ISSN 2070-0482, Mathematical Models and Computer Simulations, 2016, Vol. 8, No. 6, pp. 607–614. © Pleiades Publishing, Ltd., 2016.
Original Russian Text © N.N. Kalitkin, S.A. Kolganov, 2016, published in Matematicheskoe Modelirovanie, 2016, Vol. 28, No. 3, pp. 23–32.

Precision Approximations for Fermi–Dirac Functions of the Integer Index

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

[a] *Keldysh Institute of Applied Mathematics, Russian Academy of Sciences, Moscow, 125047 Russia*

[b] *National Research University of Electronic Technology, Zelenograd, 124498 Russia*

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

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

Received April 21, 2015

Abstract—The Fermi–Dirac functions of the integer index are widely used in electron transport problems in dense substances. Polynomial approximations are constructed for their quick calculation. A simple algorithm yielding the coefficients of such approximations based on the interpolation with a special linear-trigonometric grid is developed. It is demonstrated that this grid gives almost optimal results. For the functions of indices 1, 2, and 3, the coefficients of such interpolations ensuring the relative error of 2×10^{-16} under 9 free parameters are obtained.

Keywords: Fermi–Dirac functions, precision approximations, rational approximation, linear-trigonometric grid

DOI: 10.1134/S2070048216060090

1. THE FERMI–DIRAC FUNCTIONS

1.1. Applied Problems

The Fermi–Dirac functions arise in quantum mechanics problems for the description of the properties of a substance caused by the behavior of electrons or other fermions. Under sufficiently high densities or low temperatures, the distribution function of the electrons has the form $\{1 + \exp[(0.5p^2 - \mu)/T]\}^{-1}$, where p is the electron momentum, μ denotes the chemical potential, and T specifies temperature (in the atomic unit system where the electronic mass is 1). When solving quantum mechanics problems, researchers use different moments of the Fermi distribution that represent the convolutions of different powers of the momentum p and this distribution. In such convolutions, the role of the integration variable is played by $t = 0.5p^2/T$. Then the moments acquire the form

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

In this formula, $x = \mu/T$. Index k takes integer values for odd p and half-integer values for the even ones. The physical problems operate only the integer and half-integer indices, although the mathematical theory of these functions considers arbitrary k.

Let us mention the physical quantities corresponding to different indices. The electron density corresponds to $k = 1/2$; the kinetic energy, to $k = 3/2$; the electron conduction, to $k = 1$; the electron heat conductivity, to $k = 2$; and the electron viscosity, to $k = 3$. A series or applications leads to smaller indices (e.g., $k = -1/2$ or even $k = -3/2$), but higher indices have not been required to date.

A vast volume of literature is dedicated to the approximation of the Fermi–Dirac functions of the half-integer index. Unfortunately, no satisfactory approximations have been proposed for the Fermi–Dirac functions of the integer indices. The present paper aims at filling this gap.

1.2. Basic Properties

The theoretical properties of the Fermi–Dirac functions were studied in detail in [1–7]. Recall the properties required for further exposition. For arbitrary indices k,

$$I'_k(x) = kI_{k-1}(x). \tag{2}$$

For the ***integer*** indices k, the functions of the positive and negative arguments are connected by

$$I_k(x) = (-1)^k I_k(-x) + P_{k+1}(x), \quad x \geq 0, \tag{3}$$

where $P_{k+1}(x)$ denotes a polynomial of degree $k + 1$. For any k and negative arguments, we have the expansion

$$I_k(x) = \Gamma(k+1)\sum_{n=1}^{\infty}(-1)^{n-1}\frac{e^{nx}}{n^{k+1}}; \tag{4}$$

this series converges absolutely (yet, nonuniformly) if $x < 0$. As $x \to +\infty$, there exists an asymptotically convergent series having the dominant term $I_k(x) \to x^{(k+1)}/(k+1)$ and the next term of the infinitesimal order $O(x^{-2})$.

Only for index $k = 0$, integral (1) is taken in terms of elementary functions:

$$I_0(x) = \ln(1 + e^x). \tag{5}$$

It appears directly from (5) that

$$I_0(x) = x + I_0(-x), \quad x \geq 0; \tag{6}$$

this expression is a special case of (3). Hence, $P_1(x) = x$.

