МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ 2017 год, том 29, номер 12, стр. 134-146

ВЫЧИСЛЕНИЕ ФУНКЦИЙ ФЕРМИ-ДИРАКА ЭКСПОНЕНЦИАЛЬНО СХОДЯЩИМИСЯ КВАДРАТУРАМИ

© *2017 г.* ***Н.Н. Калиткин*[1], *С.А. Колганов*[2]**

[1] Институт прикладной математики им. М.В. Келдыша РАН, Москва
[2] Национальный исследовательский университет «МИЭТ», Зеленоград
kalitkin@imamod.ru , mkandds2012@gmail.com

Работа поддержана грантом РНФ 16-11-10001.

Для прямого вычисления функций Ферми-Дирака полуцелого индекса построены специализированные квадратурные формулы высокой точности. Показано, что зависимость погрешности от числа узлов является не степенной, а экспоненциальной. Исследованы свойства таких формул. Показано, что показатель экспоненты сходимости определяется расстоянием до ближайшего полюса подынтегрального выражения. Это обеспечивает очень быструю сходимость квадратур. Построены несложные аппроксимации функций Ферми-Дирака целых и полуцелых индексов, имеющие точность лучше 1%; они удобны для физических оценок. Попутно найдены асимптотические представления для чисел Бернулли.

Ключевые слова: функции Ферми-Дирака, полуцелые индексы, квадратуры, экспоненциальная сходимость, числа Бернулли.

CALCULATION OF THE FERMI-DIRAC FUNCTIONS WITH EXPONENTIALLY CONVERGENT QUADRATURES

***N.N. Kalitkin*[1], *S.A. Kolganov*[2]**

[1] Keldysh Institute of Applied Mathematics of Rus. Acad. of Sci., Moscow
[2] National Research University of Electronic Technology, Zelenograd
kalitkin@imamod.ru, mkandds2012@gmail.com

The special quadrature formulas of high accuracy were built for a direct calculation of the Fermi-Dirac functions of half-integer indexes. It is shown that the dependence of the error from the number of nodes is not a power law, but exponential. We investigated the properties of such formulas. It is shown that the index of this exponent is proportional to the distance between the integral segment and the nearest pole of expanded expression. This provides a very fast convergence of quadratures. The simple approximations of the Fermi-Dirac functions of integer and half-integer indexes were constructed; their accuracy was about 1%. They are convenient for the physical estimations. During the research, we found an asymptotic representation for Bernoulli numbers.

Key words: Fermi-Dirac functions, half-integer indexes, quadratures, exponential convergence, Bernoulli numbers.

1. Введение

1.1. Функции Ферми-Дирака. Функции Ферми-Дирака были предложены в связи с задачами квантовой механики. Они являются моментами фермиевского распределения,

рассматриваемого как функция импульса. Математически они определяются как [1]

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

В квантовой механике индекс k принимает только целые или полуцелые значения. Целые k соответствуют нечетным моментам импульса, а полуцелые – четным. Например, индекс $k = 1/2$ соответствует плотности фермионов, $k = 3/2$ – кинетической энергии, $k = 1$ – переносу числа частиц, т.е. электропроводности; $k = 2$ – переносу энергии, т.е. теплопроводности; $k = 3$ – вязкости.

Поскольку функции Ферми-Дирака имеют важные практические применения, их вычислению с высокой точностью посвящена обширная литература [1-10]. Для функций целых индексов аппроксимации с 16 верными десятичными знаками построены в [9], что в известной мере исчерпывает проблему при расчетах с 64-битовыми числами. Для функций полуцелых индексов опубликованные аппроксимации имеют меньшую точность. Кроме того, использование зарубежных публикаций несколько рискованно из-за возможных опечаток. Например, в [2] анонсируется 12 верных знаков; однако для индекса $k = 1/2$ по результатам расчетов видно, что имеется опечатка в 5-м или 6-м знаке. Поэтому для полуцелых индексов остается актуальным вопрос вычисления с 16 верными знаками; ему посвящена данная работа.

1.2. Связь функций разных индексов. В математической литературе рассматривают общий случай для произвольных значений k. Для значений $k \le -1$ интеграл (1) расходится. Однако для $k > 0$ справедливо проверяемое легко соотношение

$$I_k'(x) = k \cdot I_{k-1}(x). \tag{2}$$

Принимая это соотношение для меньших значений k можно доопределить функции Ферми-Дирака для $k < -1$. Запрещенными остаются только целые отрицательные значения k; в них функция Ферми-Дирака бесконечна.

1.3. Асимптотики. Функции (1) произвольного индекса k при $x < 0$ разлагаются в следующий ряд [1]:

$$I_k(x) = \Gamma(k+1) \sum_{n=1}^{\infty} (-1)^{n-1} \frac{\exp(nx)}{n^{k+1}}, \quad x < 0. \tag{3}$$

Этот ряд знакопеременный и абсолютно сходящийся, так что погрешность не превышает первого отброшенного члена. Отсюда видно, что сходимость не является равномерной: она быстрая при $x \to -\infty$ и ухудшается при $x \to -0$.

