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

Препринты ИПМ • Препринт № 235 за 2018 г.

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

Калиткин Н.Н., Колганов С.А.

Функции Ферми-Дирака. Прямое вычисление функций

Рекомендуемая форма библиографической ссылки: Калиткин Н.Н., Колганов С.А. Функции Ферми-Дирака. Прямое вычисление функций // Препринты ИПМ им. М.В.Келдыша. 2018. № 235. 29 с. doi:10.20948/prepr-2018-235
URL: http://library.keldysh.ru/preprint.asp?id=2018-235

Ордена Ленина
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Российской академии наук

Н.Н. Калиткин, С.А. Колганов

Функции Ферми-Дирака. Прямое вычисление функций

Москва — 2018

Калиткин Н.Н., Колганов С.А.

Функции Ферми-Дирака. Прямое вычисление функций.

В данной работе изложены методы прямого вычисления функций Ферми-Дирака с заданной точностью. Для функций целого индекса эта задача решается с помощью сведения функций положительного аргумента к функциям отрицательного аргумента. Для функций полуцелого индекса значения аргумента разбиваются на три области: отрицательный аргумент, где используется быстросходящийся ряд; большие положительные аргументы, где применимо асимптотическое разложение; промежуточная область, где используется прямое численное интегрирование. В последнем случае для интегрирования построены формулы с экспоненциальной, то есть очень быстрой сходимостью. Исследованы свойства таких квадратурных формул. Для вычисления интегральной функции Ферми-Дирака найден нетривиальный подход: задача записывается в виде тройного интеграла, в котором одно интегрирование выполняется аналитически, а оставшийся двойной интеграл вычисляется квадратурами с экспоненциальной сходимостью. Предложенные методы позволяют экономично вычислять функции Ферми-Дирака с относительной погрешностью 10^{-16} при любых значениях аргумента.

Ключевые слова: функции Ферми-Дирака, вычисление функций, квадратуры с экспоненциальной сходимостью

Nikolai Nikolaevich Kalitkin, Semen Andreevich Kolganov

The Fermi-Dirac functions. Direct calculation of the functions.

The paper presents methods for direct calculation of the Fermi-Dirac functions with a given accuracy. For functions of the integer index, this problem is solved with the help of formula connecting functions of positive and negative arguments. For the function of the half-integer index values of the argument are divided into three areas: negative arguments, where the fast converging series is used; large positive arguments, where asymptotic expansion is used; the intermediate region, where direct numerical integration is used. In the latter case of the constructed formula have exponential (i.e. very fast) convergence. The properties of such quadrature formulas are investigated. A nontrivial method is found for calculation of integral Fermi-Dirak function. The problem of triple integral calculation comes to calculation of double integral by quadratures with exponential convergence. These methods permit to calculate the Fermi-Dirak functions economically with relative accuracy 10^{-16} for arbitrary values of argument.

Key words: Fermi-Dirac functions, calculation of functions, exponentially converging quadratures

Работа выполнена при поддержке Российского фонда фундаментальных исследований, проект 18-01-00175.

1. Предисловие

Функции Ферми-Дирака (далее ФД) широко распространены в задачах квантовой механики. В препринте [1] было дан исторический обзор работ, посвященных этим функциям, и приведено их определение

$$I_k(x) = \int_0^\infty \frac{t^k \mathrm{d}t}{1+\exp(t-x)}, x \epsilon (-\infty;+\infty); \qquad (1)$$

здесь x – аргумент функции, k – индекс функции. В [1] также дано описание свойств этих функций и разложение их в ряды при $x \to +\infty$ и $x \to -\infty$.

В данной работе рассматриваются методы прямого вычисления функций ФД целых и полуцелых индексов, а также интегральной функции ФД с заданной точностью. Желательно вычислять функции ФД с предельно высокой точностью ε, допускаемой компьютером. Поэтому опишем точность, которую дают распространенные сейчас процессоры при вычислении с плавающей точкой.

Погрешность округления. Все операции с плавающей точкой выполняются арифметическим сопроцессором (Floating Point Unit). Его архитектура фактически не развивается с момента принятия в 1985 году стандарта IEEE-754 [2], поэтому предельная разрядность чисел остается на уровне 80 бит. Однако в наиболее распространенных математических обеспечениях для записи в память используются либо 32-битовые числа (*single precision*), либо 64-битовые числа (*double precision*). В обоих случаях не полностью используются возможности линейки процессора. В начале 2000-х годов для языка C++ была возможность использовать 80-битовые числа (*long double precision*); в настоящее время эта возможность не поддерживается и мало где сохранилась. Для суперкомпьютерных вычислений используют 128-битовые числа; но вычисления делаются программными средствами на тех же процессорах. Возможно и вычисление с произвольной разрядностью, но они также выполняются программными средствами. Например, для языка C++ это библиотека *boost::multiprecision*.

При вычислениях с плавающей точкой побитовая запись числа выглядит следующим образом: знак числа – 1 бит, двоичный порядок числа – p бит, мантисса – m бит. Порядок числа может быть положительным или отрицательным; но при записи в память к нему автоматически добавляется положительная константа, равная по модулю максимально возможному отрицательному порядку, так что в память записывается неотрицательное

число. Это эквивалентно тому, что из p разрядов порядка тратится 1 разряд на знак порядка и $p-1$ на модуль порядка.

При записи мантиссы также имеется небольшая “хитрость”. Первый разряд мантиссы всегда равен 1, поэтому при записи он отбрасывается. При считывании числа в процессор этот разряд автоматически добавляется. Такой прием позволяет фактически удлинять записываемую мантиссу на 1 бит.

Нетрудно посчитать, что максимально возможный порядок в двоичной системе есть $2^{p-1}-1$; для перевода в десятичную систему его нужно умножить на $\lg 2$. Разрядность процессора длиннее, чем считанное из памяти число, за исключением *long double*. После вычисления результат, записанный на процессоре, оказывается длиннее отведенного в памяти места. Поэтому при записи результата в память производится округление. Ошибка такого округления не превышает половины последнего отброшенного разряда. С учетом “спрятанного” первого разряда мантиссы это означает, что относительная ошибка округления есть $\varepsilon = 2^{-m-2}$.

В случае *long double* длина числа равна линейке процессора и нельзя ни спрятать первый разряд мантиссы, ни иметь лишние разряды для округления; в этом случае $\varepsilon = 2^{-m}$. В *Таблице 1* приведены предельные порядки $P = 2^{p-1}\lg 2$ и относительные ошибки единичного округления в форме $\lg\varepsilon$ для рассмотренных выше случаев вычисления. Видно, что для случая long double представление числа выбрано не вполне удачно; лучше было бы взять $p = 14, m = 65$.

Таблица 1

Компьютерные числа с плавающей точкой

точность	бит	p	m	P	$\lg\varepsilon$
float	32	8	23	±38	-7.5
double	64	11	52	±308	-16.2
long double	80	15	64	±4932	-19.3
Quadruple double	128	15	112	±4932	-34.3

Наиболее частыми являются 64-битовые вычисления, поэтому мы будем ориентироваться на точность $\varepsilon = 10^{-16}$. Вычисление *long double* кажутся заманчивыми, поскольку они повышают точность по сравнению с *double* без увеличения машинного времени. Однако разбиение числа на мантиссу и порядок было недостаточно продумано (лучше было бы перекинуть 1 бит на мантиссу), так что увеличение точности не столь значительно. Но 32-битовыми числами не стоит пренебрегать: именно такую точность дают видеокарты, позволяющие сильно ускорять вычисления за счет конвейерной реализации.

2. Функции целого индекса

Нулевой индекс. Напомним, что функции ФД $I_k(x)$ целого индекса k существуют при $k \geq 0$. При $k = 0$ функция ФД выражается через элементарные функции:

$$I_0(x) = \ln\left(1 + e^x\right). \tag{2}$$

При целых $k < 0$ функция ФД не существует, так как интеграл (1) расходится.

Отрицательные аргументы. Для произвольных, необязательно целых, индексов k существует всюду сходящийся ряд:

$$I_k(x) = 2\Gamma(k+1)\sum_{n=0}^{\infty} \frac{b_n^{(k)}}{(1+2e^{-x})^{n+1}}. \tag{3}$$

Коэффициенты этого ряда определяются несложными квадратурами

$$b_n^{(k)} = \frac{1}{\Gamma(k+1)}\int_0^{\infty} (1-2e^{-t})^n e^{-t} t^k dt; \left|b_n^{(k)}\right| \leq 1. \tag{4}$$

Этот ряд сходится при любом x не хуже, чем геометрическая прогрессия со знаменателем $g = (1+2e^{-x})^{-1}$. Эта сходимость неравномерная, и при $x \to +\infty$ становится очень медленной. Однако при $x \leq 0$ знаменатель $g \leq 1/3$, и сходимость становится достаточно быстрой. Таким образом, ряд (3) удобен для вычисления функций ФД при $x \leq 0$.

