Precise and fast computation of inverse Fermi–Dirac integral of order 1/2 by minimax rational function approximation

ARTICLE *in* APPLIED MATHEMATICS AND COMPUTATION · MAY 2015

Impact Factor: 1.55 · DOI: 10.1016/j.amc.2015.03.015

1 AUTHOR:

Toshio Fukushima

National Astronomical Observatory of …

378 PUBLICATIONS **1,384** CITATIONS

Applied Mathematics and Computation 259 (2015) 698–707

Contents lists available at ScienceDirect

Applied Mathematics and Computation

journal homepage: www.elsevier.com/locate/amc

Precise and fast computation of inverse Fermi–Dirac integral of order 1/2 by minimax rational function approximation

Toshio Fukushima

National Astronomical Observatory of Japan, Graduate University of General Sciences, 2-21-1, Ohsawa, Mitaka, Tokyo 181-8588, Japan

ARTICLE INFO

Keywords:
Fermi–Dirac integral
Function approximation
Inverse Fermi–Dirac integral
Minimax approximation
Rational function approximation

ABSTRACT

The single and double precision procedures are developed for the inverse Fermi–Dirac integral of order 1/2 by the minimax rational function approximation. The maximum error of the new approximations is one and 7 machine epsilons in the single and double precision computations, respectively. Meanwhile, the CPU time of the new approximations is so small as to be comparable to that of elementary functions. As a result, the new double precision approximation achieves the 15 digit accuracy and runs 30–84% faster than Antia's 28 bit precision approximation (Antia, 1993). Also, the new single precision approximation is of the 24 bit accuracy and runs 10–86% faster than Antia's 15 bit precision approximation.

1. Introduction

The Fermi–Dirac integral of order k and argument η is defined [1, Eq. (1.15)] as

$$F_k(\eta) \equiv \int_0^\infty \frac{x^k}{\exp(x-\eta)+1} dx. \quad (k > -1;\ -\infty < \eta < \infty) \tag{1}$$

It plays an important role in quantum statistics, especially in the solid state physics [2]. Here, x and η are the specific energy ε and the chemical potential μ normalized as

$$x \equiv \frac{\varepsilon}{k_B T}, \quad \eta \equiv \frac{\mu}{k_B T}, \tag{2}$$

where k_B is the Boltzmann constant and T is the absolute temperature. Sometimes, the integral is defined with a different normalization [3] as

$$\mathcal{F}_k(\eta) \equiv \frac{F_k(\eta)}{\Gamma(k+1)}, \tag{3}$$

where $\Gamma(s)$ is the Gamma function of argument s [4, Section 5.2]. Nevertheless, the standard form, $F_k(\eta)$, will be discussed throughout the present article.

In physical situations, needed are $F_k(\eta)$ of some integer and half integer orders [5, Table 1]. Among them, the integral of order 1/2

$$F(\eta) \equiv F_{1/2}(\eta), \tag{4}$$

E-mail address: Toshio.Fukushima@nao.ac.jp

http://dx.doi.org/10.1016/j.amc.2015.03.015
0096-3003/

is most popular because it describes N, the number density of non-relativistic fermion gas in a three dimensional space, as

$$N = N_0 F(\eta). \tag{5}$$

where N_0 is a certain normalization constant.

The computation of $F(\eta)$ when η is given has been extensively investigated since its first appearance [6]. Refer to [5] for their review up to 1982. Among them, the monumental achievement is the massive work of [1]. Meanwhile, the modern standard is `FDPOP5` , the double precision Chebyshev polynomial approximation of $F(\eta)$ [7]. It is a definite improvement of the earlier approximations with the 12 digit accuracy at most [8,9]. Recently developed is `fdlh`, a double precision minimax rational approximation of $F(\eta)$ [10]. It is of the 15 digit accuracy and runs 6 times faster than `FDPOP5`.

In practice, however, frequently required is not only the evaluation of $F(\eta)$ from η but also its inversion, namely the determination of η from $F(\eta)$ [5, Section 2.5]. This is because, in many cases, the chemical potential μ is unknown, and therefore it must be determined from the given values of N and T. Hereafter, denote by $H(u)$ the inverse function of $u \equiv F(\eta)$ for simplicity. Namely, $H(u)$ is defined as a function satisfying the relation

$$H(F(\eta)) = \eta. \tag{6}$$

Figs. 1 and 2 plot sketches of $H(u)$ corresponding for two kinds of intervals, $10^{-9} \leqslant u \leqslant 10^{3}$ and $10^{-0.9} (\approx 0.126) \leqslant u \leqslant 10^{1.3} (\approx 20.0)$, respectively. The function is positive definite and monotonically increasing with respect to u as

$$-\infty < H(u) < H(v) < +\infty. \ (0 < u < v < +\infty) \tag{7}$$

The figures tell that $H(u)$ initially grows logarithmically, then tends to be algebraic, namely increases in proportion to $u^{2/3}$. This change of the growth manner makes its precise and fast computation difficult [5, Section 4]. Refer to the pioneer work of [11] and its followers [12–17].

Currently, the best available procedures to compute $H(u)$ are two minimax rational function approximations of different accuracies developed by [9, Section 3]. Figs. 3 and 4 show that they are of 15 and 28 bit accuracies, respectively. Here, the depicted error is neither the relative nor the absolute errors but their composite defined as

$$\delta_H \equiv \frac{H^*(F(\eta)) - \eta}{\max(1, |\eta|)}, \tag{8}$$

where $H^*(u)$ denotes the approximation of $H(u)$ while $F(\eta)$ is computed by the splitting numerical quadrature method [18, Section 6.10] using a quadruple precision extension of Ooura's `intde` [19], an adaptive numerical quadrature program in the double precision environment based on the double exponential rule [20]. The reason why this error is used will be explained later.