При $x > 0$ существует следующее асимптотическое представление

$$I_k(x) \approx \frac{x^{k+1}}{k+1}\left[1 + \sum_{n=1}^{N} C_n x^{-2n}\right], \quad C_n = (k+1)k(k-1)\ldots(k-2r+2)A_n, \quad x \gg 1; \tag{4}$$

коэффициенты A_n выражаются через дзета-функцию Римана. Приведем значения первых 6 коэффициентов:

$$A_1 = \frac{\pi^2}{6}, \quad A_2 = \frac{\pi^4}{90}, \quad A_3 = \frac{\pi^6}{945}, \quad A_4 = \frac{\pi^8}{9455}, \quad A_5 = \frac{\pi^{10}}{94550}, \quad A_6 = \frac{\pi^{12}}{90}. \tag{5}$$

Представление (5) является асимптотическим. Это означает, что при возрастании N, т.е. увеличении числа членов суммы, точность сначала возрастает до некоторого номера $N(x)$, а при дальнейшем добавлении членов начинает ухудшаться. Это оптимальное $N(x)$ монотонно возрастает при $x \to +\infty$. Таким образом, представление (5) позволяет получать высокую точность при больших $x > 0$, но при умеренных x его точность невелика.

При целых $k \geq 0$ и числе членов $N \geq (k+1)/2$ среди сомножителей, связывающих C_n и A_n, появляется один нулевой. Поэтому при целом k число членов суммы ограниченно: число членов $N = \left[(k+1)/2\right]$, где квадратные скобки означают целую часть числа. В этом случае соотношение (4) путем добавления нового члена в правую часть превращается в точное соотношение:

$$I_k(x) = (-1)^k I_k(-x) + \frac{x^{k+1}}{k+1}\left[1 + \sum_{n=1}^{N} C_{2n} x^{-2n}\right], \quad k \geq 0, \quad x > 0, \quad N = [(k+1)/2]. \tag{6}$$

Оно сводит вычисление функций целого индекса положительного аргумента к функции отрицательного аргумента; последнее можно выполнить с помощью ряда (3).

Напомним также единственный случай $k = 0$, когда функция Ферми-Дирака вычисляется в элементарных функциях:

$$I_0(x) = \ln(1 + e^x). \tag{7}$$

1.4. Числа Бернулли. Мы обратили внимание, что при возрастании *n* коэффициенты A_n быстро стремятся к единице. Численный анализ этих данных позволил точно установить, что эти коэффициенты представляются в виде бесконечных сумм:

$$A_n = \sum_{p=1}^{\infty} 1/p^{2m}. \tag{8}$$

При $n > 5$ эти суммы очень быстро сходятся. Известно, что коэффициенты A_n связаны с числами Бернулли B_{2m} точным соотношением [11]

$$B_{2m} = (-1)^{m-1} \frac{(2m)!}{\pi^{2m} 2^{2m-1}} A_n. \tag{9}$$

Соотношения (8), (9) являются удобным способом вычисления чисел Бернулли. Множитель, стоящий перед A_n в (9), является главным членом асимптотики B_{2m} при $n \to \infty$.

Напомним, что при нечетном индексе все числа Бернулли равны нулю за исключением $B_1 = -1$.

2. Квадратуры для функций Ферми-Дирака

2.1. Диапазон индексов. Перечислим индексы, актуальные для физических приложений. Электронной плотности соответствует $k = 1/2$. Кинетической энергии электронов соответствует $k = 3/2$. В квантово-статистической модели атома требуются также индексы $k = -1/2$ и $k = -3/2$. Для вычисления электронной проводимости в простейшем приближении необходимо $k = 1$, для электронной теплопроводности – $k = 2$, для электронной вязкости – $k = 3$. Для уточнения этих приближений может потребоваться $k = 4$. Поэтому определим задачу так: нужны алгоритмы вычисления для целых и полуцелых индексов k в пределах $-3/2 \le k \le 4$; напомним, что для $k = -1$ функция Ферми-Дирака не вычисляется.

Для целых индексов проблема вычисления с 17 верными знаками была решена в [9]. Поэтому в данной работе построим прецизионные квадратуры только для полуцелых индексов.

2.2. Замена переменных. Квадратуры для целых k были построены в [9]. Они использовали комбинацию дробления сеток с пятиточечными формулами Гаусса. Это обеспечивало степенную сходимость $O(M^{-10})$, где M – число узлов сетки интегрирования. Однако даже такой высокий порядок точности при аргументе $x \sim 0$ требовал 128–256 интервалов сетки с пятью гауссовыми узлами на каждом интервале, т.е. всего 600–1200 расчетных точек. При дальнейшем возрастании x количество расчетных точек быстро увеличивалось. Алгоритм оказывался слишком трудоемким.

Для полуцелых индексов удалось построить несравненно более экономичные квадратуры. Для этого была предложена специфическая замена переменных в (1):

$$t = \frac{a\xi^2}{1-\xi^2}, \qquad a > 0, \quad 0 \le \xi \le 1. \tag{10}$$

Тогда (1) перепишется в следующем виде:

$$I_k(x) = 2a^{k+1}\int_0^1 \frac{\exp\left(-\dfrac{a\xi^2}{1-\xi^2}\right)\xi^{2k+1}d\xi}{\left[\exp\left(-\dfrac{a\xi^2}{1-\xi^2}\right)+\exp(-x)\right]\left(1-\xi^2\right)^{k+2}}, \quad k = -\frac{1}{2},\frac{1}{2},\frac{3}{2},\ldots, \quad x \in (-\infty;+\infty). \tag{11}$$

Обратим внимание на следующее. При $\xi = 0$, подынтегральная функция является четной, а при $\xi = 1$ она обращается в нуль со всеми своими производными из-за экспоненты от сложного аргумента в числителе. Эти свойства позволяют построить квадратуры со сверхстепенной сходимостью.