Формула (4) неудобна для непосредственного вычисления коэффициентов $b_n^{(k)}$. Более устойчив к ошибкам округления следующий рекуррентный процесс. Для индекса $k = 0$ коэффициенты вычисляются по формуле

$$b_n^{(0)} = \frac{1+(-1)^n}{2(n+1)}, n \geq 0. \tag{5}$$

Далее увеличение индекса k на единицу производится по формуле

$$b_0^{(k)} = 1; b_n^{(k)} = \frac{1}{n+1}\left(b_n^{(k-1)} + n b_{n-1}^{(k)}\right), n > 1. \tag{6}$$

Схема Горнера. Ограничимся конечным числом членов ряда (3) и запишем полученную сумму по схеме Горнера:

$$I_k(x) = 2\Gamma(k+1)g\left(b_0^{(k)} + g\left(b_1^{(k)} + g\left(b_2^{(k)} + \ldots + g\left(b_N^{(k)}\right)\ldots\right)\right)\right),$$
$$g = (1+2e^{-x})^{-1}, x \le 0; \qquad (7)$$

напомним, что для целого индекса $\Gamma(k+1) = k!$. Значения коэффициентов $b_n^{(k)}$ можно непосредственно рассчитывать по формулам (5-6) или брать из [1], где они приведены с 16-ю десятичными знаками. Такого числа знаков достаточно для вычисления с относительной точностью $\varepsilon = 10^{-16}$ (*double precision*).

Описанный алгоритм прост и экономичен. Поскольку коэффициенты $b_n^{(k)} > 0$, то нигде не возникает вычитаний, и относительные ошибки округления могут накапливаться лишь незначительно, практически не ухудшая точность.

Число членов. Нас интересует относительная погрешность. Поэтому общий множитель $2\Gamma(k+1)g$ можно откинуть. Главный член оставшегося произведения есть $b_0^{(k)} = 1$ и все остальные члены можно сравнивать непосредственно с ним. Все коэффициенты $b_n^{(k)} < 1$, положительны и убывают с увеличением номера n. Это означает, что отброшенная часть ряда сходится быстрее, чем геометрическая прогрессия с первым членом g^{N+1} и знаменателем g. Величина знаменателя возрастает от 0 до 1/3 при возрастании от $-\infty$ до 0. Поэтому сумма отброшенных членов не превышает сумму этой геометрической прогрессии $g^{N+1}/(1-g)$. Нужно, чтобы эта величина не превышала ε.

Эту оценку нетрудно немного усилить. На самом деле величину откинутых членов ряда надо сравнивать не с $b_0^{(k)} = 1$, а с суммой учтенных членов ряда. Их также можно оценить, как сумму геометрической прогрессии с первым членом 1 и знаменателем g. Тогда для сравнения вместо ε надо брать величину $\varepsilon/(1-g)$. Это дает следующую оценку:

$$N \ge \frac{\ln \varepsilon}{\ln g} - 1.$$

Эта формула дает нецелое значение N. И его надо округлить вверх до целого:

$$N = \left[\frac{\ln \varepsilon}{\ln g}\right], \qquad (8)$$

где квадратные скобки означают целую часть числа. Даже эта оценка завышена, т.к. она не учитывает заметного убывания $b_n^{(k)}$ при возрастании номера n. Однако учет этого эффекта заметно усложнил бы алгоритм, что нецелесообразно.

Наиболее медленная сходимость будет при $x=0$, когда $\varepsilon=10^{-16}$; для $g=1/3$ это дает $N=33$. При $x\to-\infty$ величина g быстро стремится к 0; при этом число членов N быстро убывает.

Положительный аргумент. Для функций ФД существует точное соотношение, связывающее значения функции при положительном и отрицательном аргументах:

$$I_k(x)=(-1)^k I_k(-x)+\frac{x^{k+1}}{k+1}\left[1+\sum_{n=1}^{1+[k/2]}(2-2^{2-2n})\frac{\zeta(2n)}{x^{2n}}\prod_{p=1}^{2n}(k+2-p)\right]. \qquad (9)$$

Здесь $\zeta(2n)$ есть дзета-функция Римана. Поскольку для $x\le 0$ алгоритм вычисления построен выше, формула (9) полностью решает вопрос вычисления функции с целым индексом. Для отрицательных аргументов алгоритм дает относительную погрешность ε. Но слагаемое $I_k(-x)$ в формуле (9) существенно меньше величины многочлена. Поэтому при $x>0$ относительная погрешность вычисления по формуле (9) будет существенно меньше ε.

Таким образом, вопрос о вычислении функции целого индекса решен исчерпывающим образом. Для удобства пользователей приведем соотношение (9) для нескольких первых индексов и значений $x\ge 0$:

$$\begin{aligned}
I_0(x)&=I_0(-x)+x,\\
I_1(x)&=-I_1(-x)+\frac{x^2}{2}+\frac{\pi^2}{6},\\
I_2(x)&=I_2(-x)+\frac{x^3}{3}+\frac{\pi^2}{3}x,\\
I_3(x)&=-I_3(-x)+\frac{x^4}{4}+\frac{\pi^2}{2}x^2+\frac{7\pi^4}{60},\\
I_4(x)&=I_4(-x)+\frac{x^5}{5}+\frac{2\pi^2}{3}x^3+\frac{7\pi^4}{15}x.
\end{aligned} \qquad (10)$$

3. Квадратуры с экспоненциальной сходимостью

Прямое вычисление функций ФД нецелого индекса требует применения квадратурных формул к интегралу (1). В общем случае такое интегрирование для получения высокой точности требует очень подробной сетки (10^5-10^6 узлов) и является чрезмерно трудоемким. Однако для полуцелых индексов k возможно кардинальное уменьшение трудоемкости при специальном преобразовании подынтегрального выражения. Можно построить квадратуры с

экспоненциальной, то есть очень быстрой сходимостью. Изложим теорию таких квадратур.

Сходимость квадратур. Рассмотрим задачу вычисления интегралов от функций $u(x)$, имеющих сколь угодно высокие непрерывные производные на отрезке интегрирования $[a,b]$ (см. напр. [3-5]). Чаще всего на практике берут равномерные или сводящиеся к равномерным сетки $\omega_N = \{x_n, n = 0,\ldots,N\}$ и используют простейшие квадратурные формулы трапеций, средних, Симпсона и т.п. Погрешность подобных формул имеет оценку $\delta < const \cdot M_p \cdot h^p = O(h^p) = O(N^{-p})$, где p есть порядок точности формулы, h – шаг интегрирования, а $M_p = \max\left|u^{(p)}(x)\right|$. Такая сходимость называется ***степенной***, поскольку погрешность выражается через степень шага. Она довольно медленная, и для получения высокой точности требуется большое N. Такие квадратуры довольно трудоемки.

Квадратуры Гаусса-Кристоффеля дают гораздо более быструю сходимость. Например, классическая формула Гаусса для интегрирования на отрезке [-1,1] с весом $\rho(x) = 1$ имеет погрешность (после упрощения факториальных множителей)

$$\delta \le \sqrt{\frac{\pi}{N}} \frac{b-a}{4} \left(e \frac{b-a}{8N} \right)^{2N} M_{2N}. \qquad (11)$$

Квадратура Эрмита для отрезка [-1,1] с весом $\rho(x) = (1-x^2)^{-1/2}$ имеет погрешность

$$\delta \le \sqrt{\frac{\pi}{N}} \left(\frac{e}{2\sqrt{2}N} \right)^{2N} M_{2N}. \qquad (12)$$

Погрешности (11-12) с точностью до логарифмически малых членов можно записать в следующем виде:

$$\delta \sim \alpha \cdot \exp(-\beta N). \qquad (13)$$

Зависимость от числа узлов является не степенной, а экспоненциальной, поэтому такую сходимость будем называть ***экспоненциальной***.

Трудоемкость подобных формул несравненно меньше, чем у квадратур со степенной сходимостью. Однако узлы и веса квадратур Гаусса-Кристоффеля найдены лишь для отдельных отрезков и весов $\rho(x)$ интегрирования. При этом только для квадратур Эрмита эти веса и узлы найдены в виде простых формул для произвольных N. Для остальных случаев узлы и веса точно вычисляются (через радикалы) лишь для $N \le 3$ или $N = 5$. Это сильно ограничивает возможности практического использования таких квадратур.

Далее покажем, что если $u(x)$ четно продолжается через обе границы отрезка, то формула трапеций на равномерной сетке дает экспоненциальную сходимость. При этом коэффициент β в экспоненте определяется расстоянием до ближайшего полюса в комплексной плоскости. Это открывает новые возможности для построения квадратур малой трудоемкости.

Случай экспоненциальной сходимости. Пусть $u^{(p)}(x)$ существуют и непрерывны на $[a;b]$ при любых p. Требуется вычислить интеграл

$$U = \int_a^b u(x)dx. \tag{14}$$

Введем равномерную сетку ω_N с $x_0 = a, x_N = b$ и воспользуемся формулой Эйлера–Маклорена, базирующейся на формуле трапеций [6,7]:

$$U_N = h(\frac{u_0}{2} + u_1 + u_2 + ... + u_{N-1} + \frac{u_N}{2}) + \sum_{p=1} (-1)^p a_p h^{2p} (u_N^{(2p-1)} - u_0^{(2p-1)}), \tag{15}$$
$$a_p \sim M_{2p-1}.$$

Если оборвать эту сумму на члене P, то первый отброшенный член будет остаточным. Его величина есть $\delta_P = O(h^{2P+2})$. В этом случае формула (15) имеет степенную сходимость.