1.3. Integer Index

In what follows, we demonstrate that the calculation and approximation of functions (1) is much simpler for the negative argument x than for the positive one. Therefore, for the functions of integer index, it suffices to construct approximations only for $x \leq 0$, taking advantage of expression (3) if $x \geq 0$. Substituting (3) into (2) yields the relationship between the polynomials

$$P'_{k+1}(x) = kP_k(x). \tag{7}$$

Since $P_1(x) = x$, the polynomials of higher degrees can be obtained by the successive integration of (7), which results in unknown integration constants. They are defined from the coincidence of the right- and left-hand sides of (3) at $x = 0$. Hence, the required functions satisfy

$$I_1(x) = \frac{x^2}{2} + 2I_1(0) - I_1(-x), \quad x \geq 0; \tag{8}$$

$$I_2(x) = \frac{x^3}{3} + 4I_1(0)x + I_2(-x), \quad x \geq 0; \tag{9}$$

$$I_3(x) = x^4/4 + 6I_1(0)x^2 + 2I_3(0) - I_3(-x), \quad x \geq 0. \tag{10}$$

Clearly, as $x \to +\infty$, the dominant term of formulas (8)–(10) is $x^{(k+1)}/(k+1)$, which matches the above asymptotics.

2. DIRECT CALCULATION OF FERMI–DIRAC FUNCTIONS

2.1. Precision

Nowadays, 64-bit computing is most widespread. In the floating point arithmetic, the relative round-off error is about 10^{-16}. Therefore, our ultimate aim lies in constructing approximations where the relative error of the Fermi–Dirac function calculation is as close to 10^{-16} as possible.

Obviously, to construct such approximations and verify their precision, one should employ the values of the Fermi–Dirac functions at some reference points that are calculated slightly more precisely. It has been impossible for us to choose 128-bit computing, as only the MATLAB package with some embedded C++ code have been available. With a careful approach, this yields up to 16 reliable decimal digits. Let us describe the computational procedure for the basic values of function (1) with $x \leq 0$, noting the key details that guarantee the required precision.

2.2. The Horner Scheme

Series (4) converges rapidly if $x \ll -1$. The convergence rate becomes increasingly worse for higher x. Since this series is absolutely convergent and alternating, the error does not exceed the first neglected term. Hence, for achieving the relative precision ε, the number of terms N in the sum must satisfy the condition

$$N|x| > -\ln\left[\varepsilon(N+1)^{k+1}\right], \quad \varepsilon = 10^{-17}, \quad x < 0. \tag{11}$$

Clearly, as $x \to -0$, the required number of terms in (4) grows rapidly. Series (4) has been applied for $x \leq -0.1$. In the case $x = -0.1$, the necessary number of terms is $N = 260$ for $k = 1$, $N = 214$ for $k = 2$, and $N = 165$ for $k = 3$.

We have verified the convergence of the series while increasing the number of terms. As this series is alternating, the convergence demonstrates a double character—17 decimal digits near the calculated number N; and it remained invariable and significant with the further growth of N.

Note an important aspect. The required precision is at the limit of the round-off errors. With the direct summation of series (4), $n = 1, 2, \ldots$, the last small terms become less than the round-off errors, and are practically not added. Such terms have to be summed in reverse order from the small terms to the large ones.

In practice, the calculations from the last terms to the first ones are implemented using Horner's scheme

$$I_k(x) = \Gamma(k+1)e^x(1/1^{(k+1)} - e^x(1/2^{(k+1)} - e^x(1/3^{(k+1)} - e^x(\ldots - e^x 1/N^{(k+1)})))). \tag{12}$$

The scheme (12) calls for the preliminary definition of N from (11).

2.3. Quadratures

For $x > -0.1$, the number of terms in the scheme (12) becomes unacceptably large. Therefore, on the interval $-0.1 < x < 0$, functions (1) are calculated using quadratures. Thus, integral (1) undergoes an appropriate transformation with the infinite integration interval being replaced by a finite one:

$$I_k(x) \approx e^x \int_0^T \frac{t^k dt}{\exp(x) + \exp(t)}. \tag{13}$$

To achieve the required precision on the interval $-0.1 < x < 0$, it is necessary to choose $T = 50$ for $k = 0$, $T = 60$ for $k = 1$, $T = 75$ for $k = 2$, and $T = 100$ for $k = 3$.

Integral (13) has been calculated by the following quadrature procedure. The segment $0 \leq t \leq T$ was splitted into N equal intervals. The Gaussian quadrature formula based on zeroes of the Legendre polynomials of degree 5 was applied on each interval. Hence, the error of this grid's Gaussian quadrature is $O(N^{-10})$. Choose a small initial partition: $N = 16$. Then N is doubled sequentially, applying the Richardson method to the resulting values (see [8, 9]). The required precision $\varepsilon = 10^{-17}$ is achieved for $N = 128$.