Выбор a. Параметр a выберем из условия, чтобы максимум подынтегрального выражения в (11) достигался в середине отрезка, т.е. при $\xi = 1/2$. Это приводит к уравнению

$$f(a) \equiv a - 3\left(k+\frac{7}{8}\right)\left[1+\exp\left(x-\frac{a}{3}\right)\right] = 0, \quad k = -\frac{1}{2},\frac{1}{2},\frac{3}{2},\ldots, \quad x \in (-\infty;+\infty). \tag{12}$$

Нетрудно показать, что для каждых x и k уравнение (12) имеет единственный корень $a(x,k)>0$. Этот корень легко находится ньютоновскими итерациями:

$$a_{s+1}=a_s-\frac{f(a_s)}{f'(a_s)},\quad a_0=3\left(k+\frac{7}{8}\right). \tag{13}$$

При указанном выборе нулевого приближения итерации всегда сходятся к корню, причем монотонно слева.

2.3. Квадратуры трапеций. Обозначим все подынтегральное выражение в (11) с добавлением множителя $2a^{k+1}$ через $f(\xi)$. Затем применим к (11) формулу Эйлера-Маклорена с неограниченным числом поправочных членов [12]. В качестве базисной формулы при этом выберем формулу трапеций с M интервалами (разумеется, на равномерной сетке). Это дает

$$\int_0^1 f(\xi)d\xi=\frac{1}{M}\left(\frac{f_0}{2}+f_1+\ldots+f_{M-1}+\frac{f_M}{2}\right)+\sum_{r=1}^{R}(-1)^r\alpha_r M^{-2r}\left(f_M^{(2r-1)}-f_0^{(2r-1)}\right). \tag{14}$$

Главный член в (14) есть формула трапеций, а поправочные члены представляют собой сумму по степеням M^{-2}. Верхний предел суммы R определяется количеством непрерывных производных $f(\xi)$. Если число непрерывных производных ограничено, то сходимость формулы Эйлера-Маклорена будет степенной по шагу сетки.

В нашем случае функция имеет непрерывные производные любых порядков, поэтому можно полагать $R=\infty$. При этом все нечетные производные на левой границе $f^{(2r-1)}(\xi_0)$ равны нулю в силу четности $f(\xi)$, а на правой границе $f^{(2r-1)}(\xi_M)$ равны нулю в силу быстрого убывания всех производных. Поэтому все члены степенной суммы (14) обращаются в нуль, и остается только формула трапеций. Отсюда можно сделать следующий вывод:

формула трапеций для* (12) *имеет сверхстепенную сходимость.

Следует выяснить, каким будет закон этой сходимости. Интуитивно ожидается, что сходимость будет очень быстрой.

2.4. Индекс $k=-3/2$. Для этого индекса интеграл (1) расходится. Поэтому воспользуемся соотношением (2) для значения $k=-1/2$ и запишем

$$I_{-3/2}=-2I'_{-1/2}=-2\int_0^\infty\frac{d}{dx}\left[\frac{1}{1+e^{t-x}}\right]\frac{dt}{\sqrt{t}}=-2\int_0^\infty\frac{e^{-t-x}dt}{\sqrt{t}\,(e^{-t}+e^{-x})^2},\quad x\in(-\infty;+\infty). \tag{15}$$

В этом интеграле также делаем замену переменных (10). Это дает

$$I_{-3/2}(x)=-4\sqrt{a}\int_0^1\frac{\exp\left(-\dfrac{a\xi^2}{1-\xi^2}-x\right)}{\left[\exp\left(-\dfrac{a\xi^2}{1-\xi^2}\right)+\exp(-x)\right](1-\xi^2)^{3/2}},\quad x\in(-\infty;+\infty). \tag{16}$$

Подынтегральная функция в (16) также четна при $\xi = 0$, а ее производные при $\xi \to 1$ быстро убывают. Поэтому для (16) формула трапеций также будет иметь сверхстепенную сходимость.

