют в расчетах. Однако по классификации жестких задач (см., например, [Хайрер, Ваннер, 1999]) она лишь A-устойчива. Поэтому в расчетах задач теплопроводности при больших градиентах решения могут возникнуть неприятности (например, нефизическое нарушение монотонности).

При $\alpha = 1$ получается так называемая обратная схема Эйлера (чисто неявная). Она L1-устойчива, что обеспечивает безусловную устойчивость и хорошее качественное поведение численного решения.

Однако она имеет невысокую точность $O(\tau)$, что препятствует ее применению.

Описанные выше схемы вещественны. Однако существует одна комплексная схема этого семейства с $\alpha = (1+i)/2$ [Розенброк, 1963]. Она обладает уникальными свойствами: точность $O(\tau^2)$, L2-устойчивость и, соответственно, безусловная устойчивость. Эта схема обладает высокой надежностью и пригодна для расчета задач с сильной жесткостью; в частности, она позволяет хорошо рассчитывать теплопроводность даже при больших градиентах решения. Именно эта схема использована в описанных ниже расчетах.

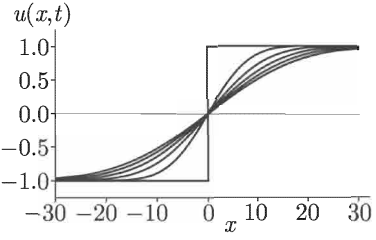

Рис. 1. Выравнивание температуры на границе двух сред; кривые соответствуют моментам времени $t = 0,\ 20,\ 40,\ 60,\ 80,\ 100$

В качестве иллюстрации возможностей метода приведем расчет задачи для уравнения теплопроводности с разрывным начальным условием

$$f_0(x) = \begin{cases} 1, & x \geqslant 0, \\ -1, & x < 0. \end{cases}$$

Такая постановка задачи моделирует выравнивание температуры на границе двух сред с разными температурами, занимающими неограниченное полупространство. Разумной постановкой является краевое условие $u(\pm\infty, t) = f_0(\pm\infty) = \pm 1$, согласованное с начальными данными. Результаты расчета для различных моментов времени приведены на рисунке 1.

Для гладкого начального профиля температуры $f_0(x) = e^{-x^2}$ известно (см., например, [Боголюбов, Кравцов, 1998]) точное решение

$$u(x,t) = \frac{\exp\{-x^2/(1+4t)\}}{\sqrt{1+4t}}.$$

Оно было использовано для тестирования работы программы. Результаты теста на сгущающихся сетках приведены на рисунке 2. Дифференциальная задача (1) была аппроксимирована нами с точностью $O(\tau^2 + N^{-2})$, поэтому сгущение временной и пространственной сеток надо проводить в одно и тоже число раз r. Наиболее удобно выбрать $r = 2$, т. к. в этом случае все узлы более редкой квазиравномерной пространственной сетки совпадают с четными узлами сетки с удвоенным числом узлов, и сравнивать между собой численные решения в этих точках особенно просто.

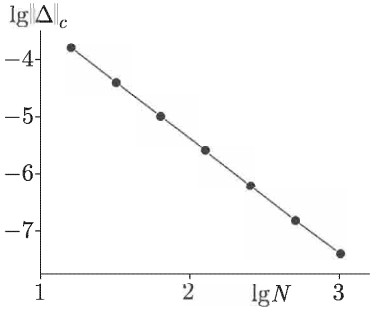

Рис. 2. Тест на сгущающихся сетках для уравнения теплопроводности на прямой

Убывание сеточной нормы C погрешности численного решения Δ с ростом числа узлов сетки в двойном логарифмическом масштабе — прямая линия, тангенс угла наклона которой $\operatorname{tg}\alpha = -1.988$ подтверждает порядок точности метода $O(N^{-2} + \tau^2)$. Сравнение проводилось в момент времени $t = 10.0$. Шаг по времени $\tau = N^{-1}$. Профили решения для моментов времени $t = 0.0, 1.0, 2.0, 5.0, 7.0, 10.0$ изображены на рисунке 3.

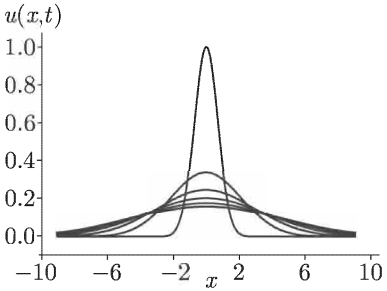

Рис. 3. Профили температуры в моменты времени $t = 0.0, 1.0, 2.0, 5.0, 7.0, 10.0$

Задачи на полупрямой. Предложенный метод позволяет также моделировать распространение краевого режима. Рассмотрим уравнение теплопроводности на полупрямой:

$$u_t = u_{xx} + f(x,t), \quad x \in [0,\infty), \quad t \in [0,T],$$
$$u(x,0) = f_0(x), \quad u(0,t) = \mu t. \tag{4}$$

Здесь также следует добавить краевое условие при $x \to +\infty$.

Построим квазиравномерную сетку, покрывающую полупрямую, например

$$x_n = c \cdot \operatorname{tg}\left(\frac{\pi n}{2N}\right), \quad 0 \leqslant n \leqslant N.$$

Модельная задача об остывании полубесконечного стержня $f_0(x) = U_0$ и $\mu t = 0$ имеет точное решение (см., например, [Боголюбов, Кравцов, 1998]), выражаемое через функцию ошибок:

$$u(x,t) = U_0 \Phi\left(\frac{x}{2\sqrt{t}}\right),$$
$$\Phi(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-\xi^2} d\xi.$$

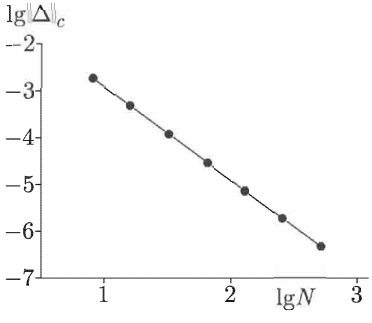

Рис. 4. Результаты теста на сгущающихся сетках для уравнения теплопроводности на полупрямой

Это точное решение было использовано для тестирования метода на сгущающихся сетках. Разумной постановкой является краевое условие $u(+\infty,t) = f_0(+\infty) = U_0$, согласованное с начальными данными.

Результаты теста показаны на рисунке 4. Убывание максимума модуля погрешности с ростом числа узлов сетки в двойном логарифмическом масштабе подтверждает порядок точности метода $O(\tau^2 + N^{-2})$.

2. Уравнение нелинейной теплопроводности и горения. Рассмотрим физический процесс инициирования горения, описываемый следующим нелинейным уравнением теплопроводности:

$$u_t = \left(u^2 u_x\right)_x + u^\beta; \quad x \in (-\infty, +\infty), \quad t \in [0, T], \qquad u(\pm\infty, t) = 0, \quad u(x, 0) = f_0(x). \tag{5}$$

В (5) коэффициент теплопроводности зависит от температуры по нелинейному закону $k \sim u^2$, а второе слагаемое в правой части описывает процесс выделения энергии при горении сплошной среды. Интенсивность горения также нелинейно зависит от температуры $\sim u^\beta$.

Одним из главных результатов конкуренции этих двух нелинейных процессов теплопередачи и тепловыделения является эффект локализации процесса горения. Свойства решения существенно различаются в случаях $\beta = 3$, $\beta > 3$, $\beta < 3$.

С точки зрения аппроксимации пространственной производной удобно уравнение записать в виде

$$u_t = \frac{1}{3}(u^3)_{xx} + u^\beta. \tag{6}$$

Аппроксимируем (6) в узлах квазиравномерной сетки (2), покрывающей всю область, где ищется решение $x \in (-\infty, +\infty)$. Применяя метод прямых, получим дифференциально-разностную систему уравнений:

$$\frac{du_n}{dt} = \frac{0.5}{x_{n+1/2} - x_{n-1/2}} \left(\frac{u_{n+1}^3 - u_n^3}{x_{n+3/4} - x_{n+1/4}} - \frac{u_n^3 - u_{n-1}^3}{x_{n-1/4} - x_{n-3/4}} \right) + u_n^\beta, \qquad -N + 1 \leqslant n \leqslant N - 1.$$

Значения u_{-N} и u_N берутся из граничных условий.

Полученную систему численно интегрируем с применением схемы Розенброка из семейства (3) с комплексным параметром $\alpha = (1 + i)/2$.

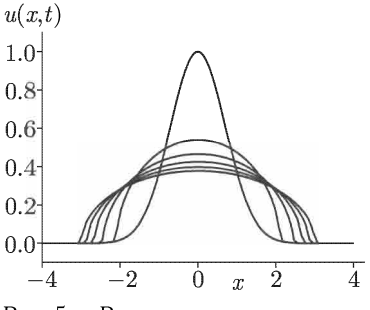

Рис. 5. Распространение тепловых волн. Линии на графике соответствуют моментам времени $t = 0, 3, 6, 9, 12, 15$

Построенный метод позволяет исследовать различные режимы горения.

Частный случай этого уравнения

$$u_t = (u^2 u_x)_x, \tag{7}$$

соответствующий распространению тепла при отсутствии горения, рассмотрен в [Самарский и др., 1987], где построено частное автомодельное решение для мгновенного точечного источника, соответствующее режиму тепловых волн: тепловое возмущение распространяется с конечной скоростью. В этом принципиальное отличие механизма нелинейной теплопро-

водности от линейного механизма передачи тепла (1). Результаты расчетов для уравнения (7), приведенные на рисунке 5, показывают, что тепловые волны распространяются не только от точечного источника. Расчет проведен для начального профиля температуры $f_0(x) = e^{-x^2}$.

Рассмотрим случай $\beta = 3$, что соответствует наличию в среде источника тепловой энергии:

$$u_t = \left(u^2 u_x\right)_x + u^3; \quad x \in (-\infty, +\infty), \quad t \in [0, T],$$
$$u(\pm\infty, t) = 0, \quad u(x, 0) = f_0(x).$$

Его автомодельное решение приведено, например, в курсе лекций [Свешников и др., 1993]:

$$u(x,t) = \frac{1}{\sqrt{T_0 - t}} \begin{cases} \dfrac{\sqrt{3}}{2} \cos\left(\dfrac{x}{\sqrt{3}}\right), & |x| < \dfrac{\pi\sqrt{3}}{2}, \\ 0, & |x| \geqslant \dfrac{\pi\sqrt{3}}{2}. \end{cases} \tag{8}$$

Это решение локализовано в области $|x| < \pi\sqrt{3}/2$, во всех точках которой температура неограниченно возрастает. Тепловая структура называется S-режимом с обострением и представляет собой стоячую температурную волну.

Результаты расчетов приведены на рисунке 6. Начальный профиль

$$f_0(x) = \begin{cases} \dfrac{\sqrt{3}}{2} \cos\left(\dfrac{x}{\sqrt{3}}\right), & |x| \leqslant \dfrac{\pi\sqrt{3}}{2}, \\ 0, & |x| > \dfrac{\pi\sqrt{3}}{2} \end{cases} \tag{9}$$

соответствует времени существования тепловой структуры $T_0 = 1.0$. Гладкость начального профиля нарушается в точках $x = \pm\pi\sqrt{3}/2$. Для сохранения общей точности расчета $O(\tau^2 + N^{-2})$ надо строить пространственную сетку таким образом, чтобы один из узлов сетки всегда попадал в точки $x = \pm\pi\sqrt{3}/2$; это можно сделать, например, выбрав в (2) масштабный коэффициент $c = \pi\sqrt{3}/2$. При $t \to T_0$ температура в центре тепловой структуры (8) стремится к бесконечности, и решение разрушается. Расчет проводился на сетке, покрывающей всю прямую $x \in (-\infty, +\infty)$. На рисунке 6 хорошо видно, что тепловая структура действительно не распространяется за пределы отрезка $x \in [-\sqrt{3}\,\pi/2, \sqrt{3}\,\pi/2]$.

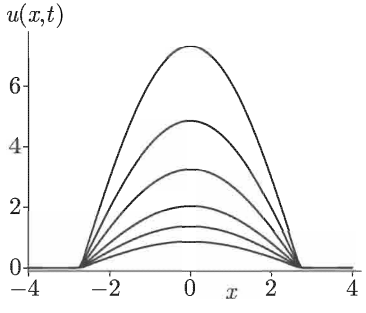

Рис. 6. Локализация процесса горения; S-режим с обострением. Линии соответствуют моментам времени $t = 0.0,\ 0.6,\ 0.82,\ 0.93,\ 0.97,\ 0.99$

Рассмотрим задачу (5) для случая $\beta = 2$. При этом тепловая структура также развивается в режиме с обострением, фронты тепловой структуры расширяются со все увеличивающейся скоростью и в момент обострения тепловая структура охватывает всю прямую, нагревая ее всюду до бесконечной температуры. Такой процесс горения называется HS-режимом. На рисунке 7 приведены результаты расчетов для HS-режима горения. Начальный профиль выбран финитным (9), таким же как и в автомодельном решении (8) для случая $\beta = 2$, но решение уже не локализуется в пространстве. Очевидно, для этого случая краевые условия надо ставить на ∞: $u(\pm\infty, t) = 0$.

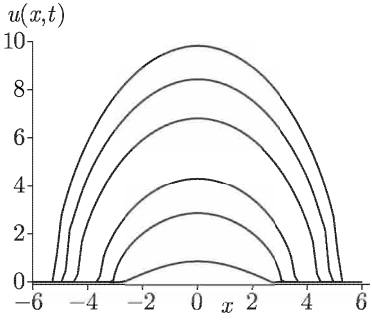

Рис. 7. Режим HS-горения. Линии соответствуют моментам времени $t = 0.0,\ 1.2,\ 1.4,\ 1.55,\ 1.6,\ 1.63$

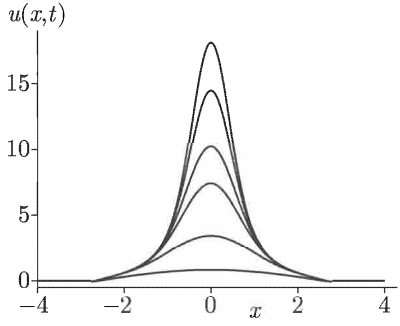

Рис. 8. LS-режим горения. Линии соответствуют моментам времени $t = 0.0,\ 0.78,\ 0.79,\ 0.7907,\ 0.791,\ 0.7911$

При $\beta = 4$ мощность источника энерговыделения при больших температурах выше, чем в режиме S-горения. Поэтому локализация проявляется сильнее. Вся энергия локализуется во все сужающейся окрестности максимума температуры. Такой процесс горения называется LS-режимом.

На рисунке 8 приведены профили температуры в LS-режиме горения для различных моментов времени. Начальный профиль температуры выбран финитным (9). В этом случае локализация понимается в эффективном смысле. В момент обострения температура неограниченно возрастает в малой (много меньше ширины начального профиля) окрестности максимума.

3. Двумерное параболическое уравнение; продольно-поперечная схема. Рассмотрим в области $-\infty < x < \infty$, $-\infty < y < \infty$ начально-краевую задачу для уравнения параболического типа

$$\frac{\partial u}{\partial t} = \Delta_2 u - u + f(x, y, t), \quad x \in (-\infty, +\infty), \quad t \in [0, T], \qquad u(x, y, 0) = f_0(x, y). \tag{10}$$

Следует также задать граничные значения $u(x, y, t)$ при $x^2 + y^2 \to \infty$. Пусть для определенности

$$\lim_{x^2+y^2\to\infty} u(x, y, t) = 0. \tag{11}$$

Для построения численного решения задачи (10) используем двумерную квазиравномерную сетку по пространственным переменным, например тангенциальную

$$\begin{aligned} x_n &= c_x \operatorname{tg}\left(\frac{\pi n}{2N}\right), \quad -N \leqslant n \leqslant N, \\ y_m &= c_y \operatorname{tg}\left(\frac{\pi m}{2M}\right), \quad -M \leqslant m \leqslant M. \end{aligned} \tag{12}$$

Масштабные коэффициенты сетки c_x, c_y и число узлов по каждой пространственной переменной $2N + 1$, $2M + 1$ могут быть различны.

Сначала для простоты выберем равномерную временную сетку $t_k = k\tau$. На нулевом временном слое значения сеточной функции известны из начального условия.

Для получения решения сеточного уравнения применим одну из лучших двумерных экономичных схем — продольно-поперечную схему [Peaceman, Rachford, 1955] точности $O(\tau^2)$:

$$\frac{1}{0.5\tau}(u_{nm}^{k+0.5} - u_{nm}^{k}) = \Lambda_{xx}[u]_{nm}^{k+0.5} + \Lambda_{yy}[u]_{nm}^{k} - u_{nm}^{k+0.5} + f_{nm}^{k+0.5},$$

$$\frac{1}{0.5\tau}(u_{nm}^{k+1} - u_{nm}^{k+0.5}) = \Lambda_{xx}[u]_{nm}^{k+0.5} + \Lambda_{yy}[u]_{nm}^{k+1} - u_{nm}^{k+0.5} + f_{nm}^{k+0.5}.$$

Здесь использована аппроксимация оператора Лапласа на квазиравномерной сетке, пригодная в неограниченной области:

$$\begin{aligned} \Delta_2 &\approx \Lambda_{xx} + \Lambda_{yy}, \\ \Lambda_{xx}[u]_{nm} &= \frac{0.5}{x_{n+1/2} - x_{n-1/2}}\left(\frac{u_{n+1,m} - u_{nm}}{x_{n+3/4} - x_{n+1/4}} - \frac{u_{nm} - u_{n-1,m}}{x_{n-1/4} - x_{n-3/4}}\right), \\ \Lambda_{yy}[u]_{nm} &= \frac{0.5}{y_{m+1/2} - y_{m-1/2}}\left(\frac{u_{n,m+1} - u_{nm}}{y_{m+3/4} - y_{m+1/4}} - \frac{u_{nm} - u_{n,m-1}}{y_{m-1/4} - y_{m-3/4}}\right). \end{aligned} \tag{13}$$

В [Самарский, 1977] поперечная схема в ограниченной области безусловно устойчива. Ниже (в п. 5) будет доказано, что и в случае неограниченной области продольно-поперечная схема безусловно устойчива.

Для тестирования данного метода в неограниченной области был проведен расчет на сгущающихся сетках. Численное решение задачи (10) для различных значений $N = M = 8$, 16, 32, 64, 128 сравнивалось с точным. Когда неоднородность в правой части уравнения (10) $f = r^q e^{-0.5r^2}(A \sin q\varphi + B \cos q\varphi)$ и начальное условие $f_0(x, y) = 0$, точ-

ное решение задачи получается при помощи интегрального преобразования Ханкеля:

$$u = (A\sin q\varphi + B\cos q\varphi)\int\limits_0^{+\infty} \frac{e^{-s^2/2}}{s^2+1}\left(1 - e^{-(s^2+1)t}\right) J_q(sr)s^{q+1}ds. \quad (14)$$

Здесь $r = (x^2+y^2)^{1/2}$ и $\varphi = \operatorname{arctg}(y/x)$ — полярные координаты на плоскости, q — неотрицательное целое число.

Мы выбирали $N = M$ и $\tau = N^{-1}$. Сгущая сетки, полученное численное решение сравнивали с точным, подсчитанным для момента времени $t = 0.25$, в сеточных нормах c и l_2.

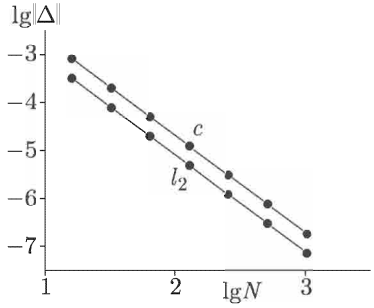

Рис. 9. Убывание погрешности численного решения двумерной начально-краевой задачи для уравнения параболического типа с ростом числа узлов квазиравномерной сетки подтверждает порядок точности метода $O(N^{-2}+M^{-2}+\tau^2)$

Убывание погрешности в обеих нормах с увеличением числа узлов полностью соответствует порядку точности метода $O(N^{-2}+M^{-2}+$ $+\tau^2)$. График этой зависимости в двойном логарифмическом масштабе — прямая с коэффициентом наклона -2 (см. рис. 9).

4. Устойчивость продольно-поперечной схемы. Для доказательства безусловной устойчивости продольно-поперечной схемы на произвольной пространственной сетке потребуется следующая лемма.

Лемма. Пусть $b_n > 0$, тогда трехдиагональная симметричная матрица D, т. ч.

$$d_{nn} = b_n + b_{n+1}, \quad d_{n-1,n} = d_{n,n-1} = -b_n, \quad 1 \leqslant n \leqslant N, \quad (15)$$

положительно определена.

Доказательство. Рассмотрим квадратичную форму

$$(x, Dx) = \sum_{n=1}^{N}(b_n + b_{n+1})x_n^2 - \sum_{n=2}^{N} b_n x_{n-1}x_n - \sum_{n=1}^{N-1} b_{n+1}x_n x_{n+1}.$$

Сдвигая на 1 индекс суммирования в последней сумме и во втором члене первой суммы, получим

$$(x, Dx) = \sum_{n=1}^{N} b_n x_n^2 + \sum_{n=2}^{N+1} b_n x_{n-1}^2 - 2\sum_{n=2}^{N} b_n x_{n-1} x_n =$$

$$= b_1 x_1^2 + b_{N+1} x_N^2 + \sum_{n=2}^{N} b_n (x_n - x_{n-1})^2 > 0$$

при любых $x \neq 0$. Лемма доказана.

Следствие. Все собственные значения матрицы D положительны, а ее собственные вектора образуют ортогональный базис.

Исследуем устойчивость продольно-поперечной схемы на квазиравномерной пространственной сетке (12) применительно к уравнению

$$w_t = w_{xx} + w_{yy} - f(x, y, t). \tag{16}$$

Погрешность решения $\delta w = \xi$ удовлетворяет однородному разностному уравнению

$$\frac{\xi^{k+1/2} - \xi^k}{0.5\tau} = \Lambda_{xx}[\xi^{k+1/2}] + \Lambda_{yy}[\xi^k],$$
$$\frac{\xi^{k+1} - \xi^{k+1/2}}{0.5\tau} = \Lambda_{xx}[\xi^{k+1/2}] + \Lambda_{yy}[\xi^{k+1}].$$

Исследуем свойства операторов Λ_{xx} и Λ_{yy}, которые аппроксимируют вторые пространственные производные. Из (13) следует, что действие этих операторов равносильно умножению сеточной функции ξ на следующие матрицы:

$$\Lambda_{xx} = XA\xi, \quad \Lambda_{yy} = \xi BY. \tag{17}$$

Здесь матрицы A, B являются симметричными трехдиагональными:

$$A_{n-1,n} = \frac{1}{(x_{n-1/4} - x_{n-3/4})}, \quad A_{n+1,n} = \frac{1}{(x_{n+3/4} - x_{n+1/4})},$$
$$A_{nn} = -A_{n-1,n} - A_{n+1,n}, \quad -N+1 \leqslant n \leqslant N-1, \tag{18}$$

$$B_{m,m-1} = \frac{1}{(y_{m-1/4} - y_{m-3/4})}, \quad B_{m,m+1} = \frac{1}{(y_{m+3/4} - y_{m+1/4})},$$
$$B_{mm} = -B_{m,m-1} - B_{m,m+1}, \quad -M+1 \leqslant m \leqslant M-1, \tag{19}$$

а матрицы X, Y — диагональными, т. ч.

$$X_{nn} = \frac{0.5}{(x_{n+1/2} - x_{n-1/2})}, \quad Y_{mm} = \frac{0.5}{(y_{m+1/2} - y_{m-1/2})},$$
$$-N+1 \leqslant n \leqslant N-1, \quad -M+1 \leqslant m \leqslant M-1. \tag{20}$$

Заметим, что матрицы $-A$ и $-B$ имеют вид (15) и в силу леммы являются положительно определенными, значит, для произвольного столбца p квадратичные формы

$$-p^{\mathrm{T}}Ap > 0 \quad \text{и} \quad -p^{\mathrm{T}}Bp > 0. \tag{21}$$

Кроме того, все собственные числа матриц $-A$ и $-B$ положительны. Диагональные матрицы X и Y содержат только положительные элементы, следовательно, определены $X^{1/2}$ и $Y^{1/2}$.

С учетом (17) продольно-поперечная схема имеет вид:

$$\frac{\xi^{k+1/2}-\xi^k}{0.5\tau} = XA\xi^{k+1/2} + \xi^k BY,$$

$$\frac{\xi^{k+1}-\xi^{k+1/2}}{0.5\tau} = XA\xi^{k+1/2} + \xi^{k+1} BY.$$

Помножив эти матричные уравнения слева на $X^{-1/2}$ и справа на $Y^{-1/2}$, преобразуем их к следующему виду:

$$\frac{X^{-1/2}\xi^{k+1/2}Y^{-1/2} - X^{-1/2}\xi^k Y^{-1/2}}{0.5\tau} =$$

$$= X^{1/2}AX^{1/2}X^{-1/2}\xi^{k+1/2}Y^{-1/2} + X^{-1/2}\xi^k Y^{-1/2}Y^{1/2}BY^{1/2},$$

$$\frac{X^{-1/2}\xi^{k+1}Y^{-1/2} - X^{-1/2}\xi^{k+1/2}Y^{-1/2}}{0.5\tau} =$$

$$= X^{1/2}AX^{1/2}X^{-1/2}\xi^{k+1/2}Y^{-1/2} + X^{-1/2}\xi^{k+1}Y^{-1/2}Y^{1/2}BY^{1/2}.$$

Вводя обозначения матриц $\theta^k = X^{-1/2}\xi^k Y^{-1/2}$, $P_x = X^{1/2}AX^{1/2}$ и $P_y = Y^{1/2}BY^{1/2}$, получаем

$$\frac{\theta^{k+1/2}-\theta^k}{0.5\tau} = P_x\theta^{k+1/2} + P_y\theta^k, \tag{22}$$

$$\frac{\theta^{k+1}-\theta^{k+1/2}}{0.5\tau} = P_x\theta^{k+1/2} + P_y\theta^{k+1}. \tag{23}$$

Матрица $-P_x = -X^{1/2}AX^{1/2}$ симметрична и положительно определена. Действительно, $P_x^{\mathrm{T}} = (X^{1/2}AX^{1/2})^{\mathrm{T}} = X^{1/2^{\mathrm{T}}}A^{\mathrm{T}}X^{1/2^{\mathrm{T}}} = X^{1/2}AX^{1/2} = P_x$. Для произвольного столбца p квадратичная форма

$$-p^{\mathrm{T}}P_x p = -p^{\mathrm{T}}X^{1/2}AX^{1/2}p = -(X^{1/2}p)^{\mathrm{T}}AX^{1/2}p > 0$$

оказалась положительной (последнее неравенство справедливо в силу (21)). Это доказывает положительную определенность матрицы $-P_x$.

Аналогично $-P_y = -Y^{1/2}BY^{1/2}$ также симметрична и положительно определена.

Значит, во-первых, все собственные значения матриц $-P_x$ и $-P_y$ положительны, во-вторых, их собственные векторы образуют ортогональный базис.

Обозначим собственные векторы и собственные значения матриц $-P_x$ и $-P_y$ как

$$-P_x a^{(q)} = \alpha_q a^{(q)} \quad \text{и} \quad -b^{(l)\mathrm{T}} P_y = \beta_l b^{(l)\mathrm{T}}. \tag{24}$$

Ортогональность собственных векторов (24) понимается в смысле скалярного произведения, а именно

$$(a^{(q)}, a^{(l)}) = \sum_n a_n^{(q)} a_n^{(l)} = 0,$$

$$(b^{(q)}, b^{(l)}) = \sum_m b_m^{(q)} b_m^{(l)} = 0$$

при $q \neq l$.

Матрицы вида $a^{(q)} b^{(l)\mathrm{T}}$ образуют удобный базис, ортогональный в смысле скалярного произведения $(U, V) = \sum\limits_m \sum\limits_n u_{nm} v_{nm}$ для сеточных функций $\theta^k = X^{-1/2} \xi^k Y^{-1/2}$, т. ч.

$$(a^{(q)} b^{(l)\mathrm{T}}, a^{(j)} b^{(p)\mathrm{T}}) = \sum_m b_m^{(l)} b_m^{(p)} \sum_n a_n^{(q)} a_n^{(j)} = 0, \quad q \neq j, \quad l \neq p.$$

Всякая матрица θ^k может быть представлена в виде ее разложения по указанному базису:

$$\theta^k = \sum_q \sum_l \theta_{ql}^k a^{(q)} b^{(l)\mathrm{T}}, \quad \xi^k = \sum_q \sum_l \theta_{ql}^k X^{1/2} a^{(q)} b^{(l)\mathrm{T}} Y^{1/2}. \tag{25}$$

Рассмотрим сеточную функцию частного вида $\theta^k = a^{(q)} \cdot b^{(l)\mathrm{T}}$, где $a^{(q)} \cdot b^{(l)\mathrm{T}}$ — произвольные собственные векторы матриц $-P_x$, $-P_y$. Используя (22) и (23), получим

$$\theta_{ql}^{k+1} = X^{-1/2} \xi_{ql}^{k+1} Y^{-1/2} = \frac{1 - 0.5\tau\alpha_q}{1 + 0.5\tau\alpha_q} \cdot \frac{1 - 0.5\tau\beta_l}{1 + 0.5\tau\beta_l} \theta_{ql}^k =$$

$$= \frac{1 - 0.5\tau\alpha_q}{1 + 0.5\tau\alpha_q} \cdot \frac{1 - 0.5\tau\beta_l}{1 + 0.5\tau\beta_l} X^{-1/2} \xi_{ql}^k Y^{-1/2}.$$

Тогда для погрешности численного решения справедлива следующая формула для перехода на следующий временной слой по продольно-поперечной схеме:

$$\xi_{ql}^{k+1} = \frac{1 - 0.5\tau\alpha_q}{1 + 0.5\tau\alpha_q} \cdot \frac{1 - 0.5\tau\beta_l}{1 + 0.5\tau\beta_l} \xi_{ql}^k. \tag{26}$$

Вычисляя норму l_2 сеточных функций

$$\|\xi\|_{l_2} = \sum_{n=-N+1}^{N-1} \sum_{m=-M+1}^{M-1} \xi_{nm}(x_{n+1/2} - x_{n-1/2})(y_{m+1/2} - y_{m-1/2})$$

в равенстве (26), получим

$$\left\|\xi_{ql}^{k+1}\right\|_{l_2} = \left|\rho_{ql}\right| \left\|\xi_{ql}^{k}\right\|_{l_2},$$

где

$$\rho_{ql} = \frac{1 - 0.5\tau\alpha_q}{1 + 0.5\tau\alpha_q} \cdot \frac{1 - 0.5\tau\beta_l}{1 + 0.5\tau\beta_l}$$

есть множитель роста одной компоненты разложения погрешности по собственному базису матриц $-P_x$, $-P_y$. В силу положительности собственных значений $\alpha_q > 0$ и $\beta_l > 0$ выполняется неравенство $|\rho_{ql}| \leqslant$ $\leqslant 1$ при любом соотношении пространственных и временных шагов. Отсюда в силу ортогональности базиса $a^{(q)}b^{(l)\,\mathrm{T}}$ следует, что в общем случае $\left\|\xi^{k+1}\right\|_{l_2} \leqslant \max\limits_{ql} \left|\rho_{ql}\right| \left\|\xi^{k}\right\|_{l_2}$. Откуда следует равномерная по N и M оценка для нормы погрешности $\left\|\xi\right\|_{l_2}$ на любом временной слое $\left\|\xi^{k}\right\|_{l_2} \leqslant \left\|\xi^{0}\right\|_{l_2}$.

