ISSN 2070-0482, Mathematical Models and Computer Simulations, 2017, Vol. 9, No. 5, pp. 554–560. © Pleiades Publishing, Ltd., 2017.
Original Russian Text © N.N. Kalitkin, S.A. Kolganov, 2017, published in Matematicheskoe Modelirovanie, 2017, Vol. 29, No. 3, pp. 42–50.

Refinements of Precision Approximations of Fermi-Dirak Functions of Integer Indices

N. N. Kalitkin[a], * and S. A. Kolganov[b], **

[a] *Keldysh Institute of Applied Mathematics, Russian Academy of Sciences, Moscow, Russia*

[b] *National Research University of Electronic Technology, Zelenograd, Russia*

**e-mail: kalitkin@imamod.ru*

***e-mail: mkandds2012@gmail.com*

Received May 4, 2016

Abstract—Fermi-Dirac functions of integer indices are broadly used in problems of electronic transport in dense substances. Polynomial approximations are constructed for their fast computation. Such coefficients are found for functions of index 1, 2, and 3, which provide an error ratio of about 2×10^{-16} with nine free parameters. In this work, we use the *boost::multiprecision* library of *C++*, which allows us to compute with any arbitrary number of digits. The precision of previously obtained relations is improved to $\sim 5 \times 10^{-18}$ and the same relation is constructed for the index $k = 4$. Also, it is shown that simple global relation consisting of a few parameters reasonably describe the order of the value of the functions for all values of the independent variable and can be used for estimations.

Keywords: Fermi–Dirac functions, precision approximations, rational approximation, estimated global approximations

DOI: 10.1134/S2070048217050052

1. FERMI–DIRAC FUNCTIONS

Fermi–Dirac functions arise in quantum mechanics problems when the properties of matter caused by the behavior of electrons (or other fermions) are described. If the density is sufficiently high or the temperature is sufficiently low, then they are various moments of the Fermi distribution and they are reduced to the following integrals:

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

The values of the index k are integers if the moment is odd and are half-integers if the moment is even. In problems of physics, only integer and are half-integer indices are needed, while the mathematical theory of those functions considers arbitrary values of k.

Various physical quantities correspond to different indices. For the density of electrons, we have $k = 1/2$, for kinetic energy, we have $k = 3/2$, for electron conductivity, we have $k = 1$, for electron heat-conductivity, we have $k = 2$, and, for electron viscosity, we have $k = 3$. In several applications, lower indices arise (e.g., $k = -1/2$ and $k = -3/2$), but higher indices have not been required to date.

Many papers are devoted to approximations of the Fermi–Dirac functions of half-integer indices (see [1–7]). For the functions of the integer indices $k = 1, 2,$ and 3, precision approximations are constructed and methods to construct them have only recently been developed [8]. For a substantial part of the range of values of the independent variable, the relative precision of $\sim 2 \times 10^{-16}$ has been obtained, which corresponds to the error of 64-bit computations. However, the detailed analysis shows that the error begins to increase for $x < -5$ and for $x \sim -10$ increases by ~4 orders. That is why the work was repeated with the use of the *boost::multiprecision* library of the *C++* language, providing computations with any number of digits. Also, a special method to compute the auxiliary independent variable was developed. Thus, we suc-

ceed in reducing the error to $\sim 5\times 10^{-18}$ in the whole range of values of x. Additionally, an approximation for the index $k = 4$ was constructed.

Also, we investigated the error of the basic smooth approximations describing the entire range $-\infty < x < +\infty$ by a unital smooth relation with only a few coefficients. Such approximations provide rough estimates of Fermi–Dirac functions and they are applicable for arbitrary noninteger indices.

2. AUXILIARY INDEPENDENT VARIABLE

Integrals (1) are not computed in elementary functions. Their asymptotic behavior qualitatively differs at the ends of the real line: $I_k(x) \approx \Gamma(k+1)e^x$ as $x \to -\infty$, while $I_k(x) \approx x^{k+1}/(k+1)$ as $x \to +\infty$. Therefore, it is hard to construct good approximations. However, a promising approach is proposed in [8]. The case where $k = 0$ is the only case, where integral (1) is computed exactly:

$$I_0(x) \equiv y(x) = \ln\left(1 + e^x\right), \quad 0 < y < +\infty. \tag{2}$$

It is easy to see that $I_k \approx \Gamma(k+1)\cdot y$ as $y \to 0$, while $I_k \approx x^{k+1}/(k+1)$ as $y \to +\infty$. It turns out that it has a asymptotic behavior at both ends of the range for y; i.e., it is a one-type asymptotic behavior. Hence, it is convenient to treat y as an auxiliary independent variable and compute functions of other indices as functions of this independent variable. This cardinally facilitates the construction of good approximations.