Meanwhile, the averaged CPU times of Antia's approximations are shown in Table 1. Here, the unit of CPU time is ns at a PC with the Intel Core i7–4600U running at 2.10 GHz clock. All the programs are coded in Fortran 90 and compiled by the Intel Visual Fortran Composer XE 2011 update 8 with the maximum optimization and executed under Windows 7 while all other programs are shut down. The results shown here are after the exclusion of the overhead time to call functions in Fortran, which amounts to 12.2 ns in the same environment. At any rate, the table shows that these approximations require

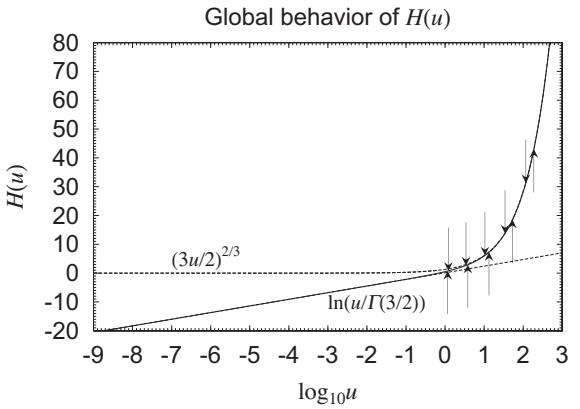

Fig. 1. Global behavior of inverse Fermi–Dirac integral of order 1/2. Plotted in the solid line is the single logarithmic curve of $H(u)$, the inverse Fermi–Dirac integral of order 1/2 defined so as to satisfy the relation, $H(F_{1/2}(\eta)) = \eta$, for the function value interval, $-20 \leqslant H(u) \leqslant 80$. Also attached are two asymptotic curves of $H(u)$ shown in broken lines; (i) $H(u) \approx \ln(u/\Gamma(3/2))$ for the limit $u \to 0$, and (ii) $H(u) \approx (3u/2)^{2/3}$ for the limit $\eta \to +\infty$. The downward and upward arrows indicate the separation points of the piecewise minimax rational approximations developed in the main text aiming to be of the single and double precision accuracies, respectively.

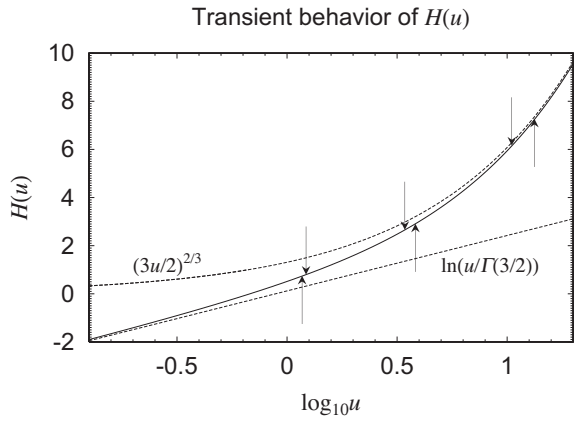

Fig. 2. Transient behavior of $H(u)$. Same as Fig. 1 but for a narrower interval such that $-0.9 \leqslant \log_{10}u \leqslant 1.3$, where the deviation from the two asymptotic forms is clearly visible.

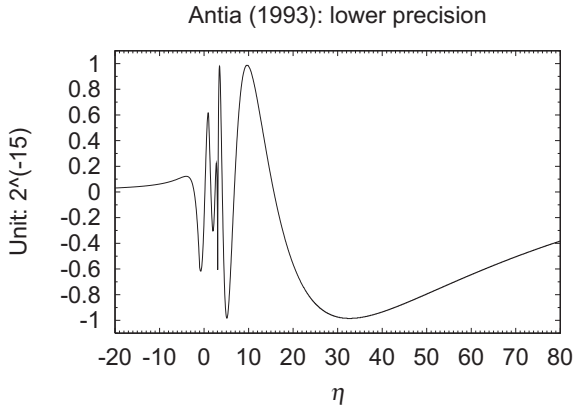

Fig. 3. Error of Antia's method: the lower precision. Plotted is the error curve of the lower precision approximation of $H(u)$ developed by [9]. The error depicted here is a sort of composite of the relative and absolute errors defined as $(H(F(\eta)) - \eta)/\max(1, |\eta|)$ while $F(\eta)$ is computed by the quadruple precision numerical quadrature. The achieved accuracy is 15 bit.

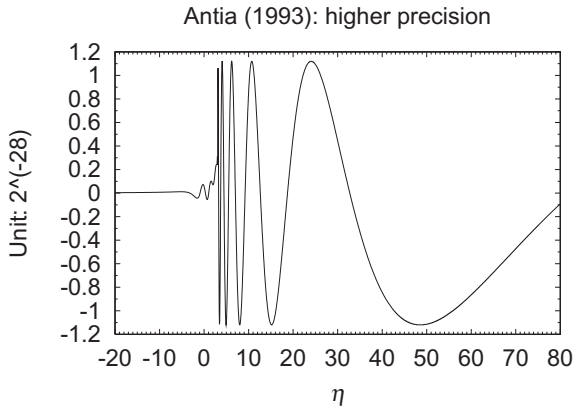

Fig. 4. Error of Antia's method: the higher precision. Same as Fig. 3 but of the higher precision approximation of [9]. This time, the achieved accuracy is 28 bit.