Пусть $u(x)$ такова, что все её **нечётные** производные на правой и левой границах одинаковы: $u^{(2p-1)}(a) = u^{(2p-1)}(b)$. Тогда в (15) сумма обращается в нуль. Оставшаяся часть квадратур является просто формулой трапеций. Из этого следуют

Утверждение 1. Пусть подынтегральная функция $u(x)$ имеет сколь угодно высокие производные, причем нечетные производные на правой и левой границах одинаковы: $u^{(2p-1)}(a) = u^{(2p-1)}(b)$. Тогда формула трапеций на равномерной сетке имеет сходимость выше степенной. ■

Частный случай. Утверждение 1 справедливо, если $u(x)$ чётно продолжается через обе границы отрезка: $u^{(2p-1)}(a) = u^{(2p-1)}(b) = 0$. ■

Таким образом, установлен класс функций, для которого формула трапеций имеет сверхстепенную сходимость. Остается найти закон этой сходимости. Проведем ее изучение на следующем тестовом примере:

$$U(q,r,c) = \int_0^\pi \frac{(c^2-1)\cdot c^r \cos(rx)}{(c^2 - 2c\cos x + 1)^q} dx, c > 1. \tag{16}$$

Параметры $r \geq 0$, $q \geq 1$ берутся целыми. Тогда подынтегральное выражение четно на обеих границах отрезка, его нечетные производные на границах

обращаются в нуль, и пример удовлетворяет требованиям Утверждения 1. При $q=1$ известно точное значение интеграла [8]:

$$U(1,r,c)=\pi. \qquad (17)$$

При $q\neq 1$ интеграл (16) не выражается через элементарные функции от параметров.

Для тщательного численного выявления закономерностей все расчеты проводились с повышенной разрядностью (45 десятичных знаков) с помощью библиотеки языка *C++ boost::multiprecision.*

Расчеты интеграла (16) при фиксированных параметрах проводились на сетках с разным числом интервалов N. Погрешность расчетов при $q=1$ определялась непосредственным сравнением с точным ответом (17). На рис.1 показана зависимость погрешности от N при $r=0$ и различных значениях c в *полулогарифмическом* масштабе. Каждому значению c соответствует своя линия погрешности. Видно, что при всех значениях c кривые погрешности в этом масштабе являются прямыми. Это означает, что погрешность подчиняется закону

$$\ln\delta_N=\alpha-\beta N, \beta=const\cdot\ln c. \qquad (18)$$

При других значениях параметров картина была аналогичной. На рис.2 показан случай $q=1, r=2$. Опять линии погрешности являются прямыми**.**

Для $q>1$ точный ответ неизвестен. В этом случае для получения значений погрешности можно воспользоваться следующими соображениями. При закономерности (18) разности значений U при возрастании N на единицу также должны ложиться на прямую в полулогарифмическом масштабе (это напоминает апостериорную оценку погрешности по методу Ричардсона для квадратур со степенной сходимостью). На рис. 3 приведены графики таких разностей для $q=2, r=1.$ Они также оказываются прямыми. Все это позволяет сделать эвристическое

Утверждение 2. При выполнении условий Утверждения 1 погрешность формулы трапеций экспоненциально зависит от числа узлов сетки N. ■

Попробуем выяснить, от чего зависит коэффициент β в (18). Он не должен зависеть от максимумов модулей каких-либо производных $u(x)$, поскольку они входят в суммы формул Эйлера-Маклорена (15) и приводят к степенной сходимости. Поэтому рассмотрим гипотезу о связи β с полюсами подынтегрального выражения. Для теста (16) подынтегральное выражение имеет полюса кратности q в точках

$$x^*=2\pi m\pm i\ln(c), -\infty<m<+\infty. \qquad (19)$$

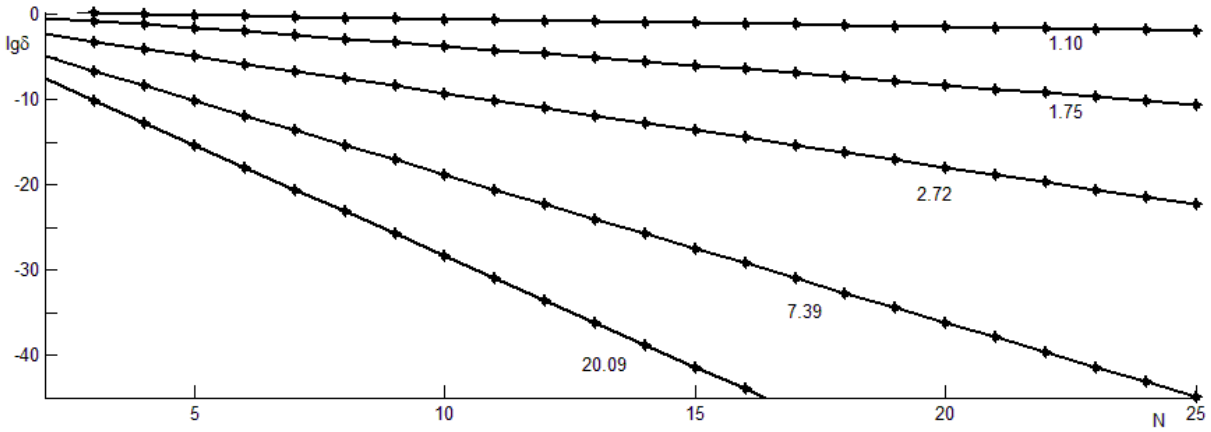

Рис.1. Погрешность квадратуры трапеций для (18) при $p=0$ и $q=1$. Цифры около линий – величины c.

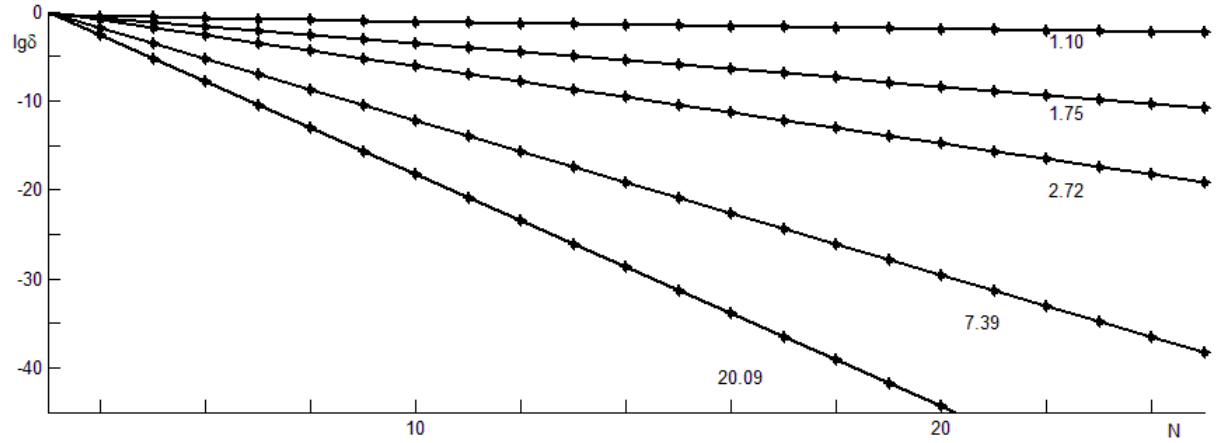

Рис.2. Погрешность квадратуры трапеций для (18) при $p=2$ и $q=1$. Цифры около линий – величины c.

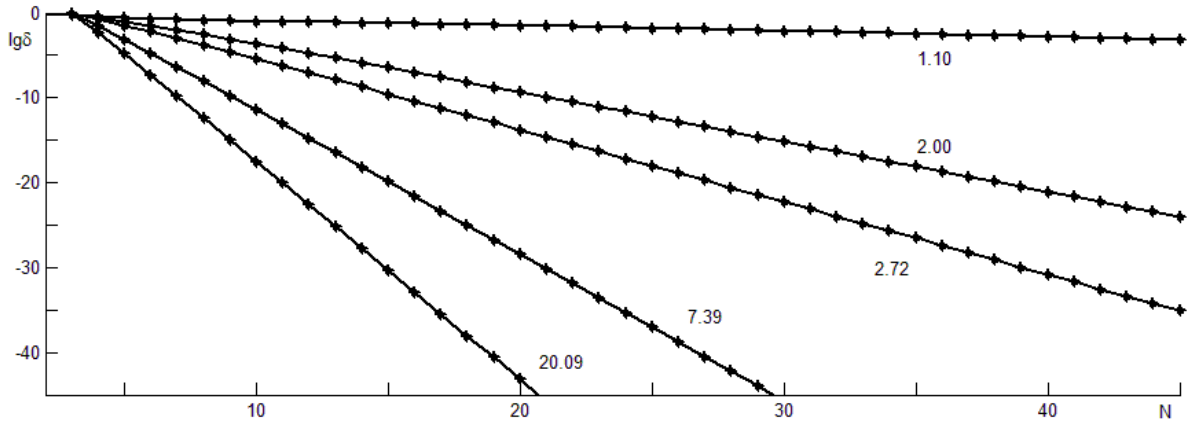

Рис.3. Погрешность квадратуры трапеций для (16) при $p=1$ и $q=2$. Цифры около линий – величины c.