However, computations carried out on the limit of the accuracy of a computer have the following peculiarity. If $x \to -\infty$, then $e^x \to 0$, and, computing $\ln(1+e^x)$ by standard computer procedures, we can lose too many significant digits. Therefore, it is reasonable to use the following special computational procedure:

$$y(x) = \sum_{n=0}^{\infty} \frac{2}{2n+1}\left(e^x/(2+e^x)\right)^{2n+1} \quad \text{for} \quad x \le 0; \tag{3}$$

the series needs to be truncated to the number of terms needed to achieve the given accuracy, and it is reasonable to use Horner's method for summing. For $x > 0$, it is reasonable to use standard computer procedures to compute $y(x)$. This computing method allows us to eliminate the growth in errors arising in [8] for $x < -5$.

3. PRECISE APPROXIMATION

In [8], the following approximating relations, including relations of polynomials of y, are proposed for the range $-\infty < x \le 0$ ($0 \le y \le \ln 2$, respectively):

$$I_k(x) \approx \Gamma(k+1)\, y \left(\frac{\sum_{n=0}^{N+1} a_n y^n}{\sum_{m=0}^{N} b_m y^m} \right)^k, \quad a_0 = 1, \quad b_0 = 1, \quad x \le 0. \tag{4}$$

To compute the coefficients a_n and b_m under the assumption that N is selected a priori, we develop a special algorithm approximately ensuring the Chebyshev alternance for extremums of the relative error (it is obvious that the relative error instead of the absolute one is important for the approximation of functions (1)). It yields the relative error $d \sim 10^{-6}$ for $N = 1$, $d \sim 10^{-10}$ for $N = 2$, and $d \sim 10^{-14}$ for $N = 3$. We would expect to obtain $d \sim 10^{-18}$ for $N = 4$; however, computations with 64-bit numbers cannot provide any analysis errors of $d \sim 10^{-16}$.

To complete this task, we executed a computation with 25 digits, using (additionally) relation (3) for the auxiliary independent variable (this also changes the far digits of the coefficients). The final values of the coefficients for $N = 4$ are presented in Table 1 with 18 digits after the comma. The coefficients for the functions of index $k = 4$ are also included in this table. These coefficients provide the relative error

Table 1. Coefficients a_n and b_m for $N = 4$

a_n, b_m	k			
	1	2	3	4
a_1	0.271511313821436278	0.226381636434069856	0.158348214538045596	0.056014879123090215
a_2	0.056266123806058763	0.053368433557479886	0.046064514990930811	0.035111795789180087
a_3	0.006742074046934569	0.006290475634079521	0.004886137910884147	0.002183438694367233
a_4	0.000516950515533321	0.000502322827445298	0.000433673330597152	0.000246486152552295
a_5	0.000019477183676577	0.000018937967508806	0.000017343561379589	0.000009222817788667
b_1	0.021511313821435284	0.038881636434069113	0.012514881204710761	−0.061172620876911286
b_2	0.023110517572972142	0.024304399874277445	0.026669340700092963	0.027996854281614683
b_3	0.000366908157736541	0.000629098532643319	0.000328543109454736	−0.000751214829430754
b_4	0.000061042440873272	0.000065701816194546	0.000082091078789006	0.000086068074714292

$d \sim 5\times10^{-18}$ for $x \le 0$. an example of the profile of the relative error is presented in Fig. 1. It is a smooth curve such that the absolute values of its extremums are approximately equal to each other, while their signs alternate.

We see that the coefficients rapidly decrease as the number increases; this is illustrated by the rapid convergence of the series in the numerator and denominator; i.e., the approximation kind has been selected successfully. Note that, in Table 1, the coefficients b_1 and b_3 for the function of index $k = 4$ are negative. Usually, negative coefficients are not desirable because they might lead to the vanishing of the denominator (if b_m is negative) or the numerator (if a_n is negative). However, the modulus of the obtained coefficients is sufficiently small to ensure that the denominator always remains positive. Thus, the negativity is negligible.