Выбор a. Параметр a также выбирается из требования максимума подынтегрального выражения при $\xi = 1/2$. Это дает уравнение

$$f(a) \equiv \frac{a}{3} - x - \ln\left(\frac{8a+27}{8a-27}\right) = 0. \tag{17}$$

В диапазоне $a > 27/8$ оно имеет единственный корень, который можно найти ньютоновскими итерациями При этом ньютоновские итерации сходятся монотонно слева. Если начальное приближение выбрано правее корня, то возможен переброс первой итерации левее допустимого предела $27/8$; при этом ньютоновский шаг первой итерации следует укорачивать. Для такого укорачивания хороший результат дает формула

$$\tilde{a}_{s+1} = \left(a_s^2 - \frac{27}{8}a_{s+1}\right) \Big/ \left(2a_s - a_{s+1} - \frac{27}{8}\right).$$

3. Экспоненциально сходящиеся квадратуры

3.1. Тестовый пример. Основное рассуждение пункта 2.3 можно обобщить следующим образом:

Утверждение 1. Пусть подынтегральная функция $u(\xi)$ имеет сколь угодно высокие производные, причем нечетные производные на правой и левой границах одинаковы: $u^{(2m-1)}(\xi_0) = u^{(2m-1)}(\xi_M)$. Тогда формула трапеций на равномерной сетке имеет сверхстепенную сходимость. ■

Для исследования этой сходимости рассмотрим пример, в котором известно точное значение интеграла в элементарных функциях:

$$U(p,q) = \int_0^{\pi} \frac{(c^2-1)c^p \cos(px)}{(c^2 - 2c\cos x + 1)^q} dx, \quad c > 1. \tag{18}$$

Параметры $p \geq 0$, $q \geq 1$ 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 $q = 1$ известен точный ответ [11]:

$$U(p,1) = \pi. \tag{19}$$

Были проведены расчеты интеграла (18) на сетке с M интервалами при разных значениях параметра. Погрешность расчетов при $q = 1$ определялась непосредственным сравнением с точным ответом (19). На рис.1 показана погрешность при $p = 0$ и различных значениях c в *полулогарифмическом* масштабе. Видно, что при всех значениях c кривые погрешности в этом масштабе являются прямыми. Это означает, что погрешность подчиняется закону.

$$\delta_M = \text{const} \cdot \exp(-\beta M), \quad \beta > 0. \tag{20}$$

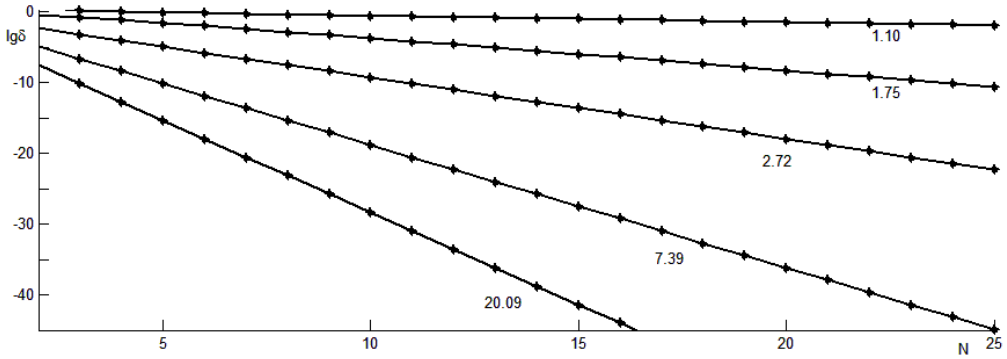

Рис.1. Погрешность квадратуры трапеций для (18) при $p=0$ и $q=1$. Цифры около линий – величины c.

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

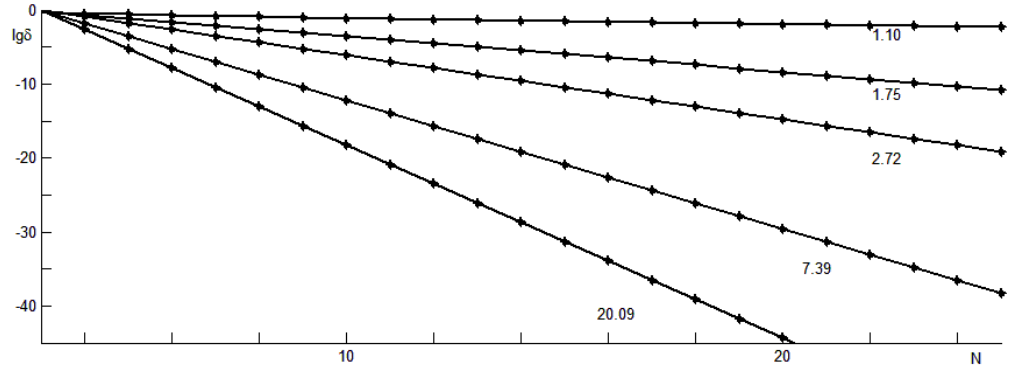

Рис.2. Погрешность квадратуры трапеций для (18) при $p=2$ и $q=1$. Цифры около линий – величины c.

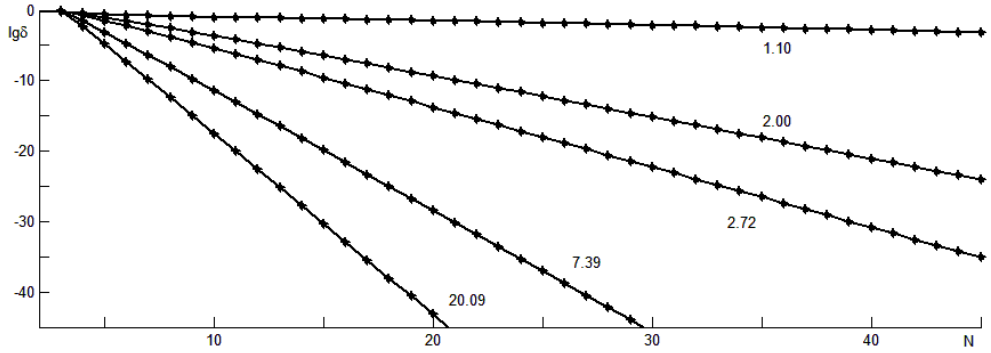

Рис.3. Погрешность квадратуры трапеций для (18) при $p=1$ и $q=2$. Цифры около линий – величины c.

