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

Препринты ИПМ • Препринт № 75 за 2020 г.

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

Белов А.А., Калиткин Н.Н., Хохлачев В.С.

Улучшенные оценки погрешности для экспоненциально сходящихся квадратур

Рекомендуемая форма библиографической ссылки: Белов А.А., Калиткин Н.Н., Хохлачев В.С. Улучшенные оценки погрешности для экспоненциально сходящихся квадратур // Препринты ИПМ им. М.В.Келдыша. 2020. № 75. 24 с. http://doi.org/10.20948/prepr-2020-75
URL: http://library.keldysh.ru/preprint.asp?id=2020-75

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

А.А. Белов, Н.Н. Калиткин, В.С. Хохлачев

УЛУЧШЕННЫЕ ОЦЕНКИ ПОГРЕШНОСТИ ДЛЯ ЭКСПОНЕНЦИАЛЬНО СХОДЯЩИХСЯ КВАДРАТУР

Москва — 2020

УДК 519.6

Белов А.А., Калиткин Н.Н., Хохлачев В.С.

Улучшенные оценки погрешности для экспоненциально сходящихся квадратур

Физикам часто требуется численно находить интегралы, причём с высокой точностью. В последние годы показано, что для некоторых практически важных классов функций возможно кардинальное увеличение точности и уменьшение трудоёмкости квадратур. В работе изложен соответствующий математический аппарат с новейшими улучшениями, что позволяет в сотни раз и более сократить трудоёмкость вычислений. Приводятся примеры физических задач, к которым он хорошо применим.

Ключевые слова: квадратура, формула трапеций, экспоненциальная сходимость

Alexander Alexandrovich Belov, Nikolay Nikolaevich Kalitkin,
Valentin Sergeevich Khokhlachev

Improved error estimates for an exponentially convergent quadratures

Physicists often need to calculate integrals numerically, and with high accuracy. In recent years, it has been shown that for some practically important classes of functions, it is possible to dramatically increase the accuracy and reduce the complexity of quadratures. The paper describes the corresponding mathematical apparatus with the latest improvements, which reduce the complexity of calculations by hundreds of times or more. Examples of physical problems to which it is well applicable are given.

Key words: quadrature, trapezoidal rule, exponential convergence

Работа поддержана грантом РФФИ 18-01-00175.

Оглавление

Введение

Прикладные задачи. В физических задачах часто требуется численно находить интегралы. Выделим два класса подынтегральных функций. Первый – периодические функции. Такие задачи возникают при разложении в ряды или интегралы Фурье; они являются некорректными [1]. Поэтому для них необходимо вычислять квадратуры с высокой точностью. Второй класс – функции, быстро убывающие на бесконечности. Приведём примеры таких задач.

1) Передача сигнала. Передать сигнал большого объёма дорого. Используют следующий подход. Сигнал разлагают в интеграл Фурье, и в памяти хранят дискретный Фурье образ. Его, в отличие от аналогового сигнала, можно программно сжать без потери информации. Сжатый сигнал передают. По нему восстанавливают Фурье образ и воспроизводят исходный сигнал.

2) Функции Ферми–Дирака. Они зависят от индекса k и параметра x:

$$I_k(x) = \int_0^\infty \frac{t^k dt}{1+\exp(t-x)}, \quad x \in (-\infty, +\infty). \tag{1}$$

Подынтегральная функции быстро убывает на бесконечности. Функции (1) были введены в [2, 3] для описания проводимости твёрдых металлов. Они являются моментами фермиевского распределения электронов. Затем эти функции были использованы для описания термодинамики плотной горячей плазмы [4]. Сейчас они используются во многих областях физики. Наиболее важны функции с полуцелыми k, но их трудно вычислять.

3) Плазменное микрополе. Хаотическое движение заряженных частиц в плазме создаёт в ней поля, флуктуирующие на расстояниях порядка межатомных. Именно этими микрополями определяются многие оптические свойства плазмы [5,6]. Распределение микрополя часто находят, представляя суммарное действие заряженных частиц интегралом Фурье, при этом Фурье образ корректируют с помощью различных модельных предположений. Первой была работа [7], в которой функция распределения равна

$$H(\beta) = \frac{2}{\pi}\beta \int_0^\infty \exp\left(-x^{3/2}\right) \sin(\beta x)\, x dx. \tag{2}$$

4) Скорости реакций. Поведение скоростей термоядерных реакций при низких температурах важно для расчёта зажигания дейтериевых мишеней. Для многих ядерных реакций измерены экспериментальные зависимости сечений от скоростей частиц $\sigma(v)$. Тогда зависимость скорости реакции от температу-

ры можно найти, усредняя поток по распределению Максвелла

$$K(T) = \langle \sigma(v)\, v \rangle = \frac{\pi}{\sqrt{m}} \left(\frac{2}{\pi T} \right)^{3/2} \int\limits_0^\infty \sigma(E)\, E \exp\left\{ -\frac{E}{T} \right\} dE,\ E = \frac{mv^2}{2}. \quad (3)$$

Подынтегральная функция быстро убывает при $v \to \infty$.

Вычисление квадратур. Для вычисления квадратур подобных типов в последние годы появился математический метод, ускоряющий вычисления в сотни раз. Оказалось, что для указанных классов функций квадратуры на равномерных сетках сходятся не по степенному закону, а по экспоненциальному. Наиболее полно данная проблема изложена в [8]; там же приведён исчерпывающий список литературы. В [9, 10] экспериментально было показано, что теоретические оценки работы [8] справедливы лишь для функций с полюсами первого порядка.

В данной работе получены следующие результаты. 1) Мажорантную оценку работы [8] можно заметно улучшить. 2) Приведенное в [8] доказательство справедливо лишь для функций с полюсами первого порядка. 3) Найдена эмпирическая оценка погрешности для функций с полюсами порядка выше первого. 4) Эмпирически показано, что зависимость от числа узлов в этих оценках является не мажорантной, а асимптотически точной. Для иллюстрации данных результатов проведены расчёты интегралов с известными точными значениями. Приведены примеры интегралов, для которых погрешность квадратурных формул зависит от шага сетки немонотонно и не описывается существующими теориями.

Экспоненциально сходящиеся квадратуры

В многочисленных учебниках по вычислительной математике строятся квадратурные формулы трапеций, средних, Симпсона и т.д. для функции $u(x)$ на сетках с шагом $h = L/N$, где L – длина отрезка интегрирования, а N – число шагов сетки. Доказывается, что для достаточно гладких функций погрешность таких квадратур есть $\mathcal{O}(M_p h^p) = \mathcal{O}(M_p N^{-p})$, где p – порядок точности конкретной формулы, а $M_p = \max \left| u^{(p)}(x) \right|$ на отрезке интегрирования. Такую сходимость называют степенной.

Квадратуры со степенной сходимостью наиболее распространены в практике вычислений, однако в последнее десятилетие выяснилось, что для некоторых классов функций эти квадратуры могут сходится много быстрее – по экспоненциальному и даже ещё более быстрым законам. В [8] дан хороший обзор предшествующей литературы и доказано несколько важных теорем о таком типе сходимости. Приведём сначала теоремы, относящиеся к интегрированию периодических функций по полному периоду.

Теорема 1 ([8], стр. 390)
Пусть $u(z)$ аналитическая в круге $|z| < R$ с $R > 1$, причём $|u(z)| \leq M_0$. Возьмём на единичной окружности равномерную сетку $z_n = e^{2\pi i n/N}, n = \overline{0, N}$. Рассмотрим на единичной окружности интеграл и квадратурную формулу трапеций:

$$I = \oint\limits_{|z|=1} u(z) \frac{dz}{iz},\ I_N = \frac{2\pi}{N} \sum_{n=1}^{N} \frac{u(z_{n-1}) + u(z_n)}{2}. \tag{4}$$

Для погрешности квадратуры справедлива оценка

$$\delta = |I - I_N| \leq \frac{2\pi M_0}{R^N - 1}. \bullet \tag{5}$$

Теорема 2 ([8], стр. 392)
Пусть $u(z)$ аналитическая в кольце $R^{-1} < |z| < R$ с $R > 1$, причём $|u(z)| \leq M_0$. Возьмём на единичной окружности равномерную сетку $z_n = e^{2\pi i n/N}$, $n = \overline{0, N}$. Рассмотрим на единичной окружности интеграл и квадратурную формулу трапеций:

$$I = \oint\limits_{|z|=1} u(z) \frac{dz}{iz},\ I_N = \frac{2\pi}{N} \sum_{n=1}^{N} \frac{u(z_{n-1}) + u(z_n)}{2}. \tag{6}$$