Recall that if $x > 0$, then the computation of the functions of the integer index is reduced to the computation of the functions of the negative independent variable according to the following relation:

$$I_k(x) = \frac{x^{k+1}}{k+1}\left[1 + \frac{\Gamma(k+2)}{\Gamma(k)}\frac{\pi^2}{6}x^{-2} + \frac{\Gamma(k+2)}{\Gamma(k-2)}\frac{7\pi^4}{360}x^{-4} + \ldots\right] + (-1)^k I_k(-x), \quad k = 0,1,2,\ldots . \tag{5}$$

The number of terms in the square brackets is finite. It is selected such that the lowest power of the polynomial multiplied by x^{k+1} is equal to 1 if k is even and is equal to 0 if k is odd. Note that $I_1(0) = \pi^2/12$ and $I_3(0) = 7\pi^4/120$.

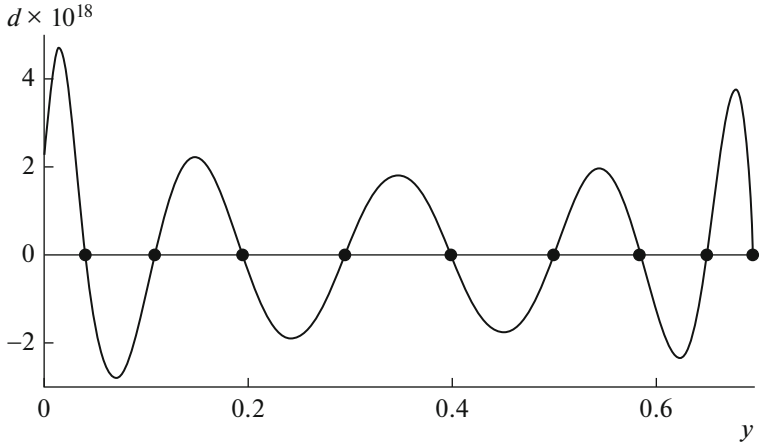

Fig. 1. Error profile for case where $k = 2$.

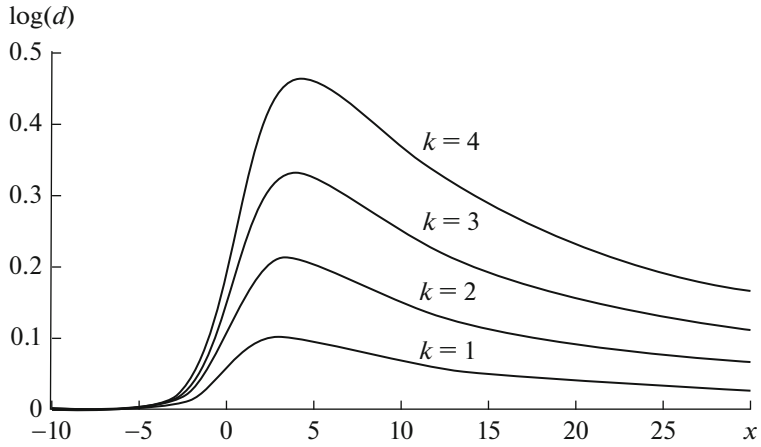

Fig. 2. Error profiles of relations (6); values of k.

The function of the negative independent variable on the right-hand side is to be computed according to approximation (4). Since $I_k(x)$ increases rapidly as the independent variable increases, it follows that the relative error for $x > 0$ is much smaller than for $x < 0$; therefore, the coefficients from Table 1 provide a relative accuracy of not worse than $\sim 5\times 10^{-18}$ everywhere.

Computing the coefficients according to the algorithm of [8], we use the requirement that the approximation of $I_k(x)$ be exact for $x = 0$. This ensures that the passage to the positive independent variable with respect to (5) is continuous at the point $x = 0$.

Remark. The use of the library *boost::multiprecision* of *C++* has the following peculiarity. The number of coefficients (including binary ones) can be used in the relations. The integer coefficients are translated into the binary code; hence, it is not necessary to specify whether they are represented with single precision or double precision. However, in *boost::multiprecision* in *C++*, they are treated as 64-bit numbers by default. Therefore, to obtain an increased number of digits, we have to describe all the coefficients (including the integers) as numbers of this system. Otherwise, only results with 16 reliable digits are obtained.

4. GLOBAL APPROXIMATIONS

For simple estimation computations, simple relations describing $I_k(x)$ continuously and smoothly in the whole range $-\infty < x < +\infty$ are useful for physicists. To be applied successfully, such relations have to describe the principal terms for both asymptotics of the function. This physically ensures that the fermions are transferred to the ideal gas for high temperatures and a completely degenerate gas for low ones. Let us present several simple relations of this kind.