2.1–2.5 times that of the double precision exponential function provided by the standard mathematical library. In terms of the computing precision, this situation is not satisfactory if compared with the forward procedures to compute $F(\eta)$ [10],

Table 1
Comparison of CPU time. Shown are the averaged CPU times to compute the inverse Fermi–Dirac integral of order 1/2. Compared are the lower and higher precision approximations given by [9], and the single and double precision approximations newly developed. Averaged are the CPU times for $2^{28} \approx 2.68 \times 10^8$ values of η evenly distributed in two intervals of argument, $-5 \leqslant \eta \leqslant 35$ and $-20 \leqslant \eta \leqslant 80$. They correspond to the integral value intervals, $5.96 \times 10^{-3} < u \equiv F(\eta) < 138$ and $1.83 \times 10^{-9} < u < 477$, respectively. The unit of CPU time is ns at a PC with the Intel Core i7–4600U running at 2.10 GHz clock. The averaged CPU time of the double precision exponential function is 23.7 ns in the same environment.

Method		Accuracy	$-5 \leqslant \eta \leqslant 35$	$-20 \leqslant \eta \leqslant 80$
Antia [9]	lower	15 bit	51.4	48.8
	higher	28 bit	60.2	57.9
New	single	24 bit	27.7	44.2
	double	50 bit	32.8	44.6

which is of the 15 digit accuracy. Therefore, this article presents new minimax rational function approximations to compute $H(u)$ aimed to be with the single and double precision accuracies, respectively.

2. Method

2.1. Functional forms to be approximated

Consider the computation of $\eta \equiv H(u)$ by a piecewise minimax approximation. In general, there is no need to regard $H(u)$ itself as the function to be approximated. Indeed, any function of it such as $\exp(\eta)$ or $1/\eta^2$ can be approximated instead as will be seen below. Thus, first of all, seek for appropriate functional forms to be approximated by rational functions.

It is well known that $u \equiv F(\eta)$ has two series expansions [1]: (i) the Maclaurin series with respect to an auxiliary variable, $z \equiv \exp(\eta)$, written as

$$u = \frac{\sqrt{\pi} z}{2}\left(1 - \frac{z}{2\sqrt{2}} + \frac{z^2}{3\sqrt{3}} - \frac{z^3}{8} + \frac{z^4}{5\sqrt{5}} - \cdots\right), \quad (\eta < 0) \tag{9}$$

and (ii) the Sommerfeld expansion [6] expressed as

$$u = \frac{2\eta^{3/2}}{3}\left(1 + \frac{\pi^2}{8}\eta^{-2} + \frac{7\pi^4}{640}\eta^{-4} + \frac{31\pi^6}{3072}\eta^{-6} + \frac{4191\pi^8}{163840}\eta^{-8} + \cdots\right). \ (\eta \gg 1) \tag{10}$$

The coefficients of these series are obtained by the following commands of Mathematica 10 [21,22]:

```
FM[z_]=Gamma[3/2]Normal[-Series[PolyLog[3/2,-z],{z,0,5}]]
FS[eta_]=(2/3) eta^(3/2)(1 + Sum[3Pochhammer[5/2-2 m,2 m-1](1-2^(1-2 m)) Zeta[2 m]
  eta^(-2 m),{m,1,4}])
```

The Sommerfeld expansion is rewritten into a simpler power series as

$$\begin{aligned} v &= w\left(1 + \frac{\pi^2}{8}w + \frac{7\pi^4}{640}w^2 + \frac{31\pi^6}{3072}w^3 + \frac{4191\pi^8}{163840}w^4 + \cdots\right)^{-4/3} \\ &= w\left(1 - \frac{\pi^2}{6}w + \frac{7\pi^4}{720}w^2 - \frac{163\pi^6}{12960}w^3 - \frac{47317\pi^8}{1555200}w^4 - \cdots\right). \end{aligned} \tag{11}$$

where

$$v \equiv \left(\frac{2}{3u}\right)^{4/3}, \tag{12}$$

is an alternative argument to be used when $u \gg 1$ and

$$w \equiv \frac{1}{\eta^2}, \tag{13}$$

is another auxiliary variable to be used when $\eta \gg 1$.

These series are inverted with respect to z and w, and then finally $\eta \equiv H(u)$ is expressed by means of the power series expansion as

$$\eta = \ln z = \ln\left(\frac{2u}{\sqrt{\pi}} + \frac{\sqrt{2}u^2}{\pi} + \frac{2\left(9 - 4\sqrt{3}\right)u^3}{9\pi\sqrt{\pi}} + \frac{\left(36 + 45\sqrt{2} - 40\sqrt{6}\right)u^4}{18\pi^2} + \frac{\left(2375 + 1350\sqrt{2} - 2100\sqrt{3} - 288\sqrt{5}\right)u^4}{225\pi^2\sqrt{\pi}} + \cdots\right), \ (u \ll 1) \tag{14}$$

$$\eta = \frac{1}{\sqrt{w}} = 1\Big/\sqrt{v + \frac{\pi^2 v^2}{6} + \frac{11\pi^4 v^3}{240} + \frac{179\pi^6 v^4}{6480} + \frac{37649\pi^8 v^5}{777600} + \cdots}. \quad (v \ll 1) \tag{15}$$