Для погрешности квадратуры справедлива оценка

$$\delta = |I - I_N| \leq \frac{4\pi M_0}{R^N - 1}. \bullet \tag{7}$$

Теорема 3 ([8], стр. 394)
Пусть $u(z)$ аналитическая в полуплоскости $\operatorname{Im} z > -l$, где $l > 0$, причём $u(z)$ имеет период 2π и $|u(z)| \leq M_0$. Введём на полном периоде вещественной оси равномерную сетку $x_n = 2\pi n/N, n = \overline{0, N}$. Рассмотрим интеграл и квадратурную формулу трапеций:

$$I = \int\limits_0^{2\pi} u(x)\, dx,\ I_N = \frac{2\pi}{N} \sum_{n=1}^{N} \frac{u(x_{n-1}) + u(x_n)}{2}. \tag{8}$$

Для погрешности квадратуры справедлива оценка

$$\delta = |I - I_N| \leq \frac{2\pi M_0}{e^{lN} - 1}. \bullet \tag{9}$$

Теорема 4 ([8], стр. 396)
Пусть $u(z)$ аналитическая в полосе $|\operatorname{Im} z| < l$, где $l > 0$, причём $u(z)$ имеет

период 2π и $|u(z)| \leq M_0$. Введём на полном периоде вещественной оси равномерную сетку $x_n = 2\pi n/N, n = \overline{0,N}$. Рассмотрим интеграл и квадратурную формулу трапеций:

$$I = \int_0^{2\pi} u(x)\,dx,\ I_N = \frac{2\pi}{N}\sum_{n=1}^{N}\frac{u(x_{n-1}) + u(x_n)}{2}. \tag{10}$$

Для погрешности квадратуры справедлива оценка

$$\delta = |I - I_N| \leq \frac{4\pi M_0}{e^{lN} - 1}. \bullet \tag{11}$$

Следствие ([8], стр. 398)

На практике периодическая функция может иметь произвольный период T. Если функция аналитична в полупространстве $\mathrm{Im}\, z > -l$, то в этом случае формулы (8), (9) принимают следующий вид:

$$I = \int_0^{T} u(x)\,dx,\ I_N = \frac{T}{N}\sum_{n=1}^{N}\frac{1}{2}\left[u\left(T\frac{n-1}{N}\right) + u\left(T\frac{n}{N}\right)\right], \tag{12}$$

$$\delta = |I - I_N| \leq \frac{TM_0}{\exp(2\pi lN/T) - 1}. \tag{13}$$

Если функция $u(z)$ аналитична в полосе $|\mathrm{Im}\, z| < l$, то в оценке погрешности δ (13) надо увеличить числитель в 2 раза.

В [8] приведены доказательства теорем 1 – 2 с помощью разложения в ряды Лорана. Теорема 3 сводится к теореме 1, а теорема 4 сводится к теореме 2 с помощью замены переменных $z = e^{ix}$.

Комментарии. Обсудим приведённые теоремы.

1° Оригинальная формулировка приведённых выше теорем в [8] была несколько иной. В них записывалась квадратура правых прямоугольников

$$I_N = \frac{2\pi}{N}\sum_{n=1}^{N} u(z_n). \tag{14}$$

Мы записали вместо неё квадратуру трапеций по следующим соображениям.

В теоремах 3 – 4 функция $u(x)$ рассматривается на полном периоде. В теоремах 1 – 2 функция $u(z)$, рассматриваемая вдоль замкнутой кривой, также является периодической. В силу периодичности $u_0 = u_N$. Легко убедиться, что в этом случае формулы трапеций и прямоугольников оказываются эквивалентными.

Более того, для интеграла по полному периоду выбор начальной точки безразличен. Поэтому можно выбрать отрезок $[h/2, 2\pi + h/2]$; тогда квадратура (14) становятся формулой средних.

Таким образом, для интеграла по полному периоду квадратурные формулы прямоугольников, средних и трапеций на равномерной сетке оказываются эквивалентны. Теоремы 1 – 4 доказывают, что сходимость этих квадратурных формул на равномерной сетке при интегрировании по полному периоду функции имеет не степенной характер.

2° Проанализируем подробнее поведение погрешности в теоремах 3 – 4. Если $lN \lesssim 0.2$, то $\delta \sim (lN)^{-1}$; тогда погрешность степенная с порядком точности $p = 1$. Если же $lN \gtrsim 2$, то погрешность $\delta \sim \exp(-lN)$, то есть убывание погрешности не степенное, а экспоненциальное. Оно гораздо быстрее степенного.

Поясним смысл величины l: это наименьшее расстояние от всех особых точек функции $u(z)$ в комплексной плоскости до ближайшей точки отрезка интегрирования. С учётом периодичности функции это просто наименьшее из всех расстояний особых точек от вещественной оси. Переход от степенной зависимости к экспоненциальной происходит в нешироком диапазоне $0.2 \lesssim lN \lesssim 2$. Задачи с $l \ll 1$ являются очень трудными, но в практике вычислений они появляются редко. Гораздо чаще l не мало, а тогда уже при небольших N сходимость становится экспоненциальной.

Погрешность квадратур в теоремах 1 – 2 заменой $R^N = e^{N \ln(R)}$ сводится к погрешности квадратур в теоремах 3 – 4. При этом вместо величины l стоит $\ln(R)$.

Таким образом, зависимость от N в (5), (7), (9), (11) является обычно не степенной, а экспоненциальной. Такая сходимость гораздо быстрее степенной и имеет качественно другой характер. Для функций, удовлетворяющих требованиям любой из теорем 1 – 4, трудоемкость квадратур очень сильно уменьшается. Это указывает на большую практическую ценность данных теорем.

3° Мы провели тщательный анализ доказательств теорем, приведенных в [8]. Было обнаружено, что в этих доказательствах делалось неявное предположение о том, что все особые точки $u(z)$ вне области её аналитичности являются полюсами первого порядка. Поэтому возникает вопрос: каково будет поведение погрешности квадратур при других типах особых точек $u(z)$ – например, полюсах высоких порядков или существенно особых точках? Далее мы исследуем эту проблему.

4° В теореме 1 $u(z)$ внутри круга $|z| < R$ имеет только одну особую точку – полюс первого порядка при $z = 0$. Поэтому интеграл определяется вычетом: $I = 2\pi\, u|_{z=0}$. Казалось бы, это позволяет точно найти интеграл в теореме 3. Однако переход к вещественной оси определяется соотношением $z = e^{ix}$, то есть $x = -i \ln(z)$ и точке $z = 0$ нельзя сопоставить никакое значение x. Тем самым искомый вычет определить невозможно. Поэтому точного решения для

интеграла (8) получить отсюда не удаётся.

5° Отметим ещё один случай, имеющий практическую ценность. Пусть функция $u\,(x)$ интегрируется на отрезке $[0,T]$, причём все её нечётные производные на обоих концах отрезка обращаются в нули: $u^{(2p-1)}\,(0)=u^{(2p-1)}\,(T)\,,\ p=\overline{1,\infty}$. Такую функцию можно чётно продолжить за обе границы отрезка, при этом она становится периодической с периодом $2T$. В этом случае применимы теоремы 3 – 4.

6° Мажорирующая константа. Пусть функция постоянна: $u(z)\equiv C$. Очевидно, для такой функции формула трапеций точна. Поэтому погрешности (5), (7), (9), (11) для функции $u(z)-C$ таковы же, как для функции $u(z)$. Однако константа M_0 для функции $u(z)-C$ будет другой.

Выберем такую константу C, чтобы минимизировать величину M_0 для новой функции. Для вещественной функции $u(x)$ этот выбор очевиден:

$$C=\frac{1}{2}(\max u\,(x)+\min u\,(x)). \tag{15}$$

Это дает

$$M=\frac{1}{2}\left(\max u\,(x)-\min u\,(x)\right). \tag{16}$$

Легко видеть, что всегда $M\le M_0$. Если же $u(x)>0$, причем $\max u\,(x)\,/\,\min u\,(x)$ близко к 1, то $M\ll M_0$.

Аналогичную оценку можно построить для комплексной функции вещественного аргумента x. Рассмотрим значения функции только на вещественном отрезке $x\in[0,2\pi]$. Положим

$$C=\frac{1}{2}\left(\max\mathrm{Re}\,u(x)+\min\mathrm{Re}\,u(x)\right)+i\frac{1}{2}\left(\max\mathrm{Im}\,u(x)+\min\mathrm{Im}\,u(x)\right). \tag{17}$$

Тогда получим оценку

$$M=\frac{1}{2}\sqrt{\left(\max\mathrm{Re}\,u(x)-\min\mathrm{Re}\,u(x)\right)^2+\left(\max\mathrm{Im}\,u(x)-\min\mathrm{Im}\,u(x)\right)^2}. \tag{18}$$

Для вещественных функций оценка (18) переходит в (16). Заметим, что авторы [8] считали свою оценку M_0 неулучшаемой.