Наименьшее расстояние между каким-либо из полюсов и ближайшей к нему точкой отрезка интегрирования есть $\ln(c)$.

Предварительный просмотр графиков показал, что наклон $\beta \sim \ln(c)$. Для тщательного анализа на рис.4 показано отношение $\beta / \ln(c)$ в зависимости от c для нескольких значений $q = 1,2$ и $r = 0,1,2$. Видно, что для полюса первого порядка ($q = 1$) это отношение с высокой точностью не зависит ни от r, ни от c. Для полюсов второго порядка ($q = 2$) появляется очень слабая зависимость, причем она линейна по c; при $c \to 0$ эти зависимости стремятся к тому постоянному значению, которое справедливо для $q = 1$.

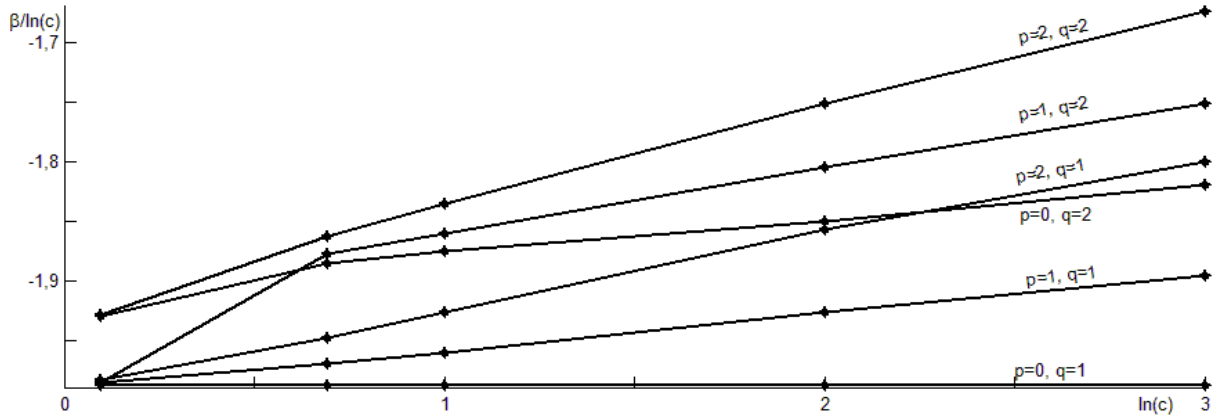

Рис.4. Зависимость $\beta / \ln(c)$ от величины c для различных p и q.

Это позволяет сделать

Утверждение 3. Наклон β в (18) с хорошей точностью пропорционален расстоянию от отрезка интегрирования до ближайшего полюса интегрируемой функции в комплексной плоскости. ■

Практические рекомендации. 1. При использовании квадратурных формул со степенной сходимостью удобно сгущать сетки по N последовательно вдвое. Это позволяет использовать обычную процедуру Ричардсона для получения априорной асимптотически точной оценки погрешности. Такое сгущение экономично, поскольку суммарный объем всех расчетов лишь вдвое превышает объем расчетов на последней сетке [4].

Для квадратур с экспоненциальной сходимостью (18) также можно пользоваться процедурой Ричардсона, если сгущать сетки не вдвое, а каждый раз увеличивая N на 1. При этом будет получаться асимптотически точная апостериорная оценка погрешности. Однако такое сгущение сеток экономически невыгодно, поскольку суммарный объем вычислений будет в $\sim N/2$ раз больше, чем расчет на последней сетке.

Поэтому в практических расчетах удобнее увеличивать N в 2 раза. Из (18) нетрудно получить, что при этом $\delta_{2N} \sim \delta_N^2$. Такой закон убывания напоминает сходимость ньютоновских итераций вблизи простого корня: число верных десятичных знаков приблизительно удваивается с увеличением N в 2 раза.

Поэтому на практике останавливаются на такой сетке $2N$, когда отклонение от результата на предыдущей сетке становится меньше $\varepsilon^{2/3}$, где ε – ошибка единичного округления компьютера.

2. Для формулы трапеций на равномерной сетке полезен следующий прием, вдвое уменьшающий трудоемкость вычислений. На сетке с N узлами и шагом h формула трапеций имеет вид

$$U_N = h(\frac{u_0}{2} + u_1 + ... + u_{N-1} + \frac{u_N}{2}). \qquad (20)$$

При удвоении сетки все узлы предыдущей сетки становятся четными узлами новой сетки, и заново вычислять значения функций в них не надо. Достаточно найти значения функции в новых (нечетных) узлах и вычислить

$$U_{2N} = \frac{1}{2}U_N + \frac{h}{2}(u_1 + u_3 + u_5 + ... + u_{2N-1}), \qquad (21)$$

где нечетные индексы относятся к узлам новой сетки.

4. Предельные выражения для функций полуцелого индекса

Отрицательные аргументы. При $x \le 0$ функции полуцелого индекса при $k \ge -1/2$ легко вычисляются с помощью того же всюду сходящегося ряда (3), что и функции целого индекса. Справедливыми остаются и схема Горнера (7), и формула (8) для расчета числа членов N. Коэффициенты $b_n^{(k)}$ для полуцелых индексов k определяются общей формулой (5), но практически рассчитываются повышением индекса k по рекуррентной процедуре (6). Для начала такого расчета необходимо предварительно найти коэффициенты для индекса $k = -1/2$. Остановимся на этом подробнее.

Коэффициенты $\boldsymbol{b_n^{(-1/2)}}$. Для их вычисления рассмотрим определение коэффициентов (4), полагая в нем $k = -1/2$. Сделаем замену переменных $t = \tau^2$. Искомые коэффициенты примут вид:

$$b_n^{(-1/2)} = \frac{2}{\Gamma(1/2)}\int_0^\infty (1 - 2e^{-\tau^2})^n e^{-\tau^2} d\tau, \Gamma(1/2) = \sqrt{\pi}. \qquad (22)$$

Скобка под интегралом не превышает 1 по модулю. Все подынтегральное выражение является четной функцией от τ, так что все его нечетные производные на левом конце интегрирования ($\tau = 0$) равны 0. На правом

конце интегрирования ($\tau \to +\infty$) вся подынтегральная функция со всеми производными быстро стремится к 0 из-за экспоненты, стоящей вне скобки.

Поэтому можно взять достаточно большой верхний предел интегрирования T, чтобы пренебречь отброшенной частью интервала интегрирования и пренебречь значением производных на правой границе. Тогда на отрезке интегрирования $0 \le \tau \le T$ формула трапеций на равномерной сетке $\omega_N = \{\tau_n = Tn/N, n = 0,...,N\}$ будет иметь экспоненциальную сходимость. Остается выбрать значение T и N, обеспечивающие точность ε.

Если интегрировать (22) от 0 до T, то отброшенный "хвост" не превышает величины $\pi^{-1/2} T^{-1} \exp(-T^2)$. Уже при $T = 6.5$ это составляет $2 \cdot 10^{-19}$. Для простоты дальнейшего расчета удобно взять $T = 8$. Это лишь незначительно увеличивает объем вычислений и обеспечивает точность $\sim 10^{-29}$. Вклад высоких производных в формуле Эйлера-Маклорена при $\tau = T$ также пренебрежимо мал. Такой точности достаточно даже для *long double precision*.

Сеточный расчет на отрезке $0 \le \tau \le 8$ начинаем с одного интервала: $N = 1$. Далее последовательно удваиваем число N. Удвоение прекращаем, когда разность результатов на 2 последних сетках становится меньше $\varepsilon^{2/3}$; для *double precision* это составляет $10^{-10} - 10^{-11}$. Требуемая точность для коэффициентов с $n = 1-4$ достигается уже при $N = 64$; при увеличении n число узлов N возрастает, но даже для 33-его коэффициента составляет всего $N = 256$ узлов. Напомним, что трудоемкость сеточных вычислений можно вдвое сократить с помощью формулы (21).

Величины коэффициентов $b_n^{(-1/2)}$ с 17 десятичными знаками приведены в Таблице 2.