The inverted series of (i) z in terms of u and (ii) w in terms of v are obtained by the following commands of Mathematica 10, respectively:

```
HM[u_]=Normal[InverseSeries[Series[FM[z],{z,0,5}]/Gamma[3/2],x]]/.x->2u/Sqrt[Pi]
HS[v_]=1/Normal[InverseSeries[Series[FS[1/Sqrt[w]]^(-4/3),{w,0,5}],x]]/.x->(3/2)^(4/3) v
```

At any rate, the above two limiting forms and the behavior of $H(u)$ shown in Figs. 1 and 2 suggest a splitting of the whole interval of u into two or more sub intervals as

$$H(u) \approx \begin{cases} H_0(u) \equiv \ln\left(uR_0(u)\right), & (0 < u \leqslant u_0) \\ H_j(u) \equiv R_j(u), & \left(u_{j-1} < u \leqslant u_j;\ j = 1(1)J\right) \\ H_S(u) \equiv \sqrt{R_S(v)/v}, & \left(u_J < u < +\infty\right) \end{cases} \tag{16}$$

where (i) J is a certain non negative integer, (ii) R_j for $j = 0(1)J$ and R_S are rational functions, and (iii) u_j for $j = 0(1)J$ are certain positive numbers.

The new forms are noticeably different from the existing approximations of $H(u)$ [11–17]. For example, [11] used a form

$$H(u) \approx \ln\left(\frac{y}{1 - y/4}\right), \tag{17}$$

where

$$y \equiv \frac{2u}{\sqrt{\pi}}. \tag{18}$$

This is of the same form as our first one. Nevertheless, it is used for all positive values of u. Also, [13] selected another form

$$H(u) \approx \ln y + uP(u), \tag{19}$$

where $P(u)$ is a degree 3 polynomial, the coefficients of which were determined by inverting the Maclaurin series, Eq. (9). Further, [14] adopted the same form as Eq. (19) but assumed $P(u)$ to be a linear function and tuned its two coefficients so as to decrease the approximation errors as much as possible. On the other hand, [15] assumed the form

$$H(u) \approx \ln y + u[S(u)]^{-1/4}, \tag{20}$$

and proposed two approximations of $S(u)$ as a linear function and a simple irrational function containing $\sqrt{u}$, respectively. These forms are applied to all the values of u.

Meanwhile, [9, Eq. (6)] adopted a splitting into two regions in computing $H(u)$. His first function is of the same form as our first one. However, his second function is significantly different from our last one as

$$H(u) \approx u^{2/3} R_A\left(u^{2/3}\right), \tag{21}$$

where $R_A(u)$ is a rational function. This form did not taken into account a fact that $F(\eta)/\eta^{3/2}$ is expanded as a power series of not $1/\eta$ but $1/\eta^2$, which is the excellent feature of the Sommerfeld expansion [6].

At any rate, before going further, consider the proper error of approximation. Denote by $\Delta\eta$ the absolute approximation error of $H(u)$ as

$$\Delta\eta \equiv H^*(u) - H(u), \tag{22}$$

where $H^*(u)$ is the approximation of $H(u)$. As [5, Section 3] emphasized, the sensitivity relation between η and u changes from $\Delta\eta \propto (\Delta u)/u$ to $(\Delta\eta)/\eta \propto (\Delta u)/u$ when u increases from 0 to $+\infty$. In other words, neither the absolute error, $\Delta\eta$, nor the relative error, $(\Delta\eta)/\eta$, is an appropriate error to be minimized for the whole interval of u. Hereafter, adopted is a composite error, δ_H, already defined in Eq. (8). Figs. 3 and 4 have shown this error of two approximations of [9].

2.2. Determination of separation points

Now that the forms of approximation function is fixed, consider the basic problem how to specify the separation points, u_j, for $j = 0(1)J$. As [9, Section 3] stressed, a naive minimax rational function approximation such as using the middle form, $H(u) \approx R_j(u)$, faces a fatal trouble in obtaining the minimax solution when the approximation interval contains its zero. This happens when the interval contains the critical input argument

$$u_C \equiv F(0) \approx 0.678093895153101007. \tag{23}$$

Following [9], this issue is resolved by demanding u_0 to satisfy the condition, $u_C < u_0$.

On the other hand, the Sommerfeld expansion is asymptotic [3]. This means that the expansion does not converge unconditionally in general. In other words, one can not arbitrary increase the number of terms in the expansion. Indeed, there exists η_S, a minimum value of η such that an appropriately truncated Sommerfeld expansion guarantees the

Table 2
Separation points. Listed are the separation points in terms of η and $u \equiv F(\eta)$ of the single and double precision piecewise minimax rational approximations of $H(u)$. The values of η are rounded down so as to be on the safer side. The corresponding value of u are given with a few more-then-enough digits in order to avoid the unnecessary loss of information in the implementation.

Precision	j	η	$u \equiv F(\eta)$
Single	0	+0.795385	+1.21838255
	1	+2.655397	+3.43021110
	2	+6.162879	+10.5407599
	3	+13.785016	+34.3434203
	4	+31.286097	+116.810894
Double	0	+0.744703	+1.17683303804380831
	1	+2.909680	+3.82993088157949761
	2	+7.272297	+13.3854493161866553
	3	+18.500335	+53.2408277860982205
	4	+43.046736	+188.411871723022843

computing accuracy required by the given relative error tolerance, δ. In case of $F(\eta) \equiv F_{1/2}(\eta)$, its value is 12.0 and 31.6 when $\delta = 2^{-24}$ and $\delta = 10^{-15}$, respectively [23].