Замечание 1. Если применить продольно-поперечную схему к решению параболического уравнения вида (10), то в доказательстве безусловной устойчивости надо лишь заменить $-P_x = -X^{1/2}AX^{1/2}$ на $-P_x = -X^{1/2}AX^{1/2} - E$, где E — единичная матрица. Это приведет к смещению границ спектра матрицы $-P_x$ на 1 в сторону положительных значений и не нарушит условие $|\rho| \leqslant 1$.

Замечание 2. Исследование устойчивости продольно-поперечной схемы в случае симметричных операторов можно найти, например, в [Самарский, Вабищевич, 2003]. Отметим особенности рассмотренной выше задачи. Во-первых, операторы не симметричны, во-вторых, из равномерной ограниченности энергетической нормы, связанной с сеточным оператором Лапласа, не следует равномерная ограниченность нормы l_2, т. к. в неограниченной области не выполняется неравенство Фридрихса. Решающим обстоятельством, позволяющим доказать безусловную устойчивость рассмотренной схемы, является то, что операторы имеют общий собственный базис.

5. Оценка спектра. Для доказательства устойчивости продольно-поперечной схемы в неограниченной области мы воспользовались тем, что все собственные значения матрицы, аппроксимирующей вторую производную на квазиравномерной сетке, отрицательны. Однако для получения значения оптимального шага по времени, количества итераций, необходимых для выхода на заданную точность, и т. п. необходимо знать границы спектра. Точно аналитически найти эти границы не

представляется возможным, однако удается получить некоторые оценки.

Так как спектры операторов $\Lambda_{xx} = XA$ и $X^{1/2}AX^{1/2}$ совпадают и матрица $X^{1/2}AX^{1/2}$ симметрична и трехдиагональна, то оценку $\max\limits_q |\alpha_q|$ удобнее получать именно для этой матрицы, она имеет следующую структуру:

$$X^{1/2}AX^{1/2} = \begin{pmatrix} d_1 & b_2 & 0 & 0 & 0 \\ b_2 & d_2 & b_3 & 0 & 0 \\ 0 & b_3 & d_3 & b_4 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \dots & 0 & b_{2N-2} & d_{2N-2} & b_{2N-1} \\ \dots & 0 & 0 & b_{2N-1} & d_{2N-1} \end{pmatrix},$$

где

$$b_{n+N+1} = \frac{0.5}{\sqrt{x_{n+1/2} - x_{n-1/2}}\sqrt{x_{n+3/2} - x_{n+1/2}}} \frac{1}{x_{n+3/4} - x_{n+1/4}},$$

$$d_{n+N} = -\frac{0.5}{x_{n+1/2} - x_{n-1/2}} \left(\frac{1}{x_{n+3/4} - x_{n+1/4}} + \frac{1}{x_{n-1/4} - x_{n-3/4}} \right),$$

$$-N + 1 \leqslant n \leqslant N - 1.$$

Для симметричной трехдиагональной матрицы может быть найден интервал, внутри которого лежат все собственные значения [Годунов, 2002]: $\alpha \leqslant \alpha_q \leqslant \alpha_{\max}$, где

$$\alpha = \min\left\{ d_1 - |b_2|, \min_{2 \leqslant k \leqslant 2N-2} (d_k - |b_k| - |b_{k+1}|), d_{2N-1} - |b_{2N-1}| \right\},$$

$$\alpha_{\max} = \max\left\{ d_1 + |b_2|, \max_{2 \leqslant k \leqslant 2N-2} (d_k + |b_k| + |b_{k+1}|), d_{2N-1} + |b_{2N-1}| \right\}.$$

Преобразуя явные выражения для элементов матрицы $X^{1/2}AX^{1/2}$, применяя разложение в ряд Тейлора с учетом членов до порядка N^{-2} включительно, получаем

$$\alpha = \min_{-N+1 \leqslant n \leqslant N-1} \left(-\frac{N}{x'_n} \left(\frac{N}{x'_{n+1/2}} + \frac{N}{x'_{n-1/2}} \right) - \right.$$

$$\left. - \frac{N}{x'_{n+1/2}\sqrt{x'_n x'_{n+1}}} - \frac{N}{x'_{n+1/2}\sqrt{x'_n x'_{n-1}}} \right) =$$

$$= -\max_{-N+1 \leqslant n \leqslant N-1} \frac{N^2}{[x'(\eta_n)]^2} \left\{ 4 - \frac{x'''(\eta_n)}{N^2 x'(\eta_n)} + \frac{9}{4} \left(\frac{x''(\eta_n)}{N x'(\eta_n)} \right)^2 \right\}. \quad (27)$$

Максимум в выражении (27) достигается там, где шаг сетки минимален. В случае использования тангенциальной сетки (2) максимум

реализуется при $n = 0$, т. к. минимальный шаг сетки соответствует точке перегиба функции $x(\eta)$; поэтому $x''(\eta_0) = x''(0) = 0$. При этом в случае сгущения сетки к точке $x = 0$ преобразование $x(\eta)$ строго монотонно и выпукло вниз при $x > 0$, и выпукло вверх при $x < 0$ ($x'''(0) > 0$). В этом важном частном случае верна следующая оценка на нижнюю границу спектра оператора второй производной на квазиравномерной сетке в неограниченной области:

$$0 > \alpha_q > \alpha \geqslant -\frac{4N^2}{[x'(0)]^2}. \tag{28}$$

Информация о границах спектра матрицы, аппроксимирующей вторую производную на сетке, в ряде случаев помогает оптимально выбирать параметры разностной схемы и, следовательно, строить более эффективные и устойчивые методы расчетов.

§ 2. Эллиптические уравнения

1. Счет на установление. При численном решении краевых задач для эллиптических уравнений используются различные методы. Один из них — счет на установление; при этом ищется решение эволюционной задачи для параболического уравнения с тем же дифференциальным оператором по пространственным переменным до выхода на стационарный режим. Так, наряду с уравнением

$$\Delta_2 u - u - \Phi = 0 \tag{29}$$

можно рассматривать эволюционную задачу для уравнения

$$\frac{\partial w}{\partial t} = \Delta_2 w - w - \Phi. \tag{30}$$

При $t \to \infty$ решение эволюционного уравнения (30) стремится к решению стационарной задачи (29) $w \to u$.

2. Оптимальный шаг. Если параболическое уравнение используется для решения эллиптического уравнения счетом на установление, то шаг выбирают из условия выхода на стационарное решение за наименьшее число шагов. Для равномерной сетки в ограниченной области известен алгоритм выбора оптимального шага продольно-поперечной схемы. Построим такой алгоритм для неограниченной области в случае квазиравномерной сетки.

Для выбора оптимального шага в продольно-поперечной схеме нужно знать границы спектра симметричных матриц $-P_x$ и $-P_y$. Согласно (26), каждая компонента сеточного решения в выбранном базисе (25) затухает по закону

$$\xi_{ql}^{k+1} = \frac{1 - 0.5\tau\alpha_q}{1 + 0.5\tau\alpha_q} \cdot \frac{1 - 0.5\tau\beta_l}{1 + 0.5\tau\beta_l} \xi_{ql}^{k}.$$

Найдем оптимальный шаг по времени из условия наискорейшего затухания начальных данных

$$\min_{\tau} \max_{ql} \left| \frac{1 - 0.5\tau\alpha_q}{1 + 0.5\tau\alpha_q} \right| \cdot \left| \frac{1 - 0.5\tau\beta_l}{1 + 0.5\tau\beta_l} \right|.$$

Сначала выберем наименее слабо затухающую компоненту, потом выберем шаг по времени, соответствующий максимальному затуханию этой компоненты. Все собственные значения матриц $-P_x$ и $-P_y$ положительны. Пусть известны границы спектра $0 < \alpha \leqslant \alpha_q \leqslant \alpha_{\max}$ и $0 < \beta \leqslant \beta_l \leqslant \beta_{\max}$.

Дробь $(1 - 0.5\tau\alpha_q)/(1 + 0.5\tau\alpha_q)$ при $\alpha_q > 0$ монотонно убывает в пределах от 1 до -1. Поэтому своего максимального по модулю значения эта дробь достигает на границах:

$$\max_q \left| \frac{1 - 0.5\tau\alpha_q}{1 + 0.5\tau\alpha_q} \right| = \frac{1 - 0.5\tau\alpha}{1 + 0.5\tau\alpha} \quad \text{либо} \quad \max_q \left| \frac{1 - 0.5\tau\alpha_q}{1 + 0.5\tau\alpha_q} \right| = \frac{0.5\tau\alpha_{\max} - 1}{1 + 0.5\tau\alpha_{\max}}.$$

Аналогично

$$\max_l \left| \frac{1 - 0.5\tau\beta_l}{1 + 0.5\tau\beta_l} \right| = \frac{1 - 0.5\tau\beta}{1 + 0.5\tau\beta} \quad \text{либо} \quad \max_l \left| \frac{1 - 0.5\tau\beta_l}{1 + 0.5\tau\beta_l} \right| = \frac{0.5\tau\beta_{\max} - 1}{1 + 0.5\tau\beta_{\max}}.$$

Таким образом, медленнее других затухают компоненты, соответствующие границам спектра, причем с ростом τ модуль множителя роста компоненты, соответствующей нижней границе спектра, уменьшается, а верхней — наоборот, увеличивается. Выберем τ_0 таким, чтобы множители роста этих экстремальных компонент были одинаковыми:

$$\frac{1 - 0.5\tau_0\alpha}{1 + 0.5\tau_0\alpha} \cdot \frac{1 - 0.5\tau_0\beta}{1 + 0.5\tau_0\beta} = \frac{0.5\tau_0\alpha_{\max} - 1}{1 + 0.5\tau_0\alpha_{\max}} \cdot \frac{0.5\tau_0\beta_{\max} - 1}{1 + 0.5\tau_0\beta_{\max}}. \tag{31}$$

Этот шаг и будет *оптимальным*, т. к. при ином шаге по времени затухание одной из граничных компонент будет более медленным. Из соотношения (31) можно явно выразить τ_0:

$$\tau_0 = 2\sqrt{\frac{(\alpha_{\max} + \beta_{\max}) - (\alpha + \beta)}{\alpha_{\max}\beta_{\max}(\alpha + \beta) - \alpha\beta(\alpha_{\max} + \beta_{\max})}}. \tag{32}$$

Исследуем скорость выхода на стационарное решение в счете на установление с оптимальным шагом. Напомним, что в случае ограниченной области счет на установление с оптимальным шагом требует $K \sim N$ шагов. Нам приходилось решать два типа параболических уравнений: уравнение (10), которое содержит диссипативное слагаемое $-u$, и аналогичное уравнение без диссипативного члена (16). Скорости выхода на стационарное решение для этих двух случаев различны. При решении задачи на всей плоскости в отсутствие сильной анизотропии имеет смысл выбирать одинаковую квазиравномерную

сетку по x и по y. Тогда границы спектра в уравнении с диссипативным слагаемым будут

$$\alpha = \beta + 1 \quad \text{и} \quad \alpha_{\max} = \beta_{\max} + 1, \tag{33}$$

а без диссипативного члена

$$\alpha = \beta \quad \text{и} \quad \alpha_{\max} = \beta_{\max}. \tag{34}$$

Для этих двух случаев оптимальные шаги равны

$$\tau_0 = 2\sqrt{\frac{2\beta_{\max} - 2\beta}{(\beta_{\max}+1)\,\beta_{\max}\,(2\beta+1) - (\beta+1)\,\beta\,(2\beta_{\max}+1)}}$$

и

$$\tau_0 = 2\sqrt{\frac{\beta_{\max} - \beta}{\beta_{\max}^2\beta - \beta^2\beta_{\max}}}, \tag{35}$$

а максимальные по модулю множители роста составляют

$$\rho(\tau_0) = \frac{1 - 0.5\tau_0(\beta+1)}{1 + 0.5\tau_0(\beta+1)} \cdot \frac{1 - 0.5\tau_0\beta}{1 + 0.5\tau_0\beta} \quad \text{и} \quad \rho(\tau_0) = \left(\frac{1 - 0.5\tau_0\beta}{1 + 0.5\tau_0\beta}\right)^2 \tag{36}$$

соответственно. Здесь $\beta_{\max}$ и β — наибольший и наименьший по модулю собственные значения матрицы $-P_x = -X^{1/2}AX^{1/2}$ (18), (20). Эти собственные значения мы вычисляли при помощи стандартных процедур FORTRAN. Оказалось, что с ростом N — числа узлов по каждой пространственной переменной — с хорошей точностью $\beta_{\max} \sim$ $\sim N^2$ и $\beta \sim N^{-2}$. Асимптотика для большего по модулю собственного значения следует также из оценки (28). Напомним, что в случае ограниченной области границы спектра $\beta_{\max} \sim N^2$ и $\beta \sim 1$. Согласно (35) и (36), в неограниченной области

$$\tau_0 \sim N^{-1} \quad \text{и} \quad \rho(\tau_0) \sim 1 - N^{-1}$$

для уравнения с диссипативным членом, и

$$\tau_0 \sim N^{-2} \quad \text{и} \quad \rho(\tau_0) \sim 1 - N^{-2}$$

без диссипативного члена.

Если при счете на установление требуемая точность $\varepsilon \ll 1$, то необходимое для достижения этой точности число шагов по времени K можно оценить из соотношения $\rho^K(\tau_0) = \varepsilon$. Таким образом, счет на установление в неограниченной области с оптимальным шагом (32) требует следующего числа шагов по времени: при наличии диссипативного слагаемого

$$K \approx \frac{\ln\varepsilon}{\ln\rho(\tau_0)} \sim N,$$

а без диссипативного слагаемого

$$K \sim N^2.$$

Для сравнения проводился счет на установление стационарного решения с различными шагами по времени. При расчетах с оптимальным шагом (32) число шагов, необходимое для достижения требуемой точности, было в десятки раз меньше, чем при случайном, хотя и разумном на первый взгляд выборе шага. На рисунке 10 показано число шагов

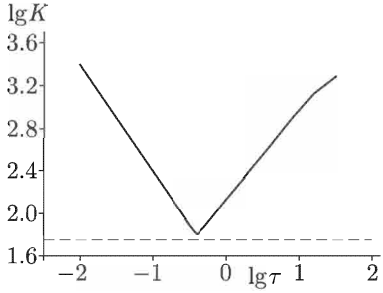

Рис. 10. Число шагов по времени K при счете на установление в зависимости от величины шага τ. Минимум $K = 64$ достигается при оптимальном шаге $\tau_0 \approx \approx 0.3957$. Горизонтальная штриховая линия показывает число шагов $K = 56$ в счете на установление с чебышевским набором шагов

по времени K, необходимое для выхода на стационарное решение с точностью $\varepsilon = 10^{-13}$ в зависимости от шага по времени τ. Расчет приведен для уравнения с диссипативным слагаемым на квазиравномерной сетке, покрывающей всю плоскость (12) и имеющей равное число узлов по обеим переменным ($N = M = 32$). В двойном логарифмическом масштабе видно, что при отклонении величины шага по времени от оптимального число шагов в счете на установление катастрофически нарастает (в данном случае согласно (35) $\tau_0 \approx 0.3957$).

Известно, что при расчетах с чебышевским набором шагов, можно снизить число шагов в счете на установление до $\sim \sqrt{N}$. Этот набор шагов по времени выбирают из условия наибольшего затухания начальных данных за фиксированное число шагов K. Оптимальным является следующий набор шагов [см., например, Самарский, 1977]:

$$\tau_k = 2\left[(\gamma_2+\gamma_1) + (\gamma_2-\gamma_1)\cos\frac{\pi(2k-1)}{2K}\right]^{-1}, \qquad K(\varepsilon) \approx \frac{1}{\ln(1/\rho)}\ln\frac{2}{\varepsilon}, \quad \rho = \frac{\sqrt{\gamma_2}-\sqrt{\gamma_1}}{\sqrt{\gamma_2}+\sqrt{\gamma_1}}. \tag{37}$$

Здесь ε — требуемая точность расчета, а γ_1, γ_2 связаны с оптимальным шагом соотношениями

$$\tau_0 = \frac{2}{\gamma_2+\gamma_1} \quad \text{и} \quad \rho(\tau_0) = \frac{\gamma_2-\gamma_1}{\gamma_2+\gamma_1}.$$

Учитывая (35) и (36), а также асимптотическое поведение границ спектра, получим для расчета с чебышевским набором шагов

$$K \approx \frac{1}{\ln(1/\rho)} \ln \frac{2}{\varepsilon} = \frac{1}{\ln\left(1+\sqrt{1-\rho^2(\tau_0)}\right) - \ln \rho(\tau_0)} \ln \frac{2}{\varepsilon}.$$

Для уравнения с диссипативным слагаемым и без него соответственно

$$K \sim \frac{\sqrt{N}}{2} \ln \frac{2}{\varepsilon} \quad \text{и} \quad K \sim \frac{N}{2} \ln \frac{2}{\varepsilon}.$$

В ограниченной области и без диссипативного слагаемого счет на установление с чебышевским набором шагов требует $K \sim \sqrt{N}$. Снижение скорости счета на установление в неограниченной области связано с тем, что в неограниченной области нижняя граница спектра $\beta \sim N^{-2}$, тогда как в ограниченной области $\beta \sim 1$.

На рисунке 10 горизонтальная линия указывает число $K = 56$ в счете на установление с чебышевским набором шагов, необходимое для достижения точности $\varepsilon = 10^{-13}$ в тестовой задаче. Для данного числа узлов пространственной сетки чебышевский набор шагов дает совсем незначительный выигрыш. Более того, в случае расчетов по продольно-поперечной схеме с переменным временным шагом прогоночные коэффициенты приходится пересчитывать на каждом новом временном слое, и выигрыша в скорости работы программы для небольшого числа узлов пространственной сетки $N < 60$ практически нет. Конечно, для квазиравномерных сеток с большим числом узлов выгоднее использовать чебышевский набор шагов. Если же решается задача типа (16), где нет диссипативного слагаемого, обеспечивающего быстрое затухание начальных данных, то счет на установление практически невозможен без чебышевского набора шагов.

В [Самарский, 1977] приведен алгоритм выбора переменного шага по Жордану, позволяющий снизить число шагов в счете на установление до $K \sim \ln N$. Однако на тех квазиравномерных сетках ($N \leqslant \leqslant 512$), которые мы использовали в расчетах, этот алгоритм дает несущественный выигрыш по сравнению с чебышевским набором шагов, но приводит к существенному усложнению программы.

§ 3. Гиперболические уравнения

1. Одномерное волновое уравнение. Неявная схема с весами. Рассмотрим одномерное уравнение колебаний с начальными данными:

$$\begin{gathered} \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + f(x,t), \\ u(x,0) = f_1(x), \quad u_t(x,0) = f_2(x); \end{gathered} \tag{38}$$

для полной постановки задачи требуется также два граничных условия, например значения $u(x,t)$ на правой и левой границах отрезка по x.

Для численного расчета уравнения (38) часто используют безусловно устойчивые неявные схемы с весами, имеющие точность $O(\tau^2 + h^2)$ (см. [Самарский, 1977]). Эти схемы пишутся на неравномерных сетках, но исследованы лишь в случае ограниченной области.

Исследуем вопрос о том, какими свойствами будут обладать такие схемы на квазиравномерных сетках.

Составим схему с весами при пространственных производных на разных временных слоях:

$$\frac{1}{\tau^2}(u_n^{k+1} - 2u_n^k + u_n^{k-1}) =$$

$$= \sigma\Lambda_{xx}[u]_n^{k+1} + (1-2\sigma)\Lambda_{xx}[u]_n^k + \sigma\Lambda_{xx}[u]_n^{k-1} + f_n^k, \quad (39)$$

$$1 \leqslant n \leqslant N-1;$$

значения u_0^{k+1} и u_N^{k+1} берутся из граничных условий. Здесь для второй пространственной производной использовано обозначение (13).

Значения u_n^1 на первом временном слое находятся с использованием начальных условий; на остальных слоях разностная схема (39) образует относительно u_n^{k+1} систему уравнений с трехдиагональной матрицей, решение которой находится методом прогонки.

Исследуем устойчивость неявной схемы с весами на квазиравномерной сетке.

Погрешность решения $\delta u = \xi$ будет удовлетворять однородному разностному уравнению

$$\frac{1}{\tau^2}(\xi_n^{k+1} - 2\xi_n^k + \xi_n^{k-1}) =$$

$$= \sigma\Lambda_{xx}[\xi]_n^{k+1} + (1-2\sigma)\Lambda_{xx}[\xi]_n^k + \sigma\Lambda_{xx}[\xi]_n^{k-1}. \quad (40)$$

Заметим, что действие оператора Λ_{xx} равносильно умножению сеточной функции ξ на следующие матрицы:

$$\Lambda_{xx}\xi = XA\xi,$$

где матрица A есть симметричная трехдиагональная, т. ч.

$$A_{n-1,n} = \frac{1}{x_{n-1/4} - x_{n-3/4}}, \quad A_{n+1,n} = \frac{1}{x_{n+3/4} - x_{n+1/4}},$$

$$A_{nn} = -A_{n-1,n} - A_{n+1,n},$$

а матрица X — диагональная

$$X_{nn} = \frac{0.5}{x_{n+1/2} - x_{n-1/2}}.$$

Свойства оператора A изучались в § 3 п. п. 3–4. В частности, было показано, что оператор $P_x = X^{1/2}AX^{1/2}$ имеет собственный базис и все собственные значения его строго меньше нуля.

Пусть $a^{(q)}$ — собственный вектор P_x, тогда $P_x a^{(q)} = \alpha_q a^{(q)}$, домножим это выражение на $X^{1/2}$:

$$\alpha_q X^{1/2} a^{(q)} = X^{1/2} P_x a^{(q)} = X^{1/2}(X^{1/2} A X^{1/2}) a^{(q)} = XAX^{1/2} a^{(q)}.$$

Таким образом, $\zeta^{(q)} = X^{1/2} a^{(q)}$ есть собственный вектор оператора $\Lambda_{xx} = XA$, а это означает, что справедлива следующая лемма.

Лемма 1. Оператор Λ_{xx} имеет собственный базис $\{\zeta^{(q)}\}$ и все его собственные значения отрицательны.

Мы можем разложить погрешность ξ^k на каждом временном слое по базису $\{\zeta^{(q)}\}$, т. ч.

$$\xi^k = \sum_q c_q^k \zeta^{(q)}. \tag{41}$$

Подставляя в (40) и приравнивая коэффициенты при $\zeta^{(q)}$, получаем

$$\frac{1}{\tau^2}(c_q^{k+1} - 2c_q^k + c_q^{k-1}) = \sigma\alpha_q c_q^{k+1} + (1 - 2\sigma)\alpha_q c_q^k + \sigma\alpha_q c_q^{k-1}$$

или

$$c_q^{k+1} = \frac{2 + (1 - 2\sigma)\alpha_q\tau^2}{1 - \sigma\alpha_q\tau^2} c_q^k - c_q^{k-1}.$$

Хорошо известно (см., например, [Бутузов и др., 2001]), что k-й член рекуррентно заданной таким образом последовательности представим в виде

$$\begin{gathered} c_q^k = s_q a^{k+1} + g_q b^{k+1}, \\ s_q = \frac{c_q^1 - bc_q^0}{a(a-b)}, \quad g_q = \frac{c_q^1 - ac_q^0}{b(b-a)}. \end{gathered} \tag{42}$$

Здесь a и b — корни квадратного уравнения $z^2 - pz + 1 = 0$, $p = [2 + (1 - 2\sigma)\alpha_q\tau^2]/(1 - \sigma\alpha_q\tau^2)$.

Для доказательства устойчивости схемы достаточно показать, что каждая из компонент в разложении (41) не нарастает с ростом k. Следовательно, в (42) необходимо потребовать, чтобы $|a| \leqslant 1$ и $|b| \leqslant 1$. Но $ab = 1$ и это возможно только при комплексно сопряженных корнях; при этом $|a| = |b| = 1$, т. е. при $|p| < 2$ или $-2 < [2 + (1 - 2\sigma)\alpha_q\tau^2]/(1 - \sigma\alpha_q\tau^2) < 2$. Правое неравенство выполняется автоматически, а левое сводится к $-4 < (1 - 4\sigma)\alpha_q\tau^2$. Если $\sigma \geqslant 1/4$, то справа стоит неотрицательное число и неравенство выполнено. Если $\sigma < 1/4$, необходимо потребовать, чтобы

$$\max_q |\alpha_q|\tau^2 \leqslant \frac{1}{1/4 - \sigma}.$$

Таким образом, справедлива следующая теорема.

Теорема 1. Если $\sigma \geqslant 1/4$, то схема (40) безусловно устойчива, если $\sigma < 1/4$, то схема условно устойчива при

$$\tau < \frac{1}{\sqrt{(1/4-\sigma)\max\limits_{q}|\alpha_q|}}.$$

Замечание. Труднопроверяемое условие для случая $\sigma < 1/4$ в теореме 1 может быть заменено на более простое: схема условно устойчива при

$$\tau < \frac{1}{\sqrt{(1/4-\sigma)|\alpha|}},$$

где α определено формулой (28). Если порождающее преобразование $x(\eta)$ строго монотонно и выпукло вниз при $x > 0$ и выпукло вверх при $x < 0$, то схема условно устойчива при

$$\frac{2\tau N}{x'(0)} < \frac{1}{\sqrt{(1/4-\sigma)}}.$$

Гладкие и обобщенные тестовые решения. Мы протестировали неявную схему с весами для задачи (38) на трех типах решений:

1) гладкие начальные данные

$$f_1(x) = e^{-x^2}, \quad f_2(x) = 0, \quad f(x,t) = 0;$$

2) негладкие начальные данные

$$f_1(x) = \begin{cases} 2-|x|, & |x| < 2, \\ 0, & |x| \geqslant 2, \end{cases} \qquad f_2(x) = 0, \quad f(x,t) = 0;$$

3) моделирование влияния точечного источника мгновенного действия (фундаментальное решение), т. е.

$$f(x,t) = \delta(x)\delta(t-b).$$

Тестовые расчеты. В схему с весами входит параметр σ. Были проведены численные расчеты с различными значениями σ, чтобы определить влияние этого параметра на отклонение Δ сеточного решения от точного в различных нормах. Результаты расчетов в нормах c и l_2 представлены на рисунках 11 и 12. Оказалось, что это влияние зависит также от соотношения шагов по времени и по пространству. На квазиравномерной сетке в качестве аналога числа Куранта æ можно взять отношение шага по времени к минимальному шагу по пространству.

Эти расчеты позволяют сделать следующие выводы: ошибка в норме c и в норме l_2 слабо зависит от σ (этот же вывод следует из анализа невязки), при маленьких числах Куранта ошибка монотонно растет

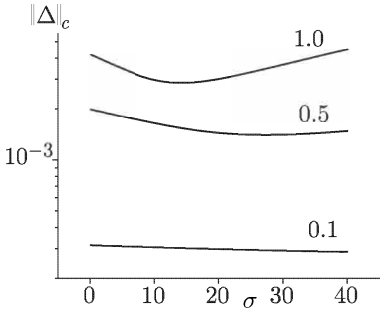

Рис. 11. Отклонение сеточного решения от точного в норме c, в зависимости от параметра σ при различных числах Куранта. Значения æ указаны около кривых

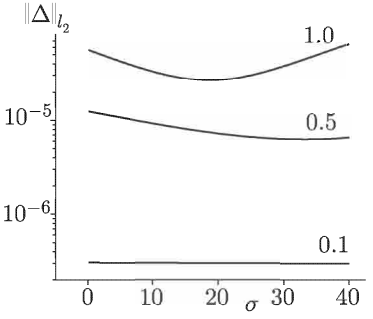

Рис. 12. Отклонение сеточного решения от точного в норме l_2, в зависимости от параметра σ при различных числах Куранта. Значения æ указаны около кривых

с ростом σ, при больших может иметь минимум при некотором $\sigma \neq \neq 0$. По-видимому, наиболее удобное значение $\sigma = 1/2$, т. к. схема (40) имеет наиболее простой вид.

Задача акустики и схема Розенброка. Уравнение второго порядка в задаче об одномерных колебаниях

$$u_{tt} = u_{xx} + f(x,t),$$
$$u(x,0) = f_1(x), \quad u_t(x,0) = f_2(x), \quad u(\pm\infty, t) = 0,$$

можно заменить эквивалентной ему парой уравнений первого порядка, введя потенциалы скоростей и правой части:

$$v(x,t) = \int_0^x u_t(\xi,t)d\xi, \quad F(x,t) = \int_0^x f(\xi,t)d\xi.$$

Результирующая система уравнений первого порядка принимает следующий вид:

$$u_t = v_x, \quad v_t = u_x + F(x,t),$$

$$u(x,0) = f_1(x), \quad v(x,0) = \int_0^x f_2(\xi)d\xi, \quad u(\pm\infty, t) = 0.$$

Будем решать ее одностадийным методом Розенброка.

Введем квазиравномерную сетку, покрывающую всю бесконечную прямую; отнесем значения смещения $u(x_n,t) = u_n$ к целым узлам сетки, а значения потенциала скорости $v(x_{n+1/2},t) = v_{n+1/2}$ и правой части $F(x_{n+1/2},t) = F_{n+1/2}$ — к полуцелым узлам. Тогда пространственная аппроксимация на квазиравномерной сетке системы акусти-

ческих уравнений приводит к следующей системе обыкновенных дифференциальных уравнений:

$$\frac{du_n}{dt} = \frac{v_{n+1/2} - v_{n-1/2}}{x_{n+1/2} - x_{n-1/2}}, \quad -N+1 \leqslant n \leqslant N-1,$$

$$\frac{dv_{n+1/2}}{dt} = 0.5\frac{u_{n+1} - u_n}{x_{n+3/4} - x_{n+1/4}} + F_{n+1/2}, \quad -N \leqslant n \leqslant N-1.$$

Введем новый вектор неизвестных $\mathbf{w}$ размерности $4N-1$, составленный из двух искомых сеточных функций $\{u_n\}$ и $\{v_{n+1/2}\}$:

$$w_{2n} = u_n, \quad -N+1 \leqslant n \leqslant N-1,$$

$$w_{2n+1} = v_{n+1/2}, \quad -N \leqslant n \leqslant N-1.$$

Для него получим систему обыкновенных дифференциальных уравнений

$$\frac{dw}{dt} = G(w, x, t),$$

для решения которой используем одностадийную схему Розенброка [Rosenbrock, 1953]:

$$\begin{aligned} \widehat{w} &= w + \tau \mathrm{Re} k, \\ (E - \alpha\tau G_w)k &= G. \end{aligned} \tag{43}$$

Здесь E — единичная матрица, G_w — матрица Якоби, в которой в данном случае отличны от нуля только верхняя и нижняя кодиагонали:

$$\begin{aligned} [G_w]_{2n,2n+1} &= -[G_w]_{2n,2n-1} = \frac{1}{x_{n+1/2} - x_{n-1/2}}, \\ [G_w]_{2n+1,2n+2} &= -[G_w]_{2n+1,2n} = \frac{0.5}{x_{n+3/4} - x_{n+1/4}}. \end{aligned} \tag{44}$$

Как сказано в § 1, при $\alpha = 0.5$ эта схема обладает $L1$-устойчивостью, а при $\alpha = 0.5 + i0.5$ — $L2$-устойчивостью; в обоих случаях она имеет второй порядок точности по времени.