Для $q>1$ точный ответ неизвестен. В этом случае для получения кривых погрешности можно воспользоваться следующими соображениями. При закономерности (20)

разности значений U при возрастании M 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 $q=2,\ p=1$. Они также оказываются прямыми.

Все это позволяет сделать эвристическое

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

Попробуем выяснить, от чего зависит коэффициент β в (20). Он не должен зависеть от максимумов модулей каких-либо производных $f(\xi)$, поскольку они входят в сумму (14) и приводят к степенной сходимости. Поэтому рассмотрим гипотезу о связи β с полюсами подынтегрального выражения. Для теста (18) подынтегральное выражение имеет полюса кратности q в точках

$$x^* = \pi \pm i\ln(c), \quad l = 0, \pm 1, \pm 2, \ldots . \tag{21}$$

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 $p = 0,1,2$. Видно, что для полюса первого порядка $(q=1)$ это отношение с высокой точностью не зависит ни от p, 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 $q=1$.

Пренебрегая этой слабой зависимостью, сделаем эвристическое

Утверждение 3. Наклон β в (20) с хорошей точностью пропорционален расстоянию от отрезка интегрирования до ближайшего полюса подынтегрального выражения в комплексной плоскости. ■

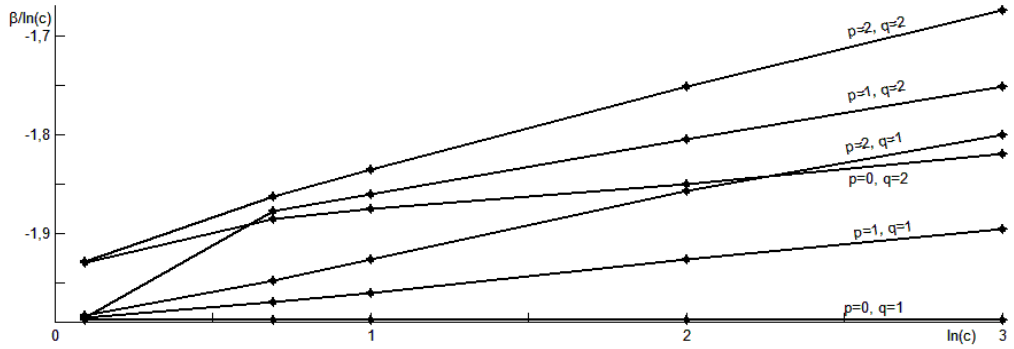

Рис.4. Зависимость $\beta / \ln(c)$ от величины c для различных p и q.

Практические рекомендации. При использовании квадратурных формул со степенной сходимостью удобно сгущать сетки по M последовательно вдвое. Это позволяет использовать обычную процедуру Ричардсона для получения априорной асимптотиче-

ски точной оценки погрешности. Такое сгущение экономично, поскольку суммарный объем всех расчетов лишь вдвое превышает объем расчетов на последней сетке [6].

Для квадратур с экспоненциальной сходимостью (20) также можно пользоваться процедурой Ричардсона, если сгущать сетки не вдвое, а каждый раз увеличивая M на 1. При этом будет получаться асимптотически точная апостериорная оценка погрешности. Однако такое сгущение сеток экономически невыгодно, поскольку суммарный объем вычислений будет в $\sim M/2$ раз больше, чем расчет на последней сетке.

Поэтому в практических расчетах удобнее увеличивать M в 2 раза. Из (20) нетрудно получить, что при этом $\delta_{2M} \sim \delta_M^2$. Такой закон убывания напоминает сходимость ньютоновских итераций вблизи простого корня: число верных десятичных знаков приблизительно удваивается с увеличением M в 2 раза. Поэтому на практике останавливаются на такой сетке $2M$, когда отклонение от интеграла на предыдущей сетке становится меньше половины стандартной ошибки округления компьютера.

3.2. О формулах Гаусса-Кристоффеля. Формулами наивысшей алгебраической точности являются формулы Гаусса-Кристоффеля. Приведем некоторые случаи, когда известна теоретическая оценка их погрешности. Например, классическая формула Гаусса для интегрирования на отрезке $[-1;1]$ с весом $\rho(x)=1$ имеет погрешность

$$\delta_M < \frac{2^{2M+1}(M!)^4}{(2M+1)[(2M)!]^3} \max_{[-1,1]} \left| f^{(2M)}(x) \right| \approx \frac{1}{2}\sqrt{\frac{\pi}{M}} \left(\frac{e}{4M} \right)^{2M} \max_{[-1,1]} \left| f^{(2M)}(x) \right|; \qquad (22)$$