Однако при интегрировании по половине периода квадратуры прямоугольников, средних и трапеций уже не эквивалентны. Экспоненциальную сходимость сохраняет только формула трапеций (19). Для функции $u\,(z)$, аналитичной в полуплоскости, получаются следующие формулы:

$$I=\int\limits_0^T u\,(x)\,dx,\ I_N=\frac{T}{N}\left(\frac{1}{2}u\,(0)+\sum_{n=1}^{N-1}u\left(\frac{Tn}{N}\right)+\frac{1}{2}u\,(T)\right), \tag{19}$$

$$\delta = |I - I_N| \le \frac{TM_0}{\exp(2\pi lN/T) - 1}. \qquad (20)$$

Но в них T есть полупериод функции, а не период. Для $u(z)$ аналитичной в полосе числитель в (20) следует умножить на 2.

Полюсы первого порядка

В работах [9–11] экспоненциально сходящиеся квадратуры применялись к важной практической задаче – построению способов быстрого вычисления функций Ферми-Дирака. Для тщательного исследования погрешности квадратур традиционно используют численные тесты. В качестве теста берут интеграл и подынтегральную функцию, удовлетворяющую требованиям одной из приведённых теорем. При этом значение тестового интеграла должно быть известным, то есть выражаться через элементарные функции (в крайнем случае, через простейшие специальные функции математической физики). Тогда можно проводить серию расчётов с различным числом интервалов сетки N. При этом точное значение погрешности находится непосредственно как разность численного расчёта и точного значения тестового интеграла. Это значение следует сравнивать с теоретической оценкой погрешности. Отсюда можно делать вывод об адекватности теоретической оценки.

Тест. Для экспериментальной проверки теоремы 4 мы взяли следующий интеграл ([12], стр. 383, №3.616):

$$\int_0^{\pi} \frac{\cos(rx)\ dx}{(a^2 - 2a\cos(x) + 1)^q} = \frac{\pi}{a^r} \sum_{k=0}^{q-1} \binom{q+r-1}{k} \binom{2q-k-2}{q-1} (a^2-1)^{k-2q+1}; \quad (21)$$

здесь $q \ge 1$ и $r \ge 1$ целые, $a > 1$. Заметим, что интеграл (21) берётся на полупериоде, а подынтегральная функция является чётной относительно обеих границ. Поэтому к нему применим комментарий ***5°***.

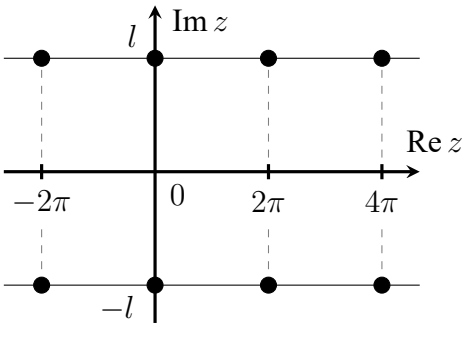

Рис. 1. Полюсы подынтегральной функции (21) при $r = 0$.

Подынтегральная функция имеет особые точки, определяемые нулями знаменателя. Это полюсы порядка q. Их положение в комплексной плоскости определяется формулой

$$z_k = \pm i \ln(a) + 2\pi k,\ -\infty < k < +\infty, \qquad (22)$$

где k целое. Особые точки лежат на двух прямых, параллельных вещественной оси (см. рис. 1) и удалённых от неё на расстояние $l = \ln(a)$. Таким образом, подынтегральная функция $u(z)$ аналитична в полосе $|\mathrm{Im}\, z| < l = \ln(a)$. Ширина этой

полосы не может быть увеличена, так как полюсы лежат на границе указанной полосы. Поэтому l есть в точности расстояние до ближайшей особой точки.

Строго говоря, числитель $\cos(rx)$, при $r > 0$ приводит к появлению особой точки при $z = \infty$. Но дополнительная погрешность, обусловленная этой особой точкой, пренебрежимо мала по сравнению с погрешностью, обусловленной полюсами. Вдобавок мы почти всюду будем рассматривать примеры с $r = 0$, где этот полюс и связанная с ним погрешность отсутствуют.

Простой полюс. В этом разделе мы ограничимся случаем $q = 1$. Для этого случая теорема 4 была строго доказана в [8].

Мы провели расчёты для двух значений расстояния особой точки от вещественной оси: когда это расстояние $l = 0.05$ много меньше периода и когда $l = 1$ сопоставимо с периодом. Расчёты проводились для всех N от 1 до 20. Рассмотрим результаты расчётов.

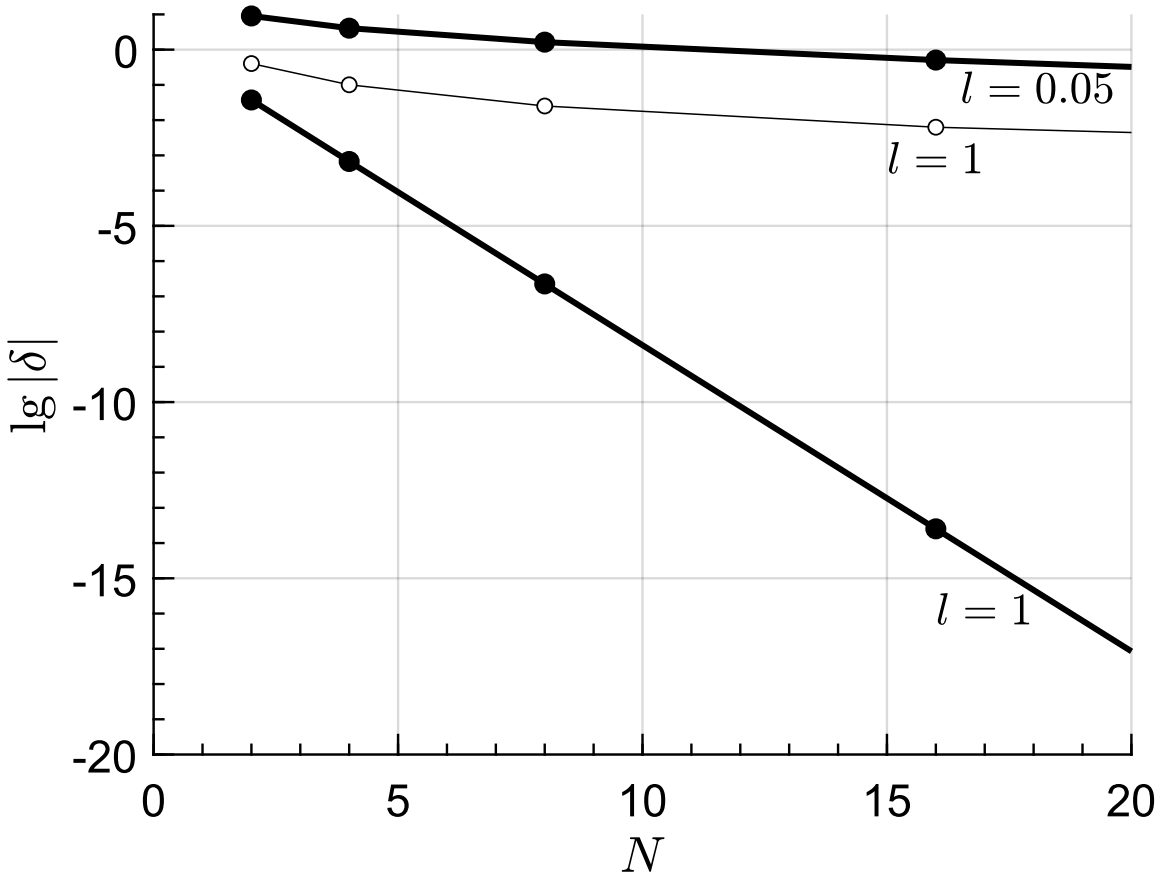

Рис. 2. Погрешность для теста (21) с $r = 0$. Около линий указаны значения l. Жирные линии – численные расчёты, тонкая линия – мажорантная оценка.

На рис. 2 чёрными маркерами показана зависимость погрешности от N. Поскольку при больших N зависимость экспоненциальная, то выбран полулогарифмический масштаб: N по оси абсцисс и $\lg|\delta|$ по оси ординат. В полулогарифмическом масштабе экспоненциальная зависимость должна изображаться

прямой линией. Действительно, график для $l = 1$ является прямой линией. Для $l = 0.05$ начало графика немного искривлено, но уже при $N \gtrsim 10$ (что соответствует $lN \gtrsim 0.5$) линия визуально становится прямой.

Видно, что при близком полюсе сходимость довольно медленная, но при неблизком полюсе сходимость очень быстрая: уже при небольшом $N = 19$ получается 16 верных знаков, то есть точность единичного округления компьютера, при расчётах с double precision.