Заметим, что в методе Розенброка приходится решать систему линейных уравнений, матрица которой в данном случае не меняется при переходе со слоя на слой. Поэтому обращать матрицу приходится всего один раз, что существенно снижает время расчета. Поскольку комплексная схема Розенброка много надежнее (см. § 1), представленные далее иллюстративные расчеты велись по ней.

Для иллюстрации возможностей метода приведем расчет волновой картины, порожденной гладким импульсом $u(x,0) = e^{-x^2}$ и импульсом, имеющим разрыв первой производной $u(x,0) = (1-|x|)\{\theta(x+1) - \theta(x-1)\}$, где $\theta(x)$ — функция Хевисайда. Числом Куранта в случае квазиравномерной сетки назовем отношение шага по времени к минимальному из шагов по пространству $æ = \tau N/c$. Все результаты приведены для числа Куранта $æ = 1$. Результаты представлены на рисунке 13 (*а*, *б*). Видно, что импульс

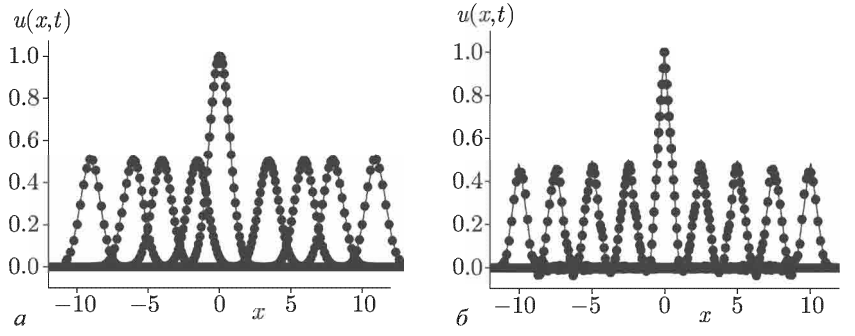

Рис. 13. Распространение гладкого (*а*) и негладкого (*б*) импульса в системе уравнений акустики. Расчет по схеме Розенброка с комплексными коэффициентами. Центральный импульс соответствует $t = 0.0$. Он распадается на два одинаковых импульса, бегущих в разные стороны. Приведенные кривые соответствуют моментам времени $t = 0.0$, 2.5, 5.0, 7.5, 10.0. Кружки — численный расчет, линия — точное решение

распространяется, точно сохраняя свою форму даже при значительном удалении от источника колебаний, что является важным достоинством предложенной схемы.

При расчетах другими схемами нарушение формы импульса визуально заметно, когда импульс смещается от источника колебаний на расстояние порядка его начальной ширины.

Количественно это свойство сохранения формы можно проиллюстрировать графиком зависимости погрешности численного решения от времени в двух сеточных нормах c и l_2 (рис. 14 и рис. 15). Погреш-

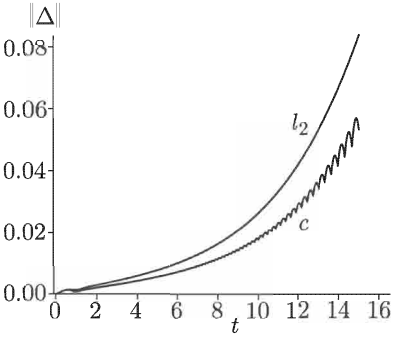

Рис. 14. Погрешность гладкого решения в различных нормах. Существенное нарастание погрешности численного решения наблюдается, когда импульс смещается в область, где сетка достаточно редкая

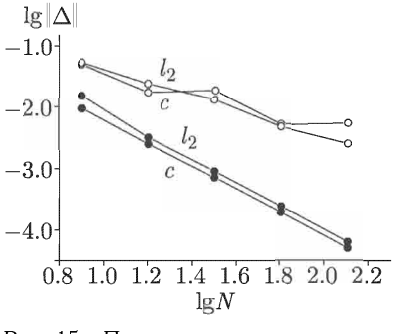

Рис. 15. Погрешность гладкого решения (•) и негладкого (○) в разных нормах. Тесты на сгущающихся сетках подтверждают второй порядок точности метода на гладких решениях. На негладком решении метод имеет первый порядок точности

ность начинает существенно нарастать, когда импульс смещается в ту область, где узлы квазиравномерной сетки расположены далеко друг от друга. Если требуется считать до больших времен, то необходимо разумно выбрать масштабный коэффициент и большее число узлов сетки.

Результаты тестов на сгущающихся сетках подтверждают порядок точности схемы $O(\tau^2 + N^{-2})$ на гладких решениях. На рисунке 15 приведен график убывания погрешности численного решения с ростом числа узлов сетки для гладкого и негладкого решения. На негладком решении метод имеет первый порядок точности. Сравнение проводилось в момент времени $t = 1.0$.

Исследование устойчивости. Погрешность численного решения в схеме Розенброка будет удовлетворять однородному уравнению

$$(E - \alpha\tau G_w)k = G_w\xi, \quad \alpha = (1+i)/2,$$
$$\widehat{\xi} = \xi + \tau \mathrm{Re} k,$$

которое можно переписать в вещественной форме:

$$\left(E - \tau G_w + \frac{\tau^2}{2}G_w^2\right)\widehat{\xi} = \xi. \tag{45}$$

Лемма 2. Собственные значения λ_q матрицы G_w — чисто мнимые, за исключением $\lambda = 0$ ($\lambda_q = i\mu_q$, $\mu_q \in \mathrm{Real}$), и соответствующий этой матрице оператор в линейном пространстве над полем комплексных чисел имеет собственный базис.

Доказательство. Удобно, используя (44), представить G_w в виде $G_w = XK$, где X — это диагональная матрица с положительными числами на диагонали, K — кососимметричная матрица, у которой на верхней кодиагонали стоят числа -1, а на нижней — числа 1. Матрица $S = X^{1/2}KX^{1/2}$ кососимметричная и имеет собственный базис, все ее собственные значения чисто мнимые за исключением $\lambda = 0$ (очевидно, iS — эрмитова). Пусть $\zeta^{(q)}$ — собственный вектор матрицы S, т. е.

$$S\zeta^{(q)} = \lambda_q\zeta^{(q)}.$$

Домножим это равенство слева на $X^{1/2}$:

$$X^{1/2}X^{1/2}KX^{1/2}\zeta^{(q)} = \lambda_q X^{1/2}\zeta^{(q)}$$

или

$$G_w X^{1/2}\zeta^{(q)} = \lambda_q X^{1/2}\zeta^{(q)},$$

тогда $X^{1/2}\zeta^{(q)}$ есть собственный вектор G_w с тем же λ_q. Так как матрица $X^{1/2}$ невырождена, то $\{X^{1/2}\zeta^{(q)}\}$ есть собственный базис G_w, что и требовалось доказать.

Дальнейшие исследования удобно провести в линейном пространстве над полем комплексных чисел. Любой вектор с действительными координатами можно интерпретировать как элемент этого линейного пространства. Поэтому, разлагая ξ в формуле (45) по собственному базису G_w:

$$\xi = \sum_q \mu_q \zeta^{(q)},$$

необходимо показать, что каждая из компонент не нарастает при переходе на следующий временной слой. Это и будет означать устойчивость схемы. Множитель роста для q-й компоненты и его модуль равны

$$\rho_q = \frac{1}{1 - i\tau\mu_q - 0.5\tau^2\mu_q^2},$$

$$|\rho_q| = \left(1 + \tau^4 \frac{\mu_q^4}{4}\right)^{-1/2} \leqslant 1.$$

Таким образом, справедлива теорема.

Теорема 2. Схема Розенброка с комплексным коэффициентом $\alpha = (1+i)/2$ для уравнений акустики безусловно устойчива.

2. Двумерное волновое уравнение. Запишем двумерную задачу для уравнения колебаний с начальными данными:

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + f(x, y, t),$$

$$u(x, y, 0) = f_1(x, y), \quad u_t(x, y, 0) = f_2(x, y);$$

в качестве граничного условия можно задать, например, $u(x, y, t)$ на границе пространственной области.

Введем двумерную квазиравномерную сетку по обеим пространственным переменным (12) и рассмотрим аналог одномерной схемы с весами:

$$\frac{1}{\tau^2}(u_{nm}^{k+1} - 2u_{nm}^k + u_{nm}^{k-1}) =$$

$$= (\Lambda_{xx} + \Lambda_{yy})\{\sigma[u]_{nm}^{k+1} + (1 - 2\sigma)[u]_{nm}^k + \sigma[u]_{nm}^{k-1}\} + f_{nm}^k;$$

при этом предполагается, что все шаги по времени τ одинаковы.

Эту схему можно переписать в следующем виде:

$$[E - \tau^2\sigma(\Lambda_{xx} + \Lambda_{yy})][u]_{nm}^{k+1} = [2E + \tau^2(1 - 2\sigma)(\Lambda_{xx} + \Lambda_{yy})][u]_{nm}^k -$$

$$- [E - \tau^2\sigma(\Lambda_{xx} + \Lambda_{yy})][u]_{nm}^{k-1} + \tau^2 f_{nm}^k.$$

Здесь на верхнем временном слое встречается трудно обратимый опе-

ратор

$$B = E - \tau^2\sigma(\Lambda_{xx} + \Lambda_{yy});$$

его можно приближенно заменить факторизованным оператором C:

$$C \equiv (E - \tau^2\sigma\Lambda_{xx})(E - \tau^2\sigma\Lambda_{xx}) =$$
$$= E - \tau^2\sigma(\Lambda_{xx} + \Lambda_{yy}) + \tau^4\sigma^2\Lambda_{xx}\Lambda_{yy} = B + O(\tau^4).$$

Поскольку $C - B = O(\tau^4)$, замена B на C не нарушает аппроксимации $O(\tau^2)$. Факторизованный оператор легко разрешается последовательными одномерными прогонками. Особенно удобен он для программирования в случае $\sigma = 0.5$.

Замечание. Факторизованные схемы применимы для любого числа пространственных переменных.

Для определения эффективного порядка точности построенного численного метода в неограниченной области были проведены тесты на сгущающихся сетках. С использованием интегрального преобразования Ханкеля можно построить (см. [Будак и др., 1980, с. 567]) точное решение двумерного уравнения колебаний в неограниченной области для частного случая радиальной симметрии:

$$f(x,y,t) = 0, \quad f_1(x,y) = \frac{1}{\sqrt{1+x^2+y^2}}, \quad f_2(x,y) = 0,$$

$$u(x,y,t) \underset{x^2+y^2\to\infty}{\longrightarrow} 0; \quad u(x,y,t) = \mathrm{Re}\frac{1}{\sqrt{(1+it)^2+x^2+y^2}}.$$

Результаты сравнения численного решения для различного числа узлов сетки приведены на рисунке 16. Каждый новый расчет соответствует удвоению числа узлов по каждой пространственной переменной и уменьшению вдвое шага по времени. Тангенс наклона прямой в двойном логарифмическом масштабе равен -2, что подтверждает точность метода $O(\tau^2 + N^{-2} + M^{-2})$.

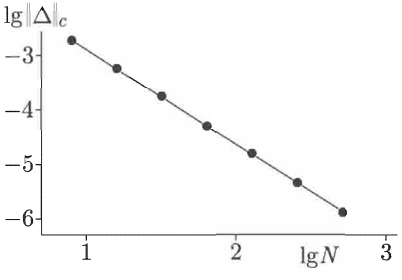

Рис. 16. Тесты на сгущающихся сетках для факторизованной схемы

В качестве иллюстраций возможностей предложенного численного метода приведем результаты расчетов для двумерного уравнения колебаний с кубической нелинейностью в неограниченной области:

$$u_{tt} = u_{xx} + u_{yy} + u^3,$$
$$u(x,y,0) = \sin(x)\sin(y)e^{-x^2-y^2}.$$

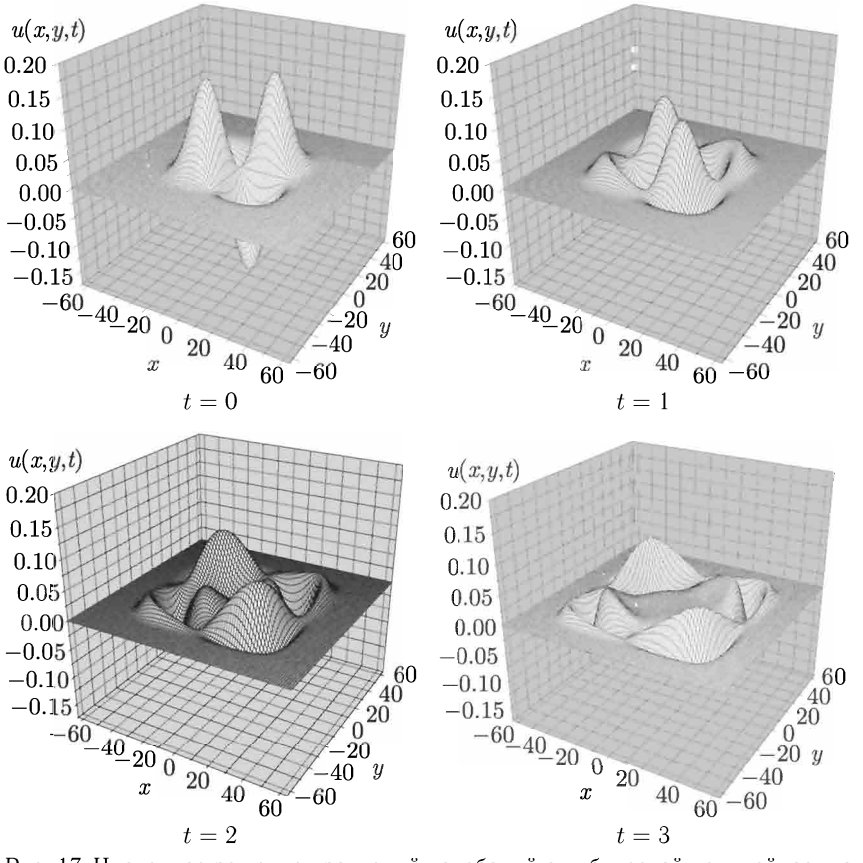

Рис. 17. Численное решение уравнений колебаний с кубической нелинейностью в неограниченной области

Решения для моментов времени $t = 0.0,\ 1.0,\ 2.0,\ 3.0$ приведены на рисунке 17.

§ 4. Уравнение нелинейного переноса

Рассмотрим задачу на всей числовой прямой для квазилинейного уравнения, в которой, как известно, даже при гладких начальных данных могут возникать разрывы решения:

$$\begin{gathered} u_t + uu_x = 0, \quad x \in (-\infty, +\infty), \\ u(x,0) = f_1(x). \end{gathered} \tag{46}$$

Предполагается, что $u(x,t)$ сохраняет знак, а граничное условие ста-

вится при $x \to -\infty$ для $u > 0$, при $x \to +\infty$ для $u < 0$. Далее предполагаем $u > 0$.

Для численного решения этой задачи сравним две схемы: комплексную схему Розенброка и схему бегущего счета. Для определенности будем решать задачу в случае $f_1(x) = 0.5 - \operatorname{arctg}(x)/\pi$. Такие начальные невозрастающие данные, как известно, приводят к разрывному решению.

Схема Розенброка. Введем квазиравномерную сетку, покрывающую прямую и аппроксимируем первую пространственную производную со вторым порядком точности:

$$\frac{du_n}{dt} = \frac{u_{n-1}^2 - u_n^2}{4(x_{n-1/4} - x_{n-3/4})}, \quad -N+1 \leqslant n \leqslant N.$$

Такая несимметричная форма записи обеспечивает лишь первый порядок аппроксимации $O(N^{-1})$, а использование дробных узлов делает формулу пригодной для неограниченной области (подробнее см. гл. III).

Для решения полученной системы обыкновенных дифференциальных уравнений применим комплексную схему Розенброка (43). Матрица системы алгебраических уравнений, возникающая в схеме Розенброка, в данном случае двухдиагональная, и ее обращение требует всего $\sim N$ арифметических действий.

Схема бегущего счета. Построим консервативную неявную схему для задачи (46) (это одностадийная вещественная схема Розенброка, с параметром $\alpha = 1.0$, обладающая L1-устойчивостью):

$$\frac{\widehat{u}_n - u_n}{\tau} = \frac{\widehat{u}_{n-1}^2 - \widehat{u}_n^2}{4(x_{n-1/4} - x_{n-3/4})}, \quad -N+1 \leqslant n \leqslant N.$$

Она формально неявная, но неизвестное значение $\widehat{u}_n$ явно выражается через значение в предыдущем узле из решения квадратного уравнения:

$$u_n^{k+1} = \left[(u_{n-1}^{k+1})^2 + \frac{4}{\tau}(x_{n-1/4} - x_{n-3/4})u_n^k + \frac{4}{\tau^2}(x_{n-1/4} - x_{n-3/4})^2\right]^{1/2} - \\ - \frac{2}{\tau}(x_{n-1/4} - x_{n-3/4}), \quad -N+1 \leqslant n \leqslant N;$$

из двух корней квадратного уравнения здесь выбран положительный. Тем самым, используя правое граничное условие, можно одно за другим вычислить $\widehat{u}_n$ по явным формулам.

Для приведенного выше гладкого монотонно убывающего начального профиля решение постепенно эволюционирует в разрывное. При этом амплитуда разрыва постепенно возрастает, асимптотически стремясь к 1. Скорость движения разрыва также постепенно возрастает, стремясь к 0.5. Это излишне сложный тест. Воспользуемся более

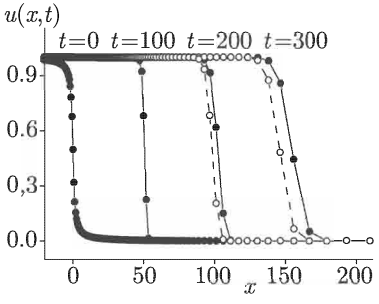

Рис. 18. Результаты расчетов для уравнения нелинейного переноса по комплексной схеме Розенброка (точки) и схеме бегущего счета (кружки)

простым тестом на полупрямой:

$$x \geqslant 0, \quad u(0,t) = 1, \quad f_1(x) = 0;$$

$$u(x,t) = \begin{cases} 0 & \text{при} \quad x > 0.5t, \\ 1 & \text{при} \quad x < 0.5t. \end{cases}$$

Такое решение есть скачок единичной амплитуды, распространяющийся вправо со скоростью 0.5.

Для сравнения этих схем приведем результаты расчетов для различных моментов времени (рис. 18). Точное решение, соответствующее данной постановке задачи, разрывно. Разрыв движется со скоростью 0.5 вправо. Обе схемы хорошо передают качественное поведение решения и удовлетворительно описывают скорость движения разрыва. При малых временах результаты, полученные по обеим схемам, практически неразличимы; существенное различие появляется при довольно больших временах (разрыв успевает пройти расстояние в 200 раз большее его высоты).

Квазиравномерная сетка по времени. Обе описанные схемы допускают использование переменного шага не только по пространству, но и по времени. Выберем квазиравномерную сетку по времени так, чтобы на фронте разрыва отношение временного и пространственного шагов было постоянным и равным 1. Соответствующие результаты приведены на рисунках 19 и 20.

Как видно из рисунков 19 и 20, схема Розенброка размазывает разрыв всего на один-два интервала сетки, тогда как у схемы бегущего счета это составляет 4–5 узлов. Численные решения приведены для

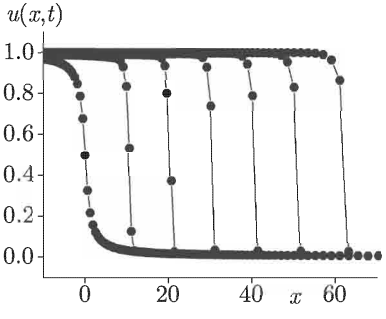

Рис. 19. Расчеты по времени комплексной схемой Розенброка для уравнения нелинейного переноса с квазиравномерной сеткой

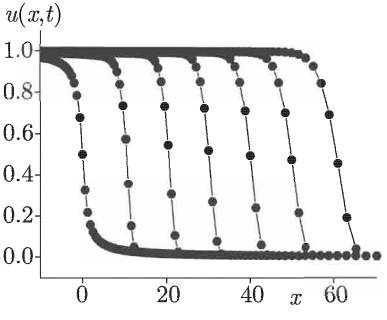

Рис. 20. Расчеты по времени схемой бегущего счета для уравнения нелинейного переноса с квазиравномерной сеткой

моментов времени $t = 0.0$, 20.54, 40.76, 60.07, 80.00, 99.86, 121.79. Но схема Розенброка немонотонная, и при неудачном выборе шага по времени (отношение шагов по времени и пространству на фронте разрыва существенно больше или меньше 1) может нарушаться монотонность. Правда, у схемы бегущего счета платой за несколько лучшее качество поведения решения является худшая точность (например, в 2–3 раза большее размазывание фронта).

При использовании квазиравномерной временной сетки шаг по времени сильно нарастает, и тем ни менее, даже при расчете разрывных решений обе схемы дают адекватный результат практически вплоть до того момента, когда разрыв достигает крайнего правого (бесконечного) интервала сетки. Отношение шага по времени к минимальному из шагов пространственной сетки при этом достигает 10000; несмотря на это, общая точность расчетов остается хорошей.

Глава VII

НЕКЛАССИЧЕСКИЕ УРАВНЕНИЯ СОБОЛЕВСКОГО ТИПА

В главе построен устойчивый численный метод решения начально-краевых задач для уравнений высокого порядка соболевского типа, возникающих при моделировании волновых процессов в средах с анизотропной дисперсией. Метод применен к различным соболевским уравнениям как в одномерном, так и в двумерном случаях. Использование квазиравномерных сеток позволило решать задачи в неограниченных областях. Проведены тесты на сгущающихся сетках с использованием специально построенных авторами точных решений.

§ 1. Математические модели, приводящие к соболевским уравнениям

Многие задачи геофизики, океанологии, физики атмосферы, физики магнитоупорядоченных структур, связанные с распространением волн в средах с сильной дисперсией, приводят к линейным дифференциальным уравнениям в частных производных, неразрешенным относительно старшей производной по времени:

$$A_0 D_t^p u + \sum_{q=0}^{p-1} A_{p-q} D_t^q u = f,$$

где A_0, $A_1, \ldots, A_p$ — линейные дифференциальные операторы по пространственным переменным. Такие уравнения часто называют уравнениями соболевского типа, т. к. именно с работы [Соболев, 1954] началось их систематическое изучение.

Исследование этих уравнений ведется в нескольких направлениях: доказательство разрешимости начально-краевых задач в различных постановках, построение асимптотики решений, спектральные задачи, численные расчеты и т.д. Приведем некоторые примеры.

1) Уравнение Соболева, описывающее малые колебания вращающейся жидкости:

$$\frac{\partial^2}{\partial t^2}\Delta_3 u + \alpha^2 \frac{\partial^2 u}{\partial x_3^2} = 0,$$

где $\Delta_3 = \dfrac{\partial^2}{\partial x_1^2} + \dfrac{\partial^2}{\partial x_2^2} + \dfrac{\partial^2}{\partial x_3^2}$ — трехмерный оператор Лапласа.

2) Уравнение

$$\frac{\partial}{\partial t}\Delta_3 u + \beta\frac{\partial}{\partial x_2}u = 0$$

появилось при изучении некоторых типов волн в тонких слоях жидкости на поверхности вращающегося глобуса. В океанологии эти процессы называют волнами Россби.

3) Основное уравнение динамики идеальной несжимаемой стратифицированной жидкости имеет вид

$$\frac{\partial^2}{\partial t^2}[\Delta_3 u - \beta^2 u] + \omega_0^2\Delta_2 u + \alpha^2\left[\frac{\partial^2 u}{\partial x_3^2} - \beta^2 u\right] = 0,$$

здесь $\Delta_2 = \dfrac{\partial^2}{\partial x_1^2} + \dfrac{\partial^2}{\partial x_2^2}$.

4) Уравнение ионно-звуковых волн в незамагниченной плазме:

$$\frac{\partial^2}{\partial t^2}(\Delta_3 u - u) + \Delta_3 u = 0.$$

Целый ряд новых уравнений, описывающих волновые процессы в различных средах с сильной дисперсией, получен в работе [Корпусов и др., 1999]. Ниже приведены некоторые из них.

5) Уравнение холодной плазмы во внешнем магнитном поле:

$$\frac{\partial^2}{\partial t^2}\left(\frac{\partial^2}{\partial t^2} + \omega_e^2\right)\left(\frac{\partial^2}{\partial t^2} + \omega_i^2\right)\Delta_2\Phi(\mathbf{x},t) + \\ + \left(\frac{\partial^2}{\partial t^2} + \omega_p^2\right)\left(\frac{\partial^2}{\partial t^2} + \omega_{B_e}^2\right)\left(\frac{\partial^2}{\partial t^2} + \omega_{B_i}^2\right)\frac{\partial^2\Phi}{\partial x_3^2} = \mathrm{div}\mathbf{F}_0.$$

6) Уравнение спиновых волн в ферритах типа “легкая плоскость”:

$$\left(\frac{\partial^2}{\partial t^2} + \omega_1^2\right)\frac{\partial^2 u}{\partial x_1^2} + \left(\frac{\partial^2}{\partial t^2} + \omega_2^2\right)\frac{\partial^2 u}{\partial x_2^2} + \left(\frac{\partial^2}{\partial t^2} + \omega_3^2\right)\frac{\partial^2 u}{\partial x_3^2} = 0.$$

Вопрос о существовании классических решений соответствующих начально-краевых задач для соболевских уравнений решается обычно с помощью развитого в последние десятилетия метода динамических потенциалов [Габов, Свешников, 1990]. Вопрос о единственности полученных решений исследуется обычно с помощью энергетического метода. Однако точные явные решения для таких задач удается получить только в небольшом классе канонических областей. Кроме того, ряд уравнений допускает обобщения, в которых присутствуют и нелинейные члены; это естественно ограничивает применение интегральных преобразований для получения явных решений. Возникает необходимость разработки эффективных численных методов решения таких задач.

§ 2. Метод квазиравномерных сеток в одномерном случае

1. Уравнение ионно-звуковых волн. Для апробации нового подхода к решению уравнений составного типа мы выбрали уравнение, описывающее ионно-звуковые волны в незамагниченной плазме:

$$\frac{\partial^2}{\partial t^2}(\Delta u - u) + \Delta u = 0. \tag{1}$$

Для этого уравнения доказаны теоремы о существовании и единственности решения начально-краевой задачи для широкого класса областей. Кроме того, известны точные решения уравнения (1), которые можно использовать для тестирования численных методов.

Начнем описание численного метода решения начально-краевых задач для уравнений соболевского типа с самого простого одномерного случая.

Задача I. Рассмотрим одномерное уравнение (1), которое является одномерным уравнением ионно-звуковых волн. Оно также возникает в одномерном случае из основного уравнения динамики идеальной несжимаемой стратифицированной жидкости и, кроме того, встречается в некоторых моделях, описывающих волновые процессы в длинных линиях. Будем решать задачу на полупрямой, на левой границе которой задан краевой режим $f(t)$. Отсутствие источников на бесконечности означает выполнение условия $\lim\limits_{x\to+\infty} u(x,t) = 0$. Дополняют постановку задачи начальные условия для функции $u(x,t)$ и ее производной по времени:

$$\begin{gathered}(u_{xx} - u)_{tt} + u_{xx} = 0; \quad x \in [0,\infty), \quad t \in [0,T];\\ u(0,t) = f(t), \quad u(+\infty,t) = 0, \quad u(x,0) = f_0(x), \quad u_t(x,0) = f_1(x).\end{gathered} \tag{2}$$

Данное дифференциальное уравнение четвертого порядка необычно тем, что при старшей производной по времени стоит неограниченный дифференциальный оператор по пространственной переменной.

Точное решение. Решение задачи (2) можно получить при помощи преобразования Лапласа. Положим для простоты начальные условия нулевыми $f_0(x) = f_1(x) \equiv 0$. Тогда образ Лапласа $U(x,p)$ искомой функции $u(x,t)$ является решением задачи:

$$\begin{gathered}p^2(U_{xx} - U) + U_{xx} = 0,\\ U(0,p) = F(p), \quad U(\infty,p) = 0.\end{gathered}$$

Отсюда легко находим

$$U(x,p) = F(p)\exp\left(-\frac{px}{\sqrt{p^2+1}}\right).$$

Таким образом, точное решение задачи (2) записывается в виде интеграла Меллина:

$$u(x,t) = \frac{1}{2\pi i} \int\limits_{\sigma-i\infty}^{\sigma+i\infty} F(p) \exp\left(-\frac{px}{\sqrt{p^2+1}}\right) e^{pt} dp. \tag{3}$$

Численное решение. Полученное точное решение (3) задачи (2) выписано в квадратурах. Аналитически вычислить интеграл (3) не представляется возможным даже для самых простых краевых режимов $f(t)$. Задача численного обращения преобразования Лапласа представляет отдельную хорошо изученную проблему. Программы для обращения преобразования Лапласа входят, например, в библиотеку IMSL стандартных программ языка FORTRAN (см., например, [Бартеньев, 2001]). Надежно эти программы работают лишь для небольших значений t, что связано с трудностью вычисления интеграла от сильно осциллирующей функции в (3). Для получения хорошей точности указанные стандартные программы обращения преобразования Лапласа требуют неприемлемо большого времени счета.

Другой подход оказался более удачным. Разобьем исходное уравнение на два. Одно будет содержать дифференцирование только по времени, другое — только по пространству. Это можно сделать, введя новую функцию

$$\Phi(x,t) = u_{xx}(x,t) - u(x,t). \tag{4}$$

Для нее задача (2) записывается в виде

$$\begin{aligned} &\Phi_{tt} + \Phi + u = 0, \\ &\Phi(x,0) = f_0''(x) - f_0(x), \\ &\Phi_t(x,0) = f_1''(x) - f_1(x). \end{aligned} \tag{5}$$

Нас интересует неограниченная область изменения пространственной переменной $x \in [0,\infty)$. Поэтому введем квазиравномерную сетку на полупрямой, порожденную, например, преобразованием, аналогичным (III.27):

$$x = x(\xi) = c \cdot \operatorname{tg}\left(\frac{\pi\xi}{2}\right), \quad \xi \in [0,1]. \tag{6}$$

В этом случае удобна следующая нумерация узлов: $x_n = x(\xi_n)$, $\xi_n = n/N$, $0 \leqslant n \leqslant N$. При конечном числе узлов N крайний узел сетки $x_N = \infty$. Тем самым правое краевое условие ставится непосредственно на бесконечности. Половина узлов сетки (6) лежит на отрезке $x_n \in [0,c]$; из этих соображений выбирают масштабный коэффициент c.