последнее упрощение получено заменой факториалов по формуле Стирлинга. Для квадратуры Эрмита на отрезке $[-1;1]$ с весом $\rho(x)=(1-x^2)^{-1/2}$ справедлива аналогичная оценка

$$\delta_M < \frac{\pi}{2^{M-1}(2M)!} \max_{[-1,1]} \left| f^{(2M)}(x) \right| \approx \sqrt{\frac{\pi}{M}} \left(\frac{e}{2\sqrt{2}M} \right)^{2M} \max_{[-1,1]} \left| f^{(2M)}(x) \right|. \qquad (23)$$

Формулы (22), (23) можно переписать в следующей форме:

$$\delta_M \approx \text{const} \cdot \max_{[-1,1]} \left| f^{(2M)}(x) \right| \cdot \exp(-\gamma_M M), \qquad (24)$$

где для формулы Гаусса-Кристоффеля $\gamma_M = 2 \cdot \ln(4M/e)$, а для формулы Эрмита $\gamma_M = 2 \cdot \ln(2\sqrt{2}M/e)$. Это внешне похоже на экспоненциальную сходимость по M.

Однако действительная картина сходимости (24) существенно сложнее. С одной стороны, показатель γ_M даже увеличивается с ростом M, то есть сходимость выглядит сверхэкспоненциальной. С другой стороны, в большом числе случаев $\max\limits_{[-1,1]} \left| f^{(2M)}(x) \right|$ может быстро увеличиваться с ростом M, причем неизвестно с какой скоростью. Все это приводит к тому, что если построить кривые сходимости аналогично рис.1-3, то они будут отличаться от прямых линий. Последнее не позволяет получить апостериорную асимптотически точную оценку погрешности. Невозможность получения апостериорной оценки погрешности является теоретическим недостатком квадратур Гаусса-Кристоффеля. Практическим же недостатком этих квадратур является то, что их узлы и веса найде-

ны лишь для нескольких случаев весов интегрирования, причем эти узлы и веса выражаются через элементарные функции лишь при очень малом $M \le 5$. Поэтому реально очень высокую точность такими формулами получить нелегко.

3.3. Сходимость для функций Ферми-Дирака. Полюсы подынтегрального выражения в (11) обусловлены нулями того сомножителя в знаменателе, который содержит экспоненты. Это дает следующие значения полюсов:

$$\xi^* = \sqrt{A/(a+A)}, \quad A = x + i\pi(1+2l), \qquad -\infty < l < +\infty; \tag{25}$$

здесь следует включать оба значения квадратного корня из комплексного выражения. Значение $A_l \to \infty$ при $l \to \infty$. Легко заметить, что при этом значения $\xi^* \to \pm 1$. Точка $\xi = +1$ является концом отрезка интегрирования. Значит, одна подпоследовательность полюсов стремится к этому концу отрезка интегрирования! Не возникнет ли при этом катастрофического ухудшения сходимости?

Расчеты показали, что ситуация несколько осложняется по сравнению с тестовым примером (18). Графики погрешности, аналогичные рис.1-3, не являются строго прямолинейными, но достаточно близки к прямым. Это можно заметить, если проводить расчеты с увеличением M на 1. Если же проводить расчеты с удвоением M, как рекомендуется делать на практике, то отклонения от прямолинейности практически незаметны. Поэтому следует проводить расчеты с удвоением сеток и остановкой по приближению к ошибкам округления компьютера, как рекомендовано в п.3.1.

Дадим правдоподобное объяснение этого. На конце отрезка $\xi = +1$ производные обращаются в нуль не просто, а очень быстро, примерно, как $\exp[-0.5a/(1-\xi)]$. Поэтому все подынтегральное выражение очень мало вблизи этого конца, и правый конец отрезка интегрирования вносит слишком малый вклад в весь интеграл.

Была составлена компьютерная программа вычисления функций Ферми-Дирака, работающая с 64-битовыми числами. Вычисления прекращались, когда относительная разность на двух последовательных сетках составляла 10^{-12}. При этом относительная погрешность результата лежала в пределах $10^{-16} - 10^{-15}$. При $x \sim 0$ число интервалов последней сетки составляло $M = 32 - 64$ для всех рассмотренных k, в то время как квадратуры формул Гаусса-Кристоффеля в [9] требовали $600 - 1200$ вычислений функции, т.е. были примерно в 20 раз более трудоемкими. Разумеется, при возрастании x следует увеличивать число узлов; но даже при $x = 50$ было достаточно $M = 500 - 1000$.

4. Глобальные аппроксимации

4.1. Трехчленная аппроксимация. В [14, 15] были предложены трехчленные аппроксимации, выражающие функции любого индекса через $I_0(x)$, которая является элементарной функцией от x. Эти аппроксимации имеют следующий вид:

$$I_k(x) \approx \frac{y}{k+1}\{[\Gamma(k+2)]^{3/k}(1+c_1 y) + y^3\}^{k/3}, \quad y \equiv I_0(x) = \ln(1+e^x). \tag{26}$$

Коэффициент c_1 подбирался из условия получения наименьшей относительной погрешности во всем диапазоне $-\infty < x < +\infty$. Опишем способ подбора этого коэффициента.

Сначала составлялась таблица значений искомой функции $I_k(x)$ на отрезке $-10 \le x \le 50$ на подробной сетке с шагом $\delta x = 0.01$. Эти значения функций вычислялись непосредственно по квадратурам с точностью не менее 4-5 верных знаков.