The same is also true for the inverse function. In fact, there exists a minimum value of u corresponding to η_S as $u_S \equiv F(\eta_S)$. The numerical value of u_S is 28.0 and 119 in the 24 bit and 15 digit computations, respectively. In order to achieve the corresponding accuracy, it is safe to set u_J, the maximum separation value, larger than this threshold value, i.e. $u_S < u_J$.

While satisfying these conditions, $u_C < u_0$ and $u_S < u_J$, the number of intermediate sub intervals, J, and the separation points, u_0 through u_J, are determined so as to realize the global minimax feature of the relative error curve when $0 < u \leqslant u_J$. Here, the word 'global' is meant to minimize the maximum relative errors of the locally minimax approximations such that they are all equal to the given value of δ. Also, in the process to obtain the global minimax approximation, the types of the rational functions are limited to be even, say of the type (N, N). This is because the even type rational functions lead to the best cost performance in general [18, Section 5.13].

Usually, the degree N is chosen as the minimum value to assure that the obtained maximum error is less than δ. Alternatively, the actual process is reversed. Namely, when N is given, u_0 is first determined such that the absolute value of the maximum approximation error in the interval $0 < u \leqslant u_0$ is equal to the given value of δ. Next, u_1 is determined such that the absolute value of the maximum approximation error in the interval $u_0 < u \leqslant u_1$ is equal to δ again. This process is repeated until $u_S < u_J$ for a certain value of J.

The rigorous equality is not necessary for realizing an almost minimax feature. Thus, by limiting to 10^{-6} the absolute accuracy in terms of not u but η, the separation points are determined as listed in Table 2.

2.3. Minimax rational approximation of inverse function

Preliminary numerical experiments to approximate $H(u)$ by rational functions of various even types (N, N) as $N = 1(1)10$ concluded that $N = 3$ and $N = 7$ result solutions requiring not so large value of J as $J = 4$ while being sufficiently accurate in the single and double precision computations, respectively. Refer to Table 3. The inverse minimax rational approximation of

Table 3
Types of rational functions. Listed are (N, M), the type of rational functions adopted by the four minimax approximations of $H(u)$: Antia's lower and higher precision approximations and the new single and double precision approximations. Notice that the separation points of the new method are expressed with only 3 significant digits. Their more precise values are given in Table 2.

Method		Accuracy	Interval	Type
Antia [9]	Lower	15 bit	$0 < u < 4$	(2,2)
			$4 \leqslant u < +\infty$	(2,2)
	Higher	28 bit	$0 < u < 4$	(4,3)
			$4 \leqslant u < +\infty$	(6,5)
New	Single	24 bit	$0 < u < 1.22$	(2,2)
			$1.22 \leqslant u < 3.43$	(3,3)
			$3.43 \leqslant u < 10.5$	(3,3)
			$10.5 \leqslant u < 34.3$	(3,3)
			$34.3 \leqslant u < 117$	(3,3)
			$117 \leqslant u < +\infty$	(2,1)
	Double	50 bit	$0 < u < 1.18$	(4,4)
			$1.18 \leqslant u < 3.83$	(7,7)
			$3.83 \leqslant u < 13.4$	(7,7)
			$13.4 \leqslant u < 53.2$	(7,7)
			$53.2 \leqslant u < 188$	(7,7)
			$188 \leqslant u < +\infty$	(3,2)

Table 4
Coefficients of minimax rational function approximation of $H(u)$: single precision. Listed are the numerical coefficients of the minimax rational function approximating $H(u)$ with the single precision accuracy. The adopted approximation form is (i) $\ln(uR_0(u))$ when $u < u_0$, (ii) $R_j(t)$ when $u_{j-1} \leqslant u < u_j$ for $j = 1, \cdots, J$ where $t \equiv \alpha_j + \beta_j u$, and (iii) $\sqrt{R_S(s)/(1-s)}$ when $u_J < u$ where $s \equiv 1 + \beta_S u^{-4/3}$. Here $R_0(u), R_j(t)$, and $R_S(s)$ are rational functions of the type (N, M) expressed such as $R_j(t) = \sum_{n=0}^{N} P_n t^n / \sum_{m=0}^{M} Q_m t^m$. The number of the intermediate intervals, J, is set as $J = 4$. The adopted types are $(2, 2)$ for $R_0(u)$, $(3, 3)$ for $R_1(t)$ through $R_4(t)$, and $(2, 1)$ for $R_S(s)$. The linear transform coefficients, α_j and β_j as well as β_S, are chosen such that t and s satisfy the standard condition, $0 \leqslant t < 1$ and $0 \leqslant s < 1$, respectively. The list also contains the values of u_j, α_j, and β_j. A few more-than-enough digits are shown so as to avoid unnecessary information loss in the implementation.

	$R_0(u)$	$R_1(t)$	$R_2(t)$	$R_3(t)$	$R_4(t)$	$R_S(s)$
P_0	+127.456123	+22.3158685	+74.0135089	+156.549383	+376.286772	+1974.50048
P_1	+30.3620672	+122.487649	+294.367987	+612.130579	+1449.93277	+144.437558
P_2	+2.29733586	+135.023156	+294.232354	+639.532875	+1487.47498	+1
P_3		+30.5460708	+64.9306737	+145.238686	+333.722383	
Q_0	+112.955041	+28.0566860	+27.8728584	+25.4019873	+27.2967945	+10.3906494
Q_1	−18.1545791	+59.9641578	+62.0649704	+59.3035998	+61.1014417	+0.669052603
Q_2	+1	+27.8629074	+27.1148810	+26.9857305	+27.1844466	
Q_3		+1	+1	+1	+1	
u_j	+1.21838255	+3.43021110	+10.5407599	+34.3434203	+116.810894	
α_j		−0.550848552	−0.482411581	−0.442839571	−0.416448071	+1
β_j		+0.452114610	+0.140636120	+0.0420121106	+0.0121259929	−111.632691

Table 5
Coefficients of minimax rational function approximation of $H(u)$: double precision. Same as Table 4 but for the first half of the approximation with the double precision accuracy. The adopted types are $(4, 4)$ for $R_0(u)$, and $(7, 7)$ for $R_1(t)$ and $R_2(t)$.