Для сравнения светлыми маркерами показана теоретическая оценка погрешности для случая $l = 1$, соответствующая обычному степенному закону сходимости формулы трапеций:

$$\delta = \frac{1}{12} M_2 T h^2, \; M_2 = \max \left| u''(x) \right|, \tag{23}$$

где максимум берётся по вещественному отрезку интегрирования. В полулогарифмическом масштабе данная кривая представляется искривлённой линией, она лежит много выше кривой экспоненциальной сходимости. Например, при $N = 19$ степенная сходимость даёт только 2 верных знака. Это наглядно демонстрирует преимущества экспоненциальной сходимости.

Исключительно интересен рис. 3. Его ордината есть логарифм отношения теоретической оценки, в которой неизвестная константа M взята единицей, к фактической погрешности $|I - I_N|$, а l есть точное расстояние от полюса до вещественной оси. Видно, что зависимость этого отношения от N оказывается горизонтальной прямой как при умеренном l, так и при очень малом. Это означает, что зависимость фактической погрешности от lN является не мажорантной и даже не асимптотически точной, а просто точной! Этот экспериментально полученный результат удивителен. Ведь если l есть точное расстояние до ближайшего полюса, то в полосе $|\mathrm{Im}\, z| < l$ $u(z)$ является неограниченной, так как полюс лежит на границе полосы. Поэтому не существует константы M, мажорирующей $|u(z)|$ в полосе. Тем не менее, теорема выполняется с некоторой константой M, хотя смысл этой константы остаётся неясным.

Практические рекомендации. Из практики расчетов известно, что при степенном характере сходимости $\delta \sim \mathcal{O}(N^{-p})$ знак реальной погрешности начиная с некоторого N_0 не меняется при дальнейшем увеличении N. Это является важным свойством, которое позволяет строить оценки точности методом Ричардсона по сгущению сеток.

В данных расчетах мы наблюдали аналогичную картину: погрешность сохраняла свой знак при сгущении сеток. Поэтому здесь также возможно апостериорное определение погрешности по расчетам на сгущающихся сетках. При этом удобно проводить расчеты не при всех N подряд, а только при последовательно удваивающихся N.

Однако сами эти оценки будут другими. Мы используем их при $lN \gg 1$, когда можно считать $\delta_N \approx \mathrm{const} \cdot \exp(-lN)$. В этом случае $\delta_{2N} \sim \mathcal{O}(\delta_N^2)$. Это

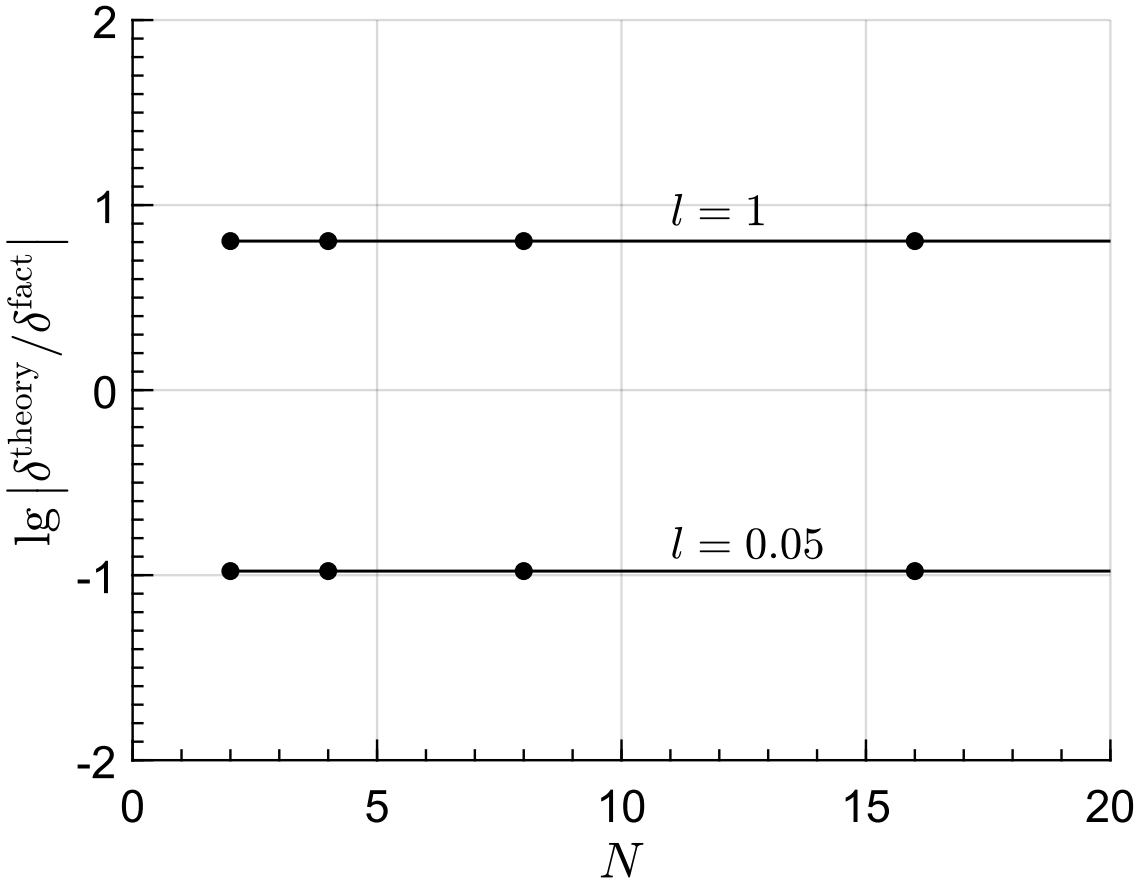

Рис. 3. Отношение теоретической оценки δ^{theory}, в которой заменено $M = 1$, к фактической погрешности δ^{fact}. Около линий указаны значения l.

напоминает ньютоновскую сходимость, которую обычно называют квадратичной (зачастую не совсем корректно говорят, что число верных знаков удваивается с удвоением N). Это позволяет сформулировать следующую практическую рекомендацию.

Если требуется получить точность ε, то остановим вычисления, как только выполнится условие $|I_N - I_{2N}| < \varepsilon^\nu$, где показатель степени выбирают в пределах $\nu \approx 0.65 \div 0.75$. Например, при ошибке единичного округления компьютера $\varepsilon = 10^{-16}$ рекомендуется брать $\varepsilon^\nu \sim 10^{-10} \div 10^{-12}$. Меньший показатель может не дать требуемой точности, больший – привести к избыточным вычислениям или отсутствию сходимости из-за ошибок округления.

Быстро осциллирующая функция. Знаменатель подынтегральной функции теста (21) положителен. Числитель при $r > 0$ является осциллирующей функцией, причём при $r \gg 1$ это быстро осциллирующая функция: на отрезке $0 \le x \le \pi$ расположено r полуволн этой функции. Вычисление интегралов от быстро осциллирующих функций само по себе является трудной задачей. В обычных сеточных методах для обеспечения удовлетворительной точности приходится ставить довольно много узлов сетки на одну полуволну.

Мы провели такие расчёты для трёх значений $r = 10,\ 50$ и 100, для сравнения к ним были присоединены расчёты для $r = 0$, когда осцилляции отсутствуют. Расстояние до полюса было выбрано очень небольшим $l = 0.05$, чтобы сделать задачу ещё более трудной.

При возрастании r само значение интеграла быстро убывает (см. табл. 1). Поэтому на графиках целесообразно отображать не саму величину погрешности, а отношение погрешности к величине интеграла. Именно такая относительная погрешность показана на рис. 4 в полулогарифмическом масштабе. Около каждой кривой написана величина r.

Таблица 1. Значения интеграла (21) при $l = 0.05$ и различных r

r	0	10	50	100
I	$29.87130579\ldots$	$18.11786280\ldots$	$2.451986094\ldots$	$0.2012712752\ldots$