Затем произвольно выбиралось некоторое значение c_1. Для него на той же сетке проводилось вычисление по аппроксимирующей формуле (26), находилось отношение приближенных значений функции к табулированным значениям и строился график относительной погрешности. Эта погрешность была близка к нулю при $x = -10$ и $x = 50$, поскольку обе асимптотики формулы (26) правильны. На отрезке график этой погрешности имел единственный нуль и два экстремума разного знака. Модули экстремумов не были равны.

Затем коэффициент c_1 вручную менялся до тех пор, пока модули обоих экстремумов не оказывались практически одинаковыми. Это означало выполнение чебышевского альтернанса, т.е. давало минимальные значения относительной погрешности и оптимальные значения коэффициента.

Найденные коэффициенты приведены в табл.1 вместе с процентными погрешностями аппроксимаций. Очевидно, полученные формулы применимы во всем диапазоне $-\infty < x < +\infty$, причем они обеспечивают передачу главных членов и левой, и правой асимптотик.

4.2. Пятичленная аппроксимация. Можно построить формулы с большим числом свободных коэффициентов для диапазона $-\infty < x < +\infty$, также правильно передающие главные члены левой и правой асимптотик:

$$I_k(x) \approx \frac{y}{k+1}\{\Gamma(k+2)^{6/k}(1 + c_1 y + c_2 y^2) + c_3 y^4 + y^6\}^{k/6}. \qquad (27)$$

Коэффициенты c_1, c_2, c_3 выбирались из различных соображений. Это давало 3 группы формул.

Таблица 1. Коэффициенты и погрешности формулы (26).

k	c	$d_{\max}(\%)$
$-3/2$		
$-1/2$	1.614	0.66
$1/2$	1.17	0.78
1	1.01	1.6
$3/2$	0.87	2.3
2	0.77	3.2
$5/2$	0.67	4.0
3	0.60	4.8
$7/2$	0.54	5.6
4	0.48	6.4

Таблица 2. Коэффициенты и погрешности формулы (27); подгоночный c_2.

k	c_2	$d_{\max}(\%)$
$-3/2$		
$-1/2$	5.77	1.09
1/2	2.4	0.54
1	1.73	0.9
3/2	1.26	1.09
2	0.98	1.3
5/2	0.77	1.40
3	0.62	1.8
7/2	0.51	1.74
4	0.42	1.9

Две асимптотики. Зададим коэффициенты c_1 так, чтобы обеспечить второй член левой асимптотики (3), а коэффициентом c_3 передадим второй член правой асимптотики (6):

$$c_1 = 3\cdot(1-2^{-k})/k, \qquad c_3 = \pi^2(k+1). \tag{28}$$

Коэффициент c_2 подберем из условия чебышевского альтернанса, как было описано в п.4.1. Найденные коэффициенты и их погрешности приведены в табл. 2.

Правая асимптотика. Передача второго члена правой асимптотики достаточно важна для физических приложений. Поэтому выберем c_3 согласно (28), а коэффициенты c_1 и c_2 подберем из условия выполнения чебышевского альтернанса. В этом случае график относительной погрешности будет иметь два нуля и три экстремума с чередующимися знаками. Ручной подбор оптимальных коэффициентов при этом сравнительно трудоемок. Однако это одноразовая работа, и проще подобрать коэффициенты вручную, чем строить компьютерный алгоритм их вычисления.

Подобранные коэффициенты и соответствующие процентные погрешности приведены в табл.3.

Наилучшая точность. Пренебрежем передачей вторых членов обеих асимптотик и подберем коэффициенты c_1, c_2, c_3 из условия чебышевского альтернанса. Здесь график относительной погрешности имеет три нуля и четыре экстремума. Подбор оптимальных коэффициентов также производился вручную. Найденные коэффициенты и соответствующие процентные погрешности приведены в табл.4.

Таблица 3. Коэффициенты и погрешности (27); подгоночные c_1, c_2.

k	c_1	c_2	$d_{\max}(\%)$
–3/2			
–1/2	1.07	7.27	0.76
1/2	1.23	2.83	0.32
1	1.15	1.99	0.45
3/2	1.03	1.47	0.58
2	0.93	1.11	0.60
5/2	0.83	0.86	0.68
3	0.75	0.69	0.70
7/2	0.67	0.56	0.74
4	0.60	0.47	0.80

Таблица 4. Коэффициенты и погрешности (27); подгоночные c_1, c_2, c_3.

k	c_1	c_2	c_3	$d_{\max}(\%)$
–3/2				
–1/2	1.846	5.430	7.166	0.28
1/2	1.44	2.47	16.58	0.14
1	1.28	1.78	21.50	0.20
3/2	1.14	1.32	26.60	0.28
2	0.99	1.02	31.42	0.30
5/2	0.87	0.801	36.513	0.41
3	0.78	0.65	41.45	0.45
7/2	0.702	0.528	46.430	0.47
4	0.63	0.44	51.59	0.50

Построенные здесь аппроксимации очень удобны для достаточно хороших оценок во многих физических приложениях.