Сетку по временной переменной для простоты возьмем равномерной:

$$t_k = k\tau, \quad 0 \leqslant k \leqslant K, \quad \tau = T/K = \text{const}.$$

Аппроксимация и алгоритм. Аппроксимируем (5) с точностью $O(\tau^2)$ на равномерной временной сетке:

$$\frac{\Phi_n^{k+1} - 2\Phi_n^k + \Phi_n^{k-1}}{\tau^2} + \Phi_n^k + u_n^k = 0; \quad 1 \leqslant n \leqslant N-1, \quad 0 \leqslant k \leqslant K-1. \tag{7}$$

О применении (7) для $k = 0$ будет сказано ниже. Полученное выражение (7) явно разрешимо относительно неизвестного значения Φ_n^{k+1} на новом временном слое.

Аппроксимируем на квазиравномерной сетке уравнение (4) с порядком точности $O(N^{-2})$ по формуле (III.66), пригодной в неограниченной области. Для неизвестных значений функции u_n^{k+1} на новом временном слое получим трехточечное уравнение

$$\Phi_n^{k+1} = \frac{0.5}{x_{n+1/2} - x_{n-1/2}} \left(\frac{u_{n+1}^{k+1} - u_n^{k+1}}{x_{n+3/4} - x_{n+1/4}} - \frac{u_n^{k+1} - u_{n-1}^{k+1}}{x_{n-1/4} - x_{n-3/4}} \right) - u_n^{k+1}, \quad 1 \leqslant n \leqslant N-1, \quad 0 \leqslant k \leqslant K-1. \tag{8}$$

Напомним, что здесь значение $x_{n+\sigma}$, $0 \leqslant |\sigma| \leqslant 1$, в дробных узлах определяется тем же преобразованием (6): $x_{n+\sigma} = x(\xi_{n+\sigma})$; $\xi_{n+\sigma} = (n+\sigma)/N$. Использование дробных узлов в аппроксимации второй производной гарантирует корректный учет краевого условия при $x_N = \infty$ (см. гл. III, § 2).

Трехточечное уравнение (8) решается методом прогонки. Таким образом, разностные уравнения (7) и (8) позволяют вычислить Φ_n^k, u_n^k во всех внутренних узлах пространственно-временной сетки $2 \leqslant k \leqslant K$, $1 \leqslant n \leqslant N-1$.

Учет начальных и краевых условий. Значения u_0^k, $0 \leqslant k \leqslant K$, определяются краевым режимом $u_0^k = f(t_k)$. Из условия равенства нулю решения $u(x,t)$ на бесконечности получаем $u_N^k = 0$, $0 \leqslant k \leqslant K$.

Сеточное решение на нулевом слое u_n^0 известно из начального условия: $u_n^0 = f_0(x_n)$. При $t = 0$ используем первое начальное условие в (5) для определения

$$\Phi_n^0 = f_0''(x_n) - f_0(x_n).$$

При $k = 0$ в выражение (7) входит Φ_n^{-1}, которое можно найти, используя второе начальное условие в (5). Аппроксимируем его, сохраняя общую точность $O(\tau^2)$:

$$\Phi_t(x_n, 0) = \frac{\Phi_n^1 - \Phi_n^{-1}}{2\tau} + O(\tau^2) = f_1''(x_n) - f_1(x_n). \tag{9}$$

Исключая Φ_n^{-1} из (9) и (7), при $k = 0$ получим явное выражение для вычисления Φ_n^1.

Погрешность построенной аппроксимации задачи I есть $O(\tau^2 + N^{-2})$.

2. Результаты расчетов. Описанный алгоритм был применен для численного решения задачи I. На рисунке 1 показаны результаты расчетов для нулевых начальных условий и краевого режима, колеблющегося с частотой $\omega = 0.3$:

$$u(0,t) = \operatorname{arctg}^2(10t)\sin(\omega t). \tag{10}$$

При частотах краевого режима $\omega < 1$ установившаяся волновая картина имеет характер бегущих волн частоты ω, распространяющихся на бесконечность. Если же $\omega > 1$, волновая картина при больших временах приобретает характер стоячих волн.

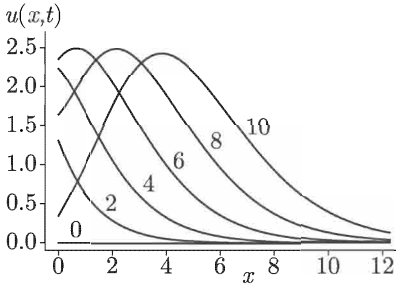

Рис. 1. Профили решения задачи I с краевым режимом (10) при $\omega < 1$, имеют характер бегущей волны, распространяющейся на бесконечность. Линии соответствуют моментам времени $t = 0, 2, 4, 6, 8, 10$

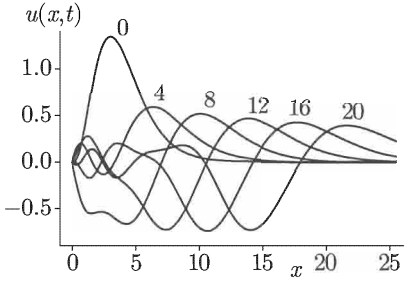

Рис. 2. Эволюция решения задачи I с начальным условием $u(x,0) = x^3e^{-x}$. Около кривых указаны моменты времени

Кроме того, расчеты проводились и для ненулевых начальных условий. На рисунке 2 показана эволюция начального профиля $u(x,0) = x^3e^{-x}$.

Тестирование метода. Для определения эффективного порядка точности построенного численного метода были проведены тесты на сгущающихся сетках. Численное решение для каждой сетки сравнивалось с точным решением (3). Интеграл Мелина в (3) вычислялся при помощи стандартной процедуры языка FORTRAN для обращения преобразования Лапласа. Так как погрешность разностной аппроксимации задачи есть $O(\tau^2 + N^{-2})$, сгущение временной и пространственной сеток нужно проводить в одно и то же число раз r. Разумно выбрать $r = 2$, т. к. в этом случае четные узлы новой пространственной сетки совпадают с узлами более редкой сетки, и в этих точках решения можно сравнивать между собой. Число узлов пространственной и временной сеток разумно выбирать так, чтобы минимальный из шагов пространственной сетки (6) $h_1 = x_1 - x_0 \approx c\pi/(2N)$ был бы равен шагу по времени $\tau = T/K$. Это дает $K \approx 2NT/(c\pi)$, и густоту сетки можно охарактеризовать единственным числом N.

Характер убывания погрешности $\Delta_n^k = u(x_n, t_k) - u_n^k$ численного решения с ростом числа узлов N отлично соответствует теоретическому порядку точности $O(\tau^2 + N^{-2})$ (рис. 3): наклон прямой в двойном логарифмическом масштабе равен -2. Результаты приведены для двух сеточных норм c и l_2, которые определяются следующим образом:

$$\|\Delta\|_c = \max_{1 \leqslant n \leqslant N} |\Delta_n^k|, \tag{11}$$

$$\|\Delta\|_{l_2} = \left[\sum_{n=1}^{N} (\Delta_n^k + \Delta_{n-1}^k)^2 (x_{n-1/4} - x_{n-3/4}) \right]^{1/2}. \tag{12}$$

Последнее выражение приближает несобственный интеграл

$$\left[\int_0^\infty \Delta^2(x,t)\,dx \right]^{1/2}$$

с точностью $O(N^{-2})$ (см. (III.14)). Нормы погрешностей в (11) и (12) свои для каждого временного слоя k. На рисунке 3 результаты приводятся для тех временных слоев k, которые соответствуют моменту времени $t = 1.0$.

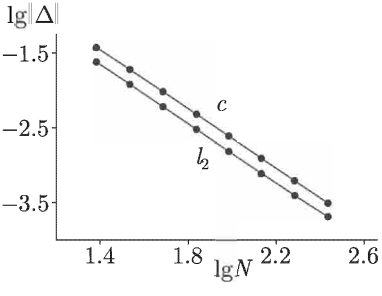

Рис. 3. Убывание погрешности численного решения с ростом числа узлов сетки N в сеточных нормах c и l_2

Результаты теста на сгущающихся сетках подтверждают, что построенный алгоритм дает возможность вычислять решение задачи I со вторым порядком точности по времени и пространству. Необходимо отметить, что для достижения заданного уровня точности ε расчет по предложенному алгоритму требует в десятки раз меньшего времени, чем вычисление интеграла в точном решении (3) с использованием стандартных подпрограмм языка FORTRAN для обращения преобразования Лапласа.

Задача II на прямой. Рассмотрим уравнение ионно-звуковых волн на прямой, дополнив его нулевыми граничными и ненулевыми начальными условиями для функции и ее производной:

$$(u_{xx} - u)_{tt} + u_{xx} = 0, \quad x \in (-\infty, \infty), \quad t \in [0, T],$$
$$u(+\infty, t) = u(-\infty, t) = 0, \quad u(x, 0) = f_0(x), \quad u_t(x, 0) = f_1(x). \tag{13}$$

Задача (13) может быть решена с помощью интегрального преобразования Фурье:

$$u(x,t)=\frac{1}{2\pi}\int\limits_{-\infty}^{+\infty}\left\{\widetilde{f}_0(p)\cos\left(\frac{pt}{\sqrt{p^2+1}}\right)+\right.$$

$$\left.+\frac{\widetilde{f}_1(p)\sqrt{p^2+1}}{p}\sin\left(\frac{pt}{\sqrt{p^2+1}}\right)\right\}\exp(ipx)dx, \quad (14)$$

$$\widetilde{f}_0(p)=\int\limits_{-\infty}^{+\infty}f_0(x)\exp(-ipx)dx, \quad \widetilde{f}_1(p)=\int\limits_{-\infty}^{+\infty}f_1(x)\exp(-ipx)dx.$$

Численное решение. Используем квазиравномерную сетку, покрывающую всю прямую $(-\infty,+\infty)$:

$$x=x(\xi)=c\cdot\operatorname{tg}\left(\frac{\pi\xi}{2}\right), \quad \xi\in[-1,1]. \quad (15)$$

В этом случае удобна следующая нумерация узлов: $x_n=x(\xi_n)$, $\xi_n=n/N$, $-N\leqslant n\leqslant N$. Крайние узлы квазиравномерной сетки, порожденной преобразованием (15), лежат в бесконечно удаленных точках, т. е. $x_{\pm N}=\pm\infty$.

Временную сетку для простоты возьмем равномерной: $t_k=\tau k$, $0\leqslant k\leqslant K$, $\tau=T/K$.

Алгоритм решения задачи II полностью аналогичен алгоритму для задачи I. Единственное отличие — в формулах (7) и (8) для нахождения решения во внутренних узлах сетки; границы изменения индекса $-N+1\leqslant n\leqslant N-1$ соответствуют сетке (15), покрывающей всю прямую.

Результаты расчетов. Для иллюстрации возможностей метода были проведены расчеты для различных начальных условий. На рисунке 4 показана эволюция начального профиля $u(x,0)=e^{-x^2}$.

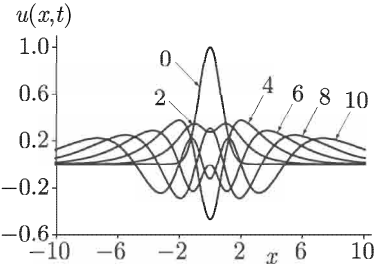

Рис. 4. Профили решения задачи II для начального условия $u(x,0)=e^{-x^2}$. Около кривых указаны соответствующие моменты времени

Тестирование метода. Для тестирования метода использовалось точное решение (14). Методика тестирования аналогична методике, описанной в задаче I, с тем лишь отличием, что при вычислении норм погрешностей следует использовать формулы

$$\|\Delta\|_c=\max_{-N\leqslant n\leqslant N}|\Delta_n^k|,$$

$$\|\Delta\|_{l_2} = \left[\sum_{n=-N+1}^{N} (\Delta_n^k + \Delta_{n-1}^k)^2 (x_{n-1/4} - x_{n-3/4})\right]^{1/2},$$

учитывающие границы изменения индексов для сетки на прямой $x \in \in (-\infty, \infty)$. Результаты теста на сгущающихся сетках подтверждают порядок точности метода $O(\tau^2 + N^{-2})$.

Постановка двумерной задачи III. Рассмотрим задачу о возбуждении двумерных ионно-звуковых волн вне круга радиуса a с граничным условием Дирихле. В полярных координатах она принимает вид

$$\begin{gathered}(\Delta_2 u - u)_{tt} + \Delta_2 u = 0,\\ r \equiv \sqrt{x^2+y^2} \in [a, \infty), \quad \varphi \equiv \operatorname{arctg}(y/x) \in [0, 2\pi],\\ u\big|_{r=a} = f(\varphi, t), \quad u\big|_{t=0} = u_t\big|_{t=0} = 0, \quad \lim_{r\to+\infty} u(r, \varphi, t) = 0;\end{gathered} \tag{16}$$

здесь $\Delta_2 = \dfrac{1}{r}\dfrac{\partial}{\partial r}\left(r\dfrac{\partial}{\partial r}\right) + \dfrac{1}{r^2}\dfrac{\partial^2}{\partial \varphi^2}$ — двумерный оператор Лапласа. Пусть функция $f(\varphi, t)$, задающая краевой режим на круге $r = a$, является дважды непрерывно дифференцируемой по обоим аргументам и 2π-периодической по переменной φ. Запишем ее разложение в ряд Фурье:

$$f(\varphi, t) = \frac{\alpha_0(t)}{2} + \sum_{q=1}^{+\infty} \alpha_q(t)\cos(q\varphi) + \beta_q(t)\sin(q\varphi). \tag{17}$$

Тогда решение задачи (16) может быть построено в виде ряда Фурье:

$$u(r, \varphi, t) = \frac{a_0(r,t)}{2} + \sum_{q=1}^{+\infty} [a_q(r,t)\cos(q\varphi) + b_q(r,t)\sin(q\varphi)], \tag{18}$$

коэффициенты которого

$$\begin{aligned} a_0(r,t) &= \frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty} \widetilde{\alpha}_0(p)\frac{K_0(pr/\sqrt{p^2+1}\,)}{K_0(pa/\sqrt{p^2+1}\,)}\exp(pt)dp,\\ a_q(r,t) &= \frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty} \widetilde{\alpha}_q(p)\frac{K_q(pr/\sqrt{p^2+1}\,)}{K_q(pa/\sqrt{p^2+1}\,)}\exp(pt)dp,\\ b_q(r,t) &= \frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty} \widetilde{\beta}_q(p)\frac{K_q(pr/\sqrt{p^2+1}\,)}{K_q(pa/\sqrt{p^2+1}\,)}\exp(pt)dp, \end{aligned} \tag{19}$$

здесь $\widetilde{\alpha}_q(p)$, $\widetilde{\beta}_q(p)$ — образы Лапласа коэффициентов Фурье $\alpha_q(t)$, $\beta_q(t)$ ряда (17); $K_q(z)$ — функция Макдональда. Ветвь корня выбирается так, чтобы $\lim\limits_{\substack{|p|\to\infty \\ \mathrm{Re}\, p>0}} \dfrac{p}{\sqrt{p^2+1}} = 1$.

Поведение решения (18), (19) при некоторых конкретных граничных условиях изучалось в работе [Альшин, Плетнер, 1994]; там, в частности, был рассмотрен вопрос о стабилизации решения при $t \to \infty$ и о существовании режима установившихся колебаний.

Непосредственной проверкой можно убедиться, что коэффициенты разложения Фурье $a_q(r,t)$, $b_q(r,t)$ в (18) удовлетворяют уравнению

$$\frac{\partial^2}{\partial t^2}\left[\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}R_q\right) - q^2\frac{R_q}{r^2} - R_q\right] + \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}R_q\right) - q^2\frac{R_q}{r^2} = 0, \quad (20)$$

$$q = 0, 1, 2, \ldots,$$

где $R_q(r,t) = a_q(r,t)$, либо $R_q(r,t) = b_q(r,t)$.

В дальнейшем будем рассматривать начально-краевую задачу для уравнения (20), дополнив его граничным условием при $r = a$, $R_q(a,t) = \alpha_q(t)$, либо $R_q(a,t) = \beta_q(t)$, нулевыми условиями при $r = \infty$, а также нулевыми начальными условиями.

Отметим, что удобнее рассматривать задачу для функции $w(r,t) = \sqrt{r}\,R_q(r,t)$, т. к. в этом случае уравнение для $w(r,t)$ не содержит первой производной по r:

$$\frac{\partial^2}{\partial t^2}\left(w_{rr} + \frac{w}{4r^2} - q^2\frac{w}{r^2} - w\right) + w_{rr} + \frac{w}{4r^2} - \frac{w}{r^2} = 0, \quad (21)$$

$$q = 0, 1, 2, \ldots$$

Введем новую функцию, аналогичную (4):

$$\Phi = w_{rr} + \frac{w}{4r^2} - q^2\frac{w}{r^2}; \quad (22)$$

подставляя (22) в (21), получим

$$\Phi_{tt} + \Phi + w = 0. \quad (23)$$

Таким образом (21) расщепилось на (22) и (23), каждое из которых содержит дифференцирование только по времени или только по пространству.

В дальнейшем алгоритм нахождения сеточного решения для задачи (22), (23) аналогичен алгоритму решения задачи I. В том лишь разница, что пространственная сетка покрывает полупрямую $r \in [a, \infty)$ и вместо сетки (6) нужно использовать квазиравномерную сетку, порожденную преобразованием

$$r = r(\xi) = a + c \cdot \mathrm{tg}\left(\frac{\pi\xi}{2}\right), \quad \xi \in [0, 1].$$

На рисунке 5 показаны профили решения задачи III для внешности круга единичного радиуса с граничным условием $u(1,\varphi,t) = \sin^2(t/4)$. При таком выборе краевого режима решение не зависит от φ. Полученное численное решение хорошо согласуется с точным решением (18).

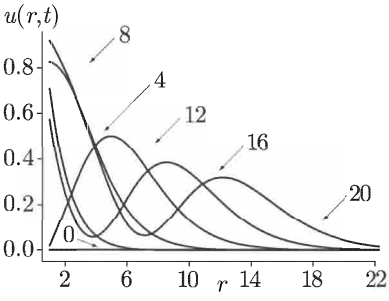

Рис. 5. Профили решения задачи III для внешности единичного круга с граничным режимом, соответствующим режиму бегущей волны. Около кривых показаны моменты времени

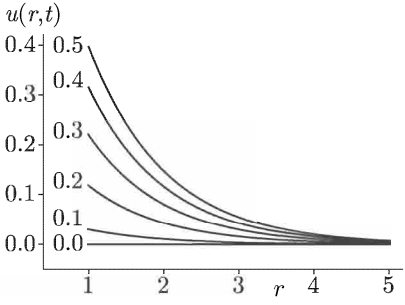

Рис. 6. Профили решения задачи IV в сферически симметричном случае для краевого режима, соответствующего стоячей волне. Около кривых указаны моменты времени

Задача IV. Метод легко обобщается для решения внешней начально-краевой задачи о возбуждении ионно-звуковых волн в случае трех пространственных переменных с краевым режимом, заданным на сфере. Исследуем частный случай, когда граничный режим, а следовательно и решение u, не зависит от углов θ и φ. В сферических координатах задача принимает вид

$$\begin{gathered}(\Delta_3 u - u)_{tt} + \Delta_3 u = 0,\\ r \in [a,\infty), \quad \varphi \in [0,2\pi], \quad \theta \in [0,\pi],\\ u\big|_{r=a} = f(t), \quad u\big|_{t=0} = u_t\big|_{t=0} = 0, \quad \lim_{r\to+\infty} u(r,\varphi,\theta,t) = 0.\end{gathered} \tag{24}$$

Здесь Δ_3 — сферический оператор Лапласа, в данном случае он совпадает со своей радиальной частью $\Delta_3 \Rightarrow \dfrac{1}{r^2}\dfrac{\partial}{\partial r}\left(r^2\dfrac{\partial}{\partial r}\right)$.

Стандартной заменой $w = r\,u$ уравнение (24) приводится к уравнению вида (2); тогда для решения задачи можно использовать алгоритм, предложенный для задачи I.

Для решения задачи IV также реализуются два различных типа волновых процессов в зависимости от частоты краевого режима $f(t)$. На рисунке 6 показаны результаты расчетов для краевого режима $u(0,t) = 0.25\,\mathrm{arctg}^2(10t)\sin(2t)$, соответствующего режиму стоячей волны.

Задача V. Рассмотрим начально-краевую задачу для нелинейного уравнения составного типа. Это уравнение описывает нелинейные волны в длинных линиях (например, телеграфных) с дисперсией [Лонгрен, 1981]. Задача ставится на прямой $(-\infty, +\infty)$:

$$(u_{xx} - u + 2u^3)_{tt} + u_{xx} = 0, \quad x \in (-\infty, \infty), \quad t \in [0, T];$$

$$u(+\infty, t) = u(-\infty, t) = 0, \quad u(x, 0) = f_0(x), \quad u_t(x, 0) = f_1(x).$$

Это уравнение имеет частное решение в виде уединенной волны

$$u(x, t) = \frac{\sqrt{1 - 1/c^2}}{\operatorname{ch}\left[\sqrt{1 - 1/c^2} \cdot (x \pm ct)\right]} \tag{25}$$

при значениях параметра $c > 1$, задающего скорость движения уединенной волны.

Для построения численного алгоритма нахождения решения будем поступать так же, как и в задаче I. Введем функцию

$$\Phi = u_{xx} - u + 2u^3.$$

Тогда для функции Φ справедлива задача, аналогичная (5):

$$\begin{gathered} \Phi_{tt} + \Phi + u - 2u^3 = 0, \\ \Phi(x, 0) = 0, \quad \Phi_t(x, 0) = 0. \end{gathered} \tag{26}$$

Надо отметить, что описанный прием расщепления уравнения составного типа на два, каждое из которых содержит дифференцирование только по временной или только по пространственной переменной, достаточно универсален; он подходит, в том числе, и для нелинейных уравнений.

Уравнение (26) аппроксимируем со вторым порядком точности и получаем явную трехслойную схему для нахождения Φ на следующем временном слое. Для нахождения u на следующем временном слое необходимо решить нелинейную краевую задачу:

$$\begin{gathered} u_{xx} - u + 2u^3 = \Phi, \\ u(-\infty, t) = u(+\infty, t) = 0. \end{gathered}$$

Ее аппроксимация на квазиравномерной сетке (15) приводит к нелинейной системе алгебраических уравнений:

$$\frac{0.5}{x_{n+1/2} - x_{n-1/2}} \left(\frac{u_{n+1}^{k+1} - u_n^{k+1}}{x_{n+3/4} - x_{n+1/4}} - \frac{u_n^{k+1} - u_{n-1}^{k+1}}{x_{n-1/4} - x_{n-3/4}} \right) - u_n^{k+1} + 2(u_n^{k+1})^3 = \Phi_n^{k+1}, \tag{27}$$

$$u_{-N}^k = u_N^k = 0, \quad -N + 1 \leqslant n \leqslant N - 1.$$

Для решения этой нелинейной системы (27) мы применяли метод Ньютона. В качестве начального приближения разумно выбирать значения функции на предыдущем временном слое; тогда для решения нелинейной системы (27) достаточно всего 3–4 итерации.

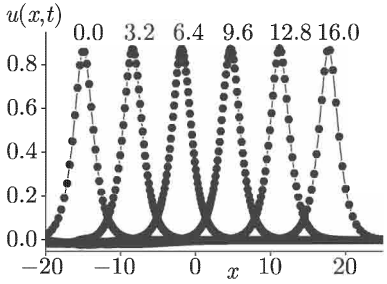

Рис. 7. Профили решения задачи V. Скорость движения уединенной волны соответствует теоретической. Около кривых указаны моменты времени. • — численное решение, сплошная линия — точное

Результаты расчетов показаны на рисунке 7. Начальный импульс

$$u(x,0) = \frac{\sqrt{1-1/c^2}}{\operatorname{ch}\left(\sqrt{1-1/c^2}\cdot(x+15)\right)}, \quad c=2,$$

сохраняя форму, распространяется вправо со скоростью $c=2$, что соответствует теории. Точками показано численное решение, линиями — точное решение (25).

§ 3. Двумерный случай

1. Двумерные ионно-звуковые волны. Будем рассматривать начально-краевые задачи в различных неограниченных по пространственным переменным двумерных областях для трех уравнений, относящихся к соболевскому типу. Рассмотрим более подробно физический смысл каждого из уравнений.

1. Уравнение ионно-звуковых волн в незамагниченной плазме:

$$(\Delta_2 u - u)_{tt} + \Delta_2 u = 0; \tag{28}$$

здесь Δ_2 — оператор Лапласа по двум пространственным переменным, u имеет смысл динамического отклонения концентрации ионов от среднего значения. Вывод (28) можно найти, например, в [Габов, 1998]. Это уравнение также возникает при рассмотрении двумерных внутренних гравитационных волн [Габов, Свешников, 1990], возникающих внутри жидкости на границе ее слоев.

Будут рассмотрены две различные постановки задачи для этого уравнения. Первая — в области, неограниченной по обеим пространственным переменным:

$$\begin{gathered}(\Delta_2 u - u)_{tt} + \Delta_2 u = 0,\\ x \in (-\infty, +\infty), \quad y \in (-\infty, +\infty), \quad t \in [0, T];\\ \lim_{\sqrt{x^2+y^2}\to\infty} u(x, y, t) = 0, \quad u\big|_{t=0} = f_0(x, y), \quad u_t\big|_{t=0} = f_1(x, y).\end{gathered} \tag{29}$$

Вторая постановка относится к полуплоскости с заданным краевым режимом:

$$\begin{gathered}(\Delta_2 u - u)_{tt} + \Delta_2 u = 0,\\ x \in (-\infty, +\infty), \quad y \in [0, +\infty), \quad t \in [0, T];\\ \lim_{\sqrt{x^2+y^2}\to\infty} u(x, y, t) = 0, \quad u\big|_{y=0} = f(x, t),\\ u\big|_{t=0} = f_0(x, y), \quad u_t\big|_{t=0} = f_1(x, y).\end{gathered} \tag{30}$$

2. В рамках модели Буссинеска, предполагающей слабую стратификацию жидкости, система уравнений Эйлера редуцируется к уравнению двумерных гравитационно-гироскопических волн, имеющему вид

$$\frac{\partial^2}{\partial t^2}\Delta_2 u + \omega_0^2 u_{xx} + \alpha^2 u_{yy} = 0, \tag{31}$$

где u — функция тока, связанная с компонентами скорости частиц жидкости $\mathbf{v} = \{v_1, v_2\}$ соотношениями $v_1 = u_y$, $v_2 = -u_x$; $\alpha/2$ — частота вращения, $\beta > 0$ — параметр стратификации, $\omega_0 = \sqrt{2\beta g}$ — частота Вейсяля–Брента.

Рассмотрим начально-краевую задачу для этого уравнения в полуплоскости $x \in (-\infty, +\infty)$, $y \geqslant x \operatorname{tg}\gamma$. Такая задача описывает волновую картину в стратифицированной и вращающейся жидкости, заполняющей сосуд с наклонным дном, здесь γ — угол наклона. После замены переменных $x = x' \cos\gamma - y' \sin\gamma$ и $y = x' \sin\gamma + y' \cos\gamma$ постановка начально-краевой задачи выглядит так:

$$\begin{gathered}\frac{\partial^2}{\partial t^2}\Delta_2 u + \lambda u_{x'x'} + \mu u_{x'y'} + \nu u_{y'y'} = 0,\\ x' \in (-\infty, +\infty), \quad y' \in [0, +\infty), \quad t \in [0, T];\\ \lim_{\sqrt{x'^2+y'^2}\to\infty} u(x', y', t) = 0, \quad u\big|_{y'=0} = f(x', t),\\ u\big|_{t=0} = f_0(x', y'), \quad u_t\big|_{t=0} = f_1(x', y').\end{gathered} \tag{32}$$

Здесь $\lambda = \alpha\cos^2\gamma + \beta\sin^2\gamma$, $\mu = (\alpha - \beta)\sin 2\gamma$, $\nu = \beta\cos^2\gamma + \alpha\sin^2\gamma$, а оператор Лапласа вычисляется в штрихованных переменных: $\Delta_2 =$

$= \dfrac{\partial^2}{\partial x'^2} + \dfrac{\partial^2}{\partial y'^2}$. Задачу удобнее решать в записи (32), где далее штрихи будут опущены.

3. Целый ряд моделей физики полупроводников, физики плазмы, гидродинамики стратифицированных и фильтрующихся жидкостей, теории "ползучести" элементов конструкций и др. приводят к неклассическим начально-краевым задачам для эволюционных уравнений в частных производных псевдопараболического типа. Наиболее полный обзор различных таких моделей можно найти в [Корпусов, Свешников, 2003]; там же приведен вывод ряда новых линейных и нелинейных волновых уравнений псевдопараболического типа, являющихся многомерными обобщениями известных одномерных уравнений типа Бенджамена–Бона–Махони и Бенджамена–Бона–Махони–Бюргерса. Мы будем рассматривать начально-краевую задачу для следующего уравнения псевдопараболического типа:

$$\begin{gathered}(\Delta_2 u - u)_t + \Delta_2 u + \beta u = 0,\\ x \in (-\infty, +\infty), \quad y \in [0, +\infty), \quad t \in [0, T];\\ \lim_{\sqrt{x^2+y^2}\to\infty} u(x, y, t) = 0, \quad u(x, y, 0) = f_0(x, y).\end{gathered} \tag{33}$$

Это уравнение содержит только первую производную по времени, что необходимо учитывать при построении разностной схемы точности $O(\tau^2)$.