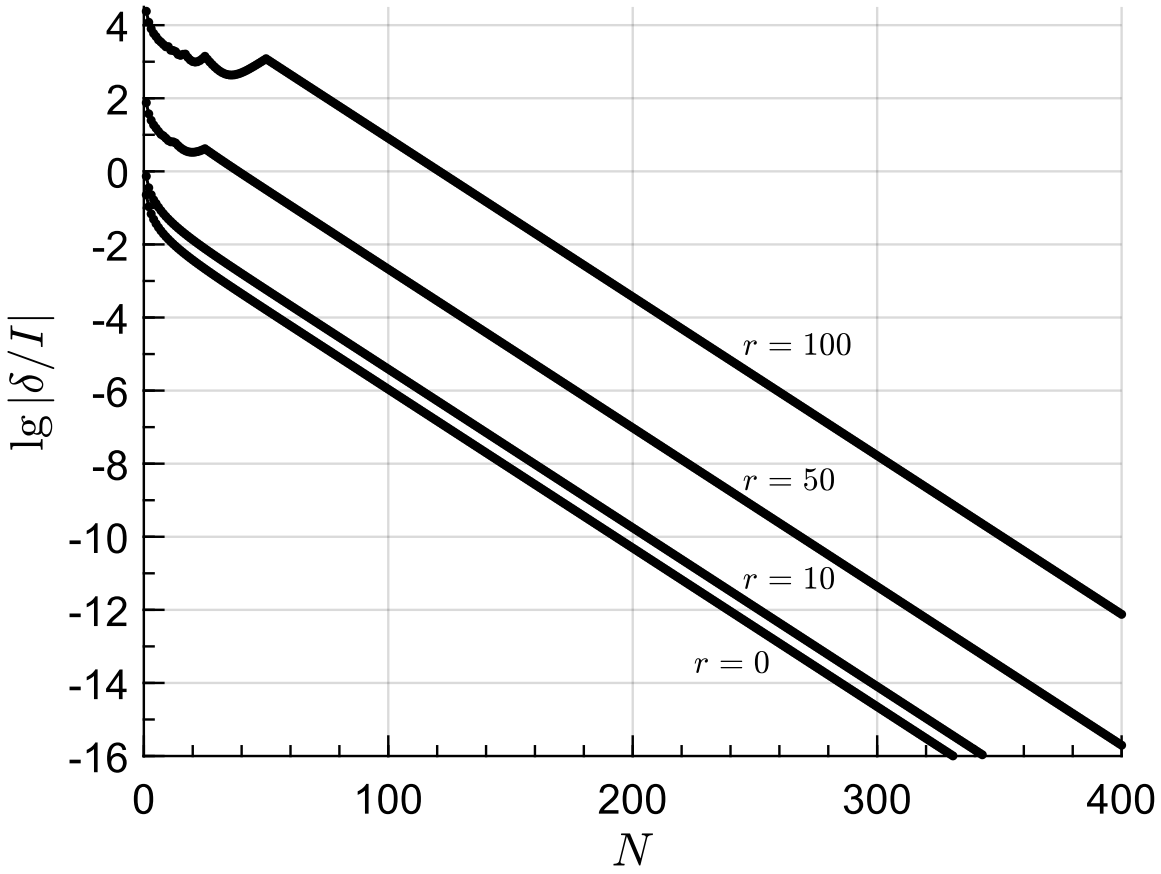

Рис. 4. Относительные погрешности для теста (21) с $l = 0.05$. Около кривых указаны значения параметра r.

График погрешности при $r = 0$ был обсуждён выше: линия слегка искривлена при малых N из-за очень мало расстояния l до полюса, а далее она прямолинейна. Погрешность компьютерного округления 10^{-16} достигается при $N \approx$

≈ 325. Начальная точка кривой лежит довольно близко к началу координат, так как при $N = 1$ относительная погрешность составляет примерно 0.23 от величины интеграла.

При $r = 10$ начальный искривлённый участок несколько длиннее, но уже при $N \approx 20$ линия становится прямой; при этом на каждую полуволну приходится лишь 2 интервала сетки. Эта прямая параллельна предыдущей, поскольку их наклон определяется одним и тем же знаменателем $\exp(-lN)$. Выход на ошибки округления здесь происходит при $N \approx 350$. Начальная точка прямой лежит почти в начале координат: там погрешность составляет примерно 0.74 от величины интеграла.

При $r = 50$ начальная точка имеет положительную ординату, так как погрешность в 76 раз превосходит само значение интеграла. Начальный участок имеет более сложное поведение, но уже при $N > 25$ линия становится прямолинейной, хотя на каждую полуволну приходится всего полинтервала сетки! Эта прямая параллельна предыдущим, и компьютерная точность достигается при $N \approx 407$, при этом на полуволне расположено 8 интервалов сетки.

При $r = 100$ начальная точка кривой лежит ещё выше, а нерегулярный участок кончается при $N = 50$; это соответствует половине интервала на одну полуволну. Далее следует прямолинейный участок с тем же наклоном. При $N = 400$ достигается высокая точность $\sim 10^{-12}$, хотя на каждую полуволну приходится 4 интервала сетки.

Таким образом, даже для быстро осциллирующих функций экспоненциальная сходимость реализуется и позволяет добиться высокой точности при небольшом числе улов сетки.

Полюсы целых порядков

Обобщение оценки погрешности. Мы провели тщательный анализ доказательства теорем 1 – 4. Удалось обнаружить, что при доказательстве теорем сделано неявное предположение о том, что ближайшая особенность является полюсом первого порядка. Поэтому мы экспериментально исследовали, какова будет погрешность для полюсов целого порядка $q > 1$. Для исследования был взят тестовый интеграл (21) при $r = 0$ (если взять $r > 0$, то как показано выше, детали поведения кривой сходимости усложняются, однако её асимптотическое поведение при $lN \gg 1$ остаётся тем же).

Численные расчеты показали, что фактическая погрешность при больших N может превышать мажорантные оценки из теорем 1 – 4. Тогда, поскольку указанные оценки мажорантны, то превышение означает, что функциональная зависимость погрешности от N должна быть несколько иной. Более тщательный анализ позволил предложить следующую эвристическую зависимость по-

грешности от N:

$$\delta = 2\pi M \frac{(lN/q + \exp(-lN/q))^{q-1}}{\exp(lN) - 1}. \tag{24}$$

Для полюса первого порядка ($q = 1$) она переходит в оценки [8]. Видно, что при $lN \gg 1$ главным членом является экспонента в знаменателе, то есть погрешность будет близка к экспоненциальной. Наличие множителя $(lN)^{q-1}$ в числителе несколько замедляет скорость сходимости, причём тем сильнее, чем больше q, то есть чем выше порядок полюса. Тем не менее сходимость очень быстрая.

Апробация оценки. Результаты расчетов тестового интеграла (21) с $l = 0.05$ и $r = 0$ показаны на рис. 5. По оси ординат отложен логарифм отношения теоретической оценки (24) к фактической погрешности расчета. Около каждой кривой указана кратность полюса q. Для сравнения приведена кривая для $q = 1$.

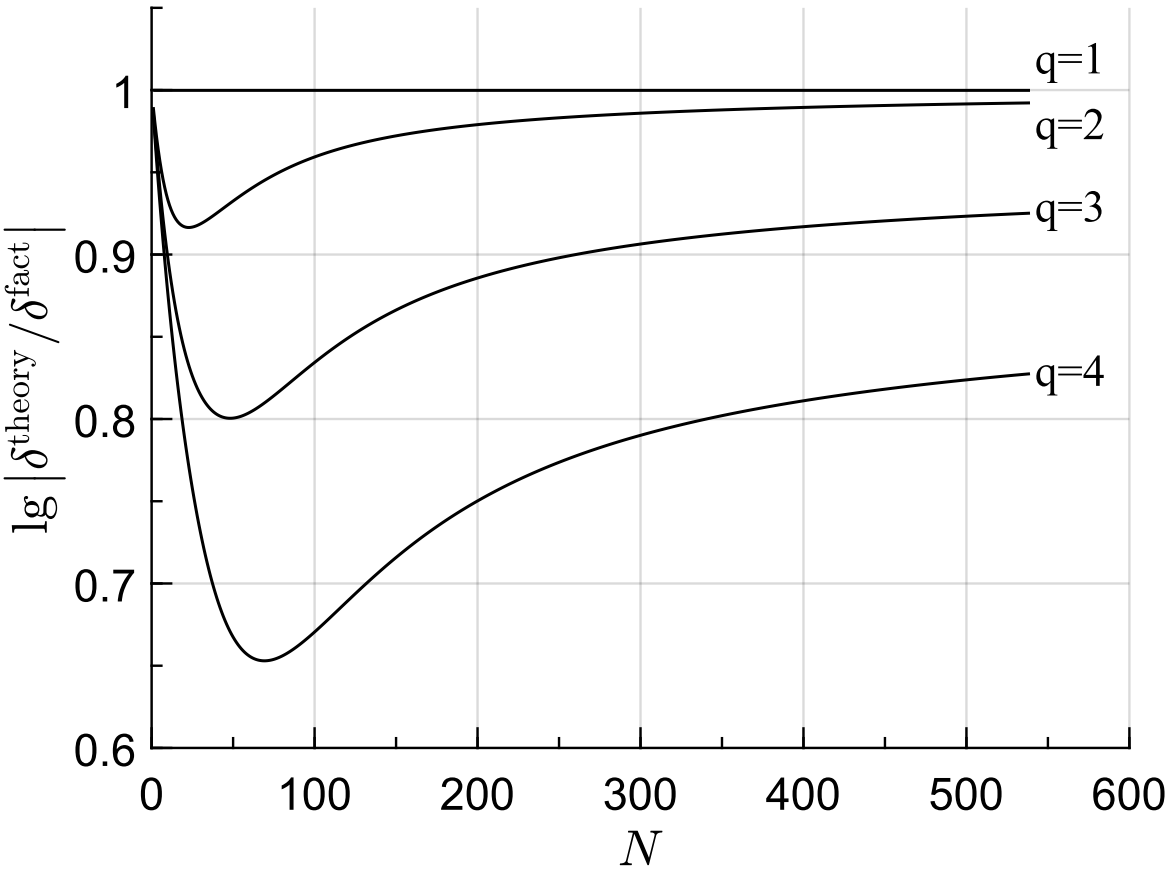

Рис. 5. Отношение эвристической оценки (24) к фактической погрешности. Около кривых указаны значения q.