There is a relevant detail here, as well. The intervals must be summed not from the left end (small t) but from the right one ($t \approx T$), since the right-hand terms are the smallest. With this precaution, the round-off error does not exceed 2×10^{-16} (the level 10^{-17} appears unachievable). At the same time, summation from the left end increases the round-off error to 10^{-15} and even further.

We have performed an additional verification of the precision. For $k = 0$, the precise expression (5) is known; the quadrature formula reproduces these values with 17 decimal digits. For $k = 1$, 2, and 3, the quadrature calculations are compared with scheme (12) with $x = -0.1$; the difference does not exceed 2×10^{-16}.

Conclusions. According to the aforesaid, the two described methods allow calculating $I_k(x)$, $x < 0$, with the relative precision of at least 2×10^{-16}. Since $I_k(x)$ increase monotonically for $k \geq 0$, their calculation using formulas (8)–(10) for $x > 0$ ensures higher relative precision (the greater x the better is the precision).

3. APPROXIMATIONS OF THE FUNCTIONS OF THE INTEGER INDEX

3.1. Type of Approximation

We use the following considerations. As $x \to -\infty$, the main term of the asymptotics is $I_k(x) \approx \Gamma(k+1)e^x$. As $x \to +\infty$, the asymptotics demonstrate not the exponential but power-type behavior, i.e., $I_k(x) \approx x^{k+1}/(k+1)$. Therefore, the variable x does not seem to be the most suitable argument of approximation.

Let us choose $y \equiv I_0(x) = \ln(1+e^x)$ as the argument. The asymptotics of this variable are $y \to e^x$ as $x \to -\infty$ and $y \to x$ as $x \to +\infty$. Such argument better fits the qualitative behavior of $I_k(x)$ on the whole interval $-\infty < x < +\infty$.

The polynomial approximation is successful only in some cases, despite its wide usage in the literature. Generally, the best results are yielded by the rational approximation, i.e., the one with the ratio of polynomials. Such an approximation reflects the different asymptotics of the source functions. In this paper, we adopt the following approximation:

$$I_k(x) \approx \Gamma(k+1)\,y\left(\frac{\sum\limits_{n=0}^{N+1} a_n y^n}{\sum\limits_{m=0}^{m=N} b_m y^m}\right)^k, \quad a_0 = 1, \quad b_0 = 1, \quad x \leq 0. \tag{14}$$

As $x \to -\infty$, approximation (14) exactly reproduces the first term of series (4). The expansion (14) with respect to the powers of e^x with $x \to -\infty$ will qualitatively resemble series (4). As $x \to +\infty$, the main term of expansion (14) is $I_k(x) \sim x^{k+1}$, but its factor differs from the exact one. However, even such a representation of the right asymptote improves the resulting precision of the approximation.

3.2. Calculation of Coefficients

As a rule, the calculation algorithms for the rational approximation coefficients are difficult in the case of minimizing some error norm. Here we confine ourselves to a simple heuristic algorithm yielding good results. Transform (14) into

$$w = \left[\frac{I_k(x)}{\Gamma(k+1)\,y}\right]^{1/k}, \quad w \approx \left(\frac{\sum\limits_{n=0}^{N+1} a_n y^n}{\sum\limits_{m=0}^{N} b_m y^m}\right). \tag{15}$$

Approximation (15) contains $(2N+1)$ free coefficients, namely, a_n, $1 \leq n \leq N+1$, and b_m, $1 \leq m \leq N$. We construct the approximation on the interval $-\infty < x \leq 0$, i.e., $0 < y \leq y_{\max} = \ln 2$. Choose $(2N+1)$ reference points $y_j : 0 < y_1 < y_2 < \ldots < y_{2N+1} = y_{\max}$. Using y_j, calculate the corresponding values x_j, $I_k(x_j)$, and w_j. Require that the approximate equality (15) becomes exact at the points y_j; i.e., we pose the interpolation problem with the chosen reference points (further called interpolation nodes). To find the free coefficients, it is necessary to solve a system of linear equations (omitted because obvious).

Note that, formally, one can choose $j = 0$ and set $y_0 = 0$. At this point, the approximate equality (15) becomes exact, as approximation (14) reproduces exactly the main term of the left asymptotics.