Алгоритм решения начально-краевой задачи. Сначала для определенности рассмотрим задачу (29). В §2 был предложен метод решения аналогичной задачи в одномерном случае. Обобщим этот метод для случая двух пространственных переменных. Разобьем уравнение (28) на два. Одно будет содержать дифференцирование только по времени, другое — только по пространству:

$$\Phi_{tt} + \Phi + u = 0, \tag{34}$$

$$\Delta_2 u - u = \Phi. \tag{35}$$

Для решения этой задачи построим двумерную квазиравномерную сетку, покрывающую всю плоскость:

$$\begin{aligned} x_n &= c_x \operatorname{tg}\left(\frac{\pi n}{2N}\right), \quad -N \leqslant n \leqslant N,\\ y_m &= c_y \operatorname{tg}\left(\frac{\pi m}{2M}\right), \quad -M \leqslant m \leqslant M.\end{aligned} \tag{36}$$

По времени для простоты возьмем равномерную сетку

$$t_k = k\tau, \quad 0 \leqslant k \leqslant K, \quad \tau = T/K = \text{const}.$$

Аппроксимируем уравнение (34) с точностью $O(\tau^2)$:

$$\frac{\Phi_{nm}^{k+1} - 2\Phi_{nm}^{k} + \Phi_{nm}^{k-1}}{\tau^2} + \Phi_{nm}^{k} + u_{nm}^{k} = 0, \tag{37}$$

$$-N+1 \leqslant n \leqslant N-1, \quad -M+1 \leqslant m \leqslant M-1, \quad 0 \leqslant k \leqslant K-1;$$

возможность использования здесь $k = 0$ обсуждалась ранее. Уравнение (35) аппроксимируем с точностью $O(N^{-2} + M^{-2})$:

$$(\Lambda_{xx} + \Lambda_{yy} - E)[u]_{nm}^{k+1} = \Phi_{nm}^{k+1}, \tag{38}$$

$$-N+1 \leqslant n \leqslant N-1, \quad -M+1 \leqslant m \leqslant M-1, \quad 0 \leqslant k \leqslant K-1.$$

Здесь использованы обозначения для операторов, аппроксимирующих вторые пространственные производные на квазиравномерной сетке в неограниченной области:

$$\Lambda_{xx}[u]_{nm}^{k} = \frac{0.5}{(x_{n+1/2} - x_{n-1/2})}\left(\frac{u_{n+1,m}^{k} - u_{nm}^{k}}{(x_{n+3/4} - x_{n+1/4})} - \frac{u_{nm}^{k} - u_{n-1,m}^{k}}{(x_{n-1/4} - x_{n-3/4})}\right), \tag{39}$$

$$\Lambda_{yy}[u]_{nm}^{k} = \frac{0.5}{(y_{m+1/2} - y_{m-1/2})}\left(\frac{u_{n,m+1}^{k} - u_{nm}^{k}}{(y_{m+3/4} - y_{m+1/4})} - \frac{u_{nm}^{k} - u_{n,m-1}^{k}}{(y_{m-1/4} - y_{m-3/4})}\right). \tag{40}$$

Учтено, что

$$[\Delta_2 u(x,y,t)]_{nm}^{k} = \Lambda_{xx}[u]_{nm}^{k} + \Lambda_{yy}[u]_{nm}^{k} + O(N^{-2} + M^{-2}). \tag{41}$$

Учет начальных условий. На нулевом временном слое значения сеточных функций u_{nm}^{0} и Φ_{nm}^{0} известны из начальных условий задачи (29):

$$u_{nm}^{0} = f_0(x_n, y_m) \quad \text{и} \quad \Phi_{nm}^{0} = \Delta_2 f_0(x_n, y_m) - f_0(x_n, y_m).$$

При переходе на первый временной слой формула (37) для $k = 0$ содержит неизвестные величины Φ_{nm}^{-1}. Согласно (34) и начальным условиям задачи (29),

$$\Phi_t(x,y,0) = \Delta_2 u_t - u_t = \Delta_2 f_1 - f_1.$$

Аппроксимируя первую производную по времени при $t = 0$ со вторым порядком точности, получим

$$\Phi_t(x_n, y_m, 0) = \frac{\Phi_{nm}^{1} - \Phi_{nm}^{-1}}{2\tau} + O(\tau^2). \tag{42}$$

Это и есть недостающее уравнение. Рассматривая (37) при $k = 0$ и (42) как систему уравнений относительно неизвестных Φ_{nm}^{-1} и Φ_{nm}^{1}, получим явную формулу для перехода на первый временной слой:

$$\Phi_{nm}^{1} = \Phi_{nm}^{0} + \tau\Phi_t^{0} - 0.5\tau^2(\Phi_{nm}^{0} + u_{nm}^{0}).$$

При $k \neq 0$ уравнения (37) явно разрешимы относительно неизвестных величин Φ_{nm}^{k+1}, и формула перехода на новый временной слой выглядит следующим образом:

$$\Phi_{nm}^{k+1} = 2\Phi_{nm}^{k} - \Phi_{nm}^{k-1} - \tau^2(\Phi_{nm}^{k} + u_{nm}^{k}).$$

Для нахождения значений u_{nm}^{k+1} на новом слое необходимо решить систему (38), которая уже неявна относительно неизвестных u_{nm}^{k+1}. Уравнение (38) является сеточным аналогом эллиптического уравнения

$$\Delta_2 u - u - \Phi = 0. \tag{43}$$

Наряду с этим уравнением рассмотрим эволюционную задачу

$$\frac{\partial w}{\partial t'} = \Delta_2 w - w - \Phi. \tag{44}$$

При $t' \to \infty$ решение эволюционного уравнения (44) стремится к решению стационарной задачи (43), т. е. $w \to u$. Поэтому вместо эллиптического уравнения (43) рассмотрим параболическое уравнение (44), стационарное решение которого и будет решением (43). Этот прием называется *счетом на установление*. Алгоритм решения параболических задач подробно будет описан далее (§ 4, п. 1).

При численном решении задачи (43) использовался следующий критерий установления стационара. Считалось, что

$$\frac{\partial w}{\partial t'} = 0, \quad \text{если} \quad \|\Delta_2 w - w - \Phi\|_c < \varepsilon. \tag{45}$$

В расчетах использовалось $\varepsilon = 10^{-14}$, что заведомо гарантировало необходимую точность расчетов. Меньшие значения ε не целесообразны, т. к. не всегда удается добиться выполнения условия (45) из-за ошибок округления. Для достижения условия (45) за минимальное число шагов был использован алгоритм выбора оптимального шага (§ 4, п. 2). При этом необходимое число действий и время работы программы снижались в десятки раз.

Тестирование точности метода. Предложенный алгоритм позволяет построить сеточное решение задачи (28). Точность предложенного метода была исследована путем сравнения с точным решением.

Если в задаче (29) выбраны начальные условия

$$u\big|_{t=0} = r^l e^{-r^2/2}[A\sin(l\varphi) + B\cos(l\varphi)], \quad u_t\big|_{t=0} = 0,$$

то решение может быть построено при помощи интегрального преобразования Ханкеля:

$$u = [A\sin(l\varphi) + B\cos(l\varphi)] \int_0^{+\infty} s^{l+1} J_l(sr) e^{-s^2/2} \cos\left(\frac{ts}{\sqrt{1+s^2}}\right) ds, \tag{46}$$

здесь r и φ — полярные координаты на плоскости, l — неотрицательное целое число.

Точное решение (46) можно использовать для тестирования предложенного метода. Для этого проводится стандартный расчет на сгущающихся сетках. Точность предложенной схемы $O(N^{-2}+M^{-2}+\tau^2)$; поэтому сетки по x, y и t сгущать надо в одно и то же число раз. Если число узлов квазиравномерной сетки (36) увеличить в два раза, то узлы новой сетки с четными номерами в точности совпадут с узлами более редкой сетки. Это позволит сравнивать решения на разных сетках. Мы провели расчеты с равным числом узлов пространственных сеток $N = M = 8,\ 16,\ 32,\ 64,\ 128$. Число узлов временной сетки подбиралось таким образом, чтобы шаг равномерной временной сетки был равен минимальному шагу пространственной сетки. Мерой погрешности служила разность точного решения (46) и численного решения в узлах сетки $\Delta^k_{nm} = u(x_n, y_m, t_k) - u^k_{nm}$.

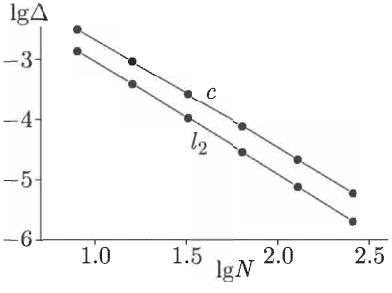

Рис. 8. Убывание погрешности численного решения начально-краевой задачи для уравнения ионно-звуковых волн с ростом числа узлов квазиравномерной сетки подтверждает порядок точности метода $O(N^{-2}+M^{-2}+\tau^2)$

На рисунке 8 показана погрешность численного решения $\|\Delta\|$ в сеточных нормах c и l_2:

$$\|\Delta\|_c = \max_{\substack{-N+1 \leqslant n \leqslant N-1 \\ -M+1 \leqslant m \leqslant M-1}} \left|\Delta^k_{nm}\right|, \tag{47}$$

$$\|\Delta\|^2_{l_2} = \sum_{n=-N+1}^{N-1} \sum_{m=-M+1}^{M-1} [\Delta^k_{nm}]^2 (x_{n+1/2} - x_{n-1/2})(y_{m+1/2} - y_{m-1/2}). \tag{48}$$

Результаты приведены для временного слоя k, соответствующего моменту времени $t = 1$. Убывание погрешности в обеих нормах с увеличением числа узлов соответствует порядку точности метода $O(N^{-2} + M^{-2} + \tau^2)$; график этой зависимости в двойном логарифмическом масштабе — прямая с коэффициентом наклона -2. Экономичность построенного метода подтверждает, что расчет численного решения по предложенному алгоритму требует на порядок меньше времени, чем вычисление с той же точностью несобственного интеграла в (46).

2. Консервативность схемы. Построенная разностная схема обладает важным практическим свойством консервативности. Пусть функции $s(x)$ и $g(x)$ заданы в узлах квазиравномерной сетки x_n, $-N \leqslant n \leqslant N$, и первая из них удовлетворяет нулевым краевым условиям $s_N = s_{-N} = 0$. Тогда для них справедлив следующий аналог интегрирования по частям:

$$\sum_{n=-N+1}^{N-1} s_n(g_{n+1/2} - g_{n-1/2}) = -\sum_{n=-N}^{N-1} 2(x_{n+3/4} - x_{n+1/4}) g_{n+1/2}[s_x]_{n+1/2}; \tag{49}$$

здесь

$$[s_x]_{n+1/2} = \frac{s_{n+1} - s_n}{2x_{n+3/4} - x_{n+1/4}}$$

есть аппроксимация (III.64) первой производной в полуцелых узлах квазиравномерной сетки, пригодная в неограниченной области.

Построенная разностная схема (37), (38) может быть записана следующим образом:

$$\frac{\Phi_{nm}^{k+1} - 2\Phi_{nm}^{k} + \Phi_{nm}^{k-1}}{\tau^2} + \Phi_{nm}^{k} + u_{nm}^{k} = 0, \tag{50}$$

$$\Phi_{nm}^{k} = \frac{[u_x]_{n+1/2,m}^{k} - [u_x]_{n-1/2,m}^{k}}{x_{n+1/2} - x_{n-1/2}} + \frac{[u_y]_{n,m+1/2}^{k} - [u_y]_{n,m-1/2}^{k}}{y_{m+1/2} - y_{m-1/2}} - u_{nm}^{k}, \tag{51}$$

$$-N+1 \leqslant n \leqslant N-1, \quad -M+1 \leqslant m \leqslant M-1, \quad 1 \leqslant k \leqslant K.$$

Подставим (51) в (50), домножим на величину $(x_{n+1/2} - x_{n-1/2}) \times$ $\times (y_{m+1/2} - y_{m-1/2})(u_{nm}^{k+1} - u_{nm}^{k-1})/(2\tau)$ и просуммируем по всем n и m. Используя дискретный аналог интегрирования по частям (49), получим

$$\frac{S^{k+1} - S^{k}}{\tau} = 0, \tag{52}$$

где

$$S^k = \frac{1}{2} \sum_{n=-N}^{N-1} \sum_{m=-M+1}^{M-1} \left(\frac{[u_x]_{n+1/2,m}^{k} - [u_x]_{n+1/2,m}^{k-1}}{\tau} \right)^2 \times$$

$$\times 2(x_{n+3/4} - x_{n+1/4})(y_{m+1/2} - y_{m-1/2}) +$$

$$+ \frac{1}{2} \sum_{n=-N+1}^{N-1} \sum_{m=-M}^{M-1} \left(\frac{[u_y]_{n,m+1/2}^{k} - [u_y]_{n,m+1/2}^{k-1}}{\tau} \right)^2 \times$$

$$\times 2(x_{n+1/2} - x_{n-1/2})(y_{m+3/4} - y_{m+1/4}) +$$

$$+ \frac{1}{2} \sum_{n=-N}^{N-1} \sum_{m=-M+1}^{M-1} [u_x]_{n+1/2,m}^{k} \cdot [u_x]_{n+1/2,m}^{k-1} \times$$

$$\times 2(x_{n+3/4} - x_{n+1/4})(y_{m+1/2} - y_{m-1/2}) +$$

$$+ \frac{1}{2} \sum_{n=-N+1}^{N-1} \sum_{m=-M}^{M-1} [u_y]^k_{n,m+1/2} \cdot [u_y]^{k-1}_{n,m+1/2} \times$$

$$\times 2(x_{n+1/2} - x_{n-1/2})(y_{m+3/4} - y_{m+1/4}) +$$

$$+ \frac{1}{2} \sum_{n=-N+1}^{N-1} \sum_{m=-M+1}^{M-1} \left(\frac{u^k_{nm} - u^{k-1}_{nm}}{\tau} \right)^2 \times$$

$$\times (x_{n+1/2} - x_{n-1/2})(y_{m+1/2} - y_{m-1/2}). \quad (53)$$

Согласно (53) величина S не изменяется при переходе на следующий временной слой. Отметим, что в непрерывном случае имеет место закон сохранения энергии

$$\mathcal{E} = \frac{1}{2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (u^2_{xt} + u^2_t + u^2_x + u^2_{yt} + u^2_y)\, dxdy = \text{const}. \quad (54)$$

Величина S в (53) аппроксимирует интеграл (54) на квазиравномерной сетке (36) с точностью $O(N^{-2} + M^{-2} + \tau^2)$. Отсюда можно было ожидать, что $S^k - S^0 = O(\tau^2)$, $1 \leqslant k \leqslant K$, однако выполняется строгое сохранение $S^k = S^0$. Таким образом, в построенной разностной схеме выполняется разностный аналог закона сохранения энергии (54). Построенная разностная схема обладает важным для адекватного описания физических процессов свойством консервативности, что было подтверждено расчетами.

3. Примеры расчета ионно-звуковых волн. В качестве иллюстрации возможностей метода был выполнен расчет волновой картины для задачи (29). Были взяты двумерная неограниченная область и начальные условия $u\big|_{t=0} = r_1 e^{-0.5r_1^2} \sin\varphi_1 + r_2^3 e^{-0.5r_2^2} \sin 3\varphi_2$, $u_t\big|_{t=0} = 0$, где величины r_1, r_2, φ_1, φ_2 выражаются через x и y с помощью соотношений $x - 3 = r_1 \cos\varphi_1$, $y = r_1 \sin\varphi_1$, $x + 3 = r_2 \cos\varphi_2$, $y = r_2 \sin\varphi_2$. Результаты расчета приведены на рисунке 9. Решение представлено в моменты времени $t = 1.0,\ 3.5,\ 7.0,\ 9.0$.

При помощи описанного метода также была решена начально-краевая задача в полуплоскости. Для этого использовалась квазиравномерная сетка, покрывающая полуплоскость $-\infty < x < \infty$, $0 \leqslant y < \infty$:

$$x_n = c_x \operatorname{tg}\left(\frac{\pi n}{2N}\right), \quad -N \leqslant n \leqslant N,$$
$$y_m = c_y \operatorname{tg}\left(\frac{\pi m}{2M}\right), \quad -M \leqslant m \leqslant M. \quad (55)$$

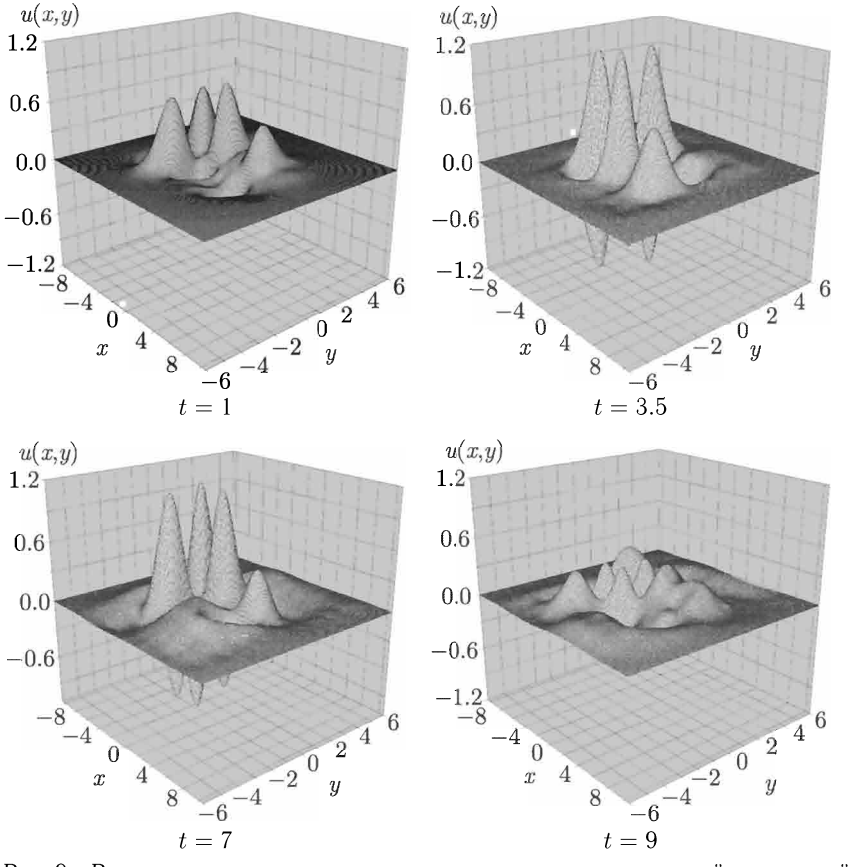

Рис. 9. Распространение ионно-звуковых волн в неограниченной двумерной области. Решение представлено в моменты времени $t = 1.0, 3.5, 7.0, 9.0$

Для задачи (30) в полуплоскости были взяты начальные условия $u(x, y, 0) = u_t(x, y, 0) = 0$ и краевой режим $u(x, 0, t) = \sin(0.5t) \times$ $\times\, (1 + x^2)^{-1}$. Результаты расчета представлены на рисунке 10 для моментов времени $t = 3.5, 7.0, 10.5, 14.0$.

4. Модификация метода для уравнения гравитационно-гироскопических волн. Предложенный метод без труда модифицируется для других уравнений составного типа. Мы применили его для уравнения гравитационно-гироскопических волн (32) и одного модельного псевдопараболического уравнения (33). Эти два примера иллюстрируют возможные видоизменения методики в случае анизотропного уравнения (32) или уравнения, содержащего лишь первую производную по времени (33).

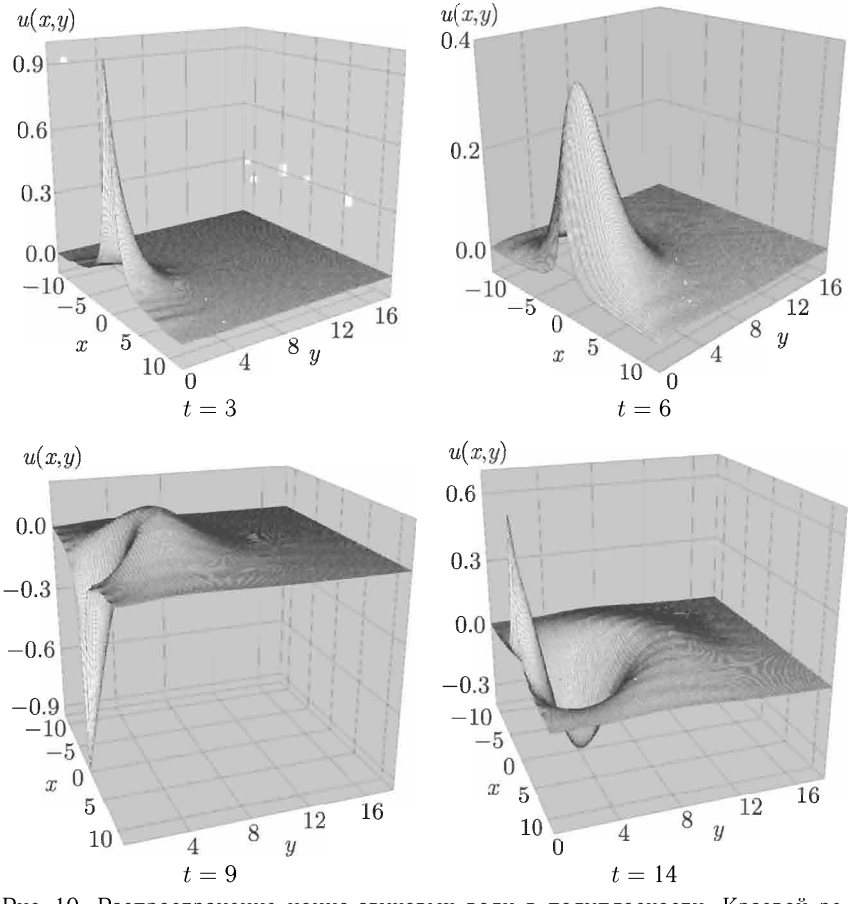

Рис. 10. Распространение ионно-звуковых волн в полуплоскости. Краевой режим $u(x,0,t) = \sin(0.5t)(1+x^2)^{-1}$. Решение представлено в моменты времени $t = 3.0,\ 6.0,\ 9.0,\ 14.0$

Рассмотрим начально-краевую задачу (32). Численный метод требует лишь небольшой модификации. Разобьем уравнение на два:

$$\Phi_{tt} + \lambda u_{xx} + \mu u_{xy} + \nu u_{xy} = 0, \tag{56}$$

$$\Delta_2 u \equiv u_{xx} + u_{yy} = \Phi. \tag{57}$$

Для решения задачи в полуплоскости используем двумерную квазиравномерную сетку (55). Временную сетку для простоты возьмем равномерной:

$$t_k = k\tau, \quad 0 \leqslant k \leqslant K, \quad \tau = \frac{T}{K} = \text{const}.$$

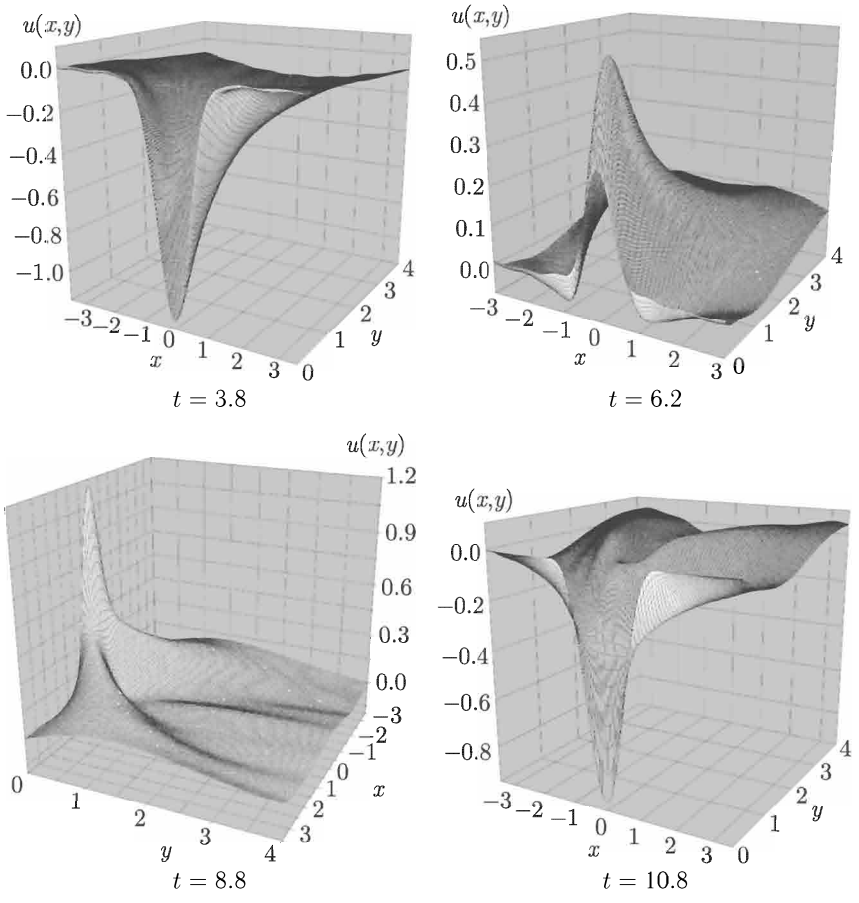

Рис. 11. Бегущая волна в стратифицированной вращающейся жидкости. Решение представлено в моменты времени $t = 3.8,\ 6.2,\ 8.8,\ 10.8$

Аппроксимируем уравнение (56) на равномерной по времени сетке с точностью $O(\tau^2)$:

$$\frac{\Phi_{nm}^{k+1} - 2\Phi_{nm}^{k} + \Phi_{nm}^{k-1}}{\tau^2} + \lambda\Lambda_{xx}[u]_{nm}^{k} + \mu\Lambda_{xy}[u]_{nm}^{k} + \nu\Lambda_{yy}[u]_{nm}^{k} = 0, \quad (58)$$

$$-N+1 \leqslant n \leqslant N-1, \quad 1 \leqslant m \leqslant M-1, \quad 0 \leqslant k \leqslant K-1.$$

Уравнение (57) аппроксимируется выражением

$$\Lambda_{xx}[u]_{nm}^{k+1} + \Lambda_{yy}[u]_{nm}^{k+1} = \Phi_{nm}^{k+1}, \quad (59)$$

$$-N+1 \leqslant n \leqslant N-1, \quad 1 \leqslant m \leqslant M-1, \quad 0 \leqslant k \leqslant K-1.$$

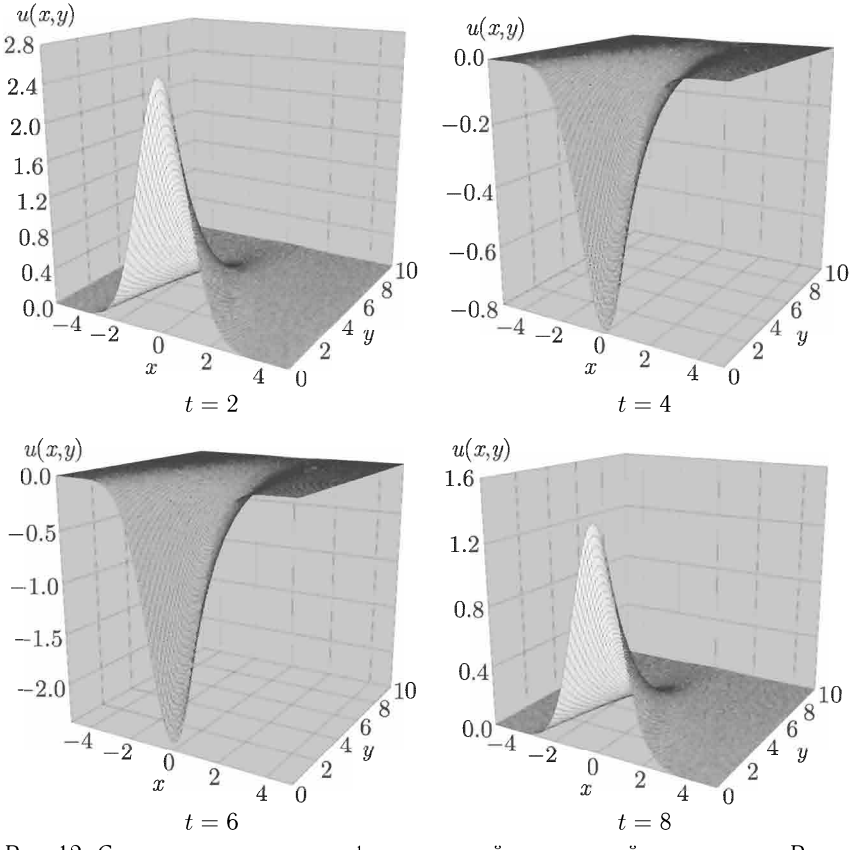

Рис. 12. Стоячая волна в стратифицированной вращающейся жидкости. Решение представлено в моменты времени $t = 2.0,\ 4.0,\ 6.0,\ 8.0$

В этой задаче необходимо построить аппроксимацию для второй смешанной производной на квазиравномерной сетке с порядком точности $O(N^{-2} + M^{-2})$:

$$\Lambda_{xy}[u]^k_{nm} = \frac{u^k_{n+1,m+1} + u^k_{n-1,m-1} - u^k_{n-1,m+1} - u^k_{n+1,m-1}}{4(x_{n+1/2} - x_{n-1/2})(y_{m+1/2} - y_{m-1/2})}, \qquad (60)$$

$$-N + 1 \leqslant n \leqslant N - 1, \quad 1 \leqslant m \leqslant M - 1, \quad 0 \leqslant k \leqslant K - 1.$$

Алгоритм построения решения полученных сеточных уравнений не претерпевает изменений. Сначала из уравнения (58) находим значение Φ^{k+1}_{nm} на следующем слое. Затем применяем счет на установление для нахождения u^{k+1}_{nm} из уравнения (60).

Известно, что в зависимости от соотношения параметров в задаче (32) возможно возбуждение двух типов волновых процессов: бегущих (рис. 11) и стоячих (рис. 12) волн.

5. Модельное псевдопараболическое уравнение. Теперь рассмотрим начально-краевую задачу для уравнения (33). Разобьем уравнение на два:

$$\Phi_t + \Phi + (1+\beta)\,u = 0, \tag{61}$$

$$\Delta_2 u - u = \Phi. \tag{62}$$

Уравнение (61) содержит лишь первую производную по времени, поэтому для получения точности $O(\tau^2)$ мы использовали схему типа предиктор-корректор:

$$\Phi_{nm}^{k+1} = \Phi_{nm}^{k} - \tau[\Phi_{nm}^{k} + (\beta+1)u_{nm}^{k}] + 0.5\tau^2[\Phi_{nm}^{k} + (\beta+1)u_{nm}^{k}]. \tag{63}$$

В остальном алгоритм построения численного решения полностью аналогичен описанному выше.

В качестве иллюстрации приведем численное решение задачи (33) для начального условия $f_0(x,y) = e^{-x^2-10y^2}\,y\sqrt{x^2+y^2}$. Учитывая то, что начальное условие затухает в направлении оси y гораздо быстрее, чем в направлении оси x, в расчетах целесообразно использовать разные масштабные коэффициенты c_x и c_y двумерной квазиравномерной сетки (36).

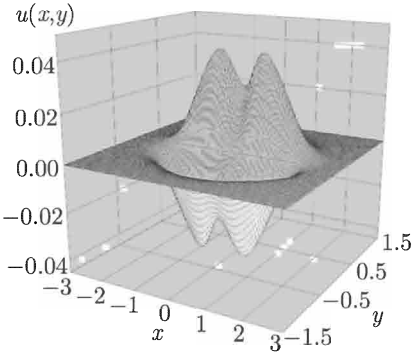

Рис. 13. Численное решение начально-краевой задачи в неограниченной области для уравнения $(\Delta_2 u - u)_t + \Delta_2 u + \beta u = 0$, $\beta = 2$. Решение представлено в момент времени $t = 0.5$

Результаты расчетов для момента времени $t = 0.5$ представлены на рисунке 13.

Глава VIII

ДВУМЕРНЫЕ ВЯЗКИЕ ТЕЧЕНИЯ

В этой главе демонстрируется использование квазиравномерных сеток для получения решения внутренних и внешних двумерных задач динамики вязкого газа с гарантированной точностью при наличии пристеночных пограничных слоев со значительными градиентами газодинамических переменных. Рассмотрены важные для практики задачи: прямая задача сопла Лаваля, а именно задача о расчете до- и сверхзвукового смешанного течения в сопле заданного контура при наличии минимального сечения (см., например, [Пирумов, Росляков, 1990]), и задача сверхзвукового обтекания гладкого затупленного тела (см., например, [Головачев, 1996]).