Видно, что линия $q = 1$ горизонтальна. Остальные линии – кривые с минимумом, переходящие в горизонтальные линии при увеличении N. Это показывает, что предложенная эвристическая оценка (24) является асимптотически точной по N. Выход на асимптоту происходит тем медленнее, чем выше по-

рядок полюса q. Из высот горизонтальных асимптот видно, что при $q = 1$ и $q = 2$ теоретическая оценка в 10 раз больше фактической погрешности. При $q = 3$ превышение составляет 8.5 раз; при $q = 4$ оно равно 6.6 раз. Такие превышения показывают, что функциональная зависимость погрешности от N, l, q подобрана достаточно удачно, но использование константы M из формулы (16) приводит к слишком большому превышению (константа M_0 даст ещё худшие результаты).

Отметим, что в этих расчетах реальная погрешность сохраняет свой знак при увеличении N, как это было для полюса кратности 1 при $r = 0$. Поэтому здесь остаются в силе те же практические рекомендации по выбору сеток и критерию окончания расчета для получения заданной точности.

Непериодические функции

Экспоненциальная сходимость

В [8] были рассмотрены не только периодические функции, но и непериодические функции на вещественной прямой. Для них была доказана следующая

Теорема 5 ([8], стр. 396)
Пусть $u(z)$ аналитическая в полосе $|\mathrm{Im}\, z| < l$, где $l > 0$. Пусть также $u(z) \rightrightarrows 0$ при $|z| \to \infty$ в данной полосе, и существует такое M, что для любого $b \in (-l, l)$

$$\int_{-\infty}^{\infty} |u(x + ib)|\, dx \le M. \tag{25}$$

Тогда квадратура трапеций на равномерной сетке существует, а для её погрешности справедлива оценка:

$$|\delta| = \left| \int_{-\infty}^{\infty} u(x)\, dx - h \sum_{k=-\infty}^{\infty} u(kh) \right| \le \frac{2M}{\exp(2\pi l/h) - 1}. \bullet \tag{26}$$

Заметим, что в теореме 5 не указывается тип особой точки. Однако приведённое в [8] доказательство справедливо лишь для особых точек, являющихся полюсами первого порядка.

Следствие. Пусть $u(x)$ является чётной и удовлетворяет условиям теоремы 5. В этом случае можно ограничиться рассмотрением интеграла на полупрямой $0 \le x < +\infty$ с соответствующим уточнением формулы трапеций

$$\left| \int_0^{\infty} u(x)\, dx - h \left[\frac{1}{2} u(0) + \sum_{k=1}^{+\infty} u(kh) \right] \right| \le \frac{M}{\exp(2\pi l/h) - 1}, \tag{27}$$

где M сохраняет то значение, которое указано в теореме 5. Это следствие естественно обобщается на случай, когда $u\,(x)$ чётна не относительно точки $x = 0$, а относительно произвольной точки.

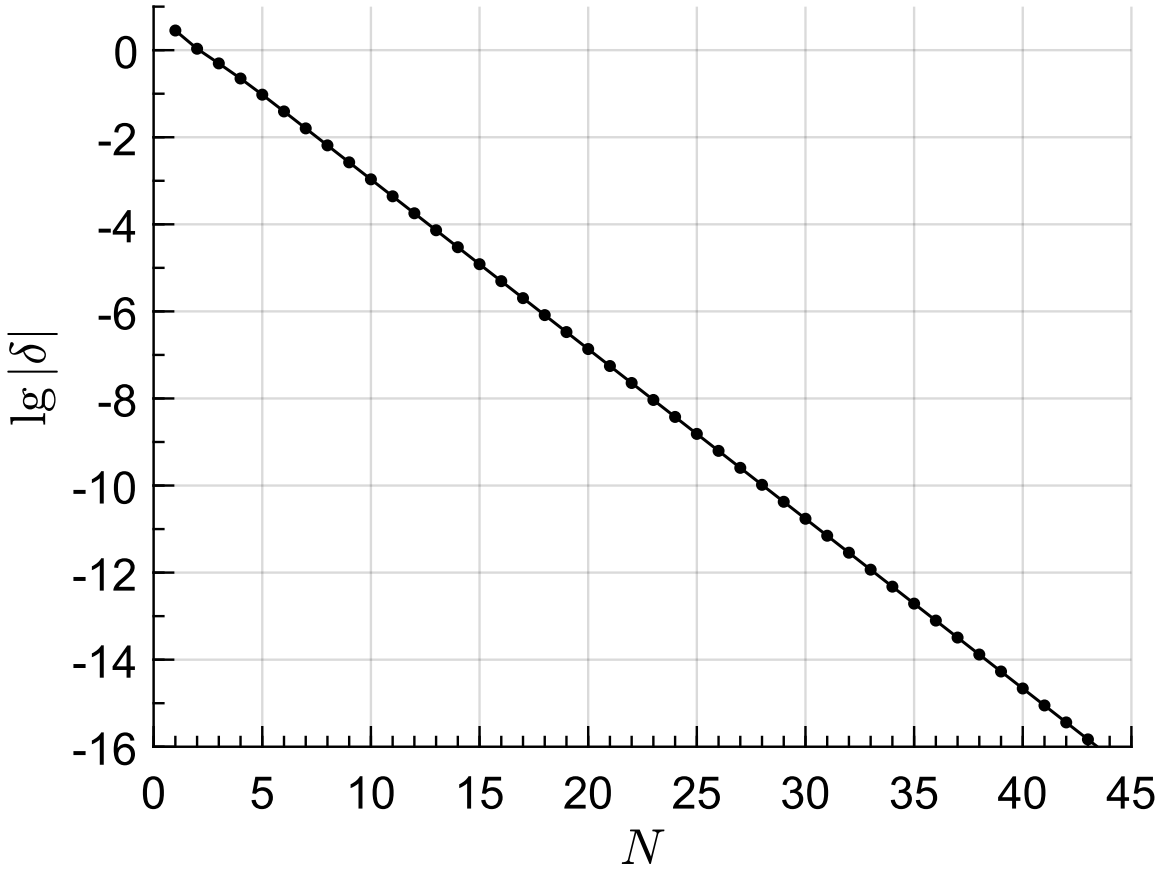

Рис. 6. Погрешность для теста (28).

Тест. Для верификации этих оценок мы взяли хороший тестовый интеграл, точное значение которого известно ([12], стр. 352, №3.466)

$$I = \int\limits_0^\infty \frac{e^{-x^2}}{1+x^2}dx = \frac{1}{2}e\pi\,(1 - \mathrm{erf}\,(1))\,. \tag{28}$$

Подынтегральная функция имеет полюсы первого порядка при $z = \pm i$. Кроме того, подынтегральная функция неограниченно растёт при $|\mathrm{Im}\,z| \to \infty$. Таким образом, функция является аналитической в полосе $|\mathrm{Im}\,z| < 1$. В суженной полосе $|\mathrm{Im}\,z| < 1-\varepsilon$ она удовлетворяет следствию из теоремы 5; однако величина константы зависит от ε, причём $M \to \infty$ при $\varepsilon \to 0$.

Суммировать по k до бесконечности невозможно. Поэтому для практической реализации теоремы 5 необходимо ограничить верхний предел суммы $k \leq N$. Это соответствует ограничению интеграла величиной $X = Nh$. При этом N

определяется как $N = X/h$. Величину X выбирают так, чтобы отброшенным "хвостом" интеграла можно было заведомо пренебречь при выбранной разрядности чисел.

Мы рассчитывали тест (28) на 64-разрядном компьютере, то есть погрешность единичного округления составляла $10^{-16} \approx e^{-36.8}$. Мы брали $X = 7$. Тогда $u(X)/u(0) < 10^{-23}$, так что отброшенным "хвостом" интеграла можно заведомо пренебречь.

Результаты численного расчёта представлены на рис. 6, на нём в полулогарифмическом масштабе изображена зависимость фактической погрешности (то есть разности суммы и точного значения интеграла) от N. Линия с высокой точностью оказывается прямой. Это неудивительно, так как расстояние до полюсов сравнительно велико $l = 1$.

Теоретическая оценка наклона этой прямой есть $-2\pi l \lg e/X \approx -0.389$; она отлично совпадает с фактическим наклоном расчётной линии.

Таким образом, численный расчёт, во-первых, хорошо подтверждает теоретическую оценку (26) теоремы 5. Во-вторых, он показывает, что в эту оценку входит не мажорантное значение $l - \varepsilon$, а точное значение расстояния до ближайшего полюса. Это согласуется с расчётами, сделанными для периодических функций.