	$R_0(u)$	$R_1(t)$	$R_2(t)$
P_0	+254870.603839626390	+489.140447310410217	+1019.84886406642351
P_1	+66722.8518750022136	+5335.07269317261966	+ 9440.18255003922075
P_2	+6881.02772176766106	+20169.0736140442509	+ 33947.6616363762463
P_3	+335.397807967219390	+35247.8115595510907	+ 60256.7280980542786
P_4	+6.66544737164926158	+30462.3668614714761	+ 55243.0045063055787
P_5		+12567.9032426128967	+24769.8354802210838
P_6		+2131.86789357398657	+4511.77288617668292
P_7		+93.6520172085419439	+211.432806336150141
Q_0	+225873.191629079972	+656.826207643060606	+350.502070353586442
Q_1	−30978.7782754284374	+4274.82831051941605	+ 2531.06296201234050
Q_2	+1906.07868101188410	+10555.7581310151498	+6939.09850659439245
Q_3	−63.6828217274155952	+12341.8742094611883	+ 9005.40197972396592
Q_4	+1	+6949.18854413197094	+5606.73612994134056
Q_5		+1692.19650634194002	+1488.76634564005075
Q_6		+129.221772991589751	+121.537028889412581
Q_7		+1	+1
u_j	+1.17683303804380831	+3.82993088157949761	+13.3854493161866553
α_j		−0.443569407329314587	−0.400808277205416960
β_j		+0.376917874490198033	+0.104651569335924949

a function, $F(\eta)$, for the interval, $\eta_L \leqslant \eta \leqslant \eta_U$, are determined by `GeneralMiniMaxApproximation` command of Mathematica Version 10, which employs the Remez's algorithm [22]. Its sample usage is

```
mmH = GeneralMiniMaxApproximation[{N[F[eta], 40], eta}, {eta, {etaL, etaU}, NH, NH}, u];
```

where `NH` $= N$. In the determination, $F(\eta)$ is given by its quadruple precision piecewise Chebyshev polynomial approximation [10]. In order to ensure the convergence of the minimax optimization process, $F(\eta)$ is evaluated with 40 working digits. Once the minimax optimization process converges, `mmH[[2, 1]]` and `mmH[[2, 2]]` provide the determined rational function and the maximum relative error, respectively.

From the viewpoint to minimize the round-off errors, the argument of rational functions is transformed such that the transformed argument is non negative definite and monotonically changes from 0 to 1 or from 1 to 0 when the original argument, u or v, moves from the lower to upper end point of the given argument interval. For example, the argument of R_j is transformed from u to a new argument t by a linear transformation as

$$t \equiv \alpha + \beta u, \tag{24}$$

where α and β are constants defined as

$$\alpha \equiv \frac{-u_L}{u_U - u_L} < 0, \quad \beta \equiv \frac{1}{u_U - u_L} > 0. \tag{25}$$

Table 6
Coefficients of minimax rational function approximation of $H(u)$: double precision, continued. Same as Table 5 but for the second half. The adopted types of the rational functions are $(7,7)$ for $R_3(t)$ and $R_4(t)$, and $(3,2)$ for $R_S(s)$.

	$R_3(t)$	$R_4(t)$	$R_S(s)$
P_0	+11885.8779398399498	+11730.7011190435638	+ 1281349.5144821933
P_1	+113220.250825178799	+99421.7455796633651	+420368.911157160874
P_2	+408524.373881197840	+327706.968910706902	+ 689.69475714536117
P_3	+695674.357483475952	+530425.668016563224	+1
P_4	+569389.917088505552	+438631.900516555072	
P_5	+206433.082013681440	+175322.855662315845	
P_6	+27307.2535671974100	+28701.9605988813884	
P_7	+824.430826794730740	+1258.20914464286403	
Q_0	+1634.40491220861182	+634.080470383026173	+6088.08350831295857
Q_1	+12218.1158551884025	+4295.63159860265838	+221.445236759466761
Q_2	+32911.7869957793233	+10868.5260668911946	+0.718216708695397737
Q_3	+38934.6963039399331	+12781.6871997977069	
Q_4	+20038.8358438225823	+7093.80732100760563	
Q_5	+3949.48380897796954	+1675.06417056300026	
Q_6	+215.607404890995706	+125.750901817759662	
Q_7	+1	+1	
u_j	+53.2408277860982205	+188.411871723022843	
α_j	−0.335850513282463787	−0.393877462475929313	+1
β_j	+0.0250907164450825724	+0.00739803415638806339	−1080.13412050984017

where $u_L \equiv F(\eta_L)$ and $u_U \equiv F(\eta_U)$. This definition of t is in order to avoid the cancellation problems as much as possible in the evaluation of the numerator and denominator polynomials by Horner's method.