§ 1. Пограничный слой в сопле

1. Газодинамическая модель. Наиболее общей кинетически и термодинамически обоснованной континуальной моделью вязких течений является система уравнений Навье–Стокса. Расчет течения в канале (сопле) на основе этой модели представляет трудоемкую процедуру, особенно для протяженных каналов и при числах Рейнольдса $\mathrm{Re} \gg \gg 1$. В последнем случае целесообразно использовать упрощенные навье-стоксовские модели, обзор которых для внутренних течений дан, например, в [Лапин, Стрелец, 1989; Рогов, Соколова, 2002, а]. Эти модели основаны на композитных уравнениях, описывающих как вязкие, так и невязкие области течений. Вычислительное преимущество упрощенных навье-стоксовских моделей заключается в эволюционном характере уравнений по продольной координате, отсчитываемой в преимущественном направлении течения, и благодаря этому в возможности использовать для их интегрирования экономичные маршевые по продольной координате численные методы. Ниже рассмотрена одна из упрощенных навье-стоксовских моделей — параболическая модель гладкого канала. Эта модель адекватно описывает течения при умеренных и больших числах Re в каналах переменного сечения с умеренной продольной кривизной контура стенки канала [Рогов, Соколова, 1995, 1997; Rogov, Sokolova, 1998].

Параболическая модель гладкого канала получается упрощением полных уравнений Навье–Стокса, записанных в криволинейных ортогональных координатах (ξ, η, ζ), адаптированных к форме стенки

канала и его оси. Благодаря этому модель гладкого канала позволяет адекватно описать процессы молекулярного переноса в пограничных слоях около искривленных стенок канала. Использование геометрически адаптированных координат дает также возможность применения для численного интегрирования уравнений данной модели разностных схем высокого порядка аппроксимации.

В случае симметричного канала адаптированная система координат связана с декартовой (x, y, z) или цилиндрической (x, y, φ) системами координат (рис. 1) следующими соотношениями:

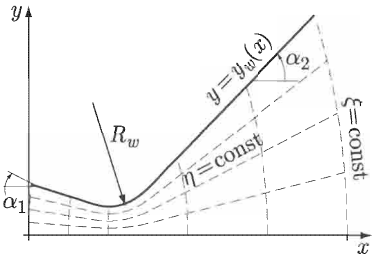

Рис. 1. Криволинейная ортогональная система координат, адаптированная к контуру стенки канала $y = y_w(x)$

$$\xi = \xi(x, y), \quad \eta = \frac{y}{y_w(x)},$$

$$\zeta = \begin{cases} z, & \text{плоский случай,} \\ \varphi, & \text{осесимметричный случай,} \end{cases} \tag{1}$$

где контур стенки канала задается гладкой функцией $y = y_w(x)$. Координата ξ является продольной и отсчитывается в направлении основного течения, η — поперечная координата, которая отсчитывается от оси канала. Здесь и далее индекс w приписан значениям величин на стенке.

Криволинейная координата $\xi = \xi(x, y)$ находится из условия совпадения координат ξ и x на оси канала $y = 0$ и условия ортогональности координатных линий $\xi = \text{const}$ и $\eta = \text{const}$, которое может быть выражено с помощью дифференциального уравнения в частных производных первого порядка:

$$\frac{y_w'(x)}{y_w(x)} \frac{\partial \xi}{\partial x} - \frac{1}{y} \frac{\partial \xi}{\partial y} = 0. \tag{2}$$

Интегрирование уравнения (2) при условии совпадения продольных координат ξ и x на оси канала дает неявное представление для функции $\xi(x, y)$:

$$\Phi(\xi) = \Phi(x) + \frac{y^2}{2}, \quad \Phi(x) = \int \frac{y_w(x)}{y_w'(x)}\, dx. \tag{3}$$

Параметры Ламе H_ξ, H_η, H_ζ ортогональной криволинейной системы координат (ξ, η, ζ) имеют вид:

$$H_\xi = \frac{y_w(\xi) y_w'(x)}{y_w(x) y_w'(\xi)} \frac{1}{\sqrt{1 + \eta^2 y_w'^2(x)}}, \quad H_\eta = \frac{y_w(x)}{\sqrt{1 + \eta^2 y_w'^2(x)}}, \quad H_\zeta = y^\nu, \tag{4}$$

где $\nu = 0$ для плоского случая и $\nu = 1$ для осесимметричного случая, а x как функция ξ и η находится из уравнения

$$\Phi(x) + \eta^2 \frac{y_w^2(x)}{2} - \Phi(\xi) = 0.$$

Функцию $x(\xi, \eta)$ можно найти также из решения следующей задачи Коши на отрезке $\eta \in [0, 1]$:

$$\frac{\partial x}{\partial \eta} = -\frac{\eta y_w(x) y'_w(x)}{1 + \eta^2 y'^2_w(x)}, \quad x = \xi \quad \text{при} \quad \eta = 0.$$

Кривизны продольных K_ξ и поперечных K_η координатных линий адаптированной системы координат даются следующими формулами:

$$\begin{aligned} K_\xi &= -\frac{1}{H_\eta}\frac{\partial \ln H_\xi}{\partial \eta} = \frac{\eta y''_w(x)}{[1 + \eta^2 y'^2_w(x)]^{3/2}}, \\ K_\eta &= -\frac{1}{H_\xi}\frac{\partial \ln H_\eta}{\partial \xi} = -\frac{y'_w(x)}{y_w(x)}\left[\frac{1}{\sqrt{1 + \eta^2 y'^2_w(x)}} - \eta y_w(x) K_\xi\right]. \end{aligned} \tag{5}$$

Запишем систему уравнений модели гладкого канала в безразмерной форме, исключив из нее плотность с помощью уравнения состояния совершенного газа:

$$\rho = \frac{\gamma p}{T}, \tag{6}$$

где ρ, p и T — плотность, статическое давление и температура газа соответственно.

В результате имеем:

уравнение неразрывности

$$\frac{\gamma H_\eta}{T}\left[p\frac{\partial u}{\partial \xi} - \frac{pu}{T}\frac{\partial T}{\partial \xi} + u\frac{\partial p}{\partial \xi}\right] = -\frac{\partial}{\partial \eta}(H_\xi \rho v) + H_\xi H_\eta K_\eta \frac{\gamma p u}{T}; \tag{7}$$

уравнение импульсов в проекции на ось ξ

$$\begin{aligned} H_\eta\left(\frac{\gamma p u}{T}\frac{\partial u}{\partial \xi} + \frac{\partial p}{\partial \xi}\right) &= -H_\xi \rho v\left(\frac{\partial u}{\partial \eta} - H_\eta K_\xi u\right) + \\ &+ \frac{1}{\mathrm{Re}_r}\left\{\frac{\partial}{\partial \eta}\left[\mu\frac{H_\xi}{H_\eta}\left(\frac{\partial u}{\partial \eta} + H_\eta K_\xi u\right)\right] - H_\xi K_\xi \mu\left(\frac{\partial u}{\partial \eta} + H_\eta K_\xi u\right)\right\}; \end{aligned} \tag{8}$$

уравнение импульсов в проекции на ось η

$$\frac{\partial p}{\partial \eta} + H_\eta K_\xi \frac{\gamma p u^2}{T} = 0; \tag{9}$$

уравнение энергии

$$\frac{\gamma H_\eta p u}{T}\left[(\gamma-1)u\frac{\partial u}{\partial \xi}+\frac{\partial T}{\partial \xi}\right]=-H_\xi \rho v\left[(\gamma-1)u\frac{\partial u}{\partial \eta}+\frac{\partial T}{\partial \eta}\right]+$$

$$+\frac{1}{\mathrm{Re}_r}\frac{\partial}{\partial \eta}\left\{\frac{H_\xi}{H_\eta}\left[(\gamma-1)\mu u\left(\frac{\partial u}{\partial \eta}+H_\eta K_\xi u\right)+\frac{\lambda}{\mathrm{Pr}}\frac{\partial T}{\partial \eta}\right]\right\}. \quad (10)$$

Здесь u и v — продольная и поперечная физические компоненты вектора скорости в адаптированной системе координат, μ и λ — коэффициенты динамической вязкости и теплопроводности газа. Уравнения выписаны для случая плоского течения в симметричном канале.

В уравнения (6)–(10) вошли безразмерные комплексы, построенные по масштабным величинам (с индексом $*$): $\gamma = c_{p*}/c_{v*}$ — отношение теплоемкостей при постоянном давлении и объеме (показатель адиабаты), $\mathrm{Re}_r = \rho_* u_* r_*/\mu_*$ — число Рейнольдса, $\mathrm{Pr} = \mu_* \mathbf{c}_{p*}/\lambda_*$ — число Прандтля. В качестве масштабных величин выбраны: для плотности и температуры — их значения на оси в начальном сечении канала (сопла); для компонент скорости — величина скорости звука, соответствующая масштабной температуре; $a_* = (\gamma R T_*)^{1/2}$, где R — газовая постоянная; для давления — величина $\rho_* u_*^2$; для переменных, имеющих размерность длины — полувысота r_* минимального (критического) сечения сопла.

Расчетная область ограничивается осью симметрии $\eta = 0$, твердой криволинейной стенкой $\eta = 1$, входным $\xi = \xi_{\mathrm{in}}$ и выходным $\xi = \xi_{\mathrm{out}}$ сечениями.

Граничные условия для решения прямой задачи сопла Лаваля в рамках параболической модели гладкого канала формулируются следующим образом.

Для уравнений (8) и (10) второго порядка относительно η задаются следующие краевые условия: прилипания на стенке $u = 0$; либо теплоизоляции стенки $\partial T/\partial \eta = 0$, либо фиксированной температуры $T = T_w$; симметрии на оси $\partial u/\partial \eta = \partial T/\partial \eta = 0$. Для уравнений (7), (9) первого порядка относительно η на стенке задается условие $v = 0$, а на оси симметрии — $v = 0$, $\varphi = 1$.

Во входном сечении, расположенном в дозвуковой области, для эволюционных по продольной координате уравнений (7), (8), (10) задаются профили u/u_a (u_a — скорость на оси), T как функции поперечной координаты η. Профиль p в этом сечении рассчитывается из уравнения (9) при его заданном значении p_a на оси.

Осевое значение скорости u_a и критический расход газа Q_{in} через сопло [Пирумов, Росляков, 1990] определяются из дополнительного условия на правой границе дозвуковой области, т. е. на звуковой линии. Это граничное условие есть условие гладкого продолжения решения через звуковую линию $\mathrm{M}_\xi = 1$, на которой вырождается эволюционная матрица коэффициентов при продольных градиентах u, T и p в подси-

стеме уравнений (8), (10) и (7):

$$H_\eta \begin{bmatrix} \gamma pu/T & 0 & 1 \\ \gamma(\gamma-1)pu^2/T & \gamma pu/T & 0 \\ \gamma p/T & -\gamma pu/T^2 & \gamma u/T \end{bmatrix}. \tag{11}$$

Детерминант этой матрицы пропорционален $u^2/T - 1 = \mathrm{M}_\xi^2 - 1$ и равен нулю при $\mathrm{M}_\xi = 1$, где $\mathrm{M}_\xi = u/T^{1/2}$ — локальное число Маха вдоль направления ξ. Эта математическая особенность уравнений модели гладкого канала, описывающей двумерные вязкие течения при умеренных и больших числах Re, аналогична особенности уравнений для одномерных течений невязкого газа (см. [Пирумов, Росляков, 1990], [Абрамович, 1991]) и приводит к ветвлению численных интегральных кривых в трансзвуковой области течения. Из этого ветвления можно определить значение критического расхода Q_{in}, которому соответствует смешанное до- и сверхзвуковое течение газа в сопле Лаваля. Для нахождения Q_{in} можно использовать либо метод дихотомии [Калиткин и др., 1999], либо более быстрый метод парабол [Рогов, 2001].

Анализ системы уравнений параболической модели гладкого канала методом характеристик показывает, что ее математический тип — параболический. Поэтому для численного интегрирования данной системы уравнений использована маршевая (эволюционная по времениподобной координате ξ) разностная схема. Производные по поперечной координате η аппроксимированы конечными разностями на трехточечном шаблоне с четвертым порядком точности, как и в [Калиткин и др., 1999]. Производные по продольной координате ξ аппроксимированы левыми разностями. При этом первый маршевый шаг осуществляется по трехстадийной схеме, использующей на каждой стадии схему первого порядка точности по ξ и повышение точности до второго порядка на последней стадии с помощью процедуры экстраполяции по Ричардсону. Дальнейшее маршевое интегрирование осуществляется по трехслойной схеме, в которой продольная производная от искомой функции f на текущем $(n+1)$-м маршевом слое аппроксимирована согласно формуле

$$\left(\frac{\partial f}{\partial \xi}\right)_{n+1} = \frac{f_{n+1}-f_n}{h_{n+1}} + \frac{h_{n+1}}{h_n+h_{n+1}}\left(\frac{f_{n+1}-f_n}{h_{n+1}} - \frac{f_n - f_{n-1}}{h_n}\right) + O(h_\xi^2), \tag{12}$$

где $h_n = \xi_n - \xi_{n-1}$. Отметим, что трехслойная маршевая схема оказалась более устойчивой по отношению к высокочастотным возмущениям, чем двухслойная схема [Калиткин и др., 1999]. Из-за особенности эволюционной матрицы (11) на звуковой линии маршевая схема регуляризуется в области смешанного до- и сверхзвукового течения [Калиткин и др., 1999].

Заметим, что к точности расчета характеристик сопел предъявляются достаточно жесткие требования. Например, хорошо известно, что первые неудачные запуски французской ракеты “Ариан” были связаны с прогаром стенок сопел двигателя. Поэтому важно надежно рассчиты-

вать тепловые потоки на стенки сопла, которые достигают максимальных значений в области горла сопла [Калиткин и др., 1997]. Выигрыш в тяге сопла на 1.5 % равносилен удвоению полезной нагрузки, которую может вывести на орбиту летательный аппарат.

Рассмотрим применение квазиравномерных сеток для расчета характеристик сопел Лаваля с гарантированной точностью. В качестве примера исследуем ламинарное смешанное до- и сверхзвуковое течение однородного совершенного газа (воздуха) в плоском сопле с заданным гиперболическим контуром. Последний характеризуется его безразмерной кривизной $K_w = 0.5$ в минимальном (критическом) сечении сопла $\xi = 0$ и углами наклона левой и правой асимптот гиперболы к линии симметрии сопла: $\alpha_{1,\lim} = 30°$, $\alpha_{2,\lim} = 20°$ (см. рис. 1). Расчет проводился от входного сечения сопла $\xi = \xi_{\text{in}} = -15$ до выходного сечения $\xi = \xi_{\text{out}} = 50$.

Течение характеризуется числом Рейнольдса $\text{Re}_r \geqslant 50$, а газ — следующими параметрами: $\text{Pr} = 0.71$, $\gamma = 1.4$, показатель степени в зависимости вязкости от температуры равен 0.76.

2. Квазиравномерные сетки. Шаг разностной сетки выбран равномерным в продольном направлении и неравномерным (со сгущением к стенке при $\text{Re}_r \gg 1$) в поперечном. Узлы сетки в поперечном направлении задаются следующими формулами:

$$\eta_m = \frac{1 - \chi\tau^2}{1 - \chi}\tau, \quad \tau = \frac{m}{N_\eta}, \quad \chi = \frac{\text{Re}_r}{3(20 + \text{Re}_r)}, \tag{13}$$

где номер узла $m = 0$ соответствует оси сопла, а номер $m = N_\eta$ — его стенке; N_η — число интервалов разностной сетки в поперечном направлении. Сетка была выбрана таким образом, что при $\text{Re}_r \to 0$ она становится равномерной, а при $\text{Re}_r \to \infty$ она сгущается вблизи стенки. Рисунок 2 демонстрирует положение узлов неравномерной по η сетки на примере профиля продольной скорости во входном сечении сопла при $\text{Re}_r = 400$ и $N_\eta = 12$ и 24. При увеличении числа интервалов N_η вдвое имеет место вложение менее подробной (грубой) сетки в более подробную (мелкую).

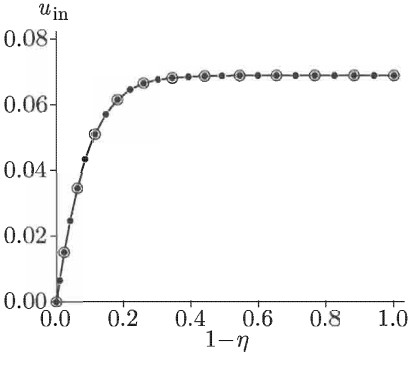

Рис. 2. Расположение узлов неравномерной сетки в поперечном направлении на примере профиля продольной скорости во входном сечении сопла. $\text{Re}_r = 400$, $\circ$ — $N_\eta = 12$, $\bullet$ — $N_\eta = 24$

Сетка (13) является квазиравномерной. Это позволяет увеличить точность решения экстраполяцией по Ричардсону при сгущении сетки в любое число раз, не обязательно целое [Калиткин, 1978].

3. Сеточная сходимость и точность схемы. Скорость сходимости численного решения к решению системы дифференциальных уравнений (7)–(10) и фактическая точность разностной схемы проверялись на вложенных сетках.

На рисунке 3 показана скорость сходимости значения критического расхода, рассчитанного на сетке с числом узлов N, к пределу $N \to \infty$. Этот предел определялся с десятью верными знаками путем расчетов на очень подробных сетках с экстраполяцией по Ричардсону (см. гл. II–III). Вычитание этого предела из расчетного Q_{in} при конкретном N давало погрешность этого расчета. Зависимость погрешности от N изображена на рисунке 3 в двойном логарифмическом масштабе.

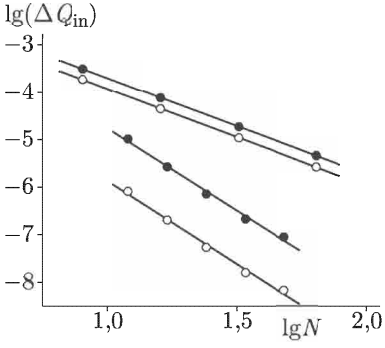

Рис. 3. Убывание погрешности при вычислении критического значения расхода Q_{in} при сгущении сетки. Кружки показывают результаты расчетов при $\text{Re}_r = 50$, точки — при $\text{Re}_r = 800$. Сплошные линии представляют линейные аппроксимации по методу наименьших квадратов результатов расчетов. Две верхние линии соответствуют сгущению по продольной координате, две нижние — по поперечной

Сгущение сетки проводилось как отдельно по продольной и поперечной координатам, так и совместно. Эффективный порядок точности при сгущении сетки в продольном направлении оказался равным 2.03 и 2.02 при $\text{Re}_r = 50$ и 800 соответственно, а при сгущении сетки в поперечном направлении — равным 4.00 и 3.96 при $\text{Re}_r = 50$ и 800. Таким образом, фактические порядки точности разностного решения близки к теоретическим порядкам аппроксимации разностной схемы. Сгущение сетки по продольной координате проводилось при фиксированном числе интервалов в поперечном направлении $N_\eta = 34$, а сгущение по поперечной координате — при фиксированном числе интервалов в продольном направлении $N_\xi = 64$.

Числа интервалов разностной сетки при ее сгущении одновременно в обоих направлениях задавались равными $(N_\xi, N_\eta) = (8, 12)$, $(16, 17)$, $(32, 24)$ и $(64, 34)$. При этом число шагов по ξ точно удваивается, а по η увеличивается в $\sqrt{2}$ раз с точностью 0.2% (несколько худшую точность 1.0% давала бы последовательность $N_\eta = 5,\ 7,\ 10,\ 14,\ \ldots$).

На рисунке 3 абсцисса N равна N_ξ в случае сгущения сетки в продольном направлении или по обоим направлениям одновременно и равна N_η в случае сгущения сетки в поперечном направлении. Верхние линии на рисунке соответствуют сгущению сетки по продольной координате. С этими линиями графически совпадают линии, соответствующие сгущению сетки по обоим координатам. Нижние линии соответствуют сгущению сетки по поперечной координате. Отношение чисел интервалов N_ξ/N_η, при которых достигается одинаковая величина невязки в значении расхода, примерно равно 4.0 при достаточно малых шагах сетки.

Рисунок 4 показывает скорость сходимости невязки в величине расхода вдоль сопла при совместном сгущении сетки по продольному и поперечному направлениям. Видно, что невязка убывает в 4 раза при уменьшении шага по ξ в 2 раза. Расчет показывает, что локальные максимумы в невязке по расходу (см. рис. 4) достигаются в горле сопла и коррелируют с локальными максимумами скорости изменения кривизны контура сопла.

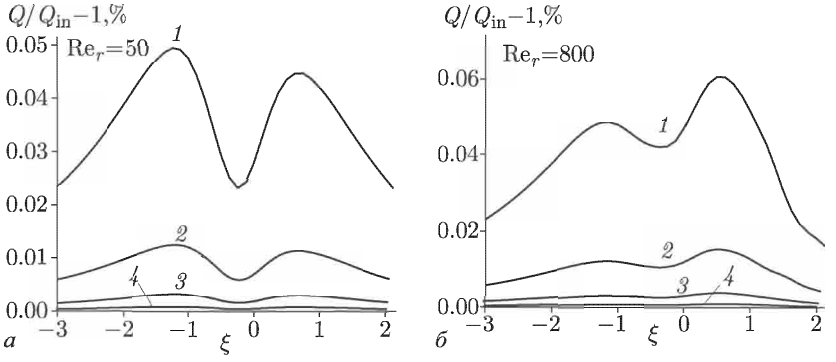

Рис. 4. Убывание погрешности в величине расхода Q вдоль сопла при сгущении сетки в продольном и поперечном направлениях одновременно; *а*) $\text{Re}_r = 50$, *б*) $\text{Re}_r = 800$. Кривые *1–4* рассчитаны при продольных шагах $h_\xi = 1/8$, $1/16$, $1/32$ и $1/64$ соответственно

Представленные расчеты выполнены на равномерной по ξ сетке. Однако с точки зрения экономии ресурсов ЭВМ более выгодны расчеты на неравномерной продольной сетке. Поскольку невязка в расходе является максимальной в области горла сопла, то в расчетах продольный шаг имеет смысл увеличивать от горла к входному и выходному участкам сопла, на которых кривизна мала и изменяется медленно. Подходящим будет следующий выбор продольного шага:

$$h_{\xi,n} = y_w^2(\xi_n)\, h_{\xi,\min}, \tag{14}$$

где $h_{\xi,\min}$ есть продольный шаг в минимальном сечении сопла, ξ_n — n-й узел разностной сетки.

§ 2. Пограничный слой при обтекании

1. Газодинамическая модель. Будем рассматривать стационарное течение вязкого, теплопроводного совершенного газа около гладкого осесимметричного или плоского затупленного тела, обтекаемого равномерным сверхзвуковым потоком под нулевым углом атаки. При умеренных и больших числах Re такое течение адекватно описывается системой уравнений полного вязкого ударного слоя [Davis, 1970]. Для задач сверхзвукового обтекания решения этой системы близки к решениям полной системы уравнений Навье–Стокса [Белоцерковский (ред.), 1974; Громов и др., 1999]. Для задач гиперзвукового обтекания затупленных тел газовым потоком с числом Маха $M_\infty > 3$ основной вклад в решение вносит гиперболо-параболическая часть системы уравнений полного вязкого ударного слоя. Эта часть данной системы уравнений составляет математическую основу модели гиперболического вязкого ударного слоя [Рогов, Соколова, 2001 и 2002, б]. Рассматривая далее течения с числами $M_\infty > 3$, в качестве определяющей модели возьмем модель гиперболического вязкого ударного слоя.

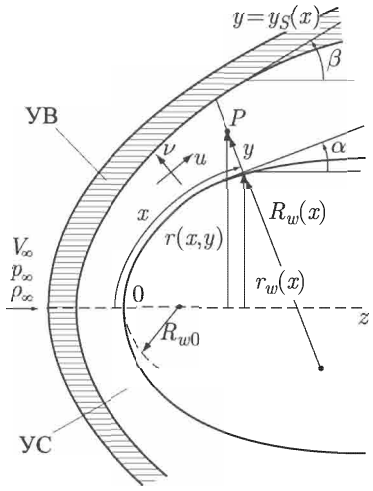

Рис. 5. Картина обтекания затупленного тела сверхзвуковым потоком вязкого газа

Модель гиперболического вязкого ударного слоя, так же как и более сложная модель полного вязкого ударного слоя, является двухслойной. Возмущенная область течения около обтекаемого тела разбивается на структуру головной ударной волны и вязкий ударный слой — область между ударной волной и поверхностью тела (см. рис. 5). В рамках этой модели навье-стоксовскую структуру ударной волны можно определить после расчета основной возмущенной области течения — ударного слоя.

Уравнения модели гиперболического вязкого ударного слоя, описывающие течение в ударном слое около плоского или осесимметричного затупленного тела, обтекаемого под нулевым углом атаки, в переменных (x, η) имеют следующий вид:

$$\frac{\partial}{\partial x}(r^\nu \rho u) - \frac{\eta y_s'}{y_s}\frac{\partial}{\partial \eta}(r^\nu \rho u) + \frac{1}{y_s}\frac{\partial}{\partial \eta}(r^\nu H_1 \rho v) = 0, \tag{15}$$

$$u\frac{\partial u}{\partial x} + \frac{1}{\rho}\left(\omega\frac{\partial p}{\partial x} + (1-\omega)\frac{p}{p_s}\frac{dp_s}{dx}\right) + \frac{V}{y_s}\frac{\partial u}{\partial \eta} - \frac{\eta y_s'}{y_s \rho}\frac{\partial p}{\partial \eta} + K_w uv =$$

$$= \frac{1}{r^\nu H_1 y_s^2 \rho} \frac{\partial}{\partial \eta} \left[r^\nu H_1^2 \mu \left(\frac{\partial u}{\partial \eta} - \frac{K_w y_s u}{H_1} \right) \right], \quad (16)$$

$$u \frac{\partial v}{\partial x} + \frac{V}{y_s} \frac{\partial v}{\partial \eta} + \frac{H_1}{y_s \rho} \frac{\partial p}{\partial \eta} - K_w u^2 = 0, \quad (17)$$

$$u \frac{\partial}{\partial x} \left[h + \frac{1}{2}(u^2 + v^2) \right] + \frac{V}{y_s} \frac{\partial}{\partial \eta} \left[h + \frac{1}{2}(u^2 + v^2) \right] =$$

$$= \frac{1}{r^\nu \rho y_s^2} \frac{\partial}{\partial \eta} \left\{ r^\nu H_1 \left[\lambda \frac{\partial T}{\partial \eta} + \mu u \left(\frac{\partial u}{\partial \eta} - \frac{K_w y_s u}{H_1} \right) \right] \right\}. \quad (18)$$

Здесь

$$r = r_w + y \cos \alpha, \quad r_w = \int_0^x \sin \alpha dx, \quad H_1 = 1 + K_w y,$$

$$y_s' = \frac{dy_s}{dx}, \quad K_w = -\frac{d\alpha}{dx}, \quad V = H_1 v - \eta y_s' u, \quad (19)$$

$$\omega = \min \left\{ 1, \ \sigma \frac{\gamma \mathrm{M}_x^2}{1 + (\gamma - 1)\mathrm{M}_x^2} \right\}, \quad 0 \leqslant \sigma < 1.$$

В этих уравнениях $\eta = y/y_s(x)$ есть нормированная поперечная координата; x, y — естественные ортогональные координаты, связанные с поверхностью обтекаемого тела (см. рис. 5); y_s — отход головной ударной волны от поверхности тела $y = 0$; r — расстояние до оси (плоскости) симметрии тела; r_w и K_w — контур поверхности обтекаемого тела и его кривизна; α — угол между касательной к поверхности и осью (плоскостью) симметрии тела; H_1 — коэффициент Ламе; u, v — касательная и нормальная к поверхности тела составляющие вектора скорости; h — удельная энтальпия газа; μ и λ — коэффициенты вязкости и теплопроводности; $\nu = 0$ или 1 для плоского или осесимметричного течения соответственно. Нижние индексы e и w приписаны значениям величин непосредственно за ударной волной и на поверхности тела. В формуле (19) коэффициент σ задан и близок к единице; γ — показатель адиабаты; M_x — определенное по продольной составляющей скорости локальное значение числа Маха.

Производная от контура ударной волны, входящая в уравнения (15)–(18), связана с ее искомым отходом y_s геометрическим соотношением

$$\frac{dy_s}{dx} = \operatorname{tg}(\beta - \alpha) H_s, \quad H_s = 1 + K_w y_s, \quad (20)$$

где β — угол между касательной к ударной волне и осью симметрии тела (см. рис. 5).

Уравнения (15)–(20) замыкаются уравнением состояния совершенного газа

$$p = \rho R T \quad (21)$$

и зависимостями энтальпии h, коэффициентов вязкости μ и теплопроводности λ от температуры.

Система уравнений модели гиперболического вязкого ударного слоя получается из системы уравнений модели полного вязкого ударного слоя в результате приближения для продольного градиента давления в уравнении продольного импульса (16) в дозвуковых областях течения [Рогов, Соколова, 2002, б]. В сверхзвуковых областях уравнения этих моделей совпадают.

В уравнениях модели гиперболического вязкого ударного слоя имеется особенность на оси симметрии течения $x = 0$. На этой оси продольная скорость u и продольные градиенты v, ρ, p равны нулю. Чтобы разрешить эту особенность, вводятся растянутая в окрестности оси симметрии продольная координата

$$\xi = \int_0^x \cos\alpha \, dx \tag{22}$$

и нормированная продольная скорость

$$\overline{u} = \frac{u}{\cos\alpha}. \tag{23}$$

Введенная координата ξ совпадает с цилиндрической координатой z в случае осесимметричного течения (см. рис. 5) или с соответствующей декартовой координатой, отсчитываемой вдоль плоскости симметрии от передней точки торможения потока (точка O на рис. 5), в случае плоского течения.