Сверхэкспоненциальная сходимость

Бесконечно удаленная особенность. Если подынтегральная функция имеет только бесконечно удаленную особую точку, то этот случай не подпадает под теорему 5: во-первых, следует полагать $l = \infty$; во-вторых, эта особая точка не является полюсом. Интуитивно можно ожидать, что сходимость будет существенно быстрее экспоненциальной.

В качестве тестового примера рассмотрим следующий интеграл, точное значение которого известно ([12], стр. 320, №3.321):

$$I = \int_0^\infty \exp(-x^2)dx = \frac{\sqrt{\pi}}{2}. \tag{29}$$

Единственная особая точка подынтегральной функции является бесконечно удалённой. При этом скорость возрастания функции $u(z)$ при $|\operatorname{Im} z| \to \infty$ почти не отличается от теста (28). Для аппроксимации этого интеграла квадратурной формулой трапеций (27) в ([8], стр. 405 (21)) была получена грубая теоретическая оценка:

$$\delta = \mathcal{O}\left(e^{-\pi^2/h^2}\right). \tag{30}$$

При уменьшении h ошибка убывает гораздо быстрее, чем в теореме 5. Поэтому такую сходимость можно называть сверхэкспоненциальной.

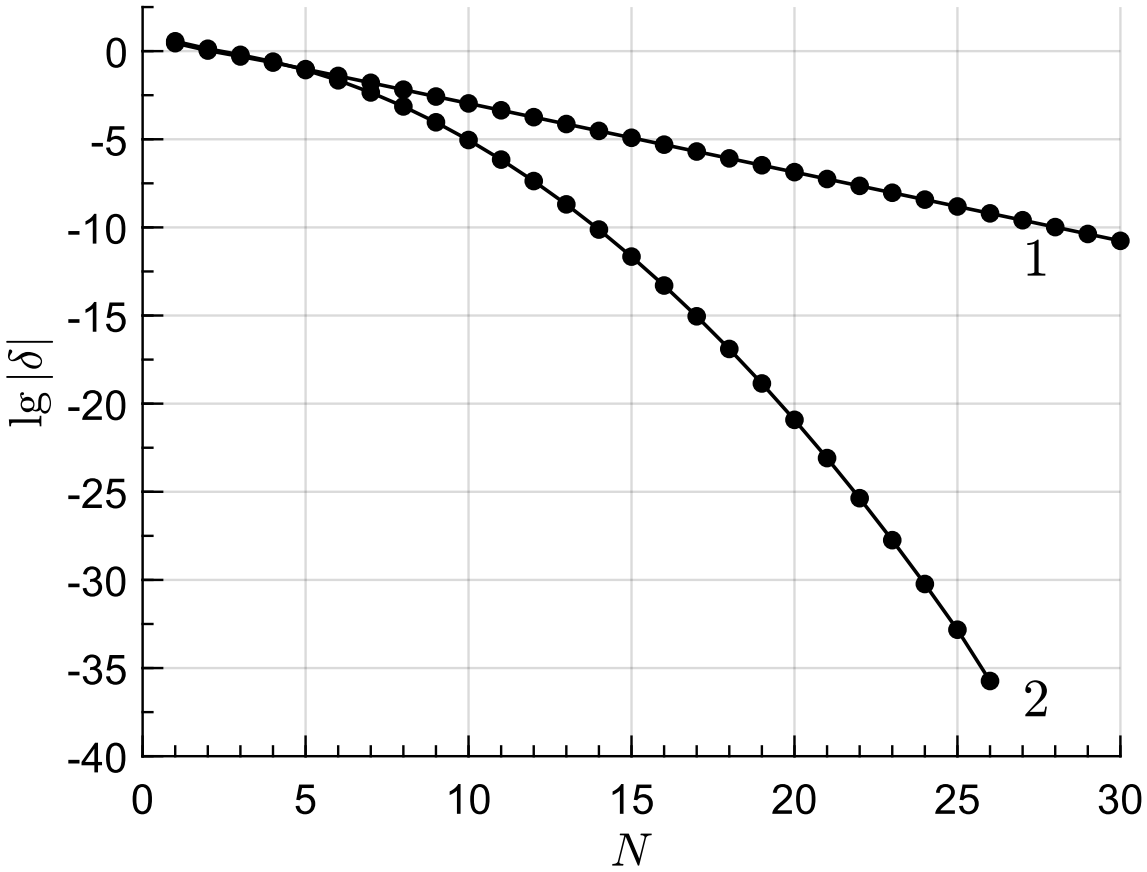

Рис. 7. Линия 1 – погрешность для теста (28), линия 2 – для теста (29).

Мы провели расчёт этого теста на тех же сетках, что и для теста (28). Зависимость фактической погрешности от числа расчётных интервалов N (тем самым – от $h \sim N^{-1}$) показана на рис. 7. Видно, что при малых N эта кривая почти совпадает с прямой, соответствующей экспоненциальной сходимости. Однако при дальнейшем возрастании N погрешность начинает убывать гораздо быстрее. Предельная точность double precision достигается уже при $N = 18$, в то время как для экспоненциальной сходимости это происходит при $N = 43$. Этот пример чётко иллюстрирует преимущества сверхэкспоненциальной сходимости перед экспоненциальной.

Сделаем частное замечание – сравним сходимость квадратурных формул с итерационными процессами решения уравнения $f(x) = 0$. Степенную сходимость $\delta \sim \mathcal{O}(h^p)$ можно сопоставить с простыми итерациями: при сгущении сетки вдвое погрешность убывает в одно и то же число раз 2^p. Экспоненциальная сходимость $\delta \sim e^{-\text{const}/h}$ аналогична Ньютоновским итерациям: $\delta_{2N} = \mathcal{O}\left(\delta_N^2\right)$. Расчет показывает, что сходимость квадратурных формул оказывается качественно быстрее степенной. Она подчиняется закону $\delta_{2N} \sim \mathcal{O}(\delta_N^2)$. Для сверхэкспоненциальной сходимости в примере (29) при удвоении сетки ошибка убывает как $\delta_{2N} = \mathcal{O}\left(\delta_N^4\right)$; такая скорость аналогична итерационно-

му процессу четвёртого порядка.

В общем случае сверхэкспоненциальная сходимость возникает лишь тогда, когда единственная особенность расположена в бесконечно удалённой точке. Однако соотношение между δ_N и δ_{2N} при этом зависит от скорости возрастания $u(z)$ при $\operatorname{Im} z \to \infty$.

Особенность на границе

Рассмотрим ещё один интеграл, не подпадающий под теорему 5 и имеющий точное решение ([12], стр. 355, №3.472):

$$I = \int_0^\infty x^2 \exp\left(-x^2 - \frac{1}{x^2}\right) dx = \frac{3\sqrt{\pi}}{4e^2}. \tag{31}$$

В этом интеграле функция имеет существенно особую точку при $x = 0$, то есть на границе отрезка интегрирования. Таким образом, ширина полосы аналитичности $u(z)$ нулевая: $l = 0$. Интеграл не подпадает под теорему 5 по двум причинам: во-первых, особая точка лежит на границе отрезка, во-вторых, эта точка существенно особая. По-видимому, первый фактор является наиболее существенным.

Интегралы с такими особенностями возникают, например, при вычислении скоростей химических или ядерных реакций по их сечению в термодинамически равновесной среде. В связи с отсутствием теоретических оценок численный расчет таких примеров становится особенно интересным. Поэтому данный случай заслуживает внимательного рассмотрения.

Пример расчета. Для интеграла (31) были проведены численные расчеты для сеток с $N = 1, 2, ..., 1000$. Было найдено точное значение погрешности $\delta = |I - I_N|$. Оказалось, что зависимость этой величины от N кардинально отличается от ранее рассмотренных случаев. Зависимость погрешности от N знакопеременная, а её амплитуда достаточно быстро убывает с увеличением N.

Графически изобразить такую погрешность достаточно сложно. Обычно для изображения погрешности, убывающей на много порядков, по ординате откладывают $\lg|\delta|$; модуль применяют потому, что заранее неизвестен знак погрешности. Этот приём хорошо работает, если знак погрешности на всех сетках одинаков (это обычно выполняется как для степенной, так и для экспоненциальной сходимости), зависимость $\lg|\delta|$ от N или $\lg(N)$ является при этом плавной кривой.

График зависимости $\lg|\delta|$ от N построен для интеграла (31) и изображён на рис. 8а, однако на нём эта зависимость оказывается не плавной. Линия фактически состоит из отдельных кусков. Каждый кусок представляет выпуклую

кривую; у него есть экстремум и быстро спадающие ”крылья”. На каждом куске погрешность имеет одинаковый знак, а на соседних кусках знаки противоположны. Провалы вниз напоминают клювы, поэтому такую кривую будем называть клювообразной. Каждый клюв соответствует прохождению погрешности через нуль.