In [8], the two-term relation is proposed. Its more convenient form is as follows:

$$I_k(x) \approx \frac{y}{(k+1)}\{[\Gamma(k+2)]^{1/k} + y\}^k. \tag{6}$$

This relation does not contain any adjustable parameters. In Fig. 2, errors of the two-term relation in form (6) are displayed for various indices. We see that if $k > 0$, then this relation overstates the value of the function everywhere. Its error is substantial. The largest increase in the error is by a factor of 1.3 for $k = 1$ and a factor of 2.9 for $k = 4$. Therefore, this relation is suitable only for very rough estimates.

In [9, 10], the three-term relation is proposed. Its more convenient form is as follows:

$$I_k(x) \approx \frac{y}{k+1}\{[\Gamma(k+2)]^{3/k}(1+cy) + y^3\}^{k/3}; \tag{7}$$

the coefficient c is selected to minimize the relative error of the approximation. For $y \to 0$, the relation expresses the principal term of the asymptotics on the left-hand side and the order of vanishing of the next

Table 2. Coefficients and errors of relation (7)

k	c	$d_{\max}(\%)$
−1/2	1.62	0.7
1/2	1.18	0.8
1	1.01	1.6
3/2	0.87	2.6
2	0.77	3.2
3	0.60	4.8
4	0.48	6.4

Table 3. Coefficients and errors of variant (a) of relation (8)

k	c_2	$d_{\max}(\%)$
1	1.73	0.9
2	0.98	1.3
3	0.62	1.8
4	0.42	1.9

term of the expansion (but not its exact coefficient) exactly. For $y \to +\infty$, it expresses the principal term of the asymptotics and the vanishing order of the next term exactly. This improves the qualitative behavior of the approximation.

Relation (7) is substantially better: using the adjustable parameter c, we can make the profile error alternate. Selecting c from the Chebyshev alternance condition, we minimize the error. The corresponding optimum values of c are provided in Table 2; also, it contains the values for the half-integer indices k taken from [9, 10]. We see that the coefficients monotonically decrease as k increases. The errors (in percentage) of the obtained relations are also provided in Table 2; they increase as the absolute value of the difference between k and zero increases. These errors are substantially smaller than the ones for relation (6). We see that relation (7) can be recommended for good estimation computations.

Figure 3 displays the error profile of relation (7) for $k = 2$. The profile shape confirms that the Chebyshev alternance is fulfilled.

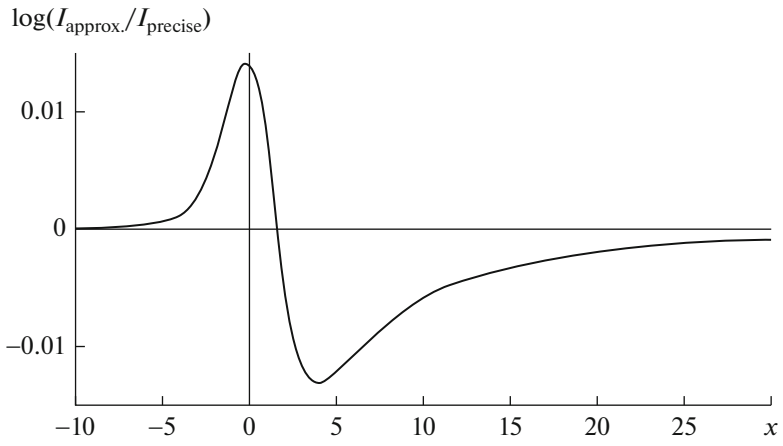

Fig. 3. Error profile of relation (7) for case where $k = 2$.

Table 4. Coefficients and errors of variant (b) of relation (8)

k	c_1	c_2	$d_{\max}(\%)$
1	1.15	1.99	0.45
2	0.93	1.11	0.60
3	0.75	0.69	0.70
4	0.60	0.47	0.80

Table 5. Coefficients and errors of variant (c) of relation (8)

k	c_1	c_2	c_3	$d_{\max}(\%)$
1	1.28	1.78	21.50	0.20
2	0.99	1.02	31.42	0.30
3	0.78	0.65	41.45	0.45
4	0.63	0.44	51.59	0.50

The five-term relation yields better results. We represent it in a form such that the terms from the first to the third terms appear as a power expansion with respect to y as $y \to 0$, while the third to the fifth terms appear as a power expansion with respect to y^{-2} as $y \to \infty$:

$$I_k(x) \approx \frac{y}{k+1}\{\Gamma(k+2)^{6/k}(1 + c_1 y + c_2 y^2) + c_3 y^4 + y^6\}^{k/6}. \tag{8}$$

To select coefficients, we can take various reasons into account. Consider three variants of this relation.