5. Заключение

1. Для функций Ферми-Дирака полуцелых индексов предложены экономичные квадратурные формулы, позволяющие вычислять значения этих функций непосредственно через интегральные определения с точностью до ошибок округления компьютера (14-15 десятичных знаков при 64-битовых вычислениях).

2. Исследованы свойства предложенных квадратурных формул. Показано, что они имеют не степенную, а гораздо более быструю экспоненциальную сходимость. Найдена связь скорости сходимости с полюсами подынтегрального выражения в комплексной плоскости.

3. Предложены несложные аппроксимации функций Ферми-Дирака. Они неограниченно дифференцируемы, имеют правильные асимптотики как при $x \to -\infty$, так и при $x \to +\infty$ и обеспечивают относительную погрешность от нескольких процентов до 0.1 % во всем диапазоне $-\infty < x < +\infty$.

СПИСОК ЛИТЕРАТУРЫ

1. *E.C. Stoner, J. McDougall.* The computation of Fermi-Dirac functions // Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences, 1938, 237(773):67–104.
2. *Jr.H.C. Thacher, W.J. Cody.* Rational chebyshev approximations for Fermi-Dirac integrals of orders -1/2, 1/2 and 3/2 // Mathematics of Computation, 1967, p.30–407.
3. *R.B. Dingle.* Asymptotic Expansions: Their Derivation and Interpretation. – London: Academic Press, 1973, 521p.
4. *P.V. Halen and D.L. Pulfrey*. Accurate, short series approximations to Fermi-Dirac integrals of order -1/2, 1/2, 1, 3/2, 2, 5/2, 3, and 7/2, // J.Appl. Phys., 1985, v.57, p.5271-5274.
5. *L.D. Cloutman.* Numerical evaluation of the Fermi-Dirac integrals // The Astrophysical Journal Supplement Series, 1989, 71:677.
6. *M. Goano.* Algorithm 745: computation of the complete and incomplete fermi-dirac integral. ACM Transactions on Mathematical Software (TOMS), 1995, 21(3):221–232.
7. *A.J. MacLeod.* Algorithm 779: Fermi-Dirac functions of order-1/2, 1/2, 3/2, 5/2 // ACM Transactions on Mathematical Software (TOMS), 1998, 24(1):1–12.
8. *Toshio Fukushima.* Precise and fast computation of Fermi-Dirac integral of integer and half integer order by piecewise minimax rational approximation // Applied Mathematics and Computation, 2015, v.259, Issue C, p.708–729.
9. *Н.Н. Калиткин, С.А. Колганов.* Прецизионные аппроксимации функций Ферми–Дирака целого индекса // Матем. моделирование, 2016, 28:3, 23–32.
   N.N. Kalitkin, S.A. Kolganov. Precision approximations for Fermi–Dirac functions of the integer index // Mathematical Models and Computer Simulations, 2016, v.8, №6, p.607-614.
10. *О.Н. Королева, А.В. Мажукин, В.И. Мажукин, П.В. Бреславский.* Аналитическая аппроксимация интегралов Ферми-Дирака полуцелых и целых порядков // Математическое моделирование, 2016, т.28, №11, с.55-63.
    O.N. Koroleva, A.V. Mazhukin, V.I. Mazhukin, P.V. Breslavskii. Analiticheskaia approksimatsiia integralov Fermi-Diraka polutselykh i tselykh poriadkov // Matematicheskoe modelirovanie, 2016, t.28, №11, s.55-63.
11. *И.С. Градштейн, И.М. Рыжик* Таблицы интегралов, сумм, рядов и произведений. 4-е издание. – М.: Физматгиз, 1963, 1100с.
    I.S. Gradshtein, I.M. Ryzhik. Tablitsy integralov, summ, riadov i proizvedenii. 4-e izdanie. – M.: Fizmatgiz, 1963, 1100s.
12. *А.А. Белов.* О коэффициентах квадратурных формул Эйлера–Маклорена // Матем. моделирование, 2013, 25:6, p.72–79.
    A.A. Belov. Coefficients of Euler-Maclaurin formulas for numerical integration // Mathematical Models and Computer Simulations, 2014, v.6, №1, p.32–37.
13. *Н.Н. Калиткин, Е.А. Альшина*. Численные методы. Кн.1. Численный анализ. – М.: Академия, 2013.
    N.N. Kalitkin, E.A. Alshina. Chislennye metody. Kn.1. Chislennyi analiz. – M.: Akademiia, 2013.
14. *Н.Н. Калиткин, И.В. Ритус.* Гладкие аппроксимации функций Ферми-Дирака // Журнал вычислительной математики и математической физики, 1986, т.26, №3, с.461-465.
    N.N. Kalitkin, I.V. Ritus. Gladkie approksimatsii funktsii Fermi-Diraka // Zhurnal vychislitelnoi matematiki i matematicheskoi fiziki, 1986, t.26, №3, s.461-465.
15. *Н.Н. Калиткин, И.В Ритус.* Гладкие аппроксимации функций Ферми-Дирака. – М., Инст. прикл. мат. АН СССР, 1981, Препринт №72, 9с.
    N.N. Kalitkin, I.V Ritus. Gladkie approksimatsii funktsii Fermi-Diraka. – M.: Inst. prikl. mat. AN SSSR, 1981, Preprint №72, 9s.

Поступила в редакцию 08.11.2016.