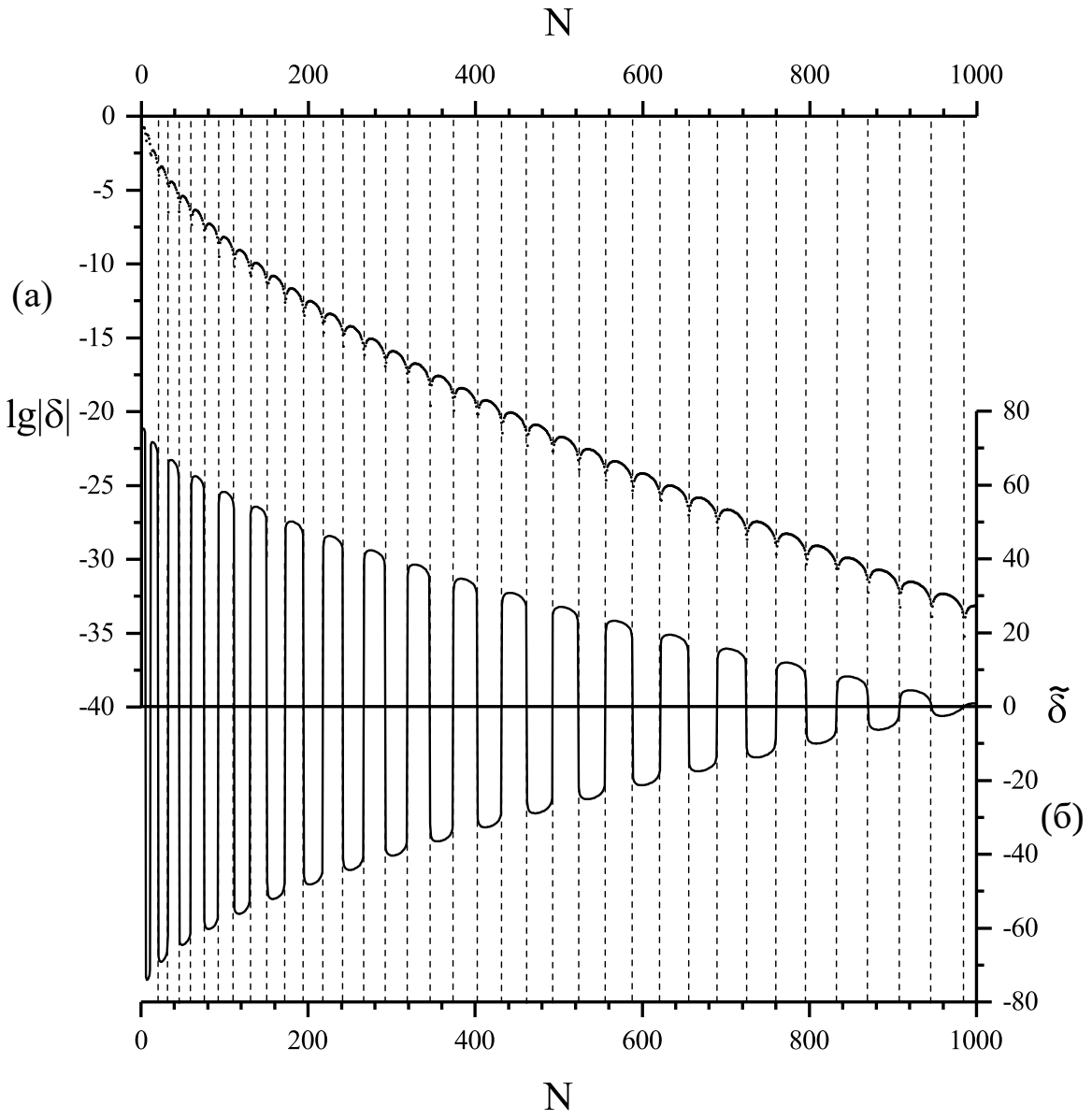

Рис. 8. Погрешность для теста (31): а – в полулогарифмическом масштабе, б – в специальном масштабе (32).

Экстремумы всех кусков монотонно убывают. Огибающая этих кусков яв-

ляется плавной монотонно убывающей кривой.

Чтобы передать зависимость δ от N с учётом смены знака, был найден специальный масштаб. По оси ординат откладывалась величина

$$\tilde{\delta} = \operatorname{arcsh}\frac{\delta}{\mu} = \operatorname{sgn}(\delta) \ln\left(\frac{|\delta|}{\mu} + \sqrt{\frac{\delta^2}{\mu^2} + 1}\right); \tag{32}$$

здесь μ – величина наименьшего (то есть последнего) из экстремумов. Отметим, что такой масштаб не является абсолютным: он привязан к ограниченному диапазону N. На рис. 8б показана погрешность в данном масштабе. Такой масштаб линейно изображает погрешность в малой окрестности нулевых значений, но логарифмически уменьшает ее вдали от нулей. Нули погрешности пунктиром снесены вверх, видно, что они точно совпадают с клювами на рис. 8а.

Такой вид погрешности, во-первых, не описывается ни одной из известных теорий. Во-вторых, к нему не применимы все известные методы апостериорного нахождения погрешности расчета. Видно, что подобные задачи требуют развития новых теоретических подходов.

Обсуждение. Проанализируем величины экстремумов на рис. 8. Отличие величин первого и последнего экстремумов составляет $\sim 10^{30}$ при изменении N от ~ 15 до ~ 1000. Это показывает, что сходимость кардинально быстрее, чем степенная, при этом огибающая этих экстремумов не слишком сильно отличается от прямой. Поэтому количественное убывание погрешности в среднем соответствует экспоненциальному характеру сходимости.

Однако наклон огибающей на рис. 8а составляет примерно 0.03. В то же время наклон прямой линии, соответствующей чисто экспоненциальной сходимости, на рис. 7 составляет 0.4. Он во много раз больше, тем не менее разница между этими двумя типами экспоненциальной сходимости много меньше, чем разница между экспоненциальной сходимостью и степенной.

Все проведённые расчёты позволяют высказать следующую гипотезу: скорость экспоненциальной сходимости определяется в первую очередь расстоянием ближайшей особой точки от отрезка интегрирования. Тип особой точки при этом имеет второстепенное значение. При этом даже в случае нахождения особой точки на отрезке интегрирования сходимость может оставаться экспоненциальной.

Заключение

Квадратуры с экспоненциальной сходимостью являются мощным инструментарием для решения физических задач. Если удается найти преобразования переменных, сводящие интегралы к указанным выше видам, то вычисления ускоряются в сотни раз.

В данной работе 1) улучшено значение константы в теоретической оценке погрешности экспоненциально сходящихся квадратур; 2) найдена эвристическая оценка погрешности для случая кратных полюсов подынтегральной функции; 3) обнаружено, что попадание особой точки на отрезок интегрирования качественно меняет поведение погрешности, что не объясняется известными теориями.

Эти результаты применены к вычислению функций Ферми–Дирака полуцелого индекса [9–11]. Замена переменных $t = \tau^2$ в (1) приводит интеграл к требуемой форме. Квадратура трапеций обеспечивает 16 верных знаков уже при $N \sim 10 - 100$, а традиционные квадратуры требуют $N \sim 10^4$.

Работа поддержана грантом РФФИ 18-01-00175.

Библиографический список

[1] Тихонов А.Н., Арсенин В.Я. Методы решения некорректных задач. М.: Наука, 1979

[2] Pauli W. // Z. Phys. 1927. V. 41, p.81-102.

[3] Sommerfeld A. // Z. Phys. 1928. V. 47, P. 1-3.

[4] Feynman R.P., Metropolis N., Teller E. // Phys. Rev. 1949. Vol. 75, P. 1561-1573.

[5] Собельман И.И. Введение в теорию атомных спектров — М.: Издательство физико-математической литературы, 1963.

[6] Грим Г. Уширение спектральных линий в плазме. — М.: Мир, 1978.

[7] Holtsmark J. // Ann. Phys. 1919. Vol. 58. P. 577-630.

[8] Trefethen L.N. and Weideman J. A. C. The exponentially convergent trapezoidal rule // SIAM Rev., 56(3):385–458, 2014. — https://doi.org/10.1137/130932132

[9] Kalitkin N.N., Kolganov S.A. // Doklady Math. 95:2 (2017). 157. — https://doi.org/10.1134/s1064562417020156

[10] Kalitkin N.N., Kolganov S.A. // Math. Models Comp. Simul. 9:5 (2017). 554. — https://doi.org/10.1134/s2070048217050052

[11] Kalitkin N.N., Kolganov S.A. // Math. Models Comp. Simul. 10:4 (2018). 472. — https://doi.org/10.1134/s2070048218040063

[12] Градштейн И.С., Рыжик И.М. Таблицы интегралов, сумм, рядов и произведений. М.: Физматгиз, 1963