(a) The coefficients c_1 and c_3 are selected to correctly express the asymptotics of the second terms of the left-hand and right-hand sides:

$$c_1 = 3 \times (1 - 2^{-k})/k, \quad c_3 = \pi^2(k+1); \tag{9}$$

the coefficient c_2 remains adjustable. This relation provides the description of the limits for the ideal and degenerate Fermi gases and the closest correction included once the ideality decreases or the degeneration is taken off. The coefficient c_3 is especially important because, using it, we can describe the heat properties of an almost degenerate gas (e.g., the heat capacity or the conductivity under low temperatures).

One adjustable coefficient c_2 can provide only one zero of the error. The coefficient c_2 is selected to satisfy the Chebyshev alternance condition. Then we can minimize the error. Table 3 presents the optimum values of c_2 and the errors of the obtained relations (in percentage). The obtained errors are between 1% and 2%, which is thrice as good as the ones for relation (7). Thus, this relation is preferable for estimation computations.

(b) For high temperatures, it is usually not as important to express the second term of the asymptotics. Therefore, the coefficients c_1 and c_2 can be treated as free parameters, while the coefficient c_3 can be preserved according to (9). This allows us to include the second zero in the error graph and to select the coefficients c_1 and c_2 to satisfy the Chebyshev alternance condition. The values of these coefficients and the corresponding values of the largest errors are given in Table 4. In this case, we see that the accuracy is between 0.5 and 1%, which is twice as good as in Table 3.

(c) Suppose that the least error is important, while the second term of the asymptotics on the right-hand side is negligible. Then all three coefficients c_1, c_2, and c_3 can be used as adjustable ones and they can be selected to satisfy the Chebyshev alternance condition. Then the error graph has three zeros. In this case the error graph will have three zeros. The corresponding values of the coefficients and errors are provided in Table 5. We see that the errors of between 0.2 and 0.5% are half the size of those in Table 4. For those simple relations, this level of accuracy is impressive.

Note that the selected values of c_3 are close to the theoretical values provided by (9). Thus, the approximation has been selected successfully.

It appears that all the proposed relations are also suitable for half-integer indices up to $k = -3/2$.

ACKNOWLEDGMENTS

This work was supported by Russian Science Foundation, project no. 16-11-10001.

REFERENCES

1. E. C. Stoner and J. McDougall, "The computation of Fermi-Dirac functions," Phil. Trans. R. Soc. London, Ser. A **237** (773), 67–104 (1938).
2. H. C. Thacher, Jr. and W. J. Cody, "Rational Chebyshev approximations for Fermi-Dirac integrals of orders -1/2, 1/2 and 3/2," Math. Comput. **21**, 30–40 (1967).
3. M. Lundstrom and R. Kim, "Notes on Fermi-Dirac integrals," arXiv:0811.0116 (2008).
4. N. N. Kalitkin, "About computation of functions the Fermi-Dirac," USSR Comput. Math. Math. Phys. **8**, 173–175 (1968).
5. L. D. Cloutman, "Numerical evaluation of the Fermi-Dirac integrals," Astrophys. J. Suppl. Ser. **71**, 677 (1989).
6. M. Goano, "Algorithm 745: computation of the complete and incomplete Fermi-Dirac integral," ACM Trans. Math. Software **21**, 221–232 (1995).
7. A. J. MacLeod, "Algorithm 779: Fermi-Dirac functions of order -1/2, 1/2, 3/2, 5/2," ACM Trans. Math. Software **24**, 1–12 (1998).
8. N. N. Kalitkin and S. A. Kolganov, "Prescision approximations for Fermi–Dirak functions of integer index," Mat. Model. **28** (3), 23–32 (2016).
9. N. N. Kalitkin and I. V. Ritus, "Smooth approximations of functions the Fermi-Dirac," KIAM Preprint No. 72 (Keldysh Inst. Appl. Math., Moscow, 1981).
10. N. N. Kalitkin and I. V. Ritus, "Smooth approximation of Fermi-Dirac functions," USSR Comput. Math. Math. Phys. **26**, 87–89 (1986).

Translated by A. Muravnik