Also, following [9], the coefficients of obtained rational functions are normalized by setting the coefficient of the highest degree of the denominator or numerator polynomial as ± 1, where the sign is selected such that the majority of the polynomial coefficients become positive. This trick saves one multiplication in the evaluation process of the rational function without degrading the computational accuracy.

Anyhow, the numerical coefficients of the approximation rational function determined by this algorithm are listed in Table 4 for the single precision approximation and in Tables 5 and 6 for the double precision approximation, respectively. Thanks to the appropriate choice of t, all the determined coefficients except some of the denominator polynomials of $R_0(u)$ are positive definite. This effectively avoids the cancellation problems. Even for the denominator polynomials of $R_0(u)$, the magnitude of the alternating coefficients decreases much more than factor 2, and therefore, there is no chance of information loss.

3. Result

Examine the computational cost and performance of the new approximations. First, the errors are measured. As described in Section 2, the two kinds of new approximations are aimed to be of the single and double precision accuracies, respectively. Figs. 5 and 6 show that the composite errors of the single precision approximation do not exceed the single precision machine epsilon. The standard minimax feature of the error curves is obvious.

Next, Fig. 7 plots the case of the double precision approximation. As long as $u \leqslant u_4 \approx 43$, the errors scatter and no clear systematic trend is seen. Meanwhile, when $u > u_4$, a slightly unbalanced distribution of errors is observed. Anyhow, Table 7

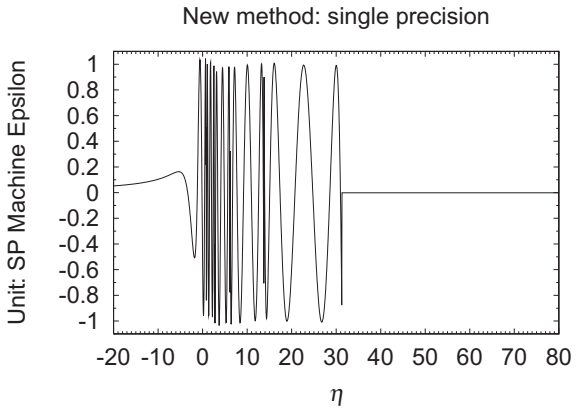

Fig. 5. Error of new method: single precision. Same as Fig. 3 but for the new method of the single precision accuracy. The achieved accuracy is 24 bit.

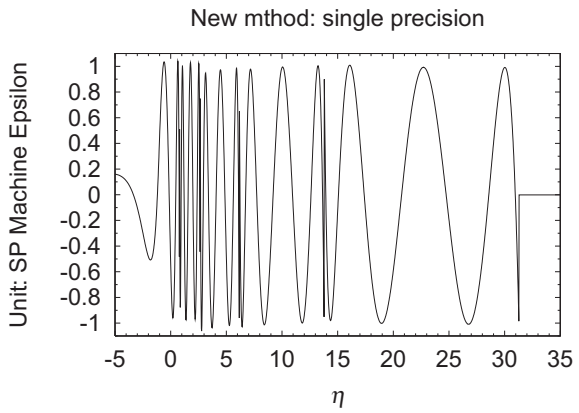

Fig. 6. Error of new method: single precision, close-up. Same as Fig. 5 but for a narrower argument interval as $-5 \leqslant \eta \leqslant 35$.

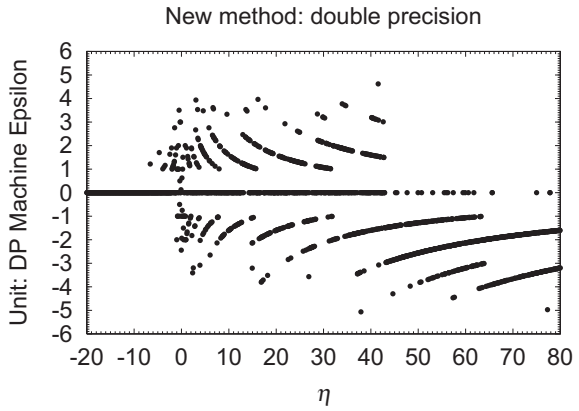

Fig. 7. Error of new method: double precision. Same as Fig. 5 but for the new method with the double precision accuracy. The errors shown here are all due to the round-off errors.

Table 7
Statistics of errors of double precision minimax approximation of $H(u)$. Listed are the mean, the sample standard deviation (SD), the maximum, and the minimum of δ_H, the composite errors of the double precision minimax approximation of $\eta \equiv H(u)$. The statistics are taken for 10^6 sample points of η evenly distributed in the interval, $[-20, 80]$. The results are expressed in the unit of the double precision machine epsilon.

Mean	SD	Max.	Min.
−0.74	1.57	7.13	−7.00

reports the statistics of these double precision errors. There exist no significant bias in the errors. The magnitude of errors is typically less than 2 machine epsilons and 7 machine epsilons at most.

Move to the aspect of computational speed. Table 1 has already compared the averaged CPU times of the two approximations of [9] and the single and double precision approximations of the new method. The averages are taken over $2^{28} \approx 2.68 \times 10^8$ values of η uniformly distributed in two domains; (i) $-20 \leqslant \eta \leqslant 80$, and (ii) $-5 \leqslant \eta \leqslant 35$. All the programs are coded in Fortran 90 and compiled by the Intel Visual Fortran Composer XE 2011 update 8 with the maximum optimization.

The new single and double precision approximations run fairly fast. For example, in the transient region, $5.96 \times 10^{-3} < u < 138$, the new single and double precision approximations require only 17 and 38 % more than that of the exponential function, respectively. As a result, they run 1.9 and 1.8 times faster than the lower and higher precision approximations of [9] which are of much lower accuracies, respectively.