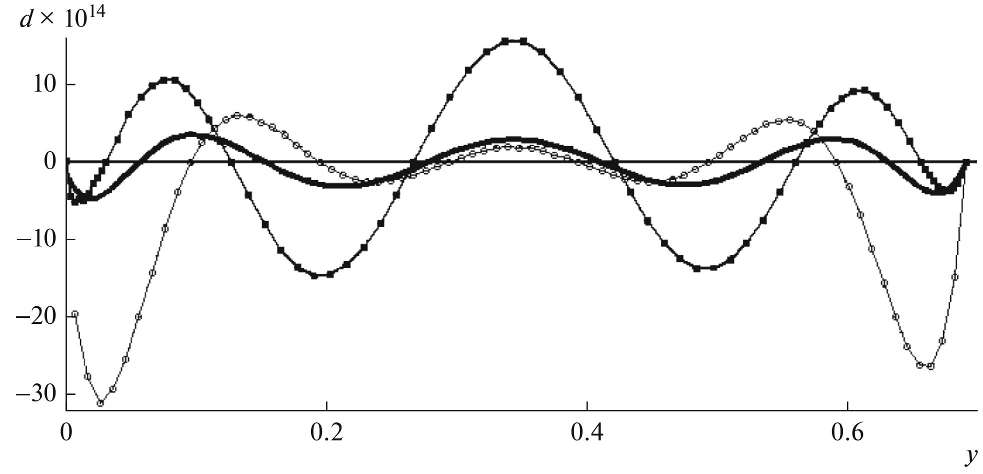

Fig. 1. Relative error d for $k = 2$ and $N = 3$. Interpolation nodes: bold line—linear trigonometric, —○— —linear, —●— —trigonometric.

3.3. Interpolation Nodes

The distribution of the chosen points strongly affects the quality of the interpolation; an inappropriate distribution can lead to negative coefficients a_n and b_m. This is dangerous, especially if the denominator or numerator vanishes within the required range of y. In the latter cases, the resulting approximation becomes absolutely unacceptable.

We have tested some ways of choosing the interpolation nodes. Let us illustrate them for $k = 2$ and $N = 3$ (which gives 7 free coefficients); for other k and N, the results were similar.

The elementary choice is the linear distribution

$$y_j = \frac{jy_{\max}}{2N+1}, \quad 0 \le j \le 2N+1; \tag{16}$$

recall that, formally, we can specify y_0, despite its noninvolvement in the calculations. For this case, Fig. 1 shows the relative error profile d. Clearly, the error vanishes at all the reference points of the interpolation, having the form of half-waves between them. The amplitudes of these half-waves are small in the middle, being increased manifold near the boundary points. This fact indicates that it is necessary to increase the distance between the interpolation nodes in the middle and reduce it near the boundaries of the segment.

The theory of approximation is well developed for polynomial, being the best in the C norm. In this theory, the distribution of the interpolation points is not calculated exactly. However, it is close to the distribution described by the trigonometric function

$$y_j = \frac{1}{2}y_{\max}\left(1-\cos\left(\frac{\pi j}{2N+1}\right)\right), \quad 0 \le j \le 2N+1. \tag{17}$$

The error profile d with the interpolation nodes (17) is also demonstrated by Fig. 1. Between the interpolation nodes, it has the form of half-waves with large amplitudes in the middle and small amplitudes near the boundary points. Therefore, for distribution (17), it is necessary to reduce the distance between the interpolation nodes in the middle and increase it near the boundaries of the segment.

Note, that this does not contradict the theoretical results for the Chebyshev approximations because they refer to the polynomial approximation, while we use the approximation by the rational functions.

For the interpolation nodes, it seems natural to design a distribution that is intermediate between the linear and trigonometric ones. Such a problem is common to the ultrafast iterative solution method for the systems of elliptic equations on a rectangular grid. Let us employ the linear-trigonometric distribution proposed in [10]:

$$y_j = \frac{y_{\max}}{2}\left[\frac{2\alpha j}{2N+1} + (1-\alpha)\left(1-\cos\left(\frac{\pi j}{2N+1}\right)\right)\right], \quad \alpha = \frac{\pi}{2+\pi}, \quad 0 \le j \le 2N+1. \tag{18}$$