Индекс $\boldsymbol{k = -3/2}$**.** При этом индексе функцию надо вычислять не по ряду (3), а почленно дифференцировать ряд для индекса $k = -1/2$. Это приводит к следующему ряду

$$I_{-3/2}(x) = -8\sqrt{\pi} e^{-x} \sum_{n=0}^{\infty} (n+1) b_n^{(-1/2)} g^{n+2}, g = (1+2e^{-x})^{-1}. \qquad (23)$$

Схема Горнера для него записывается аналогично, но немного сложнее:

$$I_{-3/2}(x) = -8\sqrt{\pi} e^{-x} g^2 \left(b_0^{(-1/2)} + g\left(2b_1^{(-1/2)} + g\left(3b_2^{(-1/2)} + \ldots \right.\right.\right.$$
$$\left.\left.\left. \ldots + g\left((N+1) b_N^{(-1/2)} \right) \ldots \right) \right) \right), x \le 0. \qquad (24)$$

Оценка числа членов также несколько усложняется из-за умножения коэффициента $b_N^{(-1/2)}$ на множитель $N+1$. Если учесть убывание самих множителей $b_N^{(-1/2)}$ с ростом N, то в пределах точности *double precision* произведение коэффициента на множитель не превышает 4 – 5. Поэтому оценку

Таблица 2

Коэффициенты $b_n^{(-1/2)}$

n	$b_n^{(-1/2)}$	n	$b_n^{(-1/2)}$
0	+1.0000000000000000	35	–0.1144223744158966
1	–0.4142135623730952	36	+0.1199715671513333
2	+0.4809739520123131	37	–0.1114746387738546
3	–0.3144374568437761	38	+0.1167365711505689
4	+0.3549697390579653	39	–0.1087452288479589
5	–0.2639146991274261	40	+0.1137478638410572
6	+0.2930522006577245	41	–0.1062083852096059
7	–0.2320601242278460	42	+0.1109758002652137
8	+0.2548090973198841	43	–0.1038424182241910
9	–0.2096150772531664	44	+0.1083954929661165
10	+0.2282719824176772	45	–0.1016289162398528
11	–0.1926940258158894	46	+0.1059858744749436
12	+0.2085041251467703	47	–0.0995521343867323
13	–0.1793427156614366	48	+0.1037289759507870
14	+0.1930573276211403	49	–0.0975985172145128
15	–0.1684564022391625	50	+0.1016093654791747
16	+0.1805638658684417	51	–0.0957563218416321
17	–0.1593570752767509	52	+0.0996137059063525
18	+0.1701927587728164	53	–0.0940153175150493
19	–0.1516023838328401	54	+0.0977304032825524
20	+0.1614065799267777	55	–0.0923665439128122
21	–0.1448897315196066	56	+0.0959493247710597
22	+0.1538405143150336	57	–0.0908021150751291
23	–0.1390041494476214	58	+0.0942615703728568
24	+0.1472371586201546	59	–0.0893150591168811
25	–0.1337881296456910	60	+0.0926592867469433
26	+0.1414090661254092	61	–0.0878991862486335
27	–0.1291232580225183	62	+0.0911355142547363
28	+0.1362160935092664	63	–0.0865489793786202
29	–0.1249185519381565	64	+0.0896840604456143
30	+0.1315511039258170	65	–0.0852595028653947
31	–0.1211027886077717	66	+0.0882993947491819
32	+0.1273306146111902	67	–0.0840263259647454
33	–0.1176193041380472		
34	+0.1234884888300876		

(8) достаточно слегка поправить, добавив в правую часть слагаемое “+1”, т.е. взять на один член больше. Даже при $x=0$ сумма будет содержать не более $N=34$ членов.

Положительные аргументы. При $x>>1$ функции целого индекса вычисляются с помощью точной формулы (9). Для функций полуцелого индекса существует сходная формула, которая является не точной, а асимптотической:

$$I_k(x)\approx\frac{x^{k+1}}{k+1}\left(1+\sum_{n=1}^{N}\frac{A_n^{(k)}}{x^{2n}}\right),A_n^{(k)}=(2-2^{2-2n})\zeta(2n)\prod_{p=1}^{2n}(k+2-p),x>>1. \quad (25)$$

Оценка погрешности формулы (25) в литературе отсутствует. Поэтому при выборе допустимого числа N руководствуются следующими соображениями. Найдем отношение двух последовательных членов. Учтем, что при $N>>1$ можно положить для оценок $\zeta(2N)\approx1$ и $2^{2-2N}\approx0$. Тогда отношение $(N+1)$-го члена к N-му примерно равно

$$q_N\equiv\frac{A_{N+1}^{(k)}}{A_N^{(k)}x^2}\approx\frac{(k+1-2N)(k+2-2N)}{x^2}\approx\left(\frac{2N-k-3/2}{x}\right)^2. \quad (26)$$

Мы рассматриваем случай $x>>1$. Будем увеличивать N. Сначала правые части в (26) будут существенно меньше 1, т.е. последовательные члены суммы быстро убывают. При этом прибавление нового члена повышает точность. Однако при дальнейшем увеличении N с некоторого момента правая часть станет больше 1, т.е. члены ряда начнут возрастать. Очевидно, в этом случае прибавление дальнейших членов ряда увеличивает погрешность. Оптимальным при заданном x следует считать то значение N, когда величина q_N в (26) немного меньше 1.

Сделаем конкретную оценку. Если $x>>1$ и $N>>1$, то величина q_N довольно медленно меняется вблизи оптимального N. Поэтому начало отброшенной части ряда близко к геометрической прогрессии со знаменателем q_N. Если положить $q_N=1/4$, то сумма такой геометрической прогрессии не превосходит 1/3 от величины последнего учтенного члена. Поэтому разумной оценкой оптимума будет

$$N_{opt}\approx\frac{1}{4}(x+2k+3). \quad (27)$$

Достигаемая при этом относительная погрешность ε_{opt} будет составлять 1/3 часть последнего оставленного члена, т.е.

$$\varepsilon_{opt} \approx \frac{1}{3} x^{-2N_{opt}} \prod_{p=1}^{2N_{opt}} |k+2-p|. \qquad (28)$$

Получить погрешность меньше ε_{opt} при данном x с помощью ряда (25) невозможно.

Возьмем в качестве ε_{opt} ошибку компьютерного округления ε и подставим ее в оценку (28). Тогда для различных k методом подбора можно найти те минимальные значения $x_{\min}$, при которых ряд (25) может обеспечить требуемую точность ε. Соответствующие значения $N_{\max}$ и $x_{\min}$ для полуцелых k приведены в Таблице 4 для трех значений ε, соответствующих точностям *single, double* и *long double precision*.

Напомним, что если $x < x_{\min}$, то ряд (25) не может обеспечить требуемую точность ε ни при каком значении N. Если же $x > x_{\min}$, то заданная точность ε обеспечивается уже при некотором $N < N_{\max}$, и можно написать оценку для соответствующего N. Однако, такая оценка довольно сложна, но даже в самом неблагоприятном случае $k = -3/2$ значение $N_{\max}$ невелико. Поэтому на практике можно суммировать ряд (25) для каждого k до своего $N_{\max}$, определяемого Таблицей 4. Само суммирование целесообразно производить по схеме Горнера:

$$\frac{1}{x^2}\left(A_1^{(k)} + \frac{1}{x^2}\left(A_2^{(k)} + \frac{1}{x^2}\left(A_3^{(k)} + \ldots + \frac{1}{x^2}\left(A_N^{(k)} \right)\ldots \right)\right)\right). \qquad (29)$$

Такая запись справедлива не только для N_{opt}, но и для произвольно заданного N.

5. Квадратуры для функций полуцелого индекса

Выше были построены формулы, по которым можно рассчитывать функции полуцелого индекса при $x \le 0$ или $x \ge x_{min}, x >> 1$ с требуемой высокой точностью. В оставшемся промежутке $0 < x < x_{min}$ можно вычислять оставшиеся функции полуцелого индекса с помощью квадратур интеграла (1). В этом случае наименее трудоемким способом будет построение квадратур с экспоненциальной сходимостью. При этом возникает два различных случая, которые опишем ниже.

Функции индекса $\boldsymbol{k \geq -1/2}$**.** Они определяются через сходящийся интеграл (1). Сделаем в интеграле (1) замену переменных $t = \tau^2$. Тогда интеграл (1) приведется к следующему виду

$$I_k(x) = 2\int_0^\infty \frac{\tau^{2k+1} \mathrm{d}\tau}{1+\exp\left(\tau^2 - x\right)}, k \geq -\frac{1}{2}. \tag{30}$$

При полуцелых $k \geq -1/2$ показатель степени в подынтегральном выражении будет целым четным неотрицательным числом. Поэтому подынтегральное выражение будет четной функцией τ, и все его нечетные производные на нижнем пределе интегрирования $\tau = 0$ обращаются в нуль. Тем самым, на нижнем пределе интегрирования удовлетворяется условие Частного случая Утверждения 1.

На верхнем пределе интегрирования подынтегральное выражение убывает как $\exp(-\tau^2)$. При этом все производные быстро стремятся к нулю. Но применять формулу Эйлера-Маклорена на равномерной сетке (15) к бесконечному интервалу невозможно. Поэтому для численного интегрирования надо обрезать интеграл (30) и положить

$$I_k(x) \approx 2\int_0^T \frac{\tau^{2k+1} \mathrm{d}\tau}{1+\exp\left(\tau^2 - x\right)}, k \geq -\frac{1}{2}. \tag{31}$$

Верхний предел T нужно выбирать так, чтобы во-первых, отброшенной частью интеграла можно было бы пренебречь, во-вторых, чтобы производные при T были бы настолько малы, чтобы их вклад в формулы Эйлера-Маклорена был пренебрежимо мал. Тогда на отрезке $0 \leq \tau \leq T$ формула Эйлера-Маклорена на равномерной сетке (15) обеспечит экспоненциальную сходимость.