4. Conclusion

By using the minimax rational function approximation, the single and double precision procedures are developed to compute $H(u)$, the inverse function of $F(\eta) \equiv F_{1/2}(\eta)$, the Fermi–Dirac integral of order 1/2. The errors of the new approximations

defined as $(H(F(\eta)) - \eta)/\max(1, |\eta|)$ is one and 7 machine epsilons at most in the single and double precision computations, respectively. On the other hand, the averaged CPU times to evaluate the new approximations is only 17–88% more than that of the exponential function provided by the standard mathematical function library. As a result, the new single precision procedure is of the 24 bit accuracy and runs 10–86% faster than the 15 bit precision approximation of [9]. Also, the new double precision procedure achieves the 15 digit accuracy and runs 30–84% faster than the 28 bit precision approximation of [9].

The Fortran 90 functions to compute the new approximations as well as their sample outputs are freely available from the following WEB site.

https://www.researchgate.net/profile/Toshio_Fukushima/.

References

[1] J. McDougall, E.C. Stoner, The computation of Fermi–Dirac functions, Phil. Trans. Royal Soc. London, Ser. A., Math. Phys. Sci. 237 (1938) 67–104.
[2] N.W. Ashcroft, N.D. Mermin, Solid State Physics, Holt, Rinehalt, and Winston, Dumfries, 1976.
[3] R. Dingle, The Fermi–Dirac integrals $\mathcal{F}_p(\eta) = (p!)^{-1} \int_0^\infty \varepsilon^p/(e^{\varepsilon-\eta} + 1)d\varepsilon$, Appl. Sci. Res. 6 (1957) 225–239.
[4] F.W.J. Olver, D.W. Lozier, R.F. Boisvert, C.W. Clark (Eds.), NIST Handbook of Mathematical Functions, Cambridge Univ. Press, Cambridge, 2010. <http://dlmf.nist.gov/>.
[5] J.S. Blakemore, Approximations for Fermi–Dirac integrals, especially the function $\mathcal{F}_{1/2}(\eta)$ used to describe electron density in a semiconductor, Solid-State Electron. 25 (1982) 1067–1076.
[6] A. Sommerfeld, Zur Elektronentheorie der Metalle auf Grund der Fermischen Statistik. I. Teil: Allgemeines, Strömungs und Austrittsvorgänge, Zeitschrift für Physik 47 (1929) 1–32.
[7] A.J. Macleod, Algorithm 779: Fermi–Dirac functions of order −1/2, 1/2, 3/2, and 5/2, ACM Trans. Math. Software 24 (1998) 1–12.
[8] W.J. Cody, H.C. Thatcher, Rational Chebyshev approximations for Fermi–Dirac integrals of orders −1/2, 1/2 and 3/2, Math. Comp. 21 (1967) 30–40.
[9] H.M. Antia, Rational function approximations for Fermi–Dirac integrals, Astrophys. J. Suppl. Ser. 84 (1993) 101–108.
[10] Fukushima, T., Precise and fast computation of Fermi–Dirac integral of integer and half integer order by piecewise minimax rational approximation, Appl. Math. Comp., submitted of publication.
[11] W. Ehrenberg, The electric conductivity of simple semiconductors, Proc. Phys. Soc. London A63 (1950) 75–76.
[12] N.G. Nilsson, An accurate approximation of the generalized Einstein relation for degenerate semiconductors, Phys. Stat. Solidi 19 (1973) K75–K78.
[13] W.B. Joyce, R.W. Dixon, Analytic approximations for the Fermi energy of an ideal Fermi gas, Appl. Phys. Lett. 31 (1977) 354–356.
[14] W.B. Joyce, Analytic approximations for the Fermi energy in (Al, Ga)As, Appl. Phys. Lett. 32 (1978) 680–681.
[15] N.G. Nilsson, Empirical approximations for the Fermi energy in a semiconductor with parabolic bands, Appl. Phys. Lett. 33 (1978) 653–654.
[16] D. Bednarczyk, J. Bednarczyk, The approximation of the Fermi–Dirac integral $\mathcal{F}_{1/2}(\eta)$, Phys. Lett. 64 (1978) 409–410.
[17] T.Y. Chang, A. Izabelle, Full range analytic approximations for Fermi energy and Fermi–Dirac integral $F_{-1/2}$ in terms of $F_{1/2}$, J. Appl. Phys. 65 (1989) 2162–2164.
[18] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes: The Art of Scientific Computing, third ed., Cambridge Univ. Press, Cambridge, 2007.
[19] T. Ooura, Numerical Integration (Quadrature) - DE Formula (Almighty Quadrature), 2006. <http://www.kurims.kyoto-u.ac.jp/ooura/intde.html>
[20] H. Takahashi, H. Mori, Double exponential formulas for numerical integration, Publ. RIMS, Kyoto Univ. 9 (1974) 721–741.
[21] S. Wolfram, The Mathematica Book, 5th ed., Wolfram Research Inc./Cambridge Univ. Press, Cambridge, 2003.
[22] Wolfram Research, Function Approximations Package Tutorial, Wolfram Research Inc., 2014. <http://reference.wolfram.com/language/FunctionApproximations/tutorial/FunctionApproximations.html>
[23] T. Fukushima, Analytical computation of generalized Fermi–Dirac integrals by truncated Sommerfeld expansions, Appl. Math. Comm. 234 (2014) 417–433.