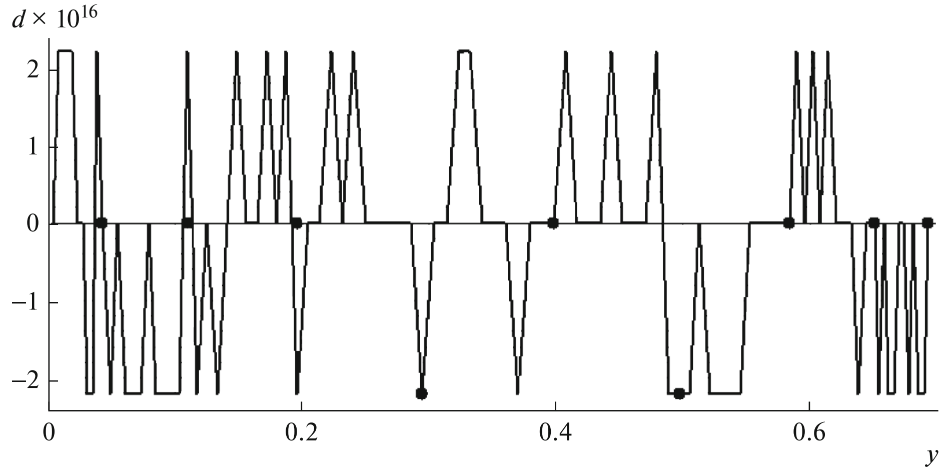

Fig. 2. Error profile for $k = 2$ and $N = 4$.

Nodes (18) were constructed for a function actually representing the ratio of polynomials of the same degree. In our case, the degrees of the polynomials in the numerator and denominator differ by 1 only; hence, their good applicability is expected here. Figure 1 illustrates the results of the calculations. Clearly, the extrema in the middle and near the boundary points almost coincide, differing merely by about 15%. Therefore, the heuristic distribution (18) can be considered unimprovable practicaly, and we may choose it for any N and k.

In addition, observe that the error of the linear-trigonometric distribution is smaller by factors of 7.5 and 4 than its counterparts for the linear and trigonometric distributions, respectively. This is a significant gain in precision.

3.4. Influence of the Number of Parameters

Now, study in detail the approximation error for the linear-trigonometric nodes. If $N \le 3$, the error profiles for all k and N have the same qualitative picture as the bold line in Fig. 1. The error vanishes at the interpolation nodes, representing the half-waves between them ($2N + 1$ totally) with almost the same extrema. In other words, the round-off errors do not affect the results.

The picture changes for $N = 4$. The relative error profile becomes chaotic with the amplitude of about 2×10^{-16} (see Fig. 2). This means that the calculations have reached the round-off errors and further increasing the number of parameters seems pointless.

The approximation error can be characterized by the C norm: $d_C = \max|d(y)|$. Table 1 shows the logarithmic values $\log(d_C)$ for different k and N. Obviously, the maximum error is weakly dependent on k, decreasing fast as a function of N. For $N = 1$, the approximation yields 6 correct decimal digits; for $N = 2$ and $N = 3$, 10 and 14 digits, respectively. For $N = 4$, the expected result is 18 digits, but the round-off errors allow reaching only about 16 digits.

Table 1. Dependence of the logarithmic error $\log(d_C)$ on the number of parameters $2N + 1$ for a linear-trigonometric distribution with different k

k	$2N + 1$			
	3	5	7	9
1	−6.17	−9.88	−13.58	−15.65
2	−5.95	−9.64	−13.36	−15.65
3	−5.84	−9.49	−13.05	−15.65

Table 2. Coefficients a_n and b_m for $N = 3$

a_n, b_m	k		
	1	2	3
a_1	0.266352676322699	0.206100260111457	0.139953685646763
a_2	0.049001409501337	0.043157971511600	0.035516724691922
a_3	0.004902384670602	0.004214705779461	0.003133713899982
a_4	0.000239283842563	0.000210745245200	0.000163991842901
b_1	0.016352676325596	0.018600260114022	−0.005879647683364
b_2	0.017135462543372	0.017896695800770	0.018804085754461
b_3	0.000164279201975	0.000196375385473	0.000046330229928

3.5. Global Approximations

It is possible to obtain approximation formulas better reproducing the asymptotic behavior as $x \to \pm\infty$. Thus, the same approximation will be applicable on the whole interval $-\infty < x < +\infty$. Of course, for higher precision such formulas require many more free parameters, and the parameter calculation procedure becomes complicated. Therefore, we give only two elementary formulas.