Вообще говоря, оценка минимального T зависит от x и k. Очевидно, что T возрастает при увеличении x или увеличении k. Нетрудно оценить отброшенную часть интеграла (31). Для не малых $x > 0$ она составляет примерно $T^{2k}\exp(x - T^2)$. Для получения относительной ошибки эту величину нужно сравнить с асимптотикой интеграла (1) $x^{k+1}/(k+1)$. Для получения относительной погрешности ε должно выполняться условие

$$e^{T^2} > \frac{(k+1)T^{2k}}{\varepsilon x^{k+1}} e^x, x >> 1; \tag{32}$$

целесообразно брать T "с запасом" для обеспечения надежности алгоритма. Но левая часть (32) быстро возрастает при увеличении T, так что на практике достаточно умеренного увеличения T.

Если определять T по значению $x_{\min}$, то такая величина будет пригодна для всех $x < x_{\min}$. Рассмотрим два крайних случая при $\varepsilon = 10^{-16}$ (*double precision*). Первый соответствует $k = -1/2, x_{\min} = 39$ (см. Таблицу 4); он дает $T > 8.4$. Второй соответствует $k = 7/2, x_{\min} = 29$; он дает $T > 8.5$. Для остальных функций полуцелых индексов $k \geq -1/2$ получаются практически такие же результаты. Поэтому для всех индексов $k \geq -1/2$ и для любых аргументов $x \leq x_{\min}$ далее будем единообразно брать **T=12**. Это создает необходимый запас надежности, включая обеспечение малости высоких производных на правой границе.

Подробнее обсудим сходимость квадратур. На рис. 5 изображена комплексная плоскость переменной интегрирования τ. Отрезок интегрирования выделен жирной линией – это вещественная положительная полуось. Показатель экспоненциальной сходимости определяется расстоянием от ближайшего полюса до промежутка интегрирования. Видно, что при $x > 0$ это расстояние не зависит от положения x и равно π; при $x < 0$ это расстояние равно $\sqrt{\pi^2 + x^2}$. Таким образом, при всех $x > 0$ скорость сходимости будет примерно одинаковой; но при $x \to -\infty$ скорость сходимости будет быстро возрастать, и потребуется существенно меньшее число узлов сетки.

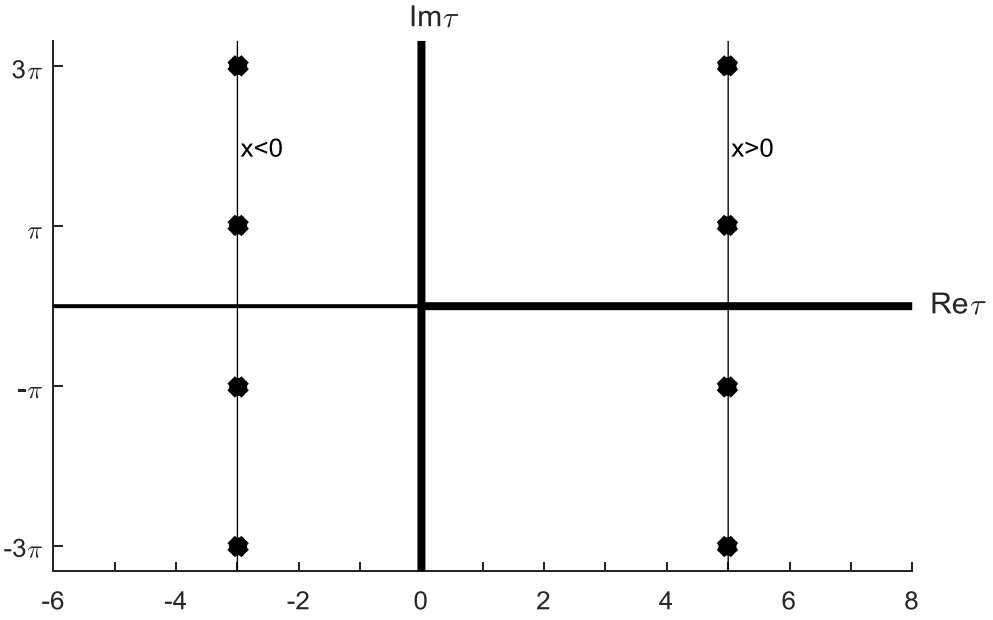

Рис.5. Полюса подынтегрального выражения для функций ФД.

Численные расчеты подтверждают эти соображения. Интеграл (31) вычисляется на отрезке [0,12] по формуле трапеций (15). Вычисление проводится с автоматическим сгущением сетки до выхода на ошибки округления. Эта процедура аналогична тому, что изложено в Практических рекомендациях в Разделе 3. При этом окончательные сетки содержат $N = 48$

или 96 узлов при $x \approx 0$ и доходят до $N = 192$ при $x \approx x_{\min}$. Действительно видно, что число узлов слабо зависит x при широких пределах изменения x.

Индекс $\boldsymbol{k = -3/2}$**.** Для этого индекса интеграл (1) оказывается расходящимся. Поэтому для вычисления функций целесообразно переопределить ее через производную функции старшего индекса и провести следующее преобразование [9]

$$I_{-3/2}(x) = -2I'_{-1/2}(x) = -2\int_0^\infty \frac{d}{dx}\left(\frac{1}{1+e^{t-x}}\right)\frac{dt}{\sqrt{t}} = \\ = -2\int_0^\infty \frac{e^{-t-x}dt}{\sqrt{t}\left(e^{-t}+e^{-x}\right)^2} = -\frac{1}{2}\int_0^\infty \left(ch\frac{t-x}{2}\right)^{-2}\frac{dt}{\sqrt{t}}. \tag{33}$$

В последнем выражении сделаем замену переменных $t = \tau^2$ и одновременно обрежем интеграл по верхнему пределу, аналогично показанному выше. Получим выражение

$$I_{-3/2}(x) = -\int_0^T \left(ch\frac{\tau^2 - x}{2}\right)^{-2} d\tau. \tag{34}$$

Подынтегральное выражение есть четная функция τ, так что при достаточно большом T квадратура трапеций на равномерной сетке для интеграла (34) будет сходиться экспоненциально. Оценку T производим аналогичным образом. Приравнивая отброшенную часть интеграла к ε-й доле главного члена асимптотики, получаем минимальную оценку

$$T^2 = x + \ln\frac{\sqrt{x}}{\varepsilon\sqrt{T}}. \tag{35}$$

Подставляя сюда наихудшее условие $x = x_{\min} = 44$ (см. Таблицу 4), находим $T \approx 9$. Для создания достаточного запаса надежности опять положим $T = 12$. Тогда квадратура трапеций с контролем экспоненциальной сходимости обеспечивает точность *double* при $N = 96$ узлах сетки для $x \approx 0$ и $N = 384$ при $x = x_{\min}$.

6. Интегральная функция

Напомним, где возникает интегральная функция ФД. При квантово-механических расчетах атома широко используется приближение Хартри–Фока, при котором в многоэлектронном уравнении Шредингера

многоэлектронная волновая функция заменяется не произведением электронных функций (как в приближении Хартри), а детерминантом, составленным из одноэлектронных функций. Это эффективно учитывает обменное взаимодействие. При этом к электростатическому взаимодействию электронов добавляется дополнительное слагаемое, называемое потенциалом обменного взаимодействия. Однако, такое приближение ещё очень сложно для численных расчетов.

Вычисления обменного потенциала в квазиклассическом приближении дает гораздо более простое выражение. Такое приближение называют приближением Слэтера, который первым написал его для нулевой температуры. При ненулевой температуре возникает следующая функция:

$$J(x) = \int_0^x [I_{-1/2}(\xi)]^2 d\xi; \tag{36}$$

ее называют интегральной функцией Ферми-Дирака. Рассмотрим методы вычисления этой функции.

Отрицательные аргументы. Функцию $J(x)$ при $x \le 0$ также удобно вычислять с помощью всюду сходящегося ряда

$$J(x) = 4\pi \sum_{n=0}^{N} c_n g^{n+2}, \tag{37}$$

где коэффициенты определяются рекуррентными формулами

$$c_0 = \frac{1}{2}; c_n = \frac{1}{n+2}\left[(n+1)c_{n-1} + \sum_{p=0}^{n} b_p^{(-1/2)} b_{n-p}^{(-1/2)} \right], n > 0. \tag{38}$$

Коэффициенты c_n с 17-ю десятичными знаками приведены в Таблице 3. Видно, что все коэффициенты не превышают по модулю c_0; коэффициенты с четными и нечетными номерами образуют две подпоследовательности, в которых они убывают по модулю. Но это убывание сравнительно медленное, так что им далее будем пренебрегать.