Уравнения (15)–(20) в новых переменных в безразмерной форме примут следующий вид:

$$g\frac{\partial}{\partial\xi}(\rho u) + \frac{\partial}{\partial\eta}(f\rho) + h\rho u = 0, \tag{24}$$

$$\begin{aligned} g\frac{\partial}{\partial\xi}(\rho u^2) + \tau y_s\left(\omega\frac{\partial p}{\partial\xi} + (1-\omega)\frac{p}{p_s}\frac{dp_s}{d\xi}\right) + \\ + \frac{\partial}{\partial\eta}\left[f\rho u - \tau\eta y_s' p - \frac{\tau H_1\mu}{\mathrm{Re}_\infty}\left(\frac{1}{y_s}\frac{\partial u}{\partial\eta} - \frac{K_w u}{H_1}\right)\right] - \\ - \frac{\tau K_w y_s\mu}{\mathrm{Re}_\infty}\left(\frac{1}{y_s}\frac{\partial u}{\partial\eta} - \frac{K_w u}{H_1}\right) + \\ + (h + \tau K_w y_s\sin\alpha)\rho u^2 + \tau K_w y_s\rho uv + (2\tau - 1)y_s' p = 0, \end{aligned} \tag{25}$$

$$g\frac{\partial}{\partial\xi}(\rho uv) + \frac{\partial}{\partial\eta}(f\rho v + \tau H_1 p) - gK_w\rho u^2 + h\rho uv - qy_s p = 0, \tag{26}$$

$$g\frac{\partial}{\partial\xi}(\rho u T^*) + h\rho u T^* +$$

$$+ \frac{\partial}{\partial\eta}\left\{ f\rho T^* - \frac{\tau H_1 \mu}{\mathrm{Re}_\infty}\left[\frac{1}{y_s\,\mathrm{Pr}}\frac{\partial T}{\partial\eta} + u\cos^2\alpha\left(\frac{1}{y_s}\frac{\partial u}{\partial\eta} - \frac{K_w u}{H_1}\right)\right]\right\} = 0, \tag{27}$$

$$T^* = T + \frac{1}{2}(u^2\cos^2\alpha + v^2),$$

$$\frac{dy_s}{d\xi} = \frac{\mathrm{tg}(\beta-\alpha)}{\cos\alpha}(1 + K_w y_s); \tag{28}$$

здесь

$$\tau = \left(\frac{r}{r_w}\right)^\nu, \quad y_s' = \frac{dy_s}{d\xi}, \tag{29}$$

$$f = \tau(H_1 v - \eta y_s' u\cos^2\alpha), \quad g = \tau y_s\cos^2\alpha,$$
$$h = (2\tau - 1)y_s'\cos^2\alpha + q y_s\sin\alpha, \quad q = \nu H_1\frac{\cos\alpha}{r_w} + \tau K_w. \tag{30}$$

В уравнениях (24)–(28) и формулах, приведенных ниже, черта над величиной u опущена. Все величины, имеющие размерность длины, отнесены к радиусу затупления R_{w0}, компоненты вектора скорости — к V_∞, плотность — к ρ_∞, давление — к $\rho_\infty V_\infty^2$, температура — к V_∞^2/c_p, коэффициент вязкости — к μ_∞. Нижние индексы ∞ и 0 приписаны значениям величин в набегающем потоке и на оси (плоскости) симметрии. При написании уравнений (24)–(28) предполагалось, что теплоемкости газа и число Прандтля Pr постоянны.

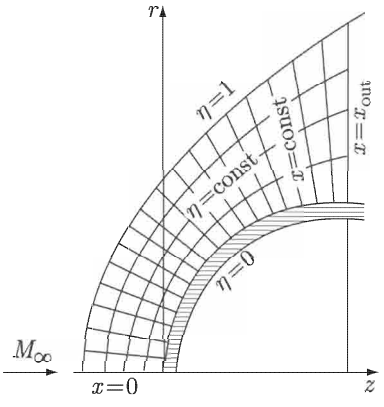

Рис. 6. Расчетная область

Расчетная область ограничивается твердой криволинейной стенкой $\eta = 0$, головной ударной волной $\eta = 1$, осью симметрии $x = 0$ и выходным сечением $x = x_{\mathrm{out}}$, расположенным в области преимущественно сверхзвукового течения (см. рис. 6).

Краевыми условиями для уравнений (24)–(28) относительно переменных u, v, T, p являются: на ударной волне (т. е. при $\eta = 1$) — три из четырех обобщенных соотношений Ренкина–Гюгонио для u_s, v_s, T_s [Головачев, 1996; Тирский, 1997]; на обтекаемой поверхности (при $\eta = 0$) — заданная температура, условия прилипания и непротекания для компонент скорости. Оставшееся (четвертое) соотношение Ренкина–Гюгонио

$$p_s = \frac{1}{\gamma \mathrm{M}_\infty^2} + \left(1 - \frac{\gamma-1}{\gamma}\frac{T_s}{p_s}\right)\sin^2\beta \tag{31}$$

связывает угол β с p_s и T_s и совместно с уравнением (28) служит для определения неизвестного отхода ударной волны y_s.

Начальные условия для искомых функций на оси (плоскости) симметрии определяются из решения системы обыкновенных дифференциальных уравнений, в которые при $\xi = 0$ вырождаются уравнения (24)–(28). При этом распределения искомых функций на оси симметрии, а также значение отхода y_{s0} зависят от величины кривизны ударной волны K_{s0}.

Кривизна ударной волны K_{s0} определяется из дополнительного условия на правой границе дозвуковой области (звуковой линии $\mathrm{M}_x = 1$). Детерминант эволюционной матрицы при продольных градиентах u, v, T, p_s в уравнениях гиперболического вязкого ударного слоя, как и детерминант матрицы при градиентах u, v, T, p в уравнениях полного вязкого ударного слоя, на звуковой линии равен нулю. Вследствие плохой обусловленности эволюционной матрицы, интегральные кривые уравнений гиперболического вязкого ударного слоя, соответствующие различным значениям K_{s0}, ветвятся в окрестности звуковой линии. Подобное поведение интегральных кривых имеет место и для уравнений, описывающих вязкое смешанное течение в сопле Лаваля (см. § 1). В случае внутренних течений аналогом величины K_{s0} является величина расхода газа Q_{in}. Аналогично существованию единственного значения критического расхода [Пирумов, Росляков, 1990], для уравнений гиперболического вязкого ударного слоя также существует некоторое "критическое" значение K_{s0}, которому соответствует единственная (предельная) интегральная кривая, которая может быть гладко продолжена за звуковую линию. Эта интегральная кривая и есть искомое решение задачи.

Анализ системы уравнений гиперболического вязкого ударного слоя методом характеристик показывает, что она имеет смешанный гиперболо-параболический тип. Для численного интегрирования данной системы уравнений использована та же маршевая разностная схема, что и для интегрирования уравнений параболической модели гладкого канала (§ 1).

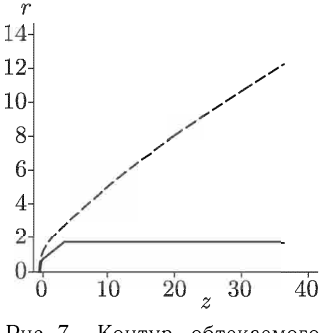

Рис. 7. Контур обтекаемого тела (сплошная линия) и головная ударная волна (штриховая линия)

Рассмотрим применение квазиравномерных сеток для расчета с гарантированной точностью характеристик взаимодействия сверхзвукового газового потока с обтекаемым телом. В качестве примера рассмотрим ламинарное обтекание длинного затупленного тела (ракеты) однородным совершенным газом (воздухом) с $\gamma = 1.4$, $\mu \sim T^{0.5}$, $\mathrm{Pr} = 0.7$; численное значение коэффициентов этих зависимостей для воздуха брались из справочников. Безразмерный контур тела изоб-

ражен на рисунке 7 сплошной кривой. Он рассчитывался по заданной гладкой монотонной функции $\alpha = \alpha(x)$. Течение характеризуется числами $M_\infty = 5$ и $\mathrm{Re}_\infty = 10^4$.

2. Квазиравномерные сетки. Расчетная область в координатах (ξ, η) является прямоугольником. Шаг разностной сетки выбран равномерным по продольной координате ξ и неравномерным по поперечной координате η (со сгущением к стенке при $\mathrm{Re}_\infty \gg 1$).

Узлы сетки в поперечном направлении задаются следующими формулами:

$$\eta_m = 1 - \frac{1 - \chi\tau^2}{1 - \chi}\tau, \quad \tau = 1 - \frac{m}{N_\eta} \geqslant 0, \quad \chi = \frac{\mathrm{Re}_0}{3(20 + \mathrm{Re}_0)}, \tag{32}$$

где номер узла $m = 0$ соответствует стенке (поверхности тела), а номер $m = N_\eta$ — головной ударной волне; здесь N_η есть число интервалов разностной сетки в поперечном направлении; число Рейнольдса Re_0 связано с числом Re_∞ формулой

$$\mathrm{Re}_0 = \frac{\mu_\infty}{\mu(T_0)}\mathrm{Re}_\infty, \tag{33}$$

где T_0 — температура адиабатически заторможенного набегающего потока. Сетка была выбрана таким образом, что при $\mathrm{Re}_0 \to 0$ она становится равномерной, а при $\mathrm{Re}_0 \to \infty$ она неограниченно сгущается вблизи стенки. Таким образом, по поперечной координате η сетка являлась квазиравномерной.

3. Сеточная сходимость и точность схемы. Рассчитанный контур головной ударной волны изображен на рисунке 7. На рисунке 8 показаны важнейшие характеристики взаимодействия высокоскоростного газового потока с обтекаемым телом — коэффициенты трения C_f и конвективной теплопередачи C_H, которые определяются формулами

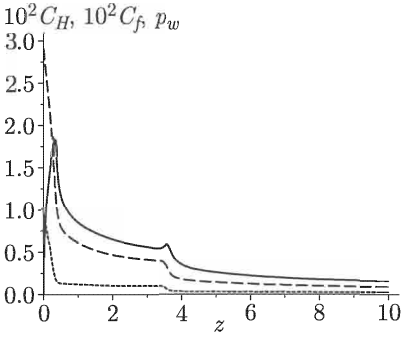

Рис. 8. Величины вдоль поверхности обтекаемого тела: давление p_w — пунктирная, $C_f \times 10^2$ — штриховая, $C_H \times 10^2$ — сплошная линия

$$C_f = \frac{2\tau_w}{\rho_\infty V_\infty^2}, \qquad C_H = \frac{q_w}{\rho_\infty V_\infty (H_\infty - h_w)}, \tag{34}$$

где τ_w — вязкое трение на стенке, q_w — тепловой поток в стенку, $H = h + (u^2 + v^2)/2$ — полная удельная энтальпия газа. Из рисунка видно, что значения давления на поверхности и теплового потока в стенку достигают максимума в передней критической точке тела.

На рисунках 9–11 показаны скорости сходимости значений искомых величин при сгущении сеток к точным значениям; последние определяются экстраполяцией по Ричардсону значений, полученных для двух самых подробных сеток. Сгущение разностной сетки проводилось по продольной и поперечной координате как раздельно, так и совместно. Сгущение в продольном направлении проводилось на последовательности сеток с $N_\xi = N_x = 24$, 34, 48 и 68; здесь $N_x = 1/h_x$, а h_x — постоянный шаг по переменной x. По поперечной координате η сгущение проводилось на последовательности сеток с $N_\eta = 24$, 34, 48, 68 и 96. Таким образом, число шагов при сгущении сетки отдельно по разным направлениям увеличивается в $\sqrt{2}$ раз с точностью 0.2 %. Количество интервалов разностной сетки при ее сгущении совместно в обоих направлениях задавалось равным $(N_\xi, N_\eta) = (24, 48)$, $(34, 57)$, $(48, 68)$ и $(68, 81)$. При таком сгущении количество интервалов сетки по ξ увеличивалось примерно в $\sqrt{2}$ раз, а по η — примерно в $\sqrt[4]{2}$.

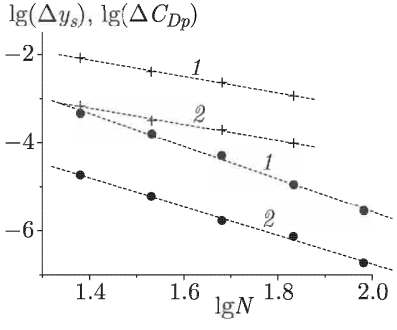

Рис. 9. Убывание погрешности при вычислении значения y_s ($x = 35$) (цифра *1* около линий) и коэффициента сопротивления давления C_{Dp} (цифра *2*) при сгущении продольной сетки (две верхние линии) и поперечной сетки (нижние линии). Кружки и крестики — численные расчеты, линии — аппроксимации зависимостей

На рисунках 9–11 абсцисса N равна N_ξ в случае сгущения сетки в продольном направлении или по обоим направлениям одновременно и равна N_η в случае сгущения сетки в поперечном направлении. Штриховые линии на рисунках 9–11 представляют линейные аппроксимации по методу наименьших квадратов результатов расчетов, показанных точками.

На рисунке 9 показана сходимость значений отхода ударной волны y_s при $x = 35$ и коэффициента сопротивления давления C_{Dp} при измельчении сетки. Верхние линии *1* и *2* на рисунке 9 соответствуют сгущению сетки по продольной координате. С этими линиями графически совпадают линии, соответствующие сгущению сетки по обеим координатам. Нижние линии *1* и *2* на данном рисунке соответствуют сгущению сетки по поперечной координате. Эффективный порядок точ-

ности при сгущении сетки в продольном направлении оказался равным 1.87 и 1.83 для y_s и C_{Dp} соответственно, а при сгущении сетки в поперечном направлении — равным 3.70 и 3.25 для вышеуказанных величин.

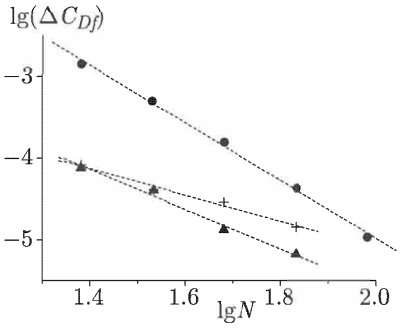

Рис. 10. Убывание погрешности при вычислении значения коэффициента сопротивления трения C_{Df} при сгущении сетки

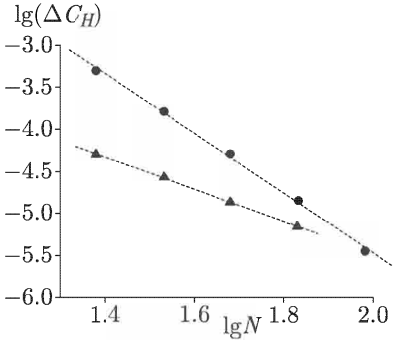

Рис. 11. Убывание погрешности при вычислении значения коэффициента теплообмена C_H в передней критической точке тела при сгущении сетки

На рисунке 10 показана сходимость значения коэффициента сопротивления трения C_{Df} при измельчении сетки. На этом рисунке кружочки показывают сеточную сходимость по η, крестики — по ξ, а треугольники — по направлениям ξ и η одновременно. Эффективный порядок точности схемы при сгущении сетки по η оказался равным 3.52, по ξ — 1.60, а по ξ и η одновременно — 2.40.

Рисунок 11 иллюстрирует сходимость значения коэффициента теплообмена C_H в передней критической точке тела при сгущении сетки. На этом рисунке кружочки показывают сеточную сходимость по направлению η, а треугольники — по направлениям ξ и η одновременно. Эффективный порядок точности схемы при сгущении сетки по переменной η оказался равным 3.55, а по ξ и η одновременно — 1.93.

Глава IX

ИНТЕГРАЛЬНЫЕ УРАВНЕНИЯ

Уравнения, содержащие неизвестную функцию под знаком интеграла, часто возникают в различных областях науки. Хороший обзор различных аналитических и численных методов решения интегральных уравнений можно найти в книге [Манжиров, Полянин; 1999]. В этой главе рассмотрен метод квадратур. Сеточный метод позволяет, используя процедуру сгущения сеток, производить вычисления с контролем точности и проводить рекуррентное уточнение решения по методу Ричардсона, существенно повышая точность расчетов практически без увеличения объема вычислений; при этом можно использовать как равномерные, так и квазиравномерные сетки, в том числе и в неограниченной области.

§ 1. Метод квадратур

1. Алгоритм. Рассмотрим уравнение Фредгольма II-го рода:

$$u(x) - \lambda \int_a^b K(x,s)u(s)ds = f(x), \quad a \leqslant x \leqslant b. \tag{1}$$

Идея метода квадратур заключается в замене интеграла в уравнении (1) некоторой приближенной квадратурной формулой:

$$\int_a^b F(x)dx \approx \sum_{n=1}^{N} c_n F(x_n), \tag{2}$$

где x_n — узлы, а c_n — веса квадратурной формулы.

Введем в квадрате $[a \leqslant x \leqslant b,\ a \leqslant s \leqslant b]$ одинаковые сетки по обеим координатам: x_n, s_n, $1 \leqslant n \leqslant N$; в качестве узлов этих сеток возьмем узлы выбранной квадратурной формулы. Если выбрана квадратурная формула трапеций, то узлами будут граничные точки интервалов; если используется квадратурная формула средних — середины интервалов. Заменяя интеграл в (1) по квадратурной формуле (2), приходим к

следующей неоднородной системе линейных алгебраических уравнений для определения значений неизвестной функции u_n в узлах сетки x_n, $1 \leqslant n \leqslant N$:

$$u_n - \lambda \sum_{m=1}^{N} c_m K(x_n, s_m) u_m = f_n, \quad 1 \leqslant n \leqslant N. \tag{3}$$

Уравнение (3) удобно записать в матричной форме:

$$Y - \lambda A Y = F. \tag{4}$$

Здесь матрица A имеет элементы $a_{n,m} = c_m K(x_n, s_m)$, а Y и F являются вектор-столбцами, составленными из значений неизвестной функции u_n и правой части f_n уравнения (1) в узлах сетки x_n, $1 \leqslant \leqslant n \leqslant N$.

Отметим, что матрица системы (3) является, вообще говоря, плотно заполненной и неэрмитовой. Если $1/\lambda$ не совпадает ни с одним из собственных значений матрицы A, то система (3) имеет единственное решение. Система (4) будет хорошо обусловлена, если λ не будет лежать в малой окрестности какого-либо из собственных значений ядра интегрального уравнения (1).

Уравнение Вольтерра II-го рода имеет следующий вид:

$$u(x) - \lambda \int_a^x K(x, s) u(s) ds = f(x), \quad a \leqslant x \leqslant b. \tag{5}$$

Его можно интерпретировать как уравнение Фредгольма II-го рода, полагая $K(x, s) = 0$, если $x < s$; при этом матрица A системы (4) является треугольной, и система (4) решается обратным ходом метода Гаусса.

Выбор квадратурной формулы. Если $K(x, s)$ и $f(x)$ имеют достаточное число непрерывных производных, то можно выбирать квадратурные формулы (2) высокого порядка точности. Однако экономичнее выбрать более простые квадратурные формулы (трапеций или средних) и, проводя сгущение сетки, уточнять решение по методу Ричардсона. Для уточнения решения при сгущении сеток удобнее всего такая ситуация, когда более подробная сетка содержит все узлы более редкой. Для формулы трапеций это происходит уже при сгущении в 2 раза, а для формулы средних придется сгустить минимум в 3 раза. Таким образом, объем вычислений при сгущении сеток с использованием формулы средних больше; тем самым формула трапеций предпочтительнее.

Если область интегрирования в уравнении (1) неограничена (один или оба предела интегрирования обращаются в бесконечность), использование метода квадратур для решения интегрального уравнение также весьма эффективно. При этом целесообразно использовать квазиравно-

мерные сетки, покрывающие неограниченную область, и квадратурные формулы вида (III.62), (III.63) или (IV.13), (IV.14), пригодные для неограниченной области.

2. Тестирование метода. Тестирование метода удобно проводить на уравнениях, имеющих точные решения. Так в случае уравнения (1) можно взять вырожденное ядро

$$K(x,s)=\sum_{q=1}^{Q}\Omega_q(x)\omega_q(s). \tag{6}$$

В случае уравнения (5) удобно взять ядро, зависящее от разности аргументов:

$$u(x)-\lambda\int_0^x K(x-s)u(s)ds=f(x). \tag{7}$$

Интеграл в уравнении (7) представляет собой вольтерровскую свертку, и решение может быть найдено, например, с помощью интегрального преобразования Лапласа.

Если область интегрирования $(-\infty,+\infty)$ и ядро зависит от разности аргументов, т. е.

$$u(x)-\lambda\int_{-\infty}^{+\infty}K(x-s)u(s)ds=f(x), \tag{8}$$

тогда точное решение строится с помощью преобразования Фурье.

Тестирование метода проводилось на следующих примерах.

Пример 1. Рассмотрим уравнение

$$u(x)-\lambda\int_0^{\pi/2}\sin x\cos s\cdot u(s)ds=\sin x. \tag{9}$$

Это уравнение с вырожденным ядром, и точное решение находится с помощью несложных выкладок:

$$u(x)=\frac{2\sin x}{2-\lambda}. \tag{10}$$

Для вычисления интеграла в (9) были выбраны равномерная сетка $x_n=\pi n/(2N)$, $0\leqslant n\leqslant N$ (узлы в этом случае естественно нумеровать с нулевого), и квадратурная формула трапеций (IV.1). Точность этой квадратурной формулы есть $O(N^{-2})$. Был выполнен стандартный тест на сгущающихся сетках для $\lambda=3$. Численное решение сравнивалось с точным (10) в сеточной норме c.

Результаты теста приведены на рисунке 1, где в двойном логарифмическом масштабе показано убывание погрешности численного

решения $\|\Delta\|_c$ с удвоением числа узлов сетки $N = 8,\ 16,\ 32, \ldots, 1024$. Наклон графика погрешности в точности соответствует порядку точности $O(N^{-2})$.

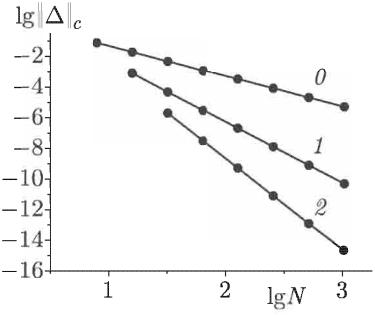

Рис. 1. Результаты теста на сгущающихся сетках для примера 1. Точки — рассчитанные значения погрешности, для наглядности они соединены линиями. Цифры около линий означают: *0* — расчет по исходной квадратурной формуле, *1* — первое экстраполяционное уточнение по Ричардсону, *2* — второе экстраполяционное уточнение по Ричардсону

Результаты, полученные по описанному выше алгоритму, помечены на рисунке 1 цифрой *0*. Было так же проведено экстраполяционное уточнение решения по методу Ричардсона, которое требует всего нескольких арифметических действий и практически не увеличивает время общего расчета. Результаты первого (цифра *1*) и второго (цифра *2*) уточнения по Ричардсону также представлены на рисунке 1. Погрешность первого уточненного решения имеет порядок $O(N^{-4})$, второго — $O(N^{-6})$, что полностью соответствует теории. Уже второе уточнение дает возможность достичь машинной точности 10^{-16} при числе узлов сетки $N = 1024$, тогда как расчет по исходной квадратурной формуле трапеций для достижения такой точности потребовал бы неприемлемо большого числа узлов сетки $N \sim 10^8$. Решение системы алгебраических уравнений такой размерности потребовало бы огромных вычислительных ресурсов и при наиболее распространенной сейчас разрядности вычислительной техники привело бы к накопленной ошибке округления, сравнимой с самим решением.

Пример 2. Рассмотрим уравнение

$$u(x) - \lambda \int_0^x \cos(x-s)u(s)ds = \sin x. \tag{11}$$

Применив преобразование Лапласа, получим

$$\widetilde{u}(p)\left(1 - \lambda\frac{p}{p^2+1}\right) = \frac{1}{p^2+1},$$

где

$$\widetilde{u}(p) = \int\limits_0^{+\infty} e^{-px} u(x)dx, \quad \widetilde{u}(p) = \frac{1}{p^2 - \lambda p + 1}.$$

Обращая преобразование Лапласа, получим точное решение задачи (11):

$$u(x) = \begin{cases} \dfrac{e^{\lambda x/2} \sin\left(x\sqrt{1-(\lambda/2)^2}\right)}{\sqrt{1-(\lambda/2)^2}}, & |\lambda| < 2, \\ xe^{\lambda x/2}, \quad |\lambda| = 2, \\ \dfrac{e^{\lambda x/2} \sin\left(x\sqrt{(\lambda/2)^2-1}\right)}{\sqrt{(\lambda/2)^2-1}}, & |\lambda| > 2. \end{cases}$$

Для численного решения этого уравнения на отрезке $x \in [0, T]$ выберем квадратурную формулу трапеций на равномерной сетке $x_n = nT/N$, $0 \leqslant n \leqslant N$. Элементы матрицы A в формуле (4) будут иметь следующий вид:

1) если $n = m = 0$, то $a_{00} = 0$;

2) если $n > 0$ и $m = 0$ или $m = n$, то $a_{nm} = \dfrac{T}{2N}\cos(x_n - x_m)$;

3) если $n > 0$ и $0 < m < n$, то $a_{nm} = \dfrac{T}{N}\cos(x_n - x_m)$;

4) если $n > 0$ и $n < m$, то $a_{nm} = 0$.

Матрица системы (4) в этом случае будет нижней треугольной, и система (4) легко решается обратным ходом метода Гаусса. Необходимо следить за тем, чтобы диагональные элементы этой матрицы не были

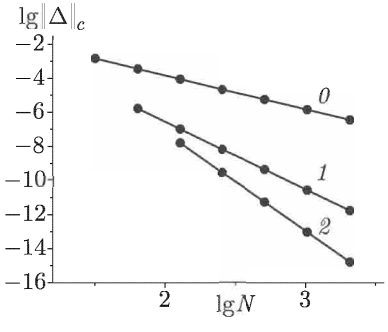

Рис. 2. Результаты теста на сгущающихся сетках для примера 2. Точки — рассчитанные значения погрешности, для наглядности они соединены линиями. Цифры около линий означают: *0* — расчет по исходной квадратурной формуле, *1* — первое экстраполяционное уточнение по Ричардсону, *2* — второе экстраполяционное уточнение по Ричардсону

бы равны или близки к нулю. Такая ситуация возможна, если при фиксированном N абсолютное значение λ достаточно велико. Однако для любого λ найдется такое N, начиная с которого система будет всегда хорошо обусловлена.

Были проведены расчеты на сетках с числом узлов сетки $N = 32, 64, \ldots, 2048$ при различных значениях λ. Для $\lambda = -2$ было проведено экстраполяционное уточнение решения по методу Ричардсона. Результаты приведены на рисунке 2: показано убывание погрешности численного решения в сеточной норме c при удвоении числа узлов. Цифра *0* около линии соответствует основному расчету, цифра *1* — первое уточнение по Ричардсону, цифра *2* — второе уточнение по Ричардсону. Углы наклонов графиков погрешности в двойном логарифмическом масштабе подтверждают точность основного расчета $O(N^{-2})$, первое экстраполяционное уточнение улучшает точность до $O(N^{-4})$, второе — до $O(N^{-6})$, что полностью соответствует теории. Отметим, что на самой густой сетке (2048) ошибка основного расчета составляет $\sim 10^{-7}$; двукратное уточнение по Ричардсону снижает погрешность практически до уровня машинной точности $\sim 10^{-15}$.

Пример 3. Рассмотрим уравнение

$$u(x) - \lambda \int\limits_{-\infty}^{+\infty} e^{-|x-s|} u(s)ds = e^{-|x|}. \tag{12}$$

Используя преобразование Фурье, в случае $\lambda < 1/2$ нетрудно получить

$$u(x) = \frac{e^{-\sqrt{1-2\lambda}\,|x|}}{\sqrt{1-2\lambda}}. \tag{13}$$

Для численного решения уравнения (12) будем использовать квазиравномерную сетку $x_n = c \cdot \operatorname{tg}(\pi n/(2N))$, $-N \leqslant n \leqslant N$, покрывающую всю область $x \in (-\infty, +\infty)$. Заменяя интеграл в (12) по квадратурной формуле трапеций (IV.14), пригодной для неограниченной области, получим

$$u_n - \lambda \sum_{m=-N+1}^{N} (e^{-|x_n - s_m|} u_m + e^{-|x_n - s_{m-1}|} u_{m-1})(s_{m-1/4} - s_{m-3/4}) = \\ = e^{-|x_n|}, \quad -N \leqslant n \leqslant N. \tag{14}$$

Однако покажем, что уравнения (14) со значениями $n = \pm N$ отщепляются от системы и явно решаются. В самом деле, при этих индексах экспоненты в сумме (14) обращаются в нуль, и остается $u_{\pm N} = f_{\pm N} = 0$. Поэтому можно считать, что индекс в (14) изменяется в переделах $-N + 1 \leqslant n \leqslant N - 1$.

Учитывая, что в краевых точках сетки ядро интегрального уравнения обращается в нуль, перепишем (14) в удобном для численного решения виде:

$$u_n - \lambda \sum_{m=-M+1}^{M-1} e^{-|x_n - s_m|} u_m (s_{m-1/4} - s_{m-3/4} + s_{m+3/4} - s_{m+1/4}) = \\ = e^{-|x_n|}, \quad -N+1 \leqslant n \leqslant N-1.$$

Ядро интегрального уравнения (12) имеет разрыв производной при $x = s$. При использовании формулы трапеций узел квадратурной формулы всегда попадает в точку разрыва производной ядра интегрального уравнения. Это равносильно тому, что область интегрирования в (12) разбивается на две, в каждой из которых подынтегральная функция бесконечно гладкая.

Использование специальной квазиравномерной сетки позволяет не только получать теоретический порядок точности $O(N^{-2})$, но проводить экстраполяционное уточнение по Ричардсону, каждый раз увеличивая порядок точности на 2. Тесты на сгущающихся сетках подтвердили эффективность предложенного подхода. На рисунке 3 показано отклонение численного решения от точного (13) в сеточной норме c для основного расчета (цифра *0* около линии), первого (цифра *1*) и второго (цифра *2*) экстраполяционного уточнения по методу Ричардсона. Снова уже второе уточнение позволяет выйти на уровень машинной точности при числе узлов сетки $N = 2048$.

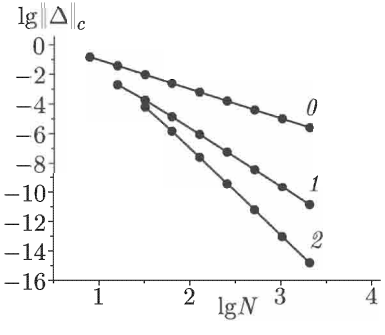

Рис. 3. Результаты теста на сгущающихся сетках для примера 3.Точки — рассчитанные значения погрешности, для наглядности они соединены линиями. Цифры около линий означают: *0* — расчет по исходной квадратурной формуле, *1* — первое экстраполяционное уточнение по Ричардсону, *2* — второе экстраполяционное уточнение по Ричардсону

Сделаем замечание. Уточнение по Ричардсону соответствовало теории потому, что на всех сетках полное число интервалов $2N$ было четным. При этом область интегрирования фактически разбивалась на две независимые подобласти $x_n \leqslant s_m$ и $x_n \geqslant s_m$, а в каждой из этих подоб-

ластей ядро бесконечно гладкое. Если менять индекс n в пределах $0 \leqslant n \leqslant N$ ($x_0 = -\infty$, $x_N = +\infty$) и выбирать нечетное N, то порядок точности может оказаться существенно хуже теоретического.

§ 2. Обтекание тела потоком стратифицированной жидкости

1. Модель. В классической гидродинамике идеальной несжимаемой жидкости одной из основных задач является задача об обтекании тел плоскопараллельным потоком. В случае однородной жидкости такие задачи сводятся к внешним краевым задачам для уравнения Лапласа. Рассмотрим задачу о двумерном обтекании тел плоскопараллельным потоком стратифицированной жидкости в случае, когда параметр $\varepsilon^2 = gl/V_0^2$ мал; здесь V_0 — скорость потока на бесконечности, l — характерный размер тела, g — ускорение силы тяжести. Величина ε^2 равна обратному числу Фруда $1/\mathrm{Fr}$.