The first formula correctly reproduces the main terms of the asymptotics as $x \to \pm\infty$:

$$I_k(x) \approx \Gamma(k+1)\, y(1 + a_1 y)^k, \quad a_1 = [\Gamma(k+2)]^{-1/k}. \tag{19}$$

The second formula reproduces an additional term of the asymptotic expansions as $x \to \pm\infty$:

$$\begin{gathered} I_k(x) \approx \Gamma(k+1)\, y \left[\frac{1 + a_1 y + a_2 y^2}{1 + b_1 y}\right]^k, \quad 0 < y < +\infty \; (-\infty < x < +\infty); \\ a_1 = [\Gamma(k+2)]^{-1/k}, \quad b_1 = a_1 - \frac{1}{2k}\left(1 - 2^{-k}\right), \quad a_2 = a_1 b_1. \end{gathered} \tag{20}$$

Formulas (19) and (20) can be used for any $k > 0$, not necessarily an integer or half-integer (except $k = 0$). Moreover, all the coefficients a_1, a_2, and b_1 turn out to be positive.

Table 3. Coefficients a_n and b_m for $N = 4$

a_n, b_m	k		
	1	2	3
$10a_1$	3.126028287472988	2.588025680820918	1.751249480400745
$10a_2$	0.673008212829461	0.601284498924688	0.484611862591945
$10a_3$	0.087798043423074	0.077052021557577	0.054886614994638
$10a_4$	0.007222414330882	0.006416284842287	0.004875355489602
$10a_5$	0.000295873218273	0.000259595076916	0.000201815238332
$10b_1$	0.626028287472659	0.713025680820707	0.292916147067307
$10b_2$	0.238723363198067	0.249854915262277	0.266194049997825
$10b_3$	0.010727527758408	0.012101958452386	0.006435803052724
$10b_4$	0.000687107172921	0.000728669232953	0.000833646424907
$I_k(0)$	0.82246703342411465	1.8030853547393952	5.6821969769834864

4. RECOMMENDED APPROXIMATIONS

In the final analysis, we recommend approximation (14) for the Fermi–Dirac functions of the integer index k. The computational intensiveness of this formula is almost independent of N. We therefore consider two sets of coefficients. The first set for $N = 3$ is given by Table 2; here the coefficients a_n and b_m are presented with 15 decimal digits, yielding a relative error about 10^{-14}. The second set is combined in Table 3, yielding a relative error about 10^{-16} (which corresponds to 16 significant decimal digits). As the feasible precision of Matlab is 15 decimal digits, the values of the coefficients are multiplied by 10.

Recall that $a_0 = b_0 = 1$. Due to this fact, the coefficients a_n and b_m decrease fast for higher indices. Such behavior of the coefficients is reasonable, testifying to the successful choice of approximation (14).

Once again, note that Tables 2 and 3 should be used if $x \le 0$ and $y \le \ln 2$. For the arguments $x > 0$, apply formulas (8)–(10) with the quantities $I_k(0)$. The latter should be chosen with precision up to the round-off error; their calculation by the 15-digit coefficients is undesirable. These values can be found in Table 3.

5. CONCLUSIONS

1. For the Fermi–Dirac functions of integer indices $k = 1, 2, 3$, this paper has designed simple approximations with the error equal to the round-off error.

2. This paper has also developed a heuristic method for constructing rational approximations based on the interpolation with a special linear-trigonometric grid. The method yields an almost unimprovable error and so it is valuable by itself in the mathematical sense.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research, project nos. 14-01-00161 and 16-31-00062.

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., 30–40 (1967).
3. M. Kim and R. Lundstrom, “Notes on Fermi-Dirac integrals,” arXiv:0811.0116 (2008).
4. N. N. Kalitkin, “About computation of functions the Fermi-Dirak,” Zh. Vychisl. Mat. Mat. Fiz. **8** (1), 173–175 (1968).
5. L. D. Cloutman, “Numerical evaluation of the Fermi-Dirac integrals,” Astrophys. J. Suppl. Ser. **71** (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, *Numerical Methods* (Fizmatlit, Moscow, 1978) [in Russian].
9. N. N. Kalitkin and E. A. Al’shina, *Numerical Methods,* Vol. 1: *Numerical Analysis* (Akademiya, Moscow, 2013) [in Russian].
10. N. N. Kalitkin and A. A. Belov, “Analogue of the Richardson method for logarithmically converging time-marching,” Dokl. Math. **88**, 596–600 (2013).

Translated by A. Mazurov