Поэтому для численного расчета с относительной точностью ε достаточно выбрать верхний предел суммы (37) по формуле (8), а само вычисление производить по схеме Горнера:

$$J(x) = 4\pi g^2 \left(c_0 + g\left(c_1 + g\left(c_2 + \ldots + g\left(c_N \right)\ldots \right) \right) \right), x \le 0. \tag{39}$$

Поскольку все коэффициенты $c_n > 0$, то в схеме фактические вычитания отсутствуют, так что ошибки округления при вычислениях минимальны.

Таблица 3

Коэффициенты c_n

n	c_n	n	c_n
0	0.50000000000000000	34	0.06080644154622106
1	0.05719095841793650	35	0.02155265169244273
2	0.32627341363306145	36	0.05813297459323694
3	0.05555337454026419	37	0.02083472052787313
4	0.24658846860286468	38	0.05569965933125015
5	0.05073748578641944	39	0.02016950769407704
6	0.20002927599276679	40	0.05347481198170670
7	0.04624864575737019	41	0.01955115665526985
8	0.16919747074124367	42	0.05143215013165339
9	0.04243339943502982	43	0.01897466444641222
10	0.14714269473929523	44	0.04954968098610817
11	0.03922350305150957	45	0.01843573019370941
12	0.13051578228471081	46	0.04780885458903570
13	0.03650451326932291	47	0.01793063490397057
14	0.11749388727681387	48	0.04619391091732902
15	0.03417682128363331	49	0.01745614524353493
16	0.10699573032743079	50	0.04469137079993952
17	0.03216234704554120	51	0.01700943589632719
18	0.09833733348435542	52	0.04328963488953086
19	0.03040122425575591	53	0.01658802643927440
20	0.09106388018262407	54	0.04197866475792093
21	0.02884749082202021	55	0.01618972965664492
22	0.08486060243799427	56	0.04074972707877874
23	0.02746553440048496	57	0.01581260893897318
24	0.07950240515223343	58	0.03959518675467695
25	0.02622743720392143	59	0.01545494295054539
26	0.07482385935080225	60	0.03850833836543254
27	0.02511104569452402	61	0.01511519615326141
28	0.07070053682482372	62	0.03748326787688946
29	0.02409857188256913	63	0.01479199408027981
30	0.06703696772323861	64	0.03651473843531775
31	0.02317557364450819	65	0.01448410248600719
32	0.06375862474843785	66	0.03559809547535675
33	0.02233020393904351		

Большие аргументы. Для интегральной функции асимптотический ряд при $x >> +1$ записывается в следующем виде:

$$J(x) \approx 2x^2\left[1-\frac{\pi^2}{6}(\ln x - j)\frac{1}{x^2} - 4\sum_{n=2}^{N}\frac{(n-1)C_n}{x^{2n}}\right], \qquad C_n = \sum_{q=0}^{n} A_q^{(-1/2)} A_{n-q}^{(-1/2)}, x \to +\infty. \tag{40}$$

Здесь $j = 0.76740941382814898$, а коэффициенты C_n приведены с 17 десятичными знаками в Таблице 4. Для практических вычислений сумму в формуле (40) удобно записать по схеме Горнера:

$$\frac{4}{x^2}\left(\frac{C_2}{x^2}+\left(\frac{2C_3}{x^2}+\left(\frac{3C_4}{x^2}+...+\left(\frac{(N-1)C_N}{x^2}\right)...\right)\right)\right). \tag{41}$$

Поскольку все $C_n < 0$ при $n \geq 2$, то в схеме Горнера знаки всех слагаемых одинаковы и при вычислении этой суммы ошибки округления также минимальны.

Таблица 4

Коэффициенты C_n

n	C_n	n	C_n
0	+1.00000000000000000	7	-1924202279.42978835
1	-0.82246703342411309	8	-376996608458.572022
2	-3.38226010534730559	9	-96469021655492.7344
3	-56.7486676763200464	10	-31243036135798104.0
4	-2076.43981697169329	11	-12492545181655248896.0
5	-133516.623919083009	12	-6044381261816933646336.0
6	-13363920.4954685569		

Оптимальное число слагаемых в (40) и (41) определяем аналогично сказанному выше. Надо требовать, чтобы отношение первого отброшенного члена к оставленному составляло примерно 1/4. Поэтому N_{opt} для заданного x определяется из соотношения

$$x^2 \approx \frac{4NC_{N+1}}{(N-1)C_N}. \tag{42}$$

Соответственно, $x_{\min}$ и $N_{\max}$ определяются из следующей системы уравнений:

$$x^2 \approx \frac{4NC_{N+1}}{(N-1)C_N}, \frac{4(N-1)C_N}{3} \approx \varepsilon x^{2N}.$$

Разрешим каждое уравнение относительно x:

$$x(N) \approx \sqrt{\frac{4NC_{N+1}}{(N-1)C_N}}, x(N) \approx \sqrt[2N]{\frac{4(N-1)C_N}{3\varepsilon}}. \qquad (43)$$

Подставляя сюда коэффициенты C_n из Таблицы 4, вычислим значения $x(N)$ для целых значений N и различных ε. Полученные величины представлены в Таблице 5. Теперь систему (43) нетрудно приближенно решить графически. В полученном результате надо округлить нецелое N вверх до целого; соответственно надо округлить $x(N)$. Найденные значения $x_{\min}$ и $N_{\max}$ приведены Таблице 6.

Таблица 5

Значения $x(N)$ системы (43)

		N								
		11	10	9	8	7	6	5	4	3
I-е ур.		46.14	42.16	38.18	34.20	30.24	26.29	22.37	18.52	14.8
II-е ур.	$\varepsilon = 10^{-8}$	19.48	19.17	18.99	18.98	19.24	19.91	21.34	24.28	30.9
	$\varepsilon = 10^{-16}$	41.96	44.30	47.72	52.83	60.83	74.22	99.04	153.2	309
	$\varepsilon = 10^{-19}$	55.95	60.64	67.41	77.55	93.68	121.6	176.1	305.7	733

Таблица 6

Границы применимости асимптотических рядов

k	single		double		long double	
	$x_{\min}$	$N_{\max}$	$x_{\min}$	$N_{\max}$	$x_{\min}$	$N_{\max}$
-3/2	22	6	44	11	55	14
-1/2	18	5	39	10	49	13
1/2	15	5	35	10	46	13
3/2	13	5	33	10	42	12
5/2	12	5	30	10	39	12
7/2	12	6	29	10	37	12
$J(x)$	23	5	46	11		

Вычисление квадратурами. Остается незаполненным промежуток между $x=0$ и $x_{\min}=46$. Вычислять в нем интеграл (36) обычными одномерными квадратурами крайне невыгодно: такие квадратуры будут иметь только степенную сходимость, и для получения точности $\varepsilon=10^{-16}$ потребуется неприемлемо большое число узлов сетки. Однако, оказалось возможным свести задачу к экспоненциально сходящимся квадратурам.

Для этого подставим в (36) выражение $I_{-1/2}(\xi)$ через одномерный интеграл, сразу делая замену переменной интегрирования $t=\tau^2$. Квадрат такого последнего одномерного интеграла будет двойным интегралом по $d\tau d\theta$. Поэтому окончательно, вместо одномерного интеграла (36) получим тройной интеграл

$$J(x)=4\int_0^x d\xi\int_0^\infty\int_0^\infty \frac{d\tau d\theta}{\left[1+\exp(\tau^2-\xi)\right]\left[1+\exp(\theta^2-\xi)\right]}. \qquad (44)$$

Переменим порядок интегрирования, и сначала произведем интегрирование по ξ. Умножая числитель и знаменатель подынтегрального выражения на $e^{2\xi}$, преобразуем его к следующему виду:

$$\frac{d\xi}{\left[1+\exp(\tau^2-\xi)\right]\left[1+\exp(\theta^2-\xi)\right]}=\frac{e^\xi d(e^\xi)}{\left(e^\xi+e^{\tau^2}\right)\left(e^\xi+e^{\theta^2}\right)}=$$

$$=\frac{1}{e^{\tau^2}-e^{\theta^2}}\left(\frac{e^{\tau^2}}{e^{\tau^2}+e^\xi}-\frac{e^{\theta^2}}{e^{\theta^2}+e^\xi}\right)d(e^\xi).$$

Теперь интегрирование по ξ выполняется в элементарных функциях, и тройной интеграл (44) превращается в двойной:

$$J(x)=4\int_0^\infty\int_0^\infty \frac{e^{\tau^2}\ln(1+e^{x-\tau^2})-e^{\theta^2}\ln(1+e^{x-\theta^2})}{e^{\tau^2}-e^{\theta^2}}d\tau d\theta. \qquad (45)$$

Подынтегральная функция симметрична и положительна. При $\tau=\theta$ в ней возникает неопределенность типа 0/0, которая раскрывается по правилу Лопиталя:

$$\left.\frac{e^{\tau^2}\ln(1+e^{x-\tau^2})-e^{\theta^2}\ln(1+e^{x-\theta^2})}{e^{\tau^2}-e^{\theta^2}}\right|_{\tau=\theta}=\ln(1+e^{x-\tau^2})-\frac{e^x}{e^{\tau^2}+e^x}. \qquad (46)$$

Вблизи линий $\tau=\theta$ при непосредственном вычислении подынтегрального выражения (45) на разностной сетке может возникать потеря точности; в этом

случае в окрестности линии $\tau = \theta$ надо использовать уточнение выражения (46) с использованием более высоких производных числителя.