Как известно (см., например, [Габов, 1998]), в однородном поле силы тяжести поток стратифицированной жидкости описывается уравнением Дюбрейль–Жакотен, которое в безразмерных переменных и в приближении больших чисел Фруда записывается в виде

$$\Delta_2 u + \frac{1}{2}\frac{\rho_1'(u)}{\rho_1(u)}(|\nabla u|^2 - 1) = 0, \tag{15}$$

где Δ_2 есть двумерный оператор Лапласа, $u(x, y)$ имеет смысл безразмерной функции тока, а оператор ∇ означает градиент. В практически важном случае экспоненциальной стратификации по вертикальной координате $\rho_0(y) = Ae^{-2\beta y}$, $\rho_1(u) = \rho_0(lu)$, $\mu^2 = \beta l$; тогда уравнение Дюбрейль–Жакотен принимает вид

$$\Delta_2 u + \mu^2(|\nabla u|^2 - 1) = 0. \tag{16}$$

Предположим, что течение рассматриваемого плоскопараллельного потока возмущено наличием обтекаемого им препятствия Ω, ограниченного контуром Γ. На контуре выполнено условие обтекания $(\mathbf{v} \cdot \mathbf{n})\big|_\Gamma = 0$, которое эквивалентно

$$u\big|_\Gamma = u_0 = \mathrm{const}; \tag{17}$$

условие невозмущенности потока на бесконечности в безразмерных переменных имеет вид

$$u - y\big|_{r \to +\infty} = 0, \quad r = \sqrt{x^2 + y^2}\,. \tag{18}$$

Итак, задача обтекания тела потоком экспоненциально стратифицированной жидкости в приближении больших чисел Фруда сводится к решению уравнения (16) с условиями (17) и (18). Как известно [Би-

цадзе, 1978], уравнение (16) может быть линеаризовано заменой $u = -\mu^{-2}\ln v$. Для неизвестной функции $v(x, y)$ получаем уравнение

$$\Delta_2 v - \mu^4 v = 0 \tag{19}$$

с дополнительными условиями

$$v - e^{-\mu^2 y}\big|_{r\to+\infty} = 0, \quad v\big|_\Gamma = e^{-\mu^2 u_0}. \tag{20}$$

Разрешимость задачи (19), (20) следует из известных результатов теории краевых задач для эллиптических уравнений, и в случае гладкого контура Γ задача всегда разрешима в классическом смысле.

2. Редукция к интегральному уравнению. Введем функцию $w = v - e^{-\mu^2 y}$, тогда задачу (19), (20) можно переписать в виде:

$$\Delta_2 w - \mu^4 w = 0, \tag{21}$$

$$w\big|_{r\to+\infty} = 0, \tag{22}$$

$$w\big|_\Gamma = f(P_0) = e^{-\mu^2 u_0} - e^{-\mu^2 y}\big|_\Gamma. \tag{23}$$

Будем искать решение задачи (21)–(23) в виде потенциала двойного слоя

$$w(M) = -\int_\Gamma \mu(P)\frac{\partial}{\partial n_P} K_0(\mu^2 R_{MP})\, dl_P. \tag{24}$$

Здесь $M \in R^2\backslash\overline{\Omega}$, $P \in \Gamma$, n_P — внешняя нормаль к контуру Γ в точке P, K_q — функция Макдональда q-го индекса, R_{MP} — расстояние от точки M до точки P. Потенциал (24) удовлетворяет уравнению (21) в области $R^2\backslash\overline{\Omega}$ и условиям регулярности на бесконечности (22). Потенциал двойного слоя терпит разрыв при переходе через контур Γ:

$$w_e(P) = \overline{w}(P) - \pi\mu(P),$$

здесь $w_e(P)$ — предельное значение потенциала в точке P снаружи, $\overline{w}$ — прямое значение потенциала на контуре. Это приводит с учетом граничного условия (23) к интегральному уравнению Фредгольма второго рода:

$$-\int_\Gamma \mu(P)\frac{\partial}{\partial n_P} K_0(\mu^2 R_{P_0P})\, dl_P - \pi\mu(P_0) = f(P_0). \tag{25}$$

Вычисляя производную по нормали, получим

$$\int_\Gamma \mu(P) K_1(\mu^2 R_{P_0P})\mu^2 \cos\varphi\, dl_P - \pi\mu(P_0) = f(P_0); \tag{26}$$

здесь φ — угол между внутренней нормалью в точке P и вектором PP_0.

3. Построение численного решения. Применим для численного решения интегрального уравнения (26) метод квадратур, описанный в § 1. Ключевым моментом в этом методе является выбор квадратурной формулы и узлов сетки.

В качестве первого примера рассмотрим задачу обтекания потоком экспоненциально стратифицированной жидкости контура Γ, ограничивающего единичный круг $\Omega = \{r \leqslant 1\}$. В этом случае можно протестировать численный метод на точном решении. Решение такой задачи построено методом разделения переменных в [Габов, 1998]. Для функции v в задаче (19), (20) решение выглядит так:

$$v = \exp(-\mu^2 r \sin\theta) + \exp(-\mu^2 u_0) - \\ - \sum_{q=-\infty}^{+\infty} I_q(\mu^2) \frac{K_q(\mu^2 r)}{K_q(\mu^2)} \exp\left(iq\theta + i\frac{\pi q}{2}\right). \quad (27)$$

Здесь I_q — функции Инфельда, а r, θ — полярные координаты, естественным образом связанные с кругом. Для аппроксимации интеграла в (26) выберем формулу средних. На единичном круге построим равномерную сетку, т. ч.

$$x_n = \cos\left(\frac{2\pi n}{N}\right), \quad y_n = \sin\left(\frac{2\pi n}{N}\right), \quad 0 \leqslant n \leqslant N. \quad (28)$$

Очевидно, N-я точка этой сетки совпадает с нулевой.

Ядро в уравнении (26) будет непрерывным, однако производная будет иметь разрыв при $P = P_0$. При использовании квадратурной формулы средних следует выделить эту особенность:

$$\int_\Gamma \mu(P) \frac{\partial}{\partial n_P} \ln(R_{P_0P}) dl_P - \int_\Gamma \mu(P) \frac{\partial}{\partial n_P} \{K_0(\mu^2 R_{P_0P}) + \ln(R_{P_0P})\} dl_P - \\ - \pi\mu(P_0) = f(P_0). \quad (29)$$

Рассмотрим аналитическую функцию

$$\ln R_{PP_0} + i\Psi(P, P_0), \quad (30)$$

где $\Psi(P, P_0)$ есть угол, измеренный против часовой стрелки между осью OX и вектором P_0P. В силу аналитичности (30) выполнены условия Коши–Римана, и ядро в первом слагаемом (29) можно заменить на производную по касательной от функции $\Psi(P, P_0)$. Ядро во втором слагаемом уже будет гладким.

Численный расчет сводится к решению системы линейных уравнений, полученных при замене интегралов в (29) квадратурными формулами средних прямоугольников на сетке (28).

Зависимость нормы погрешности $\|\Delta\|_c$ от числа интервалов сетки N изображена на рисунке 4 в двойном логарифмическом масштабе.

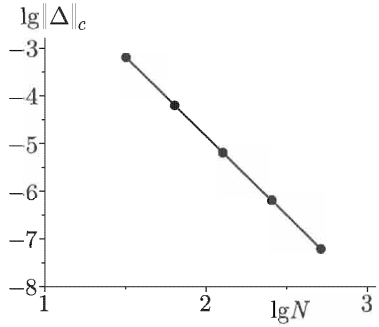

Рис. 4. Результаты теста на сгущающихся сетках для задачи обтекания единичного круга потоком стратифицированной жидкости. Наклон графика убывания погрешности с ростом числа узлов сетки указывает на точность метода $O(N^{-3})$

Расчетные точки лежат практически на прямой с наклоном -3; это соответствует точности $O(N^{-3})$, хотя использованная формула средних прямоугольников имеет точность лишь $O(N^{-2})$ для сколь угодно гладких функций. Как объяснить этот неожиданный результат?

Причина в том, что в данном случае квадратурная формула средних аппроксимирует интеграл от периодической функции, для которой формула средних может давать более высокий порядок точности. В этом случае эффект насыщения отсутствует, и порядок точности формулы средних будет тем выше, чем больше порядок гладкости функции.

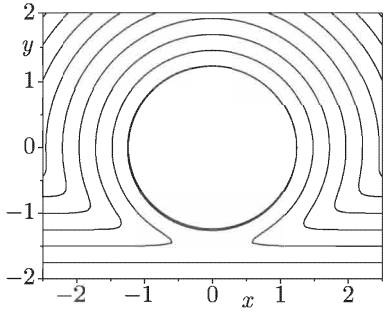

Рис. 5. Обтекание единичного круга (жирная линия) потоком стратифицированной жидкости. Показаны линии уровня функции тока $u = \text{const}$. Значение потенциала на единичном круге $u_0 = -2$

Результаты численных расчетов задачи об обтекании единичного круга потоком стратифицированной жидкости приведены на рисунке 5. Показаны линии тока $u = \text{const}$, совпадающие с траекториями движения частиц жидкости. Расчет выполнен для краевого условия (17) $u_0 = -2$.

Отметим, что предложенный подход легко обобщается для задачи обтекания системы тел потоком стратифицированной жидкости. Отличие состоит лишь в том, что приходится решать систему интегральных уравнений. На рисунке 6 приведены результат расчетов для задачи обтекания двух кругов.

Результаты решения задачи об обтекании эллипса, наклоненного к потоку, изображены на рисунке 7. Главная полуось эллипса составляет угол $3\pi/4$ с направлением невозмущенного потока.

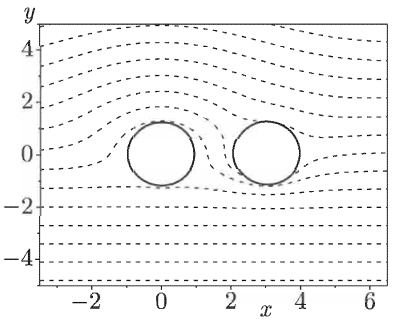

Рис. 6. Обтекание двух кругов потоком стратифицированной жидкости. Показаны линии уровня функции тока $u = \mathrm{const}$

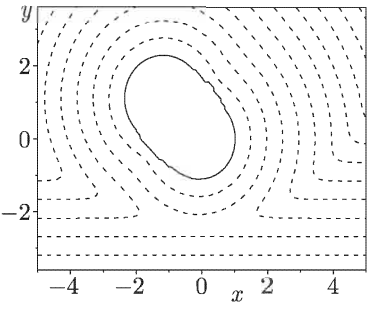

Рис. 7. Обтекание наклонного эллипса потоком стратифицированной жидкости. Показаны линии уровня функции тока $u = \mathrm{const}$

В рамках данной модели стратифицированной жидкости (в отличие от случая жидкости с постоянной плотностью $\rho = \mathrm{const}$) возможно обтекание тел с заданной ненулевой циркуляцией. Об этом свидетельствуют замкнутые линии тока на рисунках 5–7.

СПИСОК ЛИТЕРАТУРЫ

Curtis A.R. Numer. M., 1970. — V. 16, № 3.

Dormand J., Prince P.J. A family embedded Runge–Kutta formulae // J. Comp. Appl. Math. — V. 6. — P. 19–26.

Dormand J., Prince P.J. New Runge–Kutta algorithms for numerical simulation in dynamical astronomy // Celestial Mechanics. — V. 18. — P. 223–232.

Hairer E. A Runge–Kutta method of order 10 // J. Inst. Maths. Applies. — V. 21. — P. 47–59.

Kalstungen K.H., Wanner G. Computing, 1972. — V. 9, № 1.

Peaceman D.W., Rachford H.H. The numerical solution of parabolic and elliptic differential equations // J. Soc. Industr. Appl. Math., 1955. — V. 3, No 1. — P. 28–42.

Rosenbrock H.H. Some general implicit processes for the numerical solution of differential equations // Comput. J., 1963. — V. 5, No 4. — P. 329–330.

Wanner G. Bitegrationgewohnlicher Differentialghleichungen, Lie Reihen, Runge–Kutta Methoden, BI Mannheim, Htb. 831/831a, 1S2p.

Wanner G. Runge–Kutta methods with expansions in even powers of h // Computing. — V. 11. — P. 81–85.

Абрамович Г.Н. Прикладная газовая динамика. В 2 ч. Ч. 1. — 5-е изд. — М.: Наука, 1991. — 600 с.

Альшин А.Б., Альшина Е.А. Численное решение начально-краевых задач для уравнений составного типа в неограниченных областях // Ж. вычисл. матем. и матем. физ., 2002. — Т. 42, № 12. — С. 1796–1803.

Альшин А.Б., Альшина Е.А., Болтнев А.А., Качер О.А., Корякин П.В. Численное решение начально-краевых задач для уравнений соболевского типа методом квазиравномерных сеток // Ж. вычисл. матем. и матем. физ., 2004. — Т. 44, № 3. — С. 490–511.

Альшин А.Б., Альшина Е.А., Калиткин Н.Н. Численное решение гиперболических задач в неограниченной области // Матем. моделирование, 2004. — Т. 16, № 4. — С. 114–126.

Альшин А.Б., Перова Л.В. О колебаниях стратифицированной вращающейся жидкости, возбуждаемых волной, бегущей по наклонному дну // Ж. вычисл. матем. и матем. физ., 2000. — Т. 40, № 3. — С. 473–482.

Альшин А.Б., Плетнер Ю.Д. Явное решение одной начально-краевой задачи теории ионно-звуковых волн в незамагниченной плазме и его асимптотика // Ж. вычисл. матем. и матем. физ., 1994. — Т. 34, № 8–9. — С. 1323–1330.

Альшина Е.А., Калиткин Н.Н. Вычисление спектров линейных дифференциальных операторов // ДАН, 2001. — Т. 380, № 4. — С. 443–447.

Альшина Е.А., Калиткин Н.Н., Панченко С.Л. Численное решение краевых задач в неограниченных областях // Матем. моделирование, 2002. — Т. 14, № 11. — С. 10–22.

Амосов А.А., Дубинский Ю.А., Копченова Н.В. Вычислительные методы для инженеров. — М.: Изд. МЭИ, 2003.

Артемьев С.С., Демидов Г.В. Некоторые проблемы вычислительной математики. — Новосибирск: Наука, 1975. — 216 с.

Бартеньев О.В. Фортран для профессионалов. Математическая библиотека IMSL. Т. 2. — М.: Диалог-МИФИ, 2001.

Бабенко К.И. Основы численного анализа. — М.: Наука, 1986. — 452 с.

Бабенко К.И.; Брюно А.Д. Основы численного анализа. — 2-е изд., испр. и доп. — Ижевск: РХД, 2002. — 847 с.

Бахвалов Н.С. Численные методы, алгебра, обыкновенные дифференциальные уравнения. — М.: Наука, 1973.

Бахвалов Н.С., Жидков Н.П, Кобельков Г.М. Численные методы. — М.: Наука, 1987. — 598 с.

Бахвалов Н.С., Лапин А.В., Чижонков Е.В. Численные методы в задачах и упражнениях. — М.: Высшая школа, 2002. — 190 с.

Белевец П.С., Кожух И.Г. Задачник-практикум по методам математической физики. — Минск: Вышэйшая школа, 1989. — 108 с.

Белоцерковский О.М. Численное исследование современных задач газовой динамики. — М.: Наука, 1974. — 397 с.

Березин И.С., Жидков Н.П. Методы вычислений. Т.1. — М.: Физматгиз, 1959. — 464 с.

Березин И.С., Жидков Н.П. Методы вычислений. Т.2. — М.: Физматгиз, 1959. — 620 с.

Бицадзе А.В. Волны в потоке жидкости переменной плотности // Дифференциальные уравнения, 1978. — Т. 14, № 10. — С. 1053–1059.

Боголюбов А.Н., Кравцов В.В. Задачи по математической физике. — М.: Изд. МГУ, 1998.

Будак Б.М., Самарский А.А., Тихонов А.Н. Сборник задач по математической физике. — М.: Наука, 1980.

Бутузов В.Ф., Крутицкая Н.Ч., Шишкин А.А. Линейная алгебра в вопросах и задачах. — М.: Физматлит, 2001.

Виноградова И.А., Олехник С.Н., Садовничий В.А. Задачи и упражнения по математическому анализу. — М.: Высшая школа, 2002. — 190 с.

Вирт Н. Алгоритмы + структуры данных = программы. — М.: Мир, 1985. — 406 с.

Воеводин В.В. Математические модели и методы в параллельных процессах. — М.: Наука, 1986. — 259 с.

Воеводин В.В. Линейная алгебра. — М.: Наука, 1979. — 400 с.

Воеводин В.В. Математические модели и методы в параллельных процессах. — М.: Наука, 1986.

Воеводин В.В., Тыртышников Е.Е. Вычислительные процессы с теплицевыми матрицами, 1987.

Волков Е.А. Численные методы. — М.: Наука, 1987.

Габов С.А., Свешников А.Г. Линейные задачи теории нестационарных внутренних волн. — М.: Наука, 1990. — 344 с.

Габов С.А. Новые задачи математической теории волн. — М.: Наука, 1998.

Годунов С.К. Лекции по современным аспектам линейной алгебры. — Новосибирск: Научная книга, 2002.

Годунов С.К., Рябенький В.С. Введение в теорию разностных схем. — М.: Физматлит, 1962. — 340 с.

Головачев Ю.П. Численное моделирование течений вязкого газа в ударном слое. — М.: Физматлит, 1996. — 376 с.

Голосной И.О. Микрополе и теплофизические свойства неидеальной плазмы: Дис...канд. физ.-мат. наук. — ИММ РАН, 1995.

Голуб Дж., Ван Лоун Ч. Матричные вычисления. — М.: Мир, 1999. — 548 с.

Грим Г. Уширение спектральных линий в плазме. — М.: Мир, 1978. — 492 с.

Громов В.Г., Сахаров В.И., Фатеева Е.И. Численное исследование гиперзвукового обтекания затупленных тел вязким химически реагирующим газом // Изв. РАН. МЖГ, 1999. — № 5. — С. 177–186.

Дорн У., Мак-Кракен Д. Численные методы и программирование на Фортране. — М.: Мир, 1970.

Доргатт Дж., Малькольм М., Моулер К. Математические методы математических вычислений. — М.: Мир, 1980.

Дробышевич В.И., Дымников В.П., Гивин Г.С. Задачи по вычислительной математике. — М.: Наука, 1980.

Дэннис Дж., мл., Шнабель Р. Численные методы безусловной оптимизации и решения нелинейных уравнений. — М.: Мир, 1988.

Ермаков В.В., Калиткин Н.Н. Оптимальный шаг и регуляризация метода Ньютона // Ж. вычисл. матем. и матем. физ., 1981. — Т. 21, № 2. — С. 491–497.

Жидков Е.П., Пузынин И.В. Об одном методе введения параметра при решении краевых задач для нелинейных обыкновенных дифференциальных уравнений второго порядка // Ж. вычисл. матем. и матем. физ., 1967. — Т. 7, № 5. — С. 1086–1095.

Ильин В.А., Позняк Э.Г. Линейная алгебра. — М.: Наука, 1994. — 294 с.

Ибрагимов И.И. Методы интерполяции функций и некоторые их применения. — М.: Наука, 1971. — 384 с.

Калиткин Н.Н. Нелинейные разностные схемы для гиперболических уравнений // Ж. вычисл. матем. и матем. физ., 1965. — Т. 5, № 6. — С. 1107–1115.

Калиткин Н.Н. Численные методы. — М.: Наука, 1978.

Калиткин Н.Н. Об экстраполяции на сгущающихся сетках // Матем. моделирование, 1994. — Т. 6, № 1. — С. 86–98.

Калиткин Н.Н., Кузнецов Н.О., Панченко С.Л. Метод квазиравномерных сеток в бесконечной области // ДАН, 2000. — Т. 374, № 5. — С. 598–601.

Калиткин Н.Н., Кузьмина Л.В. Интерполяционные формулы для функций Ферми–Дирака // Ж. вычисл. матем. и матем. физ., 1975. — Т. 15, № 3.

Калиткин Н.Н., Рогов Б.В., Соколова И.А. Турбулизованные течения химически реагирующих газов в сопле Лаваля // ДАН, 1997. — Т. 357, № 3. — С. 339–342.

Калиткин Н.Н., Рогов Б.В., Соколова И.А. Двухстадийный маршевый расчет вязких течений через сопло Лаваля // Матем. моделирование, 1999. — Т. 11, № 7. — С. 95–117.

Корпусов М.О., Плетнер Ю.Д., Свешников А.Г. О нестационарных волнах в среде с анизотропной дисперсией // Ж. вычисл. матем. и матем. физ., 1999. — Т. 39, № 6. — С. 1006–1022.

Корпусов М.О., Свешников А.Г. Трехмерные нелинейные эволюционные уравнения псевдопараболического типа в задачах математической физики // Ж. вычисл. матем. и матем. физ., 2003. — Т. 43, № 12. — С. 1835–1869.

Коллатц Л. Численные методы решения дифференциальных уравнений. — М.: ИИЛ, 1953. — 508 с.

Коллатц Л. Функциональный анализ и вычислительная математика. — М.: Мир, 1969. — 264 с.

Кормен Т., Лейзерсон Ч., Ривест Р. Алгоритм: построение и анализ. — М.: МЦНМО, 1999. — 960 с.

Крылов А.Н. Лекции о приближенных вычислениях. — М.: Наука, ГИТТЛ, 1954.

Крылов В.И., Бобков В.В., Монастырный П.И. Начала теории вычислительных методов Т. 1. Дифференциальные уравнения. — Минск: Наука и техника, 1982. — 286 с.

Крылов В.И., Бобков В.В., Монастырный П.И. Начала теории вычислительных методов Т. 2. Интерполирование и интегрирование. — Минск: Наука и техника, 1983. — 287 с.

Крылов В.И., Бобков В.В., Монастырный П.И. Начала теории вычислительных методов Т. 3. Интегральные уравнения, некорректные задачи и улучшение сходимости. — Минск: Наука и техника, 1984. — 263 с.

Крылов В.И., Бобков В.В., Монастырный П.И. Начала теории вычислительных методов Т. 4. Линейная алгебра и нелинейные уравнения. — Минск: Наука и техника. — 1985. — 279 с.

Крылов В.И., Бобков В.В., Монастырный П.И. Начала теории вычислительных методов Т. 5. Уравнения в частных производных. — Минск: Наука и техника, 1986. — 311 с.

Ландау Л.Д., Лифшиц Е.М. Квантовая механика. — М.: Физматлит, 1963.

Ланцош К.А. Практические методы прикладного анализа. — М.: ГИФМЛ, 1961. — 524 с.

Лапин Ю.В., Стрелец М.Х. Внутренние течения газовых смесей. — М.: Наука, 1989. — 368 с.

Лебедев В.И. Функциональный анализ и вычислительная математика. — М.: Физматлит, 2000. — 295 с.

Лонгрен К. Экспериментальные исследования солитонов в нелинейных линиях передач с дисперсией / Солитоны в действии. — М.: Мир, 1981. — С. 138–162.

Ляшко И.И., Макаров В.Л., Скоробогатько А.А. Методы вычислений. — Киев: Высшая школа, 1972.

Макаров В.Л., Сплайн В.В. Аппроксимация функций. — М.: Высшая школа, 1983.

Малышев А.Н. Введение в вычислительную работу. — Новосибирск: Наука, 1999. — 25 с.

Манжиров А.В., Полянин А.Д. Методы решения интегральных уравнений. — М.: Факториал, 1999. — 272 с.

Марчук Г.И. Методы вычислительной математики. — 3-е изд. — М.: Наука, 1989.

Марчук Г.И., Шайдуров В.В. Повышение точности решений разностных схем. — М.: Наука, 1979.

Милн В.Э. Численное решение дифференциальных уравнений. — М.: ИИЛ, 1955. — 352 с.

Моисеев Н.Н. Численные методы теории оптимального управления. — М.: Изд. МГУ, 1968. — 380 с.

Никифоров А.Ф., Новиков В.Г., Уваров В.Б. Квантово-статистические модели высокотемпературной плазмы. — М.: Наука, 2000.

Ортега Дж. Введение в параллельные и векторные методы решения линейных систем. — М.: Мир, 1991. — 367 с.

Пирумов У.Г. Численные методы. — М.: Изд. МАИ, 1998. — 188 с.

Пирумов У.Г., Росляков Г.С. Газовая динамика сопел. — М.: Наука, 1990. — 368 с.

Плис А.И., Сливина Н.А. Лабораторный практикум по высшей математике. — М.: Высшая школа, 1994. — 416 с.

Программирование на параллельных вычислительных системах. — М.: Мир, 1991.

Ракитский Ю.В., Устинов С.М., Черноруцкий И.Г. Численные методы решения жестких систем. — М.: Наука, 1979.

Рихтмайер Р.Д., Мортон К. Разностные методы решения краевых задач. — М.: Мир, 1972.

Роуч П. Вычислительная гидродинамика. — М.: Мир, 1980. — 616 с.

Рунге К. Графические методы математического анализа. — М.-Л.: Гостехиздат, 1932. — 128 с.

Рогов Б.В. Метод минимальной длины для течений с трансзвуковой бифуркацией // ДАН, 2001. — Т. 381, № 1. — С. 23–26.

Рогов Б.В., Соколова И.А. Уравнения вязких течений в гладких каналах переменного сечения // ДАН, 1995. — Т. 345, № 5. — С. 615–618.

Рогов Б.В., Соколова И.А. Об асимптотической точности приближения гладкого канала при описании вязких течений // ДАН, 1997. — Т. 357, № 2. — С. 190–194.

Рогов Б.В., Соколова И.А. Гиперболическая модель вязких смешанных течений // ДАН, 2001. — Т. 378, № 5. — С. 628–632.

Рогов Б.В., Соколова И.А. Обзор моделей вязких внутренних течений // Матем. моделирование, 2002, а. — Т. 14, № 1. — С. 41–72.

Рогов Б.В., Соколова И.А. Гиперболическое приближение уравнений Навье–Стокса для вязких смешанных течений // Изв. РАН. МЖГ, 2002, б. — № 3. — С.30–49.

Рябенький В.С. Введение в вычислительную математику. — М.: Наука, 1994. — 335 с.

Самарский А.А. Введение в численные методы. — М.: Наука, 1987. — 286 с.

Самарский А.А. Теория разностных схем. — М.: Наука, 1977.

Самарский А.А., Андреев В.Б. Разностные методы для эллиптических уравнений. — М.: Наука, 1976. — 351 с.

Самарский А.А., Вабищевич П.Н. Вычислительная теплопередача. — М.: УРСС, 2003.

Самарский А.А., Галактионов В.А., Курдюмов С.П., Михайлов А.П. Режимы с обострением в задачах для квазилинейных параболических уравнений. — М.: Наука, 1987.

Самарский А.А., Гулин А.В. Численные методы. — М.: Наука, 1989. — 430 с.

Самарский А.А., Гулин А.В. Устойчивость разностных схем. — М.: Наука, 1973.

Самарский А.А., Михайлов А.П. Математическое моделирование: Идеи, методы, примеры. — М.: Наука, 1997. — 316 с.

Самарский А.А., Николаев Е.С. Методы решения сеточных уравнений. — М.: Наука, 1978. — 591 с.

Самарский А.А., Попов Ю.П. Разностные методы решения задач газовой динамики. — 3-е изд. — М.: Наука, 1992. — 423 с.

Самко С.Г., Кильбах А.А., Марычев О.И. Интегралы и производные дробного порядка и их применение. — Минск: Наука и техника, 1987. — 688 с.

Свешников А.Г., Боголюбов А.Н., Кравцов В.В. Лекции по математической физике. — М.: Изд. МГУ, 1993.

Свешников А.Г., Тверской М.Б. О точных решениях уравнения Дюбрейль–Жакотен // Ж. вычисл. матем. и матем. физ., 1995 г. — Т. 35, № 1. — С. 151–156.

Соболев С.Л. Об одной новой задаче математической физики // Изв. АН СССР, 1954. — Т. 18, № 1. — С. 3–50.

Соболь И.М. Численные методы Монте-Карло. — М.: Наука, 1973. — 311 с.

Софронов И.Л. Точные искусственные граничные условия для некоторых задач аэродинамики и дифракции. / Автореф. дис. докт. физ.-мат. наук. — М.: ИММ РАН, 1999.

Стренг Г. Линейная алгебра и ее применение. — М.: Мир, 1980. — 454 с.

Стрелков Н.А. (Сост.) Сборник задач по численным методам. — Ярославль: Изд. Яросл., 1988.

Тихонов А.Н., Арсенин В.Я. Методы решения некорректных задач. — М.: Наука, 1979. — 285 с.

Турчак Л.И., Плотников П.В. Основы численных методов. — 2-е изд. — М.: Физматлит, 2002. — 300 с.

Тирский Г.А. Континуальные модели в задачах гиперзвукового обтекания затупленных тел разреженным газом // ПММ, 1997. — Т. 61, № 6. — С. 903–930.

Тихонов А.Н., Костомаров Н.К. Вводные лекции по прикладной математике. — М.: Наука, 1984. — 160 с.

Уилкинсон Дж.Х. Алгебраическая проблема собственных значений. — М.: Наука, 1970.

Фадеев Д.К., Фадеева В.Н. Вычислительные методы линейной алгебры. — М.: Физматиз, 1963. — 734 с.

Федоренко Р.П. Введение в вычислительную физику. — М.: Изд. МФТИ, 1994. — 526 с.

Форсайт Дж., Малькольм М., Моулер К. Машинные методы математических вычислений. — М.: Мир, 1980. — 280 с.

Хайрер Э., Ваннер Г. Решение обыкновенных дифференциальных уравнений. Жесткие и дифференциально-алгебраические задачи. — М.: Мир, 1999. — 685 с.

Хайрер Э., Нерсетт С., Ваннер Г. Решение обыкновенных дифференциальных уравнений. Нежесткие задачи. — М.: Мир, 1990. — 512 с.

Хемминг Р.В. Численные методы. — М.: Наука, 1972. — 400 с.

Холл Дж., Уатт Дж. Современные численные методы решения обыкновенных дифференциальных уравнений. — М.: Мир, 1979. — 312 с.

Шифрин Э.Г. Потенциальные и вихревые трансзвуковые течения идеального газа. — М.: ФИЗМАТЛИТ, 2001. — 320 с.

Штеттер Н. Анализ методов дискретизации для обыкновенных дифференциальных уравнений. — М.: Мир, 1978. — 452 с.

Эльсгольц Л.Э., Норкин С.Б. Введение в теорию дифференциальных уравнений с запаздывающим аргументом. — М.: Наука, 1963. — 295 с.

Научное издание

КАЛИТКИН Николай Николаевич
АЛЬШИН Александр Борисович
АЛЬШИНА Елена Александровна
РОГОВ Борис Вадимович

ВЫЧИСЛЕНИЯ НА КВАЗИРАВНОМЕРНЫХ СЕТКАХ

Редактор *Е.Н. Глебова*
Оригинал-макет: *Е.Ю. Морозов*
Оформление переплета: *А.Ю. Алехина*

ЛР №071930 от 06.07.99. Подписано в печать 10.02.05.
Формат 60×90/16. Бумага офсетная. Печать офсетная.
Усл. печ. л. 14. Уч.-изд. л. 16. Тираж 400 экз. Заказ №

Издательская фирма «Физико-математическая литература»
МАИК «Наука/Интерпериодика»
117997, Москва, ул. Профсоюзная, 90
E-mail: fizmat@maik.ru, fmlsale@maik.ru;
http://www.fml.ru

Отпечатано с готовых диапозитивов
в ППП «Типография «Наука»
121099, г. Москва, Шубинский пер., 6