Подынтегральное выражение (45) четно по τ и по θ. Тем самым, на полуосях $\tau = 0$ и $\theta = 0$ выполняются условия Утверждения 2 для экспоненциальной сходимости квадратур трапеций. При $\tau^2 >> x$ или $\theta^2 >> x$ подынтегральное выражение очень быстро убывает со всеми производными. Поэтому можно ограничить область интегрирования квадратом $0 \le \tau, \theta \le T$. Следовательно, двумерная квадратура трапеций на равномерной сетке в этом квадрате будет иметь экспоненциальную сходимость. Поскольку подынтегральное выражение симметрично, можно ограничиться интегрированием по треугольнику $0 \le \tau \le \theta \le T$, что вдвое уменьшает объем вычислений.

Обозначим подынтегральное выражение через $f(\tau, \theta)$ и введем равномерные сетки $h = T / N, \tau_n = nh, \theta_m = mh$. Тогда квадратура трапеций для треугольной области запишется с весами w_{nm}:

$$J(x) = 8h^2 \sum_{n=N}^{0} \sum_{m=N}^{n} w_{nm} f(\tau_n, \theta_m);$$
$$w_{00} = \frac{1}{8}; w_{n0} = w_{nn} = \frac{1}{2}, n > 0; w_{nm} = 1, n > 1, n > m. \qquad (47)$$

Вес на верхней границе треугольника безразличен, так как там функция и ее производные пренебрежимо малы. Суммирование в формуле (47) поставлено в обратном порядке, так как суммирование от малых членов к большим уменьшает ошибки округления.

Замечание. Вычисление экспонент является довольно трудоемкой операцией. Поэтому следует заранее вычислить значения экспонент $\exp(\tau_n^2)$, $0 \le n \le N$. Далее эти значения надо подставлять при вычислении подынтегрального выражения при соответствующих значениях τ_n и θ_m. Это примерно в $\sim N$ раз уменьшает объем вычислений. Видно, что заранее вычисленные экспоненты одинаково работают для обоих направлений интегрирования. Поскольку остальные действия являются арифметическими, то трудоемкость вычисления двумерного интеграла будет мало отличаться от трудоемкости вычисления одномерного интеграла. При таком усовершенствовании вычисление двумерного интеграла оказывается пригодным, как способ прямого вычисления функции во встроенных стандартных подпрограммах.

Для обеспечения точности $\varepsilon = 10^{-16}$ при $x = 0$ достаточно сетки $N = 96$, а вблизи $x_{\min}$ – сетки $N = 384$.

7. Некоторые результаты

Приведем некоторые результаты численных расчетов. Заметим, что при $x \to -\infty$ главный член простейшего ряда есть $I_k(x) \approx \Gamma(k+1)e^x$. Поэтому отношение $I_k(x)/\Gamma(k+1) \approx e^x$, т.е. соответственно нормированная функция ФД при любых k практически совпадают для $x \to -\infty$. Фактически, эти отношения остаются довольно близкими даже при $x = 0$. В Таблице 7 приведены значения этих нормированных функций с 16-ю десятичными знаками при $x = 0$. Видно, что эти значения слабо возрастают с увеличением k. Приведенные значения полезны как реперные точки.

Таблица 7

Реперные значения функций ФД при $x = 0$

k	$I_k(0)/\Gamma(k+1)$	k	$I_k(0)/\Gamma(k+1)$
-3/2	0.380104812609684	2	0.901542677369696
-1/2	0.604898643421630	5/2	0.927553577773948
0	0.693147180559945	3	0.947032829497246
1/2	0.765147024625408	7/2	0.961483656632978
1	0.822467033424113	4	0.972119770446909
3/2	0.867199889012184		

На Рис.6 приведены графики нормированных функций ФД. Поскольку все нормированные функции положительны и меняются в очень больших пределах, то для рисунка выбран полулогарифмический масштаб. В левой части графика при $x \to -\infty$ все нормированные функции быстро стремятся к прямой, проходящей через начало координат. В правой части они расходятся, не пересекаясь; чем больше k, тем выше лежит кривая. Для $k > -1$ каждая кривая является монотонно возрастающей; но для $k = -3/2$ кривая имеет максимум.

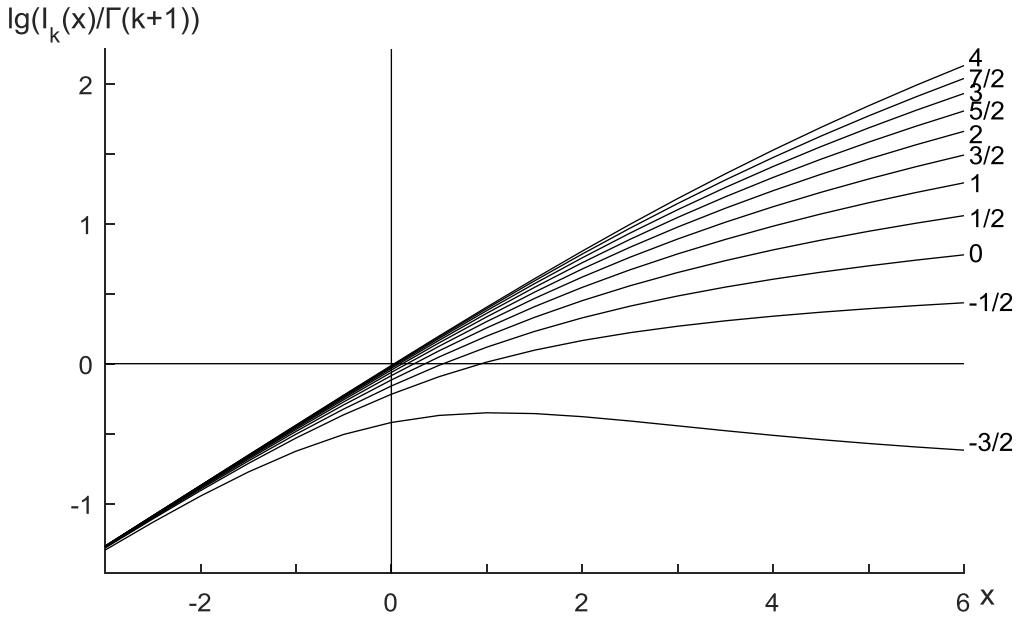

Рис.6. Нормированные функции ФД; около линий значения k.

Заметим, что для $k = -3/2$ ненормированная функция ФД отрицательна, но нормированная функция является положительной.

Библиографический список

1. Калиткин Н.Н., Колганов С.А. Функции Ферми-Дирака. I. Свойства функций., 2018, v.41, p.81-102. URL:https://link.springer.com/article/10.1007%2FBF01391920

2. IEEE Std 754-1985. IEEE Standard for Binary Floating-Point Arithmetic. URL: https://www.ime.unicamp.br/~biloti/download/ieee_754-1985.pdf

3. Бахвалов Н.С., Жидков Н.П., Кобельков Г.М. Численные методы // 3-е издание, Москва, БИНОМ. Лаборатория знаний, 2004.

4. Калиткин Н.Н. Численные методы // 2-е издание, Санкт-Петербург, «БХВ-Петербург», 2011.

5. Калиткин Н.Н., Альшина Е.А. Численные методы, книга 1, Численный анализ // Москва, «Академия», 2013.

6. Калиткин Н.Н., “Квадратуры Эйлера–Маклорена высоких порядков”, Матем. моделирование, 16:10 (2004), 64–66, URL: http://mi.mathnet.ru/rus/mm/v16/i10/p64

7. Белов А. А., О коэффициентах квадратурных формул Эйлера-Маклорена, Математическое моделирование, 25:6(2013), 72-79. URL: http://mi.mathnet.ru/rus/mm/v25/i6/p72

Belov A. A., Coefficients of Eiler-Maclaurin formulas for numerical integration. Mathematical models and computer experiments, Jan 2014, Vol 6, Issue 1, pp 32-37.

8. Градштейн И. С., Рыжик И. Б. Таблицы интегралов, сумм, рядов и произведений. 4-е издание, ФМ, Москва, 1963.

9. Калиткин Н.Н., Колганов С.А., “Вычисление функций Ферми–Дирака экспоненциально сходящимися квадратурами”, Матем. моделирование, 29:12 (2017),134–146; URL: http://mi.mathnet.ru/rus/mm/v29/i12/p134

Kalitkin N.N, Kolganov S.A., Computing the Fermi−Dirac Functions by Exponentially Convergent Quadratures, Math. Models Comput. Simul.,10:4 (2018),472–482;